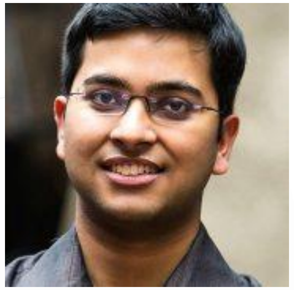
A New Frontier at the Intersection of Causality and Large Language Models

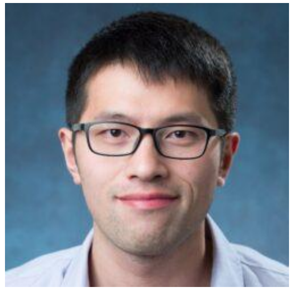
Emre Kıcıman

emrek@microsoft.com

Microsoft Research

https://arxiv.org/abs/2305.00050





Causal Reasoning and Large Language Models: Opening a New Frontier for Causality

Emre Kıcıman*
Microsoft Research
emrek@microsoft.com

Robert Ness
Microsoft Research
robertness@microsoft.com

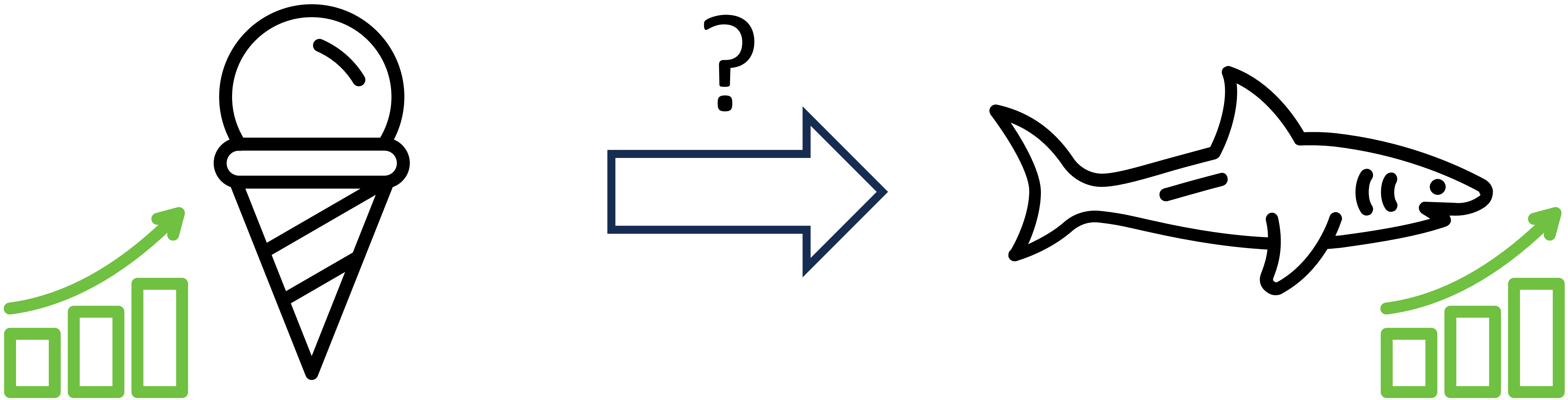
Amit Sharma
Microsoft Research
amshar@microsoft.com

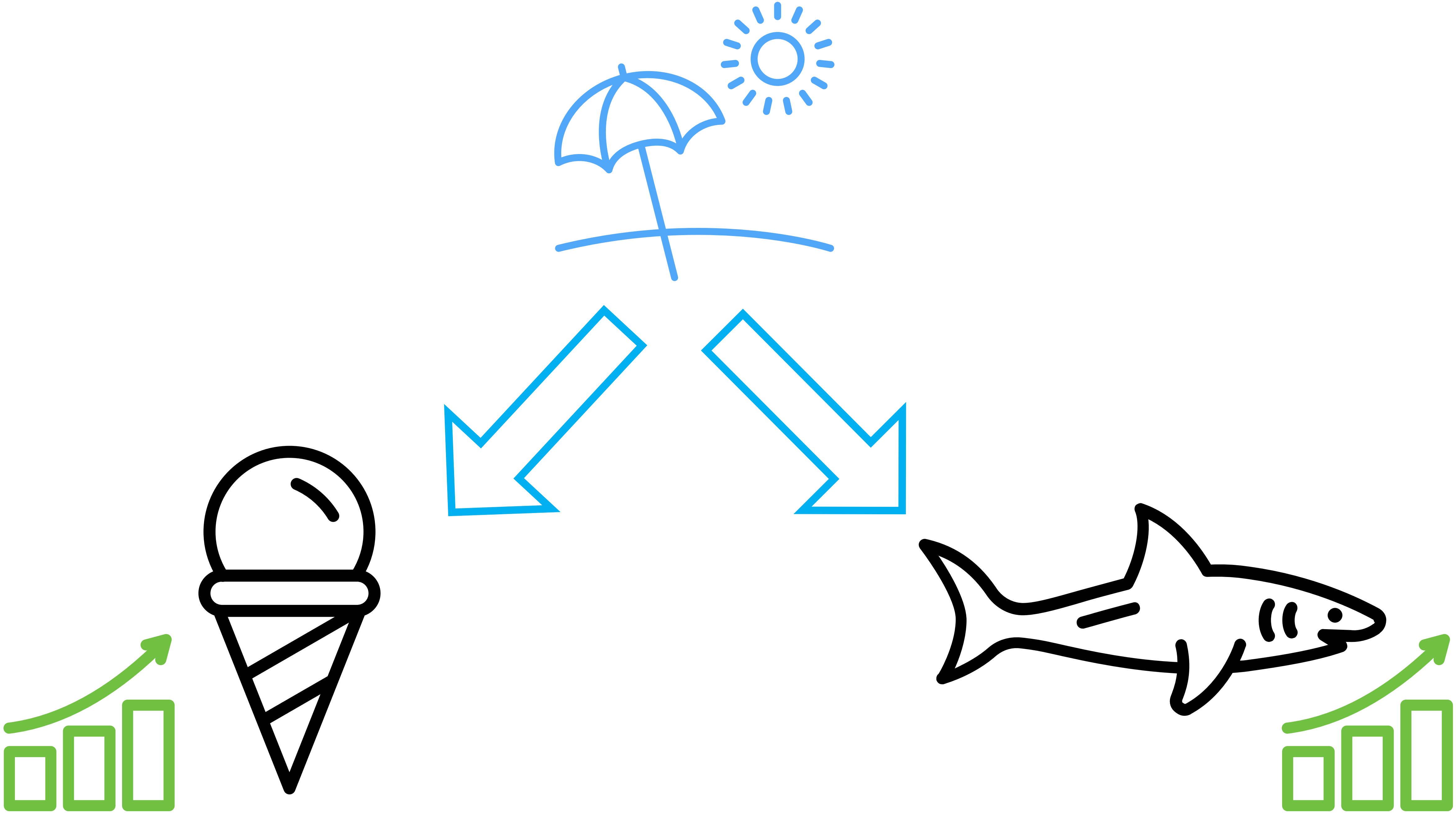
Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Working Paper May 9, 2023

