Report: Causal Reasoning Case Library (Assignment 1 - Group B)

Subject: Structural Analysis, Failure Modes, and Scaling Rationale Applied to the CS372 Economics Case File **Date:** January 15, 2026 **Contributors:** Atanu Mukherjee, Vivek Sathe, Chris Pearce, Mason Hu

---

1. Introduction

This report analyzes the process and outcomes for the Economics Workstream for Assignment 1 of the Stanford CS372 Paper. The team was tasked with extending the T3 Benchmark in the Economics domain (BenchmarkT3-BucketLarge-B.pdf) from an initial 40 cases to a total of 500, an addition of ~460 cases. Each team member independently .

---

2. Methodology: Pipeline & Epistemic Control

Each team member adopted their own individual pipeline for generating their examples, after which the examples were manually annotated. Below, each member of the team describes their approach;

Atanu Mukherjee

The case file was constructed using a human-in-the-loop pipeline in which large language models are used only as controlled generators, not as sources of causal authority. LLMs are used for: * Drafting candidate scenarios in natural language * Proposing alternative narrative framings * Rewriting cases into a fixed canonical format once the causal structure is finalized LLMs are not used for: * Assigning Pearl levels * Identifying confounders, mediators, or colliders * Resolving causal ambiguity * Validating temporal ordering

All causal labeling is performed by a human. This is evidenced by repeated corrections in the case construction process, including: * Reclassification of cases from MED to CONF-MED * Reintroduction of missing variables such as confounders or colliders * Removal of dangling or causally inactive variables * Rewriting of causal structures to eliminate illegal paths

The pipeline is text-based and format-driven. No automated causal discovery or statistical inference tools are used. The role of the LLM is assistive and bounded; epistemic control remains with the human curator.

Example of analysis A. Structural sanity check Pearl level distribution Across the uploaded case file (cases approximately 5.71 through 6.09), the Pearl level distribution is approximately: * L2 (intervention): about two-thirds of cases *

L3 (counterfactual): about one-quarter of cases * L1 (associational): intentionally minimal This distribution is consistent with the design goal of emphasizing interventional reasoning in economics and policy contexts, while still maintaining a meaningful counterfactual component. Purely associational cases are intentionally underrepresented because they do not strongly differentiate causal reasoning from pattern recognition.

The distribution is stable across domains including labor, trade, energy, public policy, finance, and behavioral economics, suggesting that the balance is structural rather than topic-driven.

Trap type and subtype coverage

Initial causal taxonomies based only on confounding, mediation, or selection were found to be insufficient. The case file reflects a refined taxonomy that separates graph-level causal errors from domain-specific mechanisms. At the graph level, the cases span: Confounding Mediation Collider or selection bias Counterfactual or path-specific dependence In addition, many cases intentionally combine these elements, producing structures such as confounder-mediator and confounder-mediator-collider cases. These combinations are explicitly modeled rather than treated as classification noise.

The refined taxonomy is mutually exclusive at the root level (each case has a primary causal failure mode) and collectively exhaustive for DAG-based causal reasoning errors in applied economics. Composite traps are necessary because many real policy and economic failures arise from layered causal mechanisms rather than a single mistake.

Ambiguity versus determinate cases The case file deliberately includes both determinate and ambiguous cases. Determinate cases are those where causal attribution can be resolved given the structure and timing. These are a minority. Ambiguous cases are those where multiple causal explanations remain valid unless additional temporal or counterfactual information is supplied. These are the majority. Ambiguity is treated as a feature rather than a flaw. The presence of hidden timestamp tests and explicit refusal conditions reinforces the norm that correct causal reasoning includes recognizing when an answer cannot be identified from available information.

Vivek Sathe

My methodology centered on developing a sophisticated **Human-AI Interaction Web Application** that evolved from basic LLM assistance to a full 4-agent agentic system. This web app became the primary workstation for generating, validating, and scaling the production of 117 high-quality economic causal reasoning cases.

Phase 1: Foundation - Human-Generated Cases as Learning Anchors

The pipeline began with fully human-generated cases that served as epistemic anchors. These initial cases established the quality baseline and provided concrete examples for LLM learning. However, the **scenario and claim generation** process revealed critical insights into the capabilities of different AI systems.

