CS372 AGI Winter 2026
Lecture 2: Causal Reasoning
Why Causality Requires More Than Data

Edward Y. Chang
Computer Science, Stanford University

January 7,2026

Lecture Outline

1. **Simpson's Paradox** — Why correlation $\neq$ causation
2. **Causal Models** — DAGs, confounders, and what we need
3. **Pearl's Ladder** — Three levels of causal reasoning
4. **LLM Benchmarks** — Impressive results, hidden problems
5. **Failure Patterns** — Four ways LLMs fail at causality

Lecture Outline

1. **Simpson's Paradox** — Why correlation $\neq$ causation
2. Causal Models — DAGs, confounders, and what we need
3. Pearl's Ladder — Three levels of causal reasoning
4. LLM Benchmarks — Impressive results, hidden problems
5. Failure Patterns — Four ways LLMs fail at causality

The Paradox: Which Job-Training Program is Better?

Overall Employment Rates:

	Program A	**Program B**
Overall Employment Rate	40%	50%

Obvious conclusion: Program B is more effective, better

The Paradox: Which Job-Training Program is Better?

Overall Employment Rates:

	Program A	**Program B**
Overall Employment Rate	40%	50%

Obvious conclusion: Program B is more effective, better

But wait... Breaking down by experience level:

	Program A	**Program B**	**Total**
Experienced	80% employed (n=200)	70% employed (n=600)	800
Entry-level	30% employed (n=800)	20% employed (n=400)	1200
Participants	1000	1000	2000

Program A is better in EVERY subgroup, yet worse overall.
How is this possible?

The Hidden Numbers: Unequal Group Sizes

	Program A	Program B	Total
Experienced	160/200 (80%)	420/600 (70%)	800
Entry-level	240/800 (30%)	80/400 (20%)	1200
Total Employed	400/1000 (40%)	500/1000 (50%)	

The asymmetry:

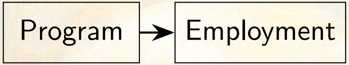
- Program A was given mostly to *entry-level* participants (800 of 1000)
- Program B was given mostly to *experienced* participants (600 of 1000)

The aggregated data mixes apples and oranges

Experience level acts as a **confounding variable** — it influences both program assignment and outcome.

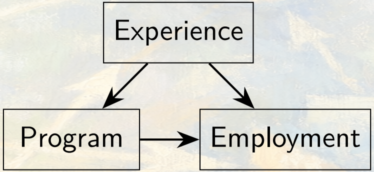
The Causal Graph Explains Everything

Wrong mental model:



If this were true, B would be better

Correct causal structure:



Experience is a confounder:

- Experience → Program (assignment policy routes entry-level to A)
- Experience → Employment (entry-level harder to employ)

The causal question: If we assign Program A vs B to the *same mix* of experience levels, which yields higher employment?
→ The subgroup analysis (controlling for experience)

The Lesson for AGI

The "correct" answer depends on the causal question:

Question	Correct Analysis
"Which program *was associated with* better outcomes?"	Aggregate (B looks better)
"Which *causes* better outcomes?"	Stratified (A is better)

Why LLMs fail here:

- Both are correct for different questions, but only one is stable under intervention
- The data alone cannot tell you which to use
- Need *causal model* to decide whether to condition on Experience

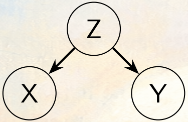
Key insight: Causal reasoning requires structure beyond the data

Lecture Outline

1. Simpson's Paradox — Why correlation $\neq$ causation
2. **Causal Models** — DAGs, confounders, and what we need
3. Pearl's Ladder — Three levels of causal reasoning
4. LLM Benchmarks — Impressive results, hidden problems
5. Failure Patterns — Four ways LLMs fail at causality

Key Causal Definitions

Confounder

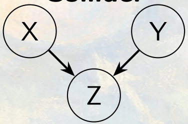


"Mixes things up"
Common cause of X and Y

Mediator

"Go-between"
On the causal pathway

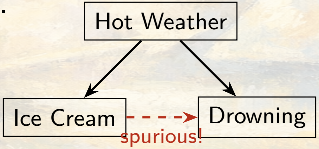
Collider



"Arrows collide into Z"
Common effect of X and Y

DAG (Directed Acyclic Graph): Arrows show cause → effect; no cycles allowed

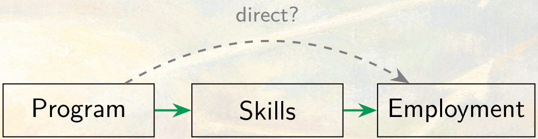
Backdoor Path: Any path from X to Y that begins with an arrow *into* X. If open, induces non-causal association.