...lly in high-stakes scenarios.
...at imply that complex causal rea-
...d in LLMs. However,
...about


SHIPPING
SENT





This happens frequently in practice

- **Night lights cause myopia in young children? [Nature 1999]**

  “Children who sleep with a night light or other artificial light in their room until the age of two have a higher incident of nearsightedness”

  Missing a confounding factor: Parents with myopia!

- **Vitamin C and E reduces heart disease? [Lancet 2001]**

  Later failed RCT replications

  Multiple confounds unaccounted: socioeconomic, behavioral, …

- **Review of 52 claims of observational studies [Young and Karr 2011]**

  None reproduced; several contradicted

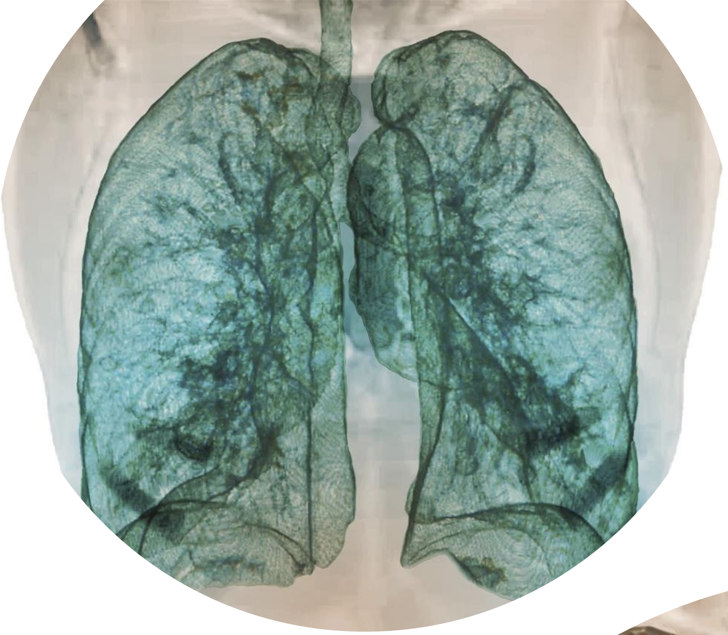
Thanks to Ronny Kohavi for collecting these and other refuted causal claims at http://bit.ly/refutedCausalClaims