Scenario and Claim Generation: Intelligent AI Systems vs. Raw LLMs

Human generation was augmented using **Gemini** and **Perplexity** as intelligent AI systems, which provided a stark contrast to raw LLM outputs:

1. **Intelligent AI Systems (Gemini & Perplexity)**:
    - **Internet Access**: Both systems could retrieve real-time economic data, recent policy changes, and current market conditions. For example, when generating a case on cryptocurrency regulation, Perplexity could cite actual SEC enforcement actions from 2024-2025.
    - **Thinking Capability**: Gemini's extended thinking mode allowed multi-step reasoning to construct economically plausible scenarios. It could reason through: "If X policy is implemented → what are the immediate effects → what are the second-order effects → what confounders exist?"
    - **Source Verification**: Perplexity provided citations to academic papers, policy reports, and economic data sources, enabling verification of factual claims in scenarios.
    - **Domain Expertise Simulation**: These systems could access and synthesize information from specialized economic domains (e.g., DeFi protocols, carbon markets) that are underrepresented in base LLM training data.
2. **Raw LLMs (Base Models)**:
    - **Hallucination Risk**: Frequently invented plausible-sounding but factually incorrect economic statistics or policy details.
    - **Surface Pattern Matching**: Generated scenarios that were structurally correct but economically shallow, lacking the nuanced mechanisms found in real-world cases.
    - **No Verification**: No ability to fact-check claims or validate economic relationships against empirical data.

Practical Example: Generating a DeFi Case

When creating **Gen1.5-VarB (Offshore Migration)**, the process revealed these differences:

- **Perplexity Query**: "What are the regulatory arbitrage mechanisms in DeFi protocols between US and offshore jurisdictions?"
    - **Result**: Retrieved actual cases of Uniswap, Compound, and Aave governance discussions about regulatory compliance, with citations to SEC guidance and CFTC statements.
- **Gemini Extended Thinking**: "Design a scenario where a DeFi protocol relocates to avoid regulation, but the intervention (relocation) is confounded by network effects."

    - **Result**: Multi-step reasoning that identified the confounder (user base concentration) and mediator (liquidity migration), producing a structurally valid CONF-MED case.
- **Raw LLM (GPT-3.5 baseline)**: "Create a DeFi regulation case."
    - **Result**: Generic scenario about "a country banning crypto" without specific mechanisms, confounders, or realistic economic structure.

Introductory Understanding of Reasoning Capabilities

This comparison provided foundational insights into AI reasoning:

1. **Tool Access Matters**: Intelligent AI systems with internet access could ground scenarios in empirical reality, reducing the need for human fact-checking by ~70%.
2. **Thinking vs. Generation**: Gemini's extended thinking mode demonstrated that **reasoning depth** (multi-step causal analysis) is distinct from **generation fluency** (producing coherent text). Raw LLMs excel at the latter but fail at the former.
3. **Verification Loop**: Perplexity's citation mechanism enabled a human-AI verification loop:
    - Human specifies the economic domain and causal structure
    - Perplexity retrieves relevant empirical cases and research
    - Human validates the causal mechanism against cited sources
    - Gemini generates the formal scenario and claim text
4. **Domain Adaptation**: For modern economic contexts (DeFi, gig economy, climate finance), intelligent AI systems could access specialized knowledge that raw LLMs lack, making them essential for generating realistic, high-quality scenarios.

Web App Interface for Incremental Construction

The web app's initial interface allowed incremental case construction with manual field entry for: - Scenario and claim formulation (augmented by Gemini/Perplexity) - Pearl level assignment (L1/L2/L3) - Variable identification (X, Y, Z with roles) - Causal structure definition - Trap type classification using a hierarchical taxonomy

This human-AI collaboration in Phase 1 established the quality baseline that informed the subsequent development of the 4-agent agentic system in Phase 2.

Phase 2: Evolution to 4-Agent Agentic System

The system evolved through multiple iterations to address scalability while maintaining epistemic control:

1. **The Architect (Agent 1)**: A Senior Causal Scientist persona with 2,500-token budget. Drafts initial causal mechanisms by exploring economic con-

cepts and theories. Generates the foundational analysis including causal pathways and potential confounders.

2. **The Auditor "Dr. Doom" (Agent 2)**: Modeled after economist **Nouriel Roubini**, this agent performs deep skeptical analysis with a 2,500-token budget. The choice of Roubini's persona was deliberate—not for pessimism, but for his rigorous approach to identifying systemic vulnerabilities and hidden risks in economic systems. The Auditor challenges the Architect's logic, identifies 2-3 critical flaws, and surfaces unobserved confounders.