Backdoor path (Ice Cream ← Hot Weather → Drowning) *looks like* causation

Mediator Example: How Programs Improve Employment

Question: How does a job training program lead to employment?



Skills as Mediator: The program's effect flows *through* skill development.

Examples of mediating skills:

- Technical certifications
- Networking / social capital
- Interview preparation
- Resume building

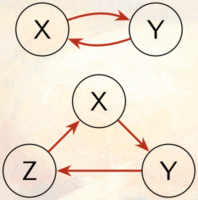
Why mediators matter:

- Do NOT adjust for mediators when estimating total effect
- Adjusting blocks the causal pathway!
- Mediators explain *how* causes work

Confounder: adjust to remove bias. Mediator: don't adjust (blocks the effect).

DAG Constraints: What's Forbidden and How to Handle It

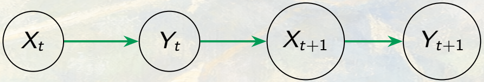
Forbidden: Cycles



Violates "Acyclic" in DAG

Real feedback loops exist (e.g., poverty $\leftrightarrow$ health)

Solution: Time-Unrolling



Feedback unrolled over time — now acyclic!

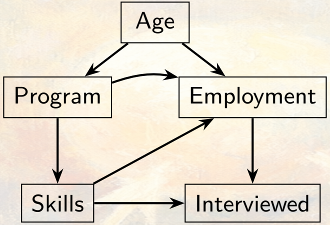
Other approaches:

- Equilibrium / steady-state models
- Dynamic Bayesian networks
- Structural equation models (SEM)

DAGs are limited but powerful. For cycles: add time or use specialized models.

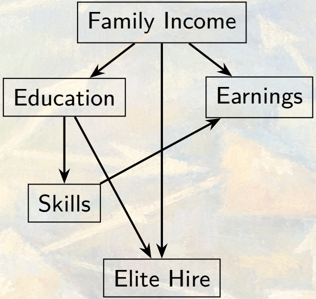
Complex DAGs: Mixing Confounders, Mediators, and Colliders

Example 1: Job Training



- Age: **Confounder**
- Skills: **Mediator**
- Interviewed: **Collider**

Example 2: Education & Earnings



- Family Income: **Confounder**
- Skills: **Mediator**
- Elite Hire: **Collider**

What “Models” Do We Need?

These are layers, not alternatives:

Domain Knowledge

↓ informs

DAG (qualitative: X $\rightarrow$ Y)

↓ identifies

Confounders List

↓ add functions

Structural Equations (Y $=$ f(X, ϵ))

↓ combines into

Full Structural Causal Model (SCM)

$\leftarrow$ **Level 2** (Intervention)
Sufficient for $P(Y|do(X))$

$\leftarrow$ **Level 3** (Counterfactual)
Required for “what if I had...”

Why We Focus on DAG

Domain Knowledge

↓

DAG

↓

... Structural Equations ... SCM

DAG is the foundational gap:

- LLMs are trained on observational prediction — no interventional semantics
- Without DAG, cannot distinguish confounder / mediator / collider
- All downstream steps depend on having the DAG first

What each level requires:

- **Level 2** (Intervention): DAG is sufficient — do-calculus works on the graph
- **Level 3** (Counterfactual): Need full SCM with structural equations

This course: How can LLMs help construct DAGs? Where do they fail?

Can LLMs Help Construct Causal Graphs?

What LLMs can do:

- Suggest candidate variables from domain knowledge
- Propose edges based on known relationships in training data
- Generate hypotheses: "Age might confound the relationship"

Can LLMs Help Construct Causal Graphs?

What LLMs can do:

- Suggest candidate variables from domain knowledge
- Propose edges based on known relationships in training data
- Generate hypotheses: "Age might confound the relationship"

What LLMs cannot do:

- Validate causal direction from data alone: requires assumptions or interventions
- Guarantee complete confounders: may propose plausible ones, but not exhaustive
- Distinguish correlation from causation: fundamentally observational training

Can LLMs Help Construct Causal Graphs?