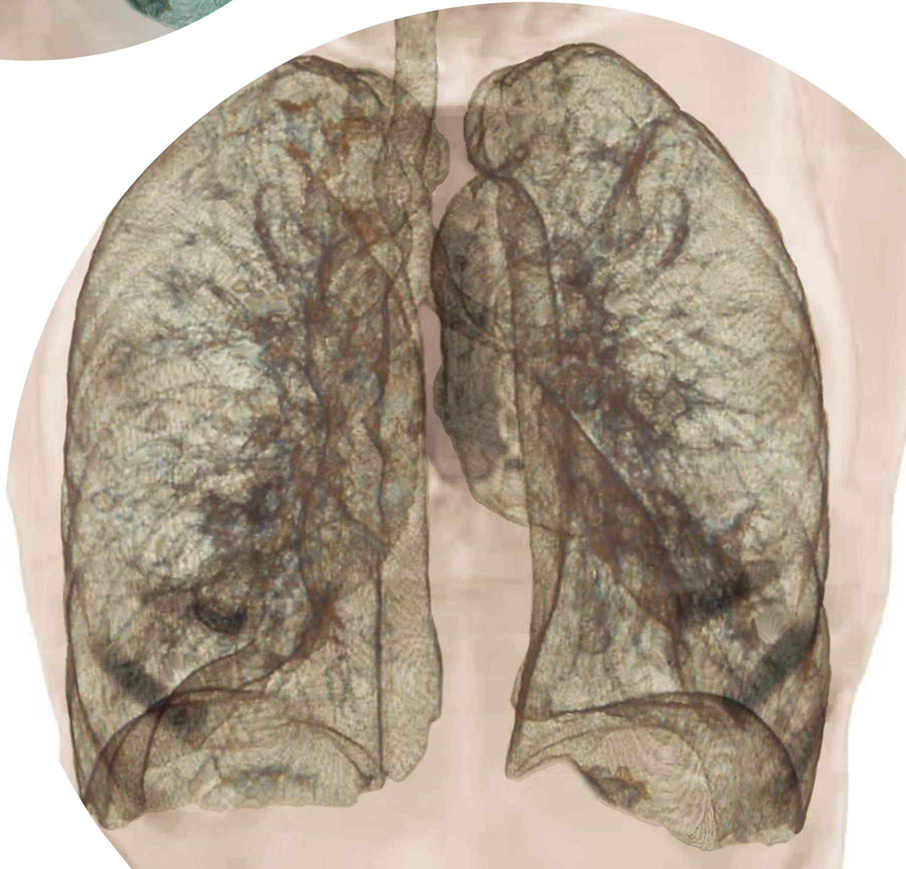
And still today

100s of chest scan COVID classifiers found unreliable

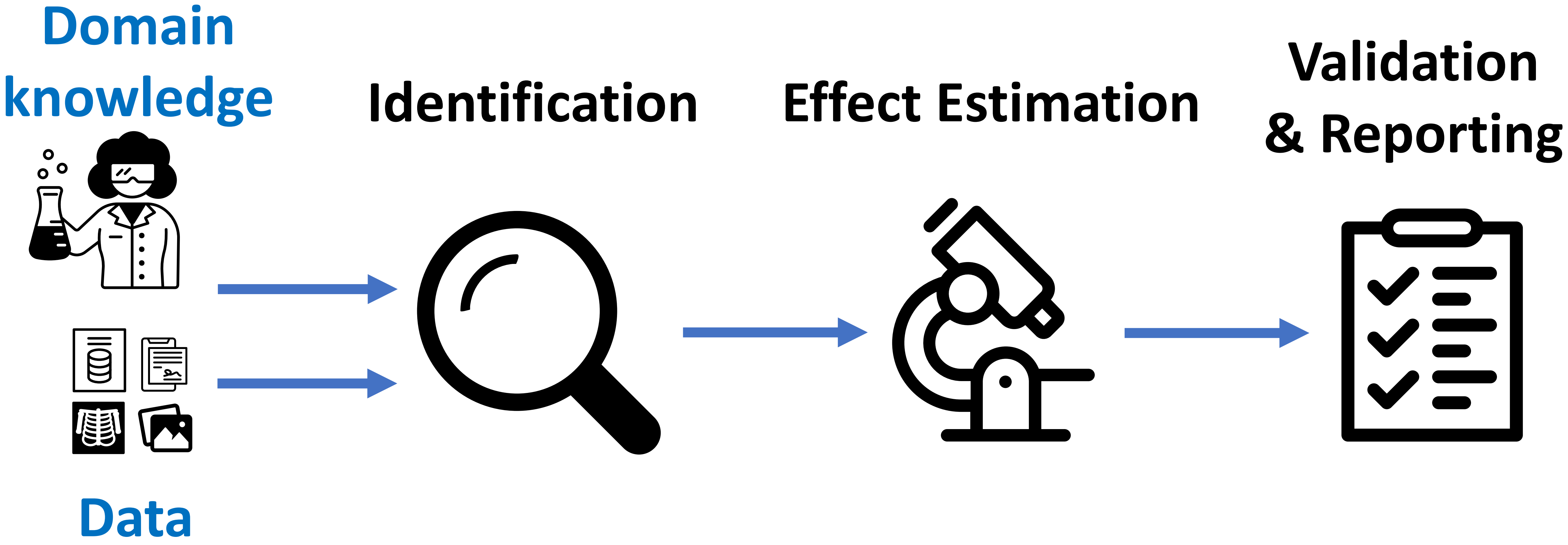
- Identified false correlates
- e.g., sitting vs lying down; pediatric scans

[Roberts et al. NMI 2021, Wynants et al. BMJ 2020]





Getting domain knowledge right is difficult

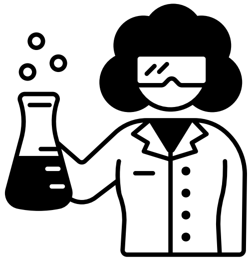


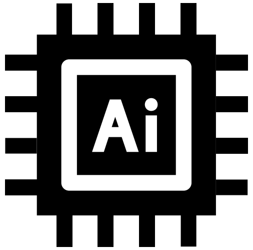
[2011.04216] DoWhy: An End-to-End Library for Causal Inference (arxiv.org)

[2305.06850] A Causal Roadmap for Generating High-Quality Real-World Evidence (arxiv.org)

This talk:
LLM reduces burden on human domain expert

Domain knowledge





Part I: LLMs and causal relationships

Part II: How else LLMs can help the end-to-end process

Part III: LLMs and causal reasoning in text

Wrapping up

Part I:
LLMs and causal relationships

Causal assumptions

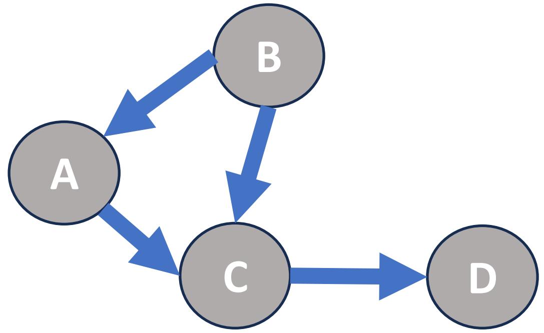
Pairwise relationships

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

Full graph recovery

Given a set of variables infer a directed acyclic graph over them.

- Infer which pairs of variables form an edge, and their direction.



Applying LLMs to pairwise causal recovery

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

Two prompts per pair

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

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

Single prompt

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

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

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

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

Tübingen Benchmark for Pairwise relationships

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

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

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

Results: LLMs recall 97% correctly

Model	Acc.	Wt. Acc.
Slope (Marx & Vreeken, 2017)	0.75	0.83
bQCD (Tagasovska et al., 2020)	0.68	0.75
PNL-MLP (Zhang & Hyvarinen, 2012)	0.75	0.73
Mosaic (Wu & Fukumizu, 2020)	83.3	81.5
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)	**0.96**	**0.97**

Data-driven causal discovery

Knowledge-based causal recovery is competitive with *or much better than* data-driven approaches

Similar results on a neuropathic pain dataset

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

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

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

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
gp4-4	78.4
gpt-4 (neuropathic pain expert)	84.3
text-davinci-003 (single prompt)	86.0
gpt-3.5-turbo (single prompt)	85.5
gpt-4 (single prompt)	**96.2**

?

Ozone $\longleftrightarrow$ Radiation

Ozone $\rightarrow$ Radiation

Radon $\rightarrow$ Ozone

Plausible answer; benchmark says is wrong

Input Pair: L5 Radiculopathy $\overset{?}{\leftrightarrow}$ Obesity

- Option A: Left L5 Radiculopathy causes Left Obesity. This relationship does not make sense because L5 radiculopathy is a nerve condition that affects the leg, while obesity is a whole-body condition related to excessive body fat. It is unlikely that a leg nerve condition would cause obesity.
- Option B: Left Obesity causes Left L5 Radiculopathy. This relationship is more plausible because obesity can put extra pressure on the spine and nerves, leading to nerve compression or irritation. It is possible that obesity could cause L5 radiculopathy.