3. **The Editor-in-Chief (Agent 3)**: Strategic adjudicator with a 7,500-token "hyper-headroom" budget to prevent truncation of complex reasoning. Resolves debates between Architect and Auditor through an **11-point professional checklist**:

   - Causal structure derivation (70-word limit)
   - Hidden structure identification (50-word limit)
   - Gold rationale synthesis (50-word limit)
   - Ground truth adjudication (50-word limit)
   - Pearl hierarchy alignment
   - Trap type/subtype classification
   - Subdomain allocation
   - Wise refusal commentary when appropriate
   - Reference identification from research literature
   - Reviewer notes (50-word limit)

4. **The JSON Drafter (Agent 4)**: Technical clerk with 5,000-token budget. Maps the Editor's strategic resolution into perfectly structured JSON, enforcing strict word limits and schema compliance through "Direct Narrative Pickup"—extracting clean, professional fields without backend-injected agent traces.

Phase 3: Iterative Refinement Through Human Judgment

The web app architecture revealed critical limitations that required constant human intervention:

- **Token Budget Optimization**: Initial implementations truncated complex reasoning. Iterative testing revealed optimal budgets: Architect (2,500), Auditor (2,500), Editor (7,500), Drafter (5,000).

- **Word vs. Token Limits**: Discovered that word limits (enforced in prompts) provided better semantic control than token limits for field-level constraints. For example, `causal_structure` limited to 70 words, `gold_rationale` to 50 words.

- **Handshake Visibility**: Implemented "Trace Persistence" using `sessionStorage` to allow re-examination of agent debates without re-running analysis. This proved essential for understanding failure modes and refining prompts.

- **Depth of Memory**: The Editor's 11-point checklist required access to both Architect and Auditor outputs simultaneously, necessitating careful prompt engineering to maintain context without exceeding model limits.

Phase 4: Challenging Cases and Human Override

The most challenging cases required extensive human judgment and back-and-forth iteration:

L3 Cases: - Gen1.4-VarD (Rent Control Reversion): Required modeling economic "hysteresis"—the path-dependent reality that removing rent control doesn't restore previous equilibrium due to structural changes (developer exit, housing stock degradation). The AI initially failed to recognize irreversible path dependencies, requiring manual correction of the counterfactual logic.

- **Gen2.1-VarA (Unicorn Survivorship)**: Tested survivorship bias in venture capital analysis. The Auditor correctly identified that analyzing only successful unicorns conditions on the outcome, but the Architect initially missed the temporal selection mechanism. Human intervention was required to clarify that the selection happens *during* the observation period.

L2 Cases: - Gen1.1-VarC (Highway Opening): Required explicit mathematical proof that X | Y due to backdoor path through Z (economic development). The web app's variable editor allowed precise definition of the DAG structure, but human judgment was essential to ensure the backdoor path was properly blocked or unblocked based on the intervention.

- **Gen8.5-VarA (Test Prep Gaming - Goodhart's Law)**: This case models the critical distinction between a *proxy* (test scores) and a *latent construct* (mathematical ability). The AI consistently failed to identify that test prep creates a direct X→Y path bypassing Z, rendering the metric invalid. Required manual insertion of references to **Goodhart's Law** and **Campbell's Law** from policy research literature.

Technical Implementation Details:

The web app (`case-review-app`) was built using: - **Backend**: Node.js/Express with real-time data persistence - **AI Integration**: Google Gemini 2.5 Flash via `@google/generative-ai` SDK - **Frontend**: Vanilla JavaScript with dynamic form validation - **Trap Taxonomy**: Hierarchical dropdown system with Pearl-level cascading (L1: 9 trap types, L2: 8 trap types, L3: 6 trap types) - **Smart Case Suggestion**: Algorithm that dynamically recalculates target gaps after each review to optimize coverage across Pearl levels, difficulty, and ground truth distributions

References and Research Integration:

The system was designed to anchor analysis in peer-reviewed research. For example: - Survivorship bias cases reference **Elton, Gruber, & Blake (1996)** on mutual fund performance - Goodhart's Law cases reference **Strathern (1997)** on audit culture - Dynamic world divergence cases reference **Arthur (1989)** on path dependence in economic systems

Outcome:

This methodology produced 117 cases with the following distribution: - **Pearl Levels**: L1 (15), L2 (77), L3 (25) - **Difficulty**: Easy (11), Medium (59), Hard (47) - **Trap Types**: Emphasized modern economic contexts (DeFi, gig economy, platform economics) and composite traps (CONF-MED, CONF-MED-COLLIDER)

The web app's 4-agent system reduced human annotation time by approximately 60% while maintaining full epistemic control through mandatory human review and endorsement of all AI-generated suggestions.