What LLMs can do:

- Suggest candidate variables from domain knowledge
- Propose edges based on known relationships in training data
- Generate hypotheses: "Age might confound the relationship"

What LLMs cannot do:

- Validate causal direction from data alone: requires assumptions or interventions
- Guarantee complete confounders: may propose plausible ones, but not exhaustive
- Distinguish correlation from causation: fundamentally observational training

A hybrid approach (roadmap preview):

1. **LLM:** Generate candidate variables and initial graph structure
2. **Causal Discovery Algorithms:** Test conditional independencies in data
3. **Human Expert**: Validate, add domain constraints, resolve conflicts
4. **Do-Calculus:** Compute $P(Y|do(X))$ from final graph (covered next segment)

Real-World Example: UC Berkeley Admissions (1973)

Aggregate data suggested gender discrimination:

	Admission Rate
Male Applicants	44%
Female Applicants	35%

Real-World Example: UC Berkeley Admissions (1973)

Aggregate data suggested gender discrimination:

	Admission Rate
Male Applicants	44%
Female Applicants	35%

But department-level analysis revealed:

- Women had *equal or higher* admission rates in most departments
- Women disproportionately applied to highly competitive majors
- Men disproportionately applied to less competitive departments

The confounder: Department choice

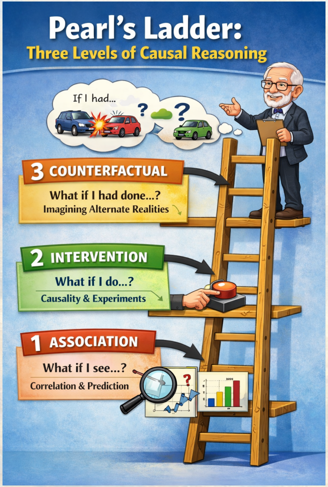
“Department”, a confounder affects both acceptance probability and application distribution.

Same paradox, different domain — the pattern is universal

Lecture Outline

1. Simpson's Paradox — Why correlation $\neq$ causation
2. Causal Models — DAGs, confounders, and what we need
3. **Pearl's Ladder** — Three levels of causal reasoning
4. LLM Benchmarks — Impressive results, hidden problems
5. Failure Patterns — Four ways LLMs fail at causality

Pearl's Ladder of Causation



Three Levels:

Level 3: Counterfactual
$P(y_x|x', y')$ — "What if I had...?"
Unsupported without SCM

Level 2: Intervention
$P(Y|do(X))$ — "What if I do...?"
Unsupported without causal graph + assumptions

Level 1: Association
$P(Y|X)$ — "What if I see...?"
✓ LLMs

Pearl & Mackenzie (2018): "The Book of Why"

Why the Levels are Provably Distinct

Causal Hierarchy (Bareinboim et al., 2020):
Level-1 observational distributions cannot identify Level-2 interventional or Level-3 counterfactual quantities without additional assumptions or experimental information. CS372 will cover techniques that make such assumptions explicit and testable.

What observation tells you:

- Among those who *ended up* in Program A, some were employed
- $P(Y{=}1 \mid T{=}\text{A}) = 0.40$

What observation cannot tell you:

- If we *assign* participants to Program A, what is $P(Y{=}1 \mid do(T{=}\text{A}))$?
- For a specific person who took A and was employed, would Y have been 1 under $do(T{=}\text{None})$?

Why the Levels are Provably Distinct

Causal Hierarchy (Bareinboim et al., 2020):
Level-1 observational distributions cannot identify Level-2 interventional or Level-3 counterfactual quantities without additional assumptions or experimental information. CS372 will cover techniques that make such assumptions explicit and testable.

What observation tells you:

- Among those who *ended up* in Program A, some were employed
- $P(Y{=}1 \mid T{=}\mathrm{A}) = 0.40$

What observation cannot tell you:

- If we *assign* participants to Program A, what is $P(Y{=}1 \mid do(T{=}\mathrm{A}))$?
- For a specific person who took A and was employed, would Y have been 1 under $do(T{=}\mathrm{None})$?

Why? In observational data, groups differ due to selection/confounding (e.g., experience level), so in general $P(Y \mid T) \neq P(Y \mid do(T))$.

Notation: T is the program variable; we reserve $P(\cdot)$ for probability.