Step 3: Choose the more likely cause-and-effect relationship
- Based on the analysis, Option B is more likely: Left Obesity causes Left L5 Radiculopathy.

<Answer>B</Answer>

Benchmark Answer:
L5 Radiculopathy causes Obesity.

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

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

Construct Validity: Is Benchmark Memorized?

SYSTEM

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

USER

pair0005 Age

GPT

pair0005 Age Length Abalone ->

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

Yes, Tübingen dataset clearly in the training dataset.

Construct Validity: What are we Measuring?

Let us model knowledge-based discovery as:

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

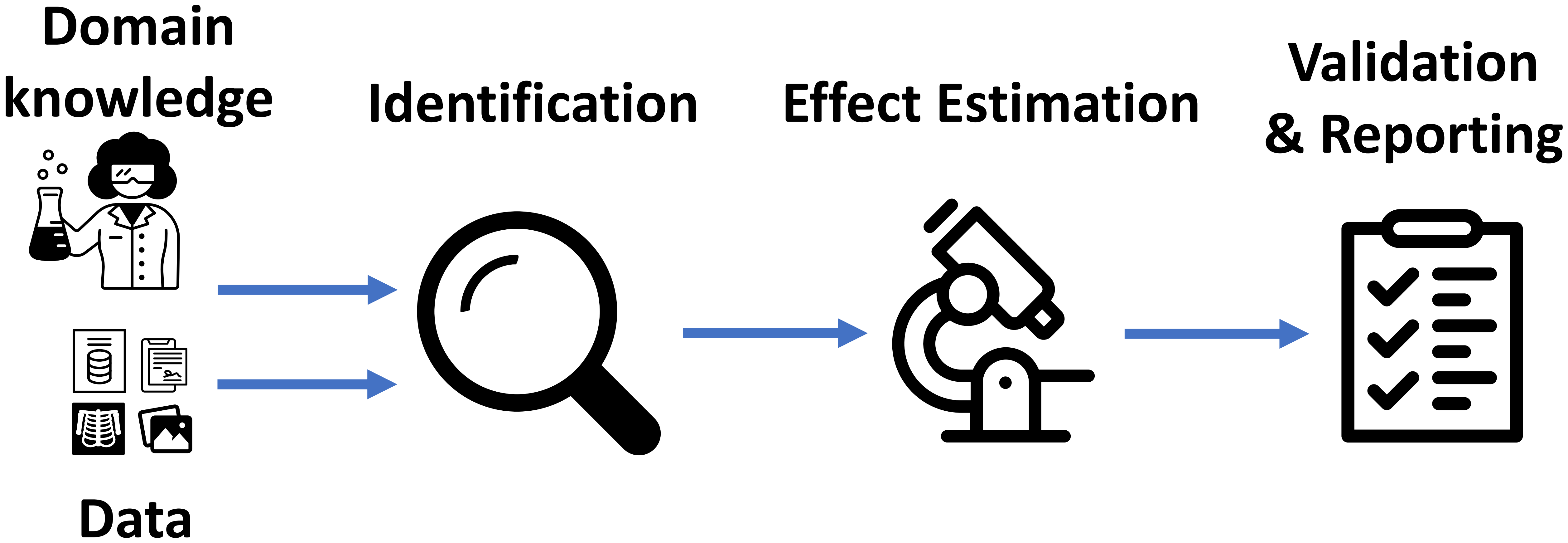
- With memorized benchmark data, we are *not* measuring $P(D)$
- We *are* measuring: how LLM can process and transform D into the necessary causal relationship Y

Part I: Takeaways

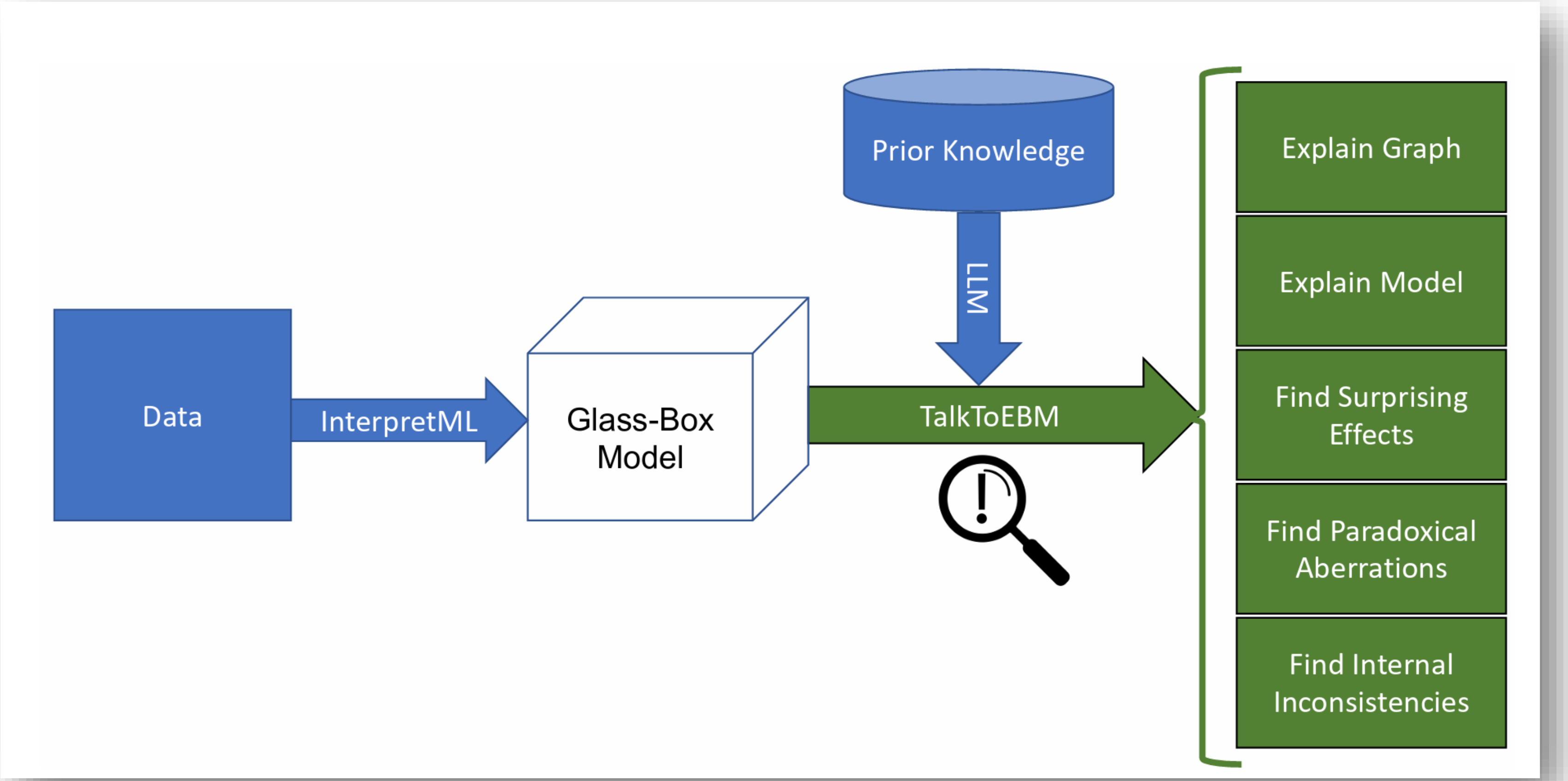
- LLMs enable knowledge-based causal discovery or recovery
  - Strong performance for pairwise causal relationships
  - Across multiple datasets in varied domains incl. medicine and climate science
  - Full graph recovery poses additional challenges (e.g., direct/indirect causes)