Chris Pearce

Overview My approach was based on using a combination of Gemini Web and Claude Code to construct and iteratively improve a series of cases. My approach was to; 1 - Supply the Benchmark Document to each as context 2 - Produce several example cases that I had produced myself and refined using Gemini as additional context 3 - Supply a series of links to economists or associated writers whose work I admire and find provocative to generate lists of ideas for topics, including; * John Cochrane * Paul Krugman * Noah Smith * Scott Alexander * Andrew Gelman * Tyler Cowan * Matt Levine * Other individual articles that have influenced my thinking recently 4 - Get Claude Code to generate an initial case 5 - Workshop the case to improve it an make it more rigorous, including cross checking with Gemini. 6 - Generate several more cases and workshop those 7 - Generate a harness for review agents to check cases to see which cases were capable of confounding the models (at least part of the time), and assist Claude to refine it's approach to case generation. Three review agent types were generated; * Naive * Skeptical * Expert 8 - Refine wording of the cases so as to be more convincing to LLMs 9 - Iterate steps 5-9 gradually to build out the full suite of cases 10 - Mannually perform annotations

Distribution Management I adopted several approaches to managing the distribution of cases through the process; * Tracking the intended Trap Type during generation, but prior to manual annotation * Monitoring difficulty by using the confusion rate of each agent

Review Agent Prompts

Naive Agent Prompt

```
You are evaluating causal claims in economics scenarios. For each case, you will see
a scenario describing a study or evidence, followed by a causal claim.
```

```
Your approach: Accept the claim as valid if the methodology described sounds
reasonable and the evidence appears to support the conclusion. Give the benefit
of the doubt to well-designed studies.
```

Skeptical Agent Prompt

```
You are a critical evaluator of causal claims. For each case, you will see a
scenario describing a study or evidence, followed by a causal claim.

Your approach: Actively look for causal reasoning flaws including:
- Confounding (omitted variables that affect both cause and effect)
- Selection bias (non-random sampling that distorts conclusions)
- Reverse causation (effect actually causes the supposed cause)
- Measurement issues (outcomes don't capture what's claimed)
- Overgeneralization (extrapolating beyond what evidence supports)
```

Expert Agent Prompt

```
You are an expert econometrician evaluating causal claims. You have deep expertise in:
- Pearl's causal hierarchy (Association, Intervention, Counterfactual)
- Rubin's potential outcomes framework and SUTVA
- Identification strategies: RCT, Difference-in-Differences, Regression Discontinuity,
  Instrumental Variables, Synthetic Control, Propensity Score Matching

For each case, evaluate whether the causal claim is justified by the evidence and
methodology described. Consider:
- Internal validity (does the design identify the causal effect?)
- External validity (does the effect generalize as claimed?)
- Mechanism validity (is the proposed mechanism supported?)
- LATE vs ATE issues (local vs average treatment effects)
- Parallel trends, exclusion restrictions, and other identifying assumptions
```

Mason Hu

My methodology was much simpler and more hands-on than Vivek's or Chris's, but still systematic in its own way.

I worked in a tight loop with the model, gradually shaping what a "good" causal case looks like. Instead of building a full agent system or automated reviewers, I focused on understanding the structure of causal reasoning itself: what counts as X, Y, and Z, what really distinguishes L1 from L2 and L3, what the core traps are and how these should appear clearly in a scenario. all while leaning towards cases that involve actually economic theory instead of factual historic events that exist in the big corpus. I kept refining the format, the wording, and the logic until each case expressed the causal issue in a way that was conceptually clean and unambiguous, not just superficially well-written.

In practice, I acted like a careful editor and curator. I generated cases, inspected them, corrected their causal logic, tightened definitions, removed misleading patterns, and enforced consistency across the entire set. Over time, this iterative process produced a standardized structure and a coherent causal language, aligning economic intuition with formal concepts from Pearl's hierarchy. The emphasis of my approach was not automation, but clarity and control: making sure every case truly tested the intended type of reasoning in a way that a human reader could follow, evaluate, and learn from.

---

3. Structural Sanity Check

Total Cases: 465 / 460 (101.1%) **Authors:** atanu, chris, mason, vivek **Output File:** BenchmarkT3_Extension_B.json

---

Pearl Level Distribution

Category	atanu	chris	mason	vivek	Total	%	Target	Target %
L1	0	15	15	15	45	9.7%	50	11%
L2	98	76	78	77	329	70.8%	308	67%
L3	18	25	23	25	91	19.6%	101	22%
Total	**116**	**116**	**116**	**117**	**465**	**100.0%**	**460**	**100%**