What Each Level Requires

Level	Name	Question Type	Requires	LLM?
1	Association	$P(Y\|X)$	Observational data	✓
2	Intervention	$P(Y\|do(X))$	Causal graph + adjustment	limited
3	Counterfactual	$P(y_x\|x', y')$	Full structural model	×

Examples:

Level	Example Question
1	"Do participants in Program A tend to get employed?"
2	"If we assign this person to Program A, what is their employment probability?"
3	"This person did Program A and got employed. Would they have without it?"

Key insight: LLMs trained on text learn Level-1 patterns. Levels 2-3 require *reasoning* about causal structure — not pattern retrieval.

Two Different Questions

Simpson's Paradox gives contradictory answers because it mixes two questions.

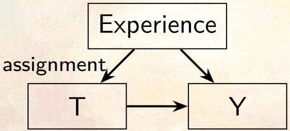
Question	Quantity
Among those who enrolled in A, what fraction were employed?	$P(Y \mid T{=}\mathrm{A})$
If we assign participants to A, what employment rate would result?	$P(Y \mid do(T{=}\mathrm{A}))$

- **After-the-fact (observational):** describes what we observed in the data, $P(Y \mid T)$.
- **Policy decision (interventional):** predicts what would happen under an assignment policy, $P(Y \mid do(T))$. This requires extra work: state assumptions, identify and adjust for confounders, and produce a defensible analysis for stakeholders or regulators.

Decision-making needs $P(Y \mid do(T))$, not just $P(Y \mid T)$.

Why $P(Y \mid T = \text{A})$ Is Confounded

The causal structure:

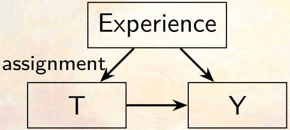


What happened:

- Program A assigned mostly to entry-level
- Program B assigned mostly to experienced

Why $P(Y \mid T = \text{A})$ Is Confounded

The causal structure:



What happened:

- Program A assigned mostly to entry-level
- Program B assigned mostly to experienced

$P(Y \mid T = \textbf{A})$ **mixes together:**

1. Effect of program on employment
2. Effect of experience on program assignment
3. Effect of experience on employment

The confounding path:

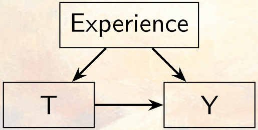
$$T \leftarrow \text{Experience} \rightarrow Y$$

This "backdoor path" creates spurious association!

The do-Operator: Graph Surgery

"$do(T = A)$" = Assign everyone to Program A, **regardless of experience**

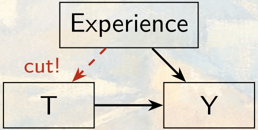
Before: Observational



Experience determines program
$P(Y|T)$ is confounded

$\Rightarrow$
surgery

After: $do(T = A)$



T is **set by intervention** (not by Experience)
Backdoor from Experience is cut

Key insight: Cutting the incoming arrow removes confounding from Experience because assignment is no longer determined by Experience.

Cut the Backdoor Path (Design)

Backdoor in our running example: $T \leftarrow E \rightarrow Y$ (E affects both T and Y)

Cut by design: change how T is assigned

- **Goal:** make T independent of E by construction
- **Methods:**
  - Randomized assignment
  - Stratified randomization (randomize within each E level)
  - Rerandomization until balance criteria are met
- **Interpretation:** closest practical version of $do(T{=}\cdot)$

Note: even with randomization, you may see 55/45 splits due to chance. That is not confounding; it is finite-sample imbalance.

One-liner: Confounding is prevented by *how we assign* T.

Cut is a property of the assignment mechanism.

Block the Backdoor Path (Analysis)

Backdoor in our running example: $T \leftarrow E \rightarrow Y$

Block by analysis: keep data, remove the spurious path

- **Goal:** compare like strata to block spurious T–Y association
- **Methods:**
  - Stratify on E and reweight to target $P(E)$
  - Matching on E (and other confounders)
  - Inverse propensity weighting
  - Regression adjustment with pre-treatment covariates
- **Limitation:** works only for measured confounders (unknown U can still bias)

One-liner: Confounding is handled by *how we adjust* in the analysis.

Block is a property of the estimation strategy given a graph.

The Backdoor Adjustment Formula

Problem: We can't actually assign everyone to Program A.
Solution: Use observed data + graph structure to *compute* $P(Y|do(T))$.