- GPT-3.5, GPT-4 have memorized Tübingen benchmark
  - Valid measurement of LLM *transformation of knowledge into a causal answer*
  - *Does not* give likelihood that arbitrary relationship has been memorized

Part II: How else might LLMs help?



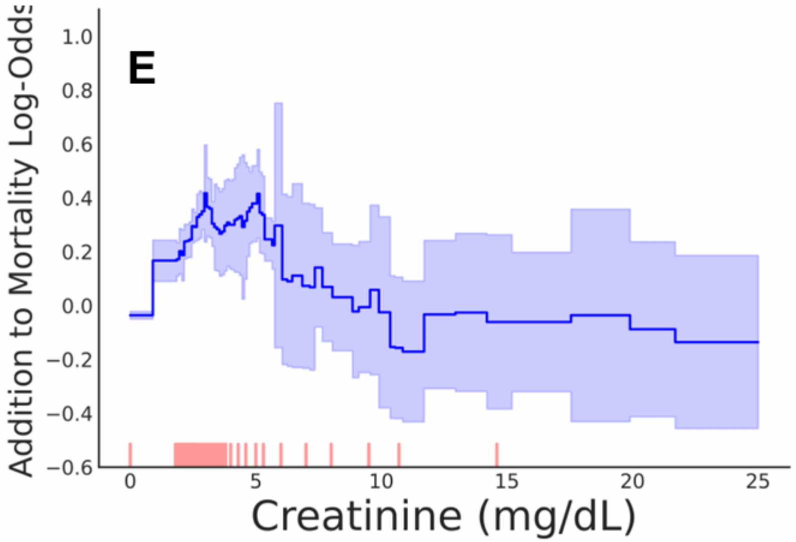
Interpreting Data with Talk2EBM



[[2308.01157] LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs](https://arxiv.org/abs/2308.01157)