L3 Ground Truth Distribution

Ground Truth	atanu	chris	mason	vivek	Total	%	Target	Target %
VALID	0	7	0	3	10	11.0%	30	30%
INVALID	0	2	0	9	11	12.1%	10	10%
CONDITIONAL	18	16	23	13	70	76.9%	60	60%
Total	**18**	**25**	**23**	**25**	**91**	**100.0%**	**101**	**100%**

Difficulty Distribution

Category	atanu	chris	mason	vivek	Total	%	Target	Target %
Easy	4	17	0	11	32	6.9%	69	15%
Medium	73	52	59	59	243	52.3%	271	59%
Hard	39	47	57	47	190	40.9%	119	26%
Total	**116**	**116**	**116**	**117**	**465**	**100.0%**	**460**	**100%**

Trap Type Distribution

Category	atanu	chris	mason	vivek	Total	%	Target	Target %
COLLIDER	4	7	0	8	19	4.1%	32	7%
CONF-MED	61	43	47	47	198	42.6%	165	36%
CONF-MED-COLLIDER	17	0	0	0	17	3.7%	0	0%
CONFOUNDING	0	0	4	0	4	0.9%	0	0%
CONF–MED–COLLIDER	9	0	0	0	9	1.9%	0	0%
COUNTERFACTUAL	18	25	23	25	91	19.6%	101	22%
MECHANISM	0	4	0	3	7	1.5%	9	2%
MED	7	0	0	0	7	1.5%	0	0%
REVERSE	0	13	18	19	50	10.8%	59	13%
SELECTION	0	24	4	15	43	9.2%	50	11%
SELECTION / COLLIDER	0	0	20	0	20	4.3%	0	0%
Total	**116**	**116**	**116**	**117**	**465**	**100.0%**	**460**	**91%**

---

4. Failure-Mode Analysis

Due to differences in the methods for generating and testing models, team members identified different issues with the various pipelines described below;

Mason Hu

Chris Pearce The cases in the supplied dataset are relatively simple and rely on fairly well known economic phenomenon (minimum wage debates, causes of immigration etc). Most state of the art models are well trained on these kinds of cases and are capable of pattern matching their way to an answer to these kinds of examples without really needing to test their reasoning skills.

In order to make the tests expose reasoning gaps more effectively I tried several approaches; 1. **Recency:** Trying to find cases that are recent enough to be after the cutoff date for state of the art LLMs. The debate about the impact of holding back third grade students for a year is one such example, motivated by an extensive debate between Andrew Gelman and other high profile bloggers over the last few weeks over the true causes of Missisipi's educational performance.

Other such examples include John Cochraine's writing on the extent of Central Bank independence, and his recent writings on 'Questions for the incoming Federal Reserve Chair'

2. **Obscurity:** Trying to find more obscure topics that have not been as extensively studied. The "Priority Alpha" case is one such example, based on a historic study into failures that the Russian Military suffered in their railway logistics chains during World War One.
3. **Authoritativeness:** Adding additional wording to make the case seem more credible, such as citing statistics and study designs intending to make the case seem more balanced and less superficial.
4. **Source Suppression:** For cases based on real life events, trying to take off names and specifics that directly identify the scenario, which would simplify the process of pattern matching for the models.

Collectively by applying these changes I was gradually able to confound the three model personas (naive, skeptica, expert) more effectively at least some of the time. This was quite an iterative process. In early iterations, claims that we assumed to be difficult to assess were easily solved by the models, but progressively this approach was able to produce results that could confound the LLMs.

An example case that confused the LLMs some of the time;

```
"scenario": "Researchers used a natural experiment when a shipping disruption randomly affec
"claim": "Supplier diversification reduces supply chain risk because the natural experiment
```

This process justifies the need to generate the additional cases, as significant effort is required to generate cases that can break the models out of simple pattern matching effort and to attempt to force deep reasoning from them. Even after this, confounder / mediator examples proved to be relatively easier for the LLMs to resolve, while collider and machanism models proved to be much more difficult.

One additional factor that justifies expanding the dataset is the sheer subjectivity of the topic area. Economics is a subject where many causalities are not clear, timing may vary and lags can be long, there can be small datasets and limited opportunity to conduct controlled experiments. Opinions vary widely on causes and effects, and as to the extent to which evidence is sufficient enough to support a claim. This is not physics or chemistry. As such, there is likely to be a lot of noise in the benchmark, where reasonable people will differ about whether claims are true or not, and as such a large sample set is probably needed in order to create a corpus that is, on balance, reasonable, even if individual claims made may be disputable.

Vivek Sathe

Qualitative testing with general-purpose reasoning models revealed several recurring failure patterns in my case set:

1. **Mediator Collapse:** Models frequently skip intermediate variables and reason directly from intervention to outcome, failing to recognize the causal chain. This was particularly evident in cases involving multi-step economic mechanisms.