Backdoor Adjustment Formula

$$P(Y|do(T = A)) = \sum_e P(Y|T = A, E = e) \cdot P(E = e)$$

Term	**Meaning**	**Source**
$P(Y \mid T{=}\mathrm{A}, E{=}e)$	Effect within each stratum	Stratified data
$P(E = e)$	Target population distribution	Overall data
$\sum_e$	Weighted average	Marginalization

Intuition: Within each experience level, there's no confounding. So we compute the effect per stratum, then average over the population.

Assumes: no unmeasured confounding, positivity (each stratum has both programs).

Step-by-Step Calculation

Step 1: Stratified effects $P(Employed|T, Experience)$

	Program A	**Program B**
Experienced	$160/200 = \mathbf{0.80}$	$420/600 = \mathbf{0.70}$
Entry-level	$240/800 = \mathbf{0.30}$	$80/400 = \mathbf{0.20}$

Step 2: Target population distribution $P(Experience)$

Experienced	$800/2000$	$= \mathbf{0.40}$
Entry-level	$1200/2000$	$= \mathbf{0.60}$

Step 3: Apply formula

$$P(Y|do(T = A)) = 0.80 \times 0.40 + 0.30 \times 0.60 = 0.32 + 0.18 = \mathbf{0.50}$$
$$P(Y|do(T = B)) = 0.70 \times 0.40 + 0.20 \times 0.60 = 0.28 + 0.12 = \mathbf{0.40}$$

The Verdict: Paradox Resolved

	Program A	**Program B**
Observational $P(Y\|T)$	40%	**50%** $\leftarrow$ looks better
Causal $P(Y\|do(T))$	**50%** $\leftarrow$ actually better	40%

Observational (confounded):

- B looks better (50% ¿ 40%)
- But B participants were mostly experienced!

Causal (adjusted):

- A is truly better (50% ¿ 40%)
- After accounting for experience

For this target population: $P(Y|do(T=A)) > P(Y|do(T=B))$

Why "Target Population" Matters

The formula uses $P(E = e)$ — but **which population's distribution**?

Same program effect per stratum, different target populations:

	% Exp'd	**% Entry**	$P(Y \mid do(T = A))$
Region 1 (tech hub)	90%	10%	$0.80 \times 0.9 + 0.30 \times 0.1 = \mathbf{0.75}$
Region 2 (rural)	10%	90%	$0.80 \times 0.1 + 0.30 \times 0.9 = \mathbf{0.35}$
Overall population	40%	60%	$0.80 \times 0.4 + 0.30 \times 0.6 = \mathbf{0.50}$

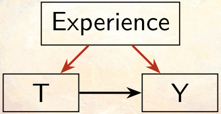
Key insight:

- Same program, same per-stratum effect
- Different overall rates due to different experience mix
- You must specify: "Effect for whom?"

Why Causal Graphs Are Essential (1): What to Adjust For

The Backdoor Criterion (Pearl): Adjust for variables that block all backdoor paths.

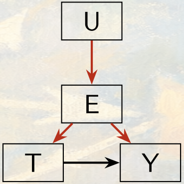
Our example:



Backdoor path: T ← E → Y

Adjust for E ✓

More complex:



Backdoor: T ← E → Y

Adjust for E (blocks path)

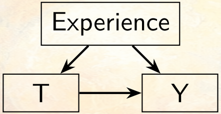
U not needed here — *but what if U affects Y directly?*

Without the graph, how would you know WHAT to adjust for?

Why Causal Graphs Are Essential (2): Confounders vs Colliders

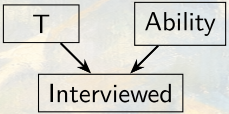
Adjusting for the wrong variable creates bias!

Confounder (adjust = good)



Adjusting **removes** bias
(Blocks backdoor path)

Collider (adjust = bad)



Adjusting **creates** bias
(Opens a path that was closed!)

Collider bias example:

- Both program participation AND high ability → get interviewed
- Among interviewed: spurious negative T–Ability correlation appears
- This is "Berkson's paradox" / selection bias

Data alone cannot tell you which is which — only the causal graph can

Summary: The do-Calculus Recipe

1. **Draw the causal graph**
   From domain knowledge — what causes what?
2. **Identify backdoor paths**
   Paths from T to Y that start with an arrow *into* T
3. **Find adjustment set**
   Variables that block ALL backdoor paths (use backdoor criterion)
4. **Apply the formula** $P(Y|do(T)) = \sum_z P(Y|T, Z = z) \cdot P(Z = z)$
5. **Interpret**
   This is the *causal effect* — what happens if we intervene

---

No causal graph $\Rightarrow$ Don't know what to adjust for $\Rightarrow$ Cannot resolve Simpson's
With causal graph $\Rightarrow$ Backdoor criterion $\Rightarrow$ Principled causal inference

Counterfactuals: The Three-Step Process (Level 3)

Question: "Person did Program A and got employed. Would they have been employed *without* the program?"