Lengerich, Bordt, Nori, Nunnally, Aphiyanaphongs, Kellis, Caruana

Interpreting Data with Talk2EBM

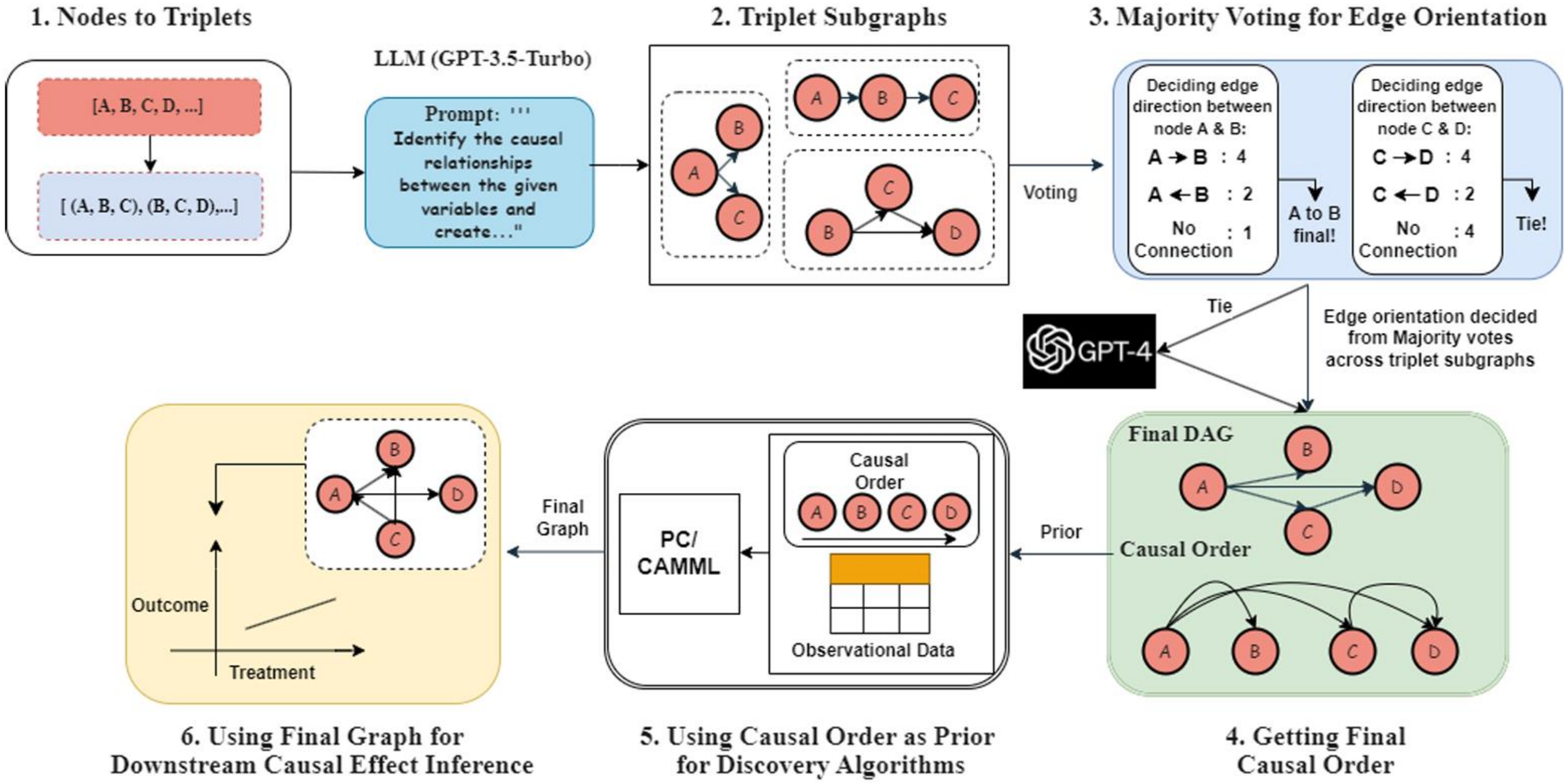


"The graph for creatinine level shows a surprising decrease in the predicted log-odds of death from pneumonia as creatinine levels increase beyond approximately 5.0 and then again after about 10.4, which is counterintuitive given that higher creatinine levels usually indicate worse health outcomes. This could potentially be due to a skewed distribution of data with fewer patients having very high creatinine levels, causing the model to learn inaccurate patterns. Alternatively, it might reflect that patients with extremely high creatinine levels receive more aggressive or specialized care, improving their outcomes. The wide confidence intervals in these ranges also suggest a high level of uncertainty in the model's predictions. [...]"

[2308.01157] LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs

Lengerich, Bordt, Nori, Nunnally, Aphiyanaphongs, Kellis, Caruana

Identification: LLMs give causal order for backdoor identification



[2310.15117] Causal Inference Using LLM-Guided Discovery (arxiv.org)

Vashishtha, Reddy, Kumar, Bachu, Balasubramanian, Sharma

Effect Estimation
LLMs aid code generation

5. Use DoWhy to create a causal model:

```
# Create a causal model using the learned graph
causal_model = CausalModel(
    data=data,
    treatment='X1',
    outcome='Y',
    graph=graph.graph_.to_string()
)

# Identify causal effect
identified_estimand = causal_model.identify_effect()
```

6. Estimate the treatment effect using EconML:

```
# Train a T-learner using linear regression
t_learner = TLearner(LinearRegression(), LinearRegression())
t_learner.fit(Y=data['Y'], T=data['X1'], X=data[['X2']])

# Estimate the treatment effect
treatment_effect = t_learner.const_marginal_effect(data[['X2']])
```

7. Validate the results using sensitivity analyses:

```
# Perform sensitivity analysis using DoWhy
refutation_results = causal_model.refute_estimate(
    identified_estimand,
    treatment_effect,
    method_name='random_common_cause'
)

print(refutation_results)
```

Validation
LLMs suggest negative controls

USER: In a causal observational study of vaccine efficacy, we are matching people who have voluntarily gotten a vaccine with individuals who have not yet done so. We are matching on attributes intended to capture all confounding effects. To help validate our analysis, what are some examples of negative controls where we might expect to see zero treatment effect?

GPT first misunderstands and gives examples of factors that not caused by the treatment:

- hair color, blood type, …

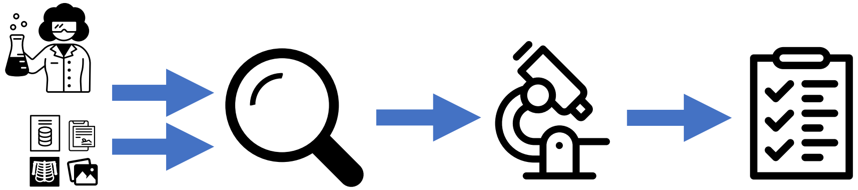
USER: In what subpopulations would we expect to see zero treatment effect on disease prevention? (for the disease being vaccinated against)

GPT then gives reasonable answers:

- Individuals with pre-existing immunity
- Individuals with specific immune deficiencies
- Nonresponders
- Individuals vaccinated post-infection

If we add that we have longitudinal infection data GPT identifies time-bound negative controls

- Pre-vaccination period
- Short time window post-vaccination



Adding to the Open Source Ecosystem for Causality

PyWhy-LLM

Python library for using LLMs in causal analysis process

Integrates with PyWhy libraries (DoWhy, EconML, …)

Work-in-progress

Latent confounders

In [9]:

```
variables = ["ice cream sales", "temperature", "cavities"]
latents = modeler.suggest_confounders(variables, treatment="ice cream sales", outcome = "shark attacks")

print(latents)
```

system — You are a helpful assistant for causal reasoning.

user — What latent confounding factors might influence the relationship between ice cream sales and shark attacks?

We have already considered the following factors ['ice cream sales', 'temperature', 'cavities']. Please do not repeat them.

List the confounding factors between ice cream sales and shark attacks enclosing the name of each factor in <conf> </conf> tags.

assistant — <conf>Beach Attendance</conf>, <conf>Season of the Year</conf>, <conf>Water Temperature</conf>, <conf>Public Holidays</conf>, <conf>Availability of Ice Cream Vendors</conf>, <conf>Shark Population</conf>, <conf>Swimming Conditions</conf>, <conf>Tourist Season</conf>.

```
['Beach Attendance', 'Season of the Year', 'Water Temperature', 'Public Holidays', 'Availability of Ice Cream Vendors', 'Shark Population', 'Swimming Conditions', 'Tourist Season']
```

https://pywhy.org/

https://github.com/py-why/pywhy-llm/

PART III:
Causal reasoning over text

(LLMs and event or actual causality)

Event/Actual Causality and Causal Judgments

Type Causality

Inference over a (sub)population

Actual Causality

Inference over a single event

- Bob has lung cancer and smokes.
  Did Bob's smoking cause his cancer?