2. **Collider Blindness:** Models consistently ignore selection bias effects, such as analyzing only surviving firms in survivorship bias scenarios without recognizing the conditioning on success.

3. **Temporal Insensitivity:** Models frequently fail hidden timestamp tests, treating causal ordering as symmetric even when the scenario explicitly specifies temporal precedence ($t_X < t_Z$). This represents a fundamental failure in understanding that correlation does not imply causation when temporal ordering is reversed.

4. **Metric Gaming (Proxy vs. Latent Construct):** In **Case Gen8.5-VarA (Test Prep Gaming)**, models often accept "Test Scores" (Y) as proof of "Math Ability" (Z), failing to identify that the intervention (Test Prep) creates a direct path to Y that bypasses Z. This "Metric Gaming" trap is essential for evaluating AI alignment in policy settings where Goodhart's Law applies—when a measure becomes a target, it ceases to be a good measure.

5. **Dynamic World Divergence Failures:** Models struggle with hysteresis effects in L3 counterfactual scenarios. In **Gen1.4-VarD (Rent Control Reversion)**, models incorrectly assume that removing a policy would restore the previous equilibrium, failing to recognize that structural changes during the policy period create irreversible path dependencies.

6. **Modern Economic Context Gaps:** Testing revealed that models trained primarily on traditional macroeconomic examples struggle with decentralized systems (DeFi) and platform economics (gig economy), suggesting the need for domain-specific training data that reflects contemporary economic structures.

Atanu Mukherjee

Using one to two representative general-purpose reasoning models in a qualitative, non-competitive manner, several recurring failure patterns emerge when these models are applied to the cases:

- **Mediator collapse**: Models often skip intermediate variables and reason directly from intervention to outcome.
- **Confounder misclassification**: Models introduce generic background variables rather than identifying the true confounder specified in the case.
- **Collider blindness**: Conditioning on selected samples (such as surviving firms, successful projects, or observed outcomes) is frequently ignored.

- **Temporal insensitivity**: Hidden timestamp tests are often failed, with models treating causal ordering as symmetric.
- **Overconfidence under ambiguity**: Models frequently provide a single causal explanation even when the case is explicitly underdetermined.

These failures occur even when surface-level economic knowledge is correct. The existing cases already expose these weaknesses, but coverage is uneven. Some trap subtypes appear only once or twice, making failure detection unstable.

Additional Observations from Case Analysis

- Infrastructure, energy, and grid economics naturally generate **confounder–mediator** and **confounder–mediator–collider** structures, making them especially valuable for causal testing.
- Many policy debates fail because of implicit conditioning on success or survival; the explicit use of collider variables in the cases makes this failure mode visible.
- Behavioral economics mechanisms fit cleanly into causal graphs when treated as mediators rather than informal explanations.
- **Explicit refusal conditions** are critical. They distinguish causal reasoning from rhetorical explanation and are an area where models consistently underperform.
- **Human oversight is essential**. Nearly every nontrivial case required correction of causal structure that an LLM would not reliably detect on its own.

Examples Supporting the Conclusions

LLMs Require Human Correction of Causal Structure

Several cases in the file required explicit human intervention to fix causal errors that a generative model would commonly make.

- **Case 5.73 (engineering firm scale and profitability)**
  The initial version omitted the time-expansion mediator and misclassified baseline capability as a mediator rather than a confounder.
  The final case explicitly separates:
  - capability as a **confounder**, and
  - time expansion as a **mediator**,
    correcting a common LLM error of collapsing both into a single explanation.
- **Case 5.78 (net metering and retail electricity prices)**
  The causal chain required splitting:
  - power procurement cost effects, and

  - fixed-cost sharing effects.
    Earlier drafts conflated these. The final case separates multiple mediators and clarifies which variables are causal versus accounting artifacts.

These corrections demonstrate why LLMs cannot be trusted to finalize causal labels without human verification.

Pearl-Level Balance Is Structural, Not Accidental

The Pearl-level distribution is reflected clearly in how cases are written.

- **L2 dominance examples**
  Cases 5.71 (interest rates and inflation), 5.73 (engineering firm scale), 5.77 (EV incentives and emissions), and 6.01–6.06 (grid reliability, pricing, and dispatch) are intervention-focused and rely on causal ordering rather than counterfactual worlds.
- **L3 counterfactual examples**
  Cases 5.82–5.87 (infrastructure lock-in, refinery upgrades, capital controls vs. floating exchange rates) explicitly ask what would have happened under an unrealized policy or technology path.
- **Minimal L1 presence**
  There are almost no cases framed as pure correlation without intervention or counterfactual structure, by design.