This requires reasoning about a **specific individual** in a **hypothetical world**.

Step 1: Abduction Infer U from evidence → **Step 2: Action** Apply $do(T = \text{None})$ → **Step 3: Prediction** Compute Y in new world

Example:

1. **Abduction:** Person did Program A, got employed $\Rightarrow$ infer their latent factors U
2. **Action:** Imagine $do(T = \text{None})$ — surgery on the graph
3. **Prediction:** With their specific U, would they have been employed? Compute $Y_{do(T=\text{None})}$

Requires full Structural Causal Model (SCM) with functional equations, not just DAG

Discussion: Can We Ever Be Certain?

Scenario: Observational study shows $P(\text{Employed}|\text{Program A}) = 0.40$

Question 1: Can you tell a participant: "You have 40% chance of employment if you enroll in Program A"?

A. Yes — the data clearly shows 40%

B. No — this assumes no confounding

C. It depends on the study design

Discussion: Can We Ever Be Certain?

Scenario: Observational study shows $P(\text{Employed}|\text{Program A}) = 0.40$

Question 1: Can you tell a participant: "You have 40% chance of employment if you enroll in Program A"?

A. Yes — the data clearly shows 40%

B. No — this assumes no confounding

C. It depends on the study design

Question 2: If correlation is **100%** (everyone in Program A got employed), can we now be certain the program works?

A. Yes — 100% is definitive proof

B. No — could still be perfect confounding

C. Only if we have a large sample size

Discussion: Can We Ever Be Certain?

Question 3: We adjusted for Experience. What if there's an **unknown confounder U** (e.g., neighborhood GDP, motivation)?

A. No problem — adjusting for Experience is enough

B. Our causal estimate could be completely wrong

C. We can never do causal inference without RCTs

Discussion: Can We Ever Be Certain?

Question 3: We adjusted for Experience. What if there's an **unknown confounder U** (e.g., neighborhood GDP, motivation)?

A. No problem — adjusting for Experience is enough
B. Our causal estimate could be completely wrong
C. We can never do causal inference without RCTs

Question 4: How can we EVER be confident about causation?

A. Randomized Controlled Trials (RCTs) — break ALL confounding by design
B. Explicit causal assumptions + sensitivity analysis
C. Multiple converging lines of evidence
D. **All of the above**

Discussion: Key Takeaways

The uncomfortable truth about unknown confounders:

- We can **never be certain** we've identified all confounders
- Domain expertise helps, but isn't foolproof
- This is why we must **state assumptions explicitly**

What we can do:

- **RCTs** — break ALL confounding (known and unknown) by design
- **Sensitivity analysis** — "How wrong could we be if U exists?"
- **Multiple evidence** — different studies, different potential confounders
- **Negative controls** — test assumptions where we know the answer

Causal inference requires humility: state your assumptions, quantify uncertainty

Lecture Outline

The Good News: LLMs Score High on Several Causal Benchmarks

Kıcıman et al. (TMLR, 2023): strong accuracy on multiple tasks/benchmarks.

Task	**Benchmark / Setup**	**GPT-4**
Pairwise causal discovery	Tübingen Cause-Effect Pairs (bivariate direction)	**97%**
Counterfactual reasoning	CRASS counterfactual query benchmark (physics/logic/common sense)	**92%**
Event causality	15 standard vignettes (necessary vs sufficient cause)	**86%**

- They also report robustness checks and generalization to newer datasets created after training cutoff.
- LLMs can help draft graphs and causal context from natural language.

So... problem solved?

The Bad News: High Scores Can Hide Fragility

Same study, same models: high average accuracy does not imply reliable causal reasoning.

What they observe:

- **Unpredictable failure modes** even on tasks where accuracy is high
- **Over-reliance on text cues:** the paper notes behavior driven by *text metadata*
- **Data neglect:** they explicitly warn that LLMs can *ignore the actual data*

Implication:

- High benchmark scores + occasional sharp failures is consistent with strong pattern completion, not a dependable causal procedure.