- A customer saw a newspaper ad and bought toys.
  What would have happened if they hadn't seen the ad?

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

Causal context is hard to formalize

- **Causal frame:** Factors relevant to causal question
- **Necessary causality:** Did cause need to happen for outcome to occur?
- **Sufficient causality:** Is cause alone enough for outcome to occur?
- **Normality:** Do events line up statistical/social/... norms?
- **Other human factors:** bias towards action, intention, epistemic, …

Necessary and sufficient causes

- **Necessary causality:**
  If an event C does not occur, then the outcome event, E, will not occur.
- **Sufficient causality:**
  If an event C occurs then the outcome event, E, will occur.
- **Robust sufficient causality:**
  … even if other contributing factors did not occur.

Sufficiency is harder, because we have to determine causal frame

Evaluation Vignettes

Kueffner (2021)

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

Novel vignettes

Vignette Type	Input Context	Event	Actor	Nec.	Suff.
Overdetermination	There is a fire in the chemistry lab. A can of water would douse the fire. Agents X and Y both spray a can of water each, dousing the fire.	fire being doused	Agent X	No	Yes
Switch	Reagent X is added to a mixture, which leads to an explosion and kills Sam. Otherwise, Reagent Y in Sam's pocket would have infected him and killed him as well.	Sam's death	Reagent X	No	Yes
Late preemption	Any of Reagent X or Reagent Y can be added to a mixture to convert it into a crystal. Reagent X is added first and the mixture turns to crystal. Reagent Y is added later and but does not mix since the crystal is already formed.	crystal formation	Reagent X	No	Yes

Necessary and Sufficient - Results

Vignette Type	Necessary	Sufficient
gpt-3.5-turbo		
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%
gpt-4		
Overdetermination	✓, ✓	✓, ✓
Switch	✓, ✓	✓, ✓
Late preemption	✓	✓
Early preemption	✓, ✓,✓	✓,
Double preemption	✓	X
Bogus preemption	✓	✓
Short circuit	X	X
Miscellaneous	✓,X, ✓,✓	✓,✓, ✓,✓
Total Accuracy	**86.6%**	**86.6%**

Vignette Type	Necessary	Sufficient
gpt-3.5-turbo		
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%
gpt-4		
Overdetermination	✓, ✓	✓, ✓
Switch	✓, ✓	X, ✓
Late preemption	✓	✓
Early preemption	✓, ✓	X
Double preemption	✓	✓
Bogus preemption	✓	✓
Short circuit	✓	✓
Miscellaneous	✓,X, ✓,✓	✓,✓, X,✓
Total Accuracy	**92.8%**	**78.5%**

GPT-3.5

GPT-4

Takeaways on text-based reasoning

GPT-4 understands scenarios, identifies necessity and sufficiency

- Not possible before

See paper for additional experiments

- Counterfactual reasoning: GPT-4: 92.44% accuracy
  - Only 6% below human baseline
- Normality: TL;DR: 70% accuracy with GPT-4

Wrapping Up

What's new with causality now

LLMs provide…

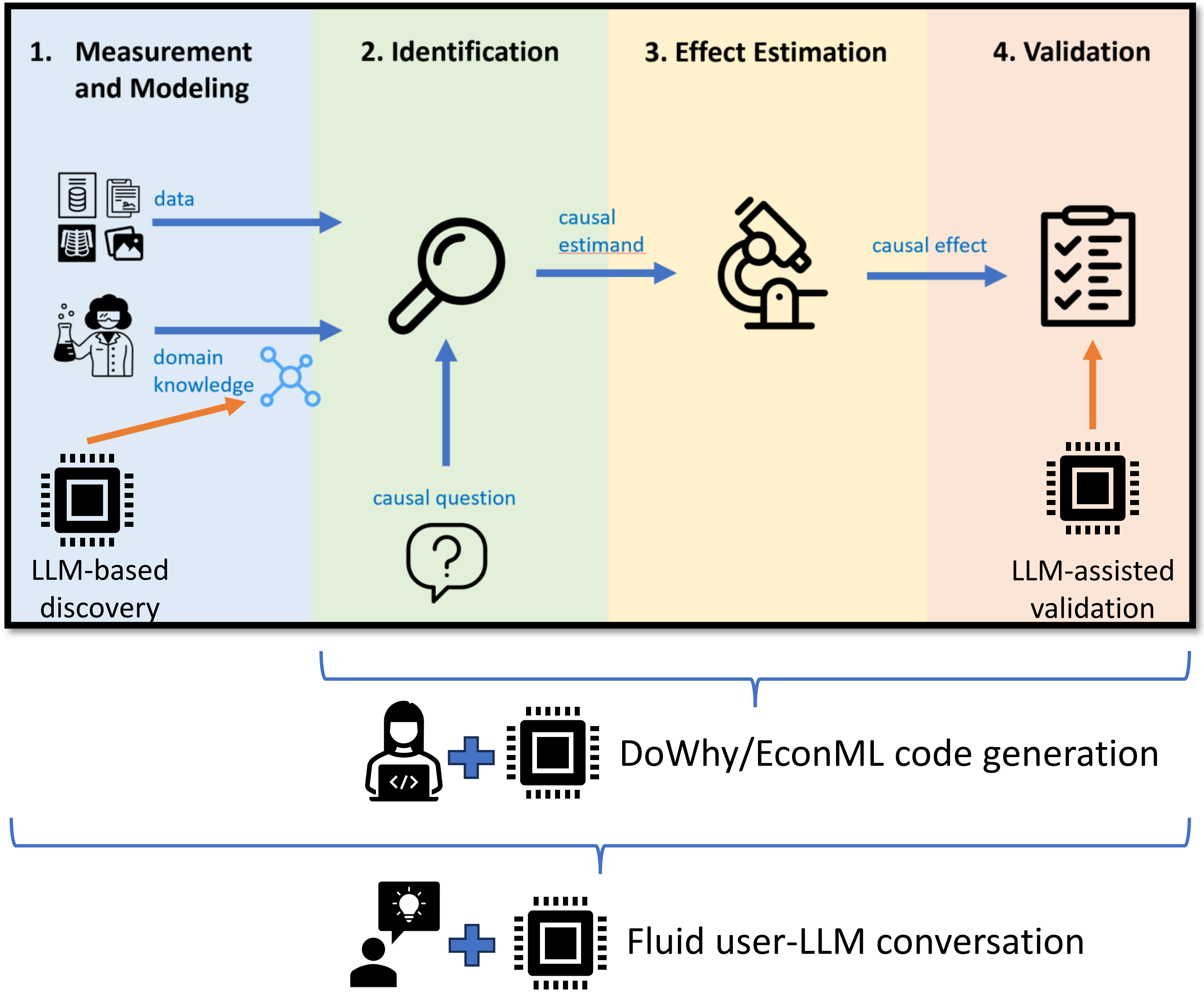
- Domain knowledge only available via human experts before
  - Provide when explicitly asked
  - Also implicitly, e.g., in background knowledge for actual causality vignettes.
- New capability to extract key primitives of text-based reasoning
  - (necessity, sufficiency, normality, etc.)
  - Possibility of system to analyze actual causality for practical scenarios