This supports the conclusion that the Pearl-level mix is intentional and maintained.

Trap Taxonomy Must Include Composite Traps

Multiple cases cannot be classified meaningfully under a single trap label.

- **Case 5.72 (grid reliability under high renewables)**
  Combines mediation (renewable penetration affecting dispatch costs) with collider bias (conditioning on "no blackout days"). Treating it as only MED would miss the selection error.
- **Case 5.80 (industrial policy success stories)**
  Conditioning on surviving firms creates a collider, while subsidies operate through investment mediation. This is a **CONF–MED–COLLIDER** case by structure, not by narrative choice.
- **Case 5.92 (financial deepening and inequality)**
  Asset price inflation is a mediator, while unequal access to finance is a

confounder. Treating it as only CONF or only MED would be structurally incorrect.

These cases justify abandoning a single-label trap taxonomy.

Ambiguity Is a Deliberate Feature

Several cases are explicitly unidentifiable without additional information, and this is surfaced in the design.

- **Case 5.71 (rate hikes and inflation)**
  Inflation decline could be due to demand destruction or supply normalization. The case cannot be resolved without timing data.
- **Case 5.619-style policing logic (crime–policing variants)**
  Higher recorded crime may reflect reporting intensity rather than actual crime. The case explicitly cannot be resolved without separating reporting from incidence.
- **Case 5.88 (tariffs and prices)**
  Tariffs raise shelf prices mechanically but reduce demand and import volume, potentially leaving equilibrium prices unchanged. Multiple paths remain valid.

These cases support the claim that ambiguity is intentional and pedagogically important.

Models Fail Systematically on Mediators, Colliders, and Timing

The structure of the cases themselves exposes common failure modes.

- **Mediator collapse examples**
  Case 5.73 (scale → profit) and Case 5.77 (EVs → emissions) both require passing through intermediate variables. Skipping these leads to wrong conclusions.
- **Collider blindness examples**
  Case 5.80 (successful firms), Case 5.98 (surviving banks), and Case 6.02 (no-blackout grid days) all involve conditioning on outcomes jointly caused by multiple variables.
- **Temporal insensitivity examples**
  Hidden timestamp tests in cases like 5.71 and 5.78 explicitly ask whether Z occurred before X. Ignoring this reverses causality.

These are not hypothetical weaknesses; they are structurally embedded stress tests.

Structurally Rich but Statistically Sparse

Many trap subtypes appear only once or twice.

Examples include:

- Net-metering fixed-cost spiral (primarily Case 5.78)
- Refinery constraint trade paradox (one counterfactual sequence, Cases 5.82–5.83)
- Engineering firm time-expansion paradox (Case 5.73 only)
- Grid reliability collider (small cluster, Cases 6.01–6.03)

This supports the conclusion that:

- Each case is dense in structure
- The corpus is too small for stable statistics
- Controlled expansion is necessary

Wise Refusal Is Essential and Nontrivial

Several cases explicitly require refusal rather than resolution.

- **Case 5.71**: Cannot assign credit to monetary policy without timing
- **Case 5.73**: Cannot conclude scale causes margin decline without contract and duration data
- **Case 5.92**: Cannot infer inequality effects without asset ownership distribution

These cases support the finding that refusal is a first-class reasoning outcome, not a failure.

Summary

The conclusions in the report are not abstract. They are directly supported by:

- Specific case structures
- Repeated correction of causal graphs
- Explicit use of mediators, confounders, and colliders
- Systematic inclusion of ambiguity and refusal conditions

---

5. Justification for Scaling the Benchmark Test Set

The current dataset of ~467 cases, while structurally rich and carefully curated, is statistically insufficient for robust AI training and evaluation. Our empirical findings from the web app testing and multi-agent analysis reveal three critical limitations that necessitate scaling:

5.1 Statistical Instability and Small Sample Effects

Several trap subtypes appear only 1-3 times in the current dataset, creating statistical instability:

- **L3 Preemption Cases**: Only 2-3 instances of “Early Preemption” vs. “Late Preemption” distinction
- **Goodhart’s Law (L1)**: Only 3-4 cases testing “Static Metric Gaming” vs. “Proxy Drift”
- **Simpson’s Paradox**: Insufficient instances to distinguish “Aggregation Bias” from “Imbalanced Group Composition”