CS372 Project: Scaling Causal Benchmarks for AGI Research

Goal: Scale the T3 Causal Benchmark from 454 $\rightarrow$ 5,000+ vignettes for rigorous algorithm evaluation.

Assignment 1 Scale to V1 → **Assignment 2** Peer Review → **Assignment 3** Quality Control → **Final Phase** Algorithm Testing

Why this matters:

- Current benchmarks (454 cases) lack statistical power for NeurIPS-level claims
- Testing RCA, UCCT, and novel algorithms requires diverse, high-quality vignettes
- You will contribute to publishable research infrastructure

T3 Benchmark: 10 categories $\times$ 3 levels (L1 Association, L2 Intervention, L3 Counterfactual)

Assignment 1: Scale T3 Benchmark (V1)

Timeline: 1 week **Deliverable:** V1 of expanded vignettes + quality analysis

Team Structure:

- Teams of 6 student, each team assigned **1 of 10 categories**
- Target: $\sim 10\times$ increase per category

Your Tasks:

1. **Analyze** existing vignettes in your assigned category (difficulty, coverage)
2. **Generate** new vignettes for L1, L2, and L3 levels
3. **Document** quality concerns, ambiguities, or gaps discovered
4. **Deliver** V1 dataset + quality analysis report

What happens next:

- **A2:** Teams swap categories for cross-review, editing, and justification
- **A3:** Quality control phase — reassigned by expertise (your major matters!)
- **Final:** LLM validation runs + test your own reasoning algorithms

Failure Pattern 1: Sensitivity to Wording

The problem: Small wording changes can change outcomes.

Evidence (Kıcıman et al., 2023):

- Redaction probing shows key causal trigger words (e.g., "changing", "causes") strongly affect accuracy.
- Even redacting seemingly minor words can hurt accuracy, suggesting sensitivity to phrasing and grammar.

Why this matters: A causal reasoner should be invariant to paraphrase when meaning is preserved. Here, behavior indicates reliance on surface cues and instruction patterns.

Mapping to Pearl: this looks like Level-1 style sensitivity, not stable Level-2 reasoning.

Failure Pattern 2: Semantic Cues Override Data

The problem: When labels or context carry strong connotations, models may follow semantics rather than the evidence.

Evidence:

- LLMs can pick an answer aligned with label meaning even when the data strongly supports the opposite conclusion.

Why this matters: If the model is not reliably using the dataset and assumptions to reason about confounding, it is not robustly answering an interventional question.

Mapping to Pearl: Level-2 requires reasoning about $P(Y \mid do(T))$, not shortcuts from text semantics.

Failure Pattern 3: No Grounded Intervention Mechanism

The problem: Correct answers on famous confounding examples do not imply a reliable ability to compute interventions.

What we observe in the literature:

- Models can look strong on benchmarks yet still show failures where they rely on non-causal textual signals and can even ignore the underlying data representation.

Core gap: Without an explicit causal model (or a tool that enforces one), the system has no guaranteed way to separate $P(Y \mid X)$ from $P(Y \mid do(X))$.

Pearl: distinguishing association from intervention is exactly the Level-1 vs Level-2 boundary.

Failure Pattern 4: Simple, Unpredictable Mistakes

The problem: Even when average accuracy is high, LLMs can make simple mistakes on specific inputs.

Example (from Kıcıman et al., 2023): necessity vs sufficiency slip

- On necessary/sufficient-cause vignettes, GPT-4 is often correct,
- but on some cases (e.g., the "short circuit" vignette), it applies the wrong principle and fails.

Why this matters:

- The mistake is not a missing fact.
- It is an inconsistency in applying the causal criterion (which principle is relevant?).

Pearl's Ladder link:

- Many vignette tasks are Level 2 to Level 3 flavored,
- Brittleness signals missing or unstable causal control, not just missing knowledge.

Summary: High Scores, Fragile Causal Control

From Kıcıman et al., 2023:

LLMs exhibit unpredictable failure modes, and accuracy depends substantially on the prompt used.

Failure Pattern	Where it shows up	Diagnosis
Brittleness to prompts	Across tasks	Sensitivity to surface form
Misread the data context	Obs vs causal settings	Prior patterns can override the dataset
No explicit intervention engine	Level 2 questions	No guaranteed $do(\cdot)$ computation
Unpredictable logical slips	Level 2/3 vignettes	Unstable application of causal criteria

Takeaway: High benchmark scores do not imply reliable causal reasoning. They can reflect partial, pattern-based competence with brittle control.