What’s not changing with causality

- Need for rigorous, well-documented, and verifiable analyses
    - Especially for high-risk and high-value tasks
    - Must ensure correctness for decision making

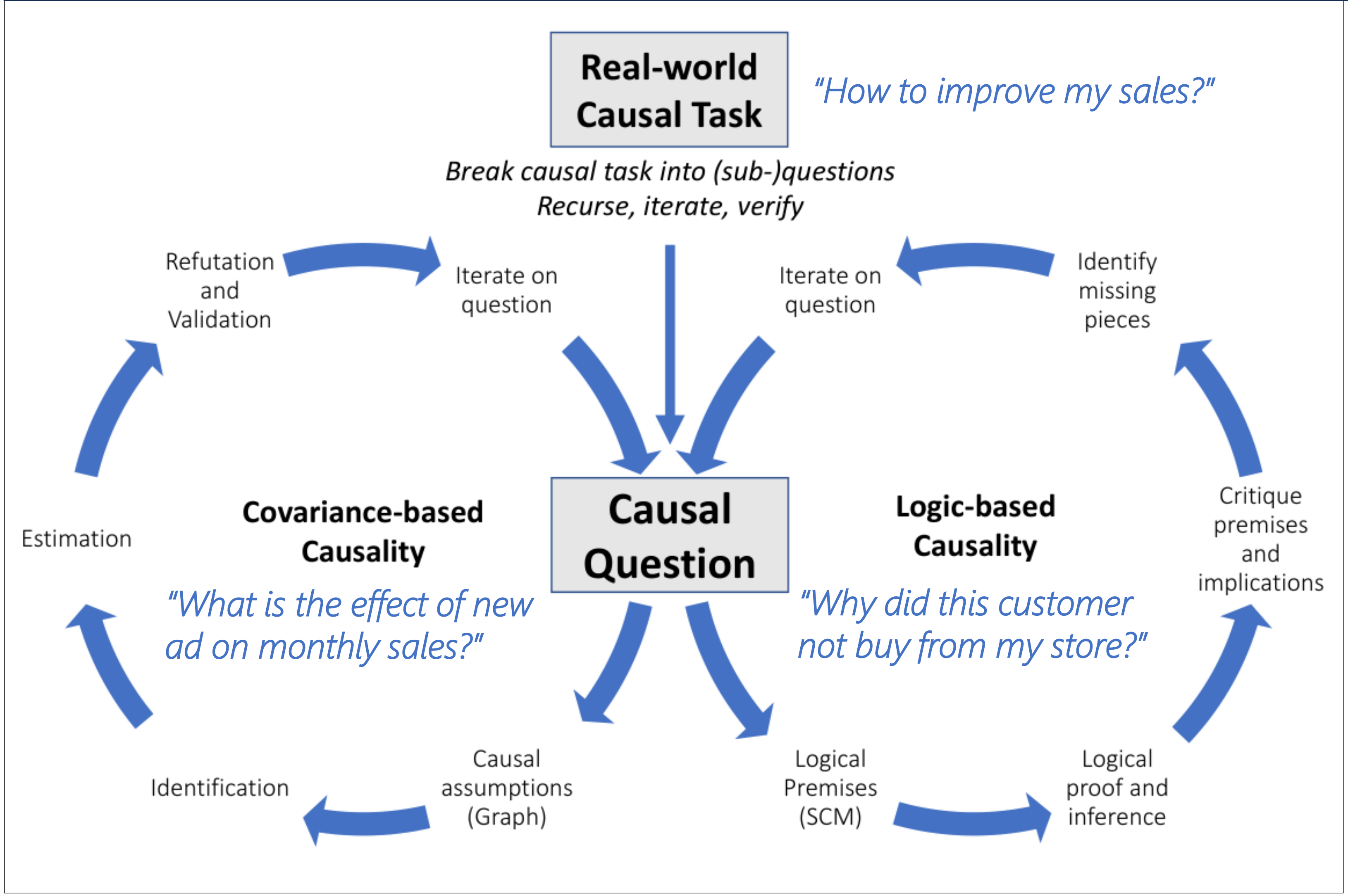
Implications for Practitioners

- Augmenting human expertise with LLMs
  - Assisting in graph creation, validation, and robustness checks
  - Case study: LLM-assisted identification of negative controls
- LLMs can enable end-to-end causal tools
  - Case study: asking LLM to generate DoWhy and EconML analysis code
- LLM as a fluid conversational interface



Many Kinds of Causality

Different tasks: Graph Discovery, Effect inference, Attribution, Prediction



Type Causality

Inference over a (sub)population

Actual Causality

Inference over a single event

Conclusion: A New Frontier for Causality

- Human domain knowledge critical for causal analysis
- LLMs mimic this capability
  - Building causal assumptions and arguments, counterfactual inference, and systematization of necessity, sufficiency, …
- Implications for practice:
  Reduce burden on human domain expert
- New research questions:
  Combining data-driven and text-based analysis?

Questions?

Emre Kıcıman, emrek@microsoft.com

https://pywhy.org/
https://arxiv.org/abs/2305.00050