This sparsity means that: 1. **Failure detection is unstable**: A model might pass by chance on 1-2 cases but fail systematically on the underlying reasoning pattern 2. **Fine-tuning is impossible**: Modern LLMs require 50-100 examples per pattern for reliable learning 3. **Evaluation metrics lack confidence intervals**: With n<10 for many subtypes, we cannot distinguish true capability from random performance

5.2 Surface Variation Sensitivity

Testing with the 4-agent system revealed that **small wording changes** can dramatically shift model outputs, even when the underlying causal structure remains identical. For example:

- **Case Gen1.1-VarC (Highway Opening)**: Changing “economic development” to “regional growth” caused the Architect to miss the confounder in 40% of trials
- **Case Gen8.5-VarA (Test Prep Gaming)**: Replacing “test scores” with “exam performance” reduced the Auditor’s detection of the proxy trap from 80% to 45%

This sensitivity indicates that models are pattern-matching on **surface features** rather than learning the **structural causal logic**. To train robust reasoning, we need: - **Controlled expansion**: Preserve the causal graph (X, Y, Z structure) while varying narrative framing - **Lexical diversity**: Same trap type expressed across different economic domains (labor, finance, trade, development) - **Temporal variations**: Same structure with different time horizons (short-term vs. long-term effects)

5.3 Modern Economic Domain Coverage

The benchmark currently under-represents contemporary economic structures that are increasingly relevant for policy AI:

- **Platform Economics**: Only 8 cases on gig economy, sharing economy, and two-sided markets
- **Decentralized Finance (DeFi)**: Only 5 cases on crypto, blockchain governance, and algorithmic stablecoins
- **Climate Economics**: Only 3 cases on carbon markets, green subsidies, and environmental externalities
- **Behavioral Nudges**: Only 4 cases on choice architecture and libertarian paternalism

These domains exhibit **unique causal patterns** not found in traditional macroeconomics: - **Network effects** and **feedback loops** in platforms - **Algorithmic governance** and **smart contract enforcement** in DeFi - **Intergenerational externalities** and **tipping points** in climate - **Bounded rationality** and **default effects** in behavioral economics

Scaling to 500+ cases allows proper representation of these modern contexts while maintaining the 67% L2 / 22% L3 / 11% L1 distribution.

5.4 Ambiguity and Wise Refusal Training

Economic causality is inherently **underdetermined** in many real-world scenarios. Our dataset intentionally includes: - **CONDITIONAL cases** (76.9% of L3): Answer depends on hidden information (timestamps, unobserved confounders) - **Wise refusal commentary**: Teaching AI to say "I cannot determine this without additional data"

However, with only 70 CONDITIONAL cases across all trap types, we lack sufficient examples to train models to: 1. **Distinguish** between "truly unknowable" vs. "difficult but solvable" 2. **Articulate** the specific missing information needed (e.g., "Need timing: did X precede Z?") 3. **Avoid** false confidence when data is insufficient

Scaling to 500 cases with 60% CONDITIONAL (300 cases) would provide the statistical power needed for this critical capability.

5.5 Recommended Scaling Strategy

Based on the web app's smart case suggestion algorithm and empirical testing:

1. **Target: 500 total cases** (current: 467 cases developed in the assignment, starter: 46 cases provided)
2. **Prioritize L3 VALID cases**: Currently at 11% (target: 30% of L3 = 30 cases, gap: 20 cases)
3. **Expand modern domains**: Add 15-20 cases in DeFi, platform economics, climate
4. **Surface variation clones**: For each "seed" case with unique structure, create 2-3 narrative variants
5. **Maintain Pearl distribution**: 67% L2, 22% L3, 11% L1

This expansion would transform the benchmark from a **structural proof-of-concept** to a **statistically robust training corpus** suitable for fine-tuning and rigorous evaluation of causal reasoning AI.

---

7. Conclusion and Observations

This has been a challenging assignment to complete. The scale of the dataset requested is significant, in addition to which we have had to contend with * A team based across four different timezones spanning 12 hours, limiting our ability to connect directly * Variations in our understanding of the required json structure as the discussions on the course Slack channel evolved the specification * Variations in our understanding of the report contents; we initially believed that the report was to be of the form of the BenchmarkT3-BucketLarge-B.pdf and so commenced by assembling a shared Google doc matching this format. We only came to understand part way through the exercise what was required in the report * Resulting from the above, we have a file that has been processed multiple times before ending in its final form, introducing the risk of transposition errors having occurred Given this, the file submitted should be considered to be in a preliminary form. The second stage peer review process is going to be necessary to ensure that the file is free from material error and that our understanding of the final submission requirements are reflected in the file structure.