

T³ Benchmark Analysis: Bucket 10 (Social Science)

1. Structural Sanity Check (A)

Pearl Level Distribution

The dataset maintains a specific distribution across Pearl's Causality Hierarchy, which is essential for testing "levels of thinking" from simple observation to hypothetical intervention and counterfactual reasoning.

- **L1 (Association): 5 cases (11.1%)**
 - *Focus:* Correlation, patterns, and base rates.
 - *Examples:* Shark attack news (10.21), role-adjusted wage gap (10.33).
- **L2 (Intervention): 30 cases (66.7%)**
 - *Focus:* The "do-calculus"—what happens if we change a variable?
 - *Examples:* Admissions paradox (10.1), police stop data (10.20), minimum wage (10.38).
- **L3 (Counterfactual): 10 cases (22.2%)**
 - *Focus:* "What if things had been different?" and unobserved states.
 - *Examples:* Determinism (10.25), attributable fraction (10.27), election rain (10.45).

Assessment: The distribution is heavily weighted toward **L2 (Intervention)**. In Social Science, this is appropriate as most policy debates center on the effects of specific interventions (e.g., tax changes, minimum wage, quotas). However, L1 is relatively thin, which might make it difficult to establish a baseline for a model's "observational" common sense.

Trap Type & Subtype Coverage

The dataset identifies six signature traps: **Simpson's Paradox**, **Ecological Fallacy**, **Selection**, **Composition Effect**, **Collider**, and **Base Rate Neglect**.

- **Exhaustiveness:** The list covers the most common statistical errors in demographics. However, the analysis of the 45 cases reveals "New" trap types not listed in the initial signature (e.g., **Goodhart's Law**, **Reverse Causality**, **Time Horizon**, and **Mismatch**). These are vital additions because social science often involves humans "gaming" the system (Goodhart) or variables that have bidirectional relationships (Reverse).
- **Mutual Exclusivity:** There is significant overlap. For instance, **Simpson's Paradox** is often a specific result of a **Composition Effect** or **Selection Bias**. The dataset handles

this well by using specific Subtypes (e.g., "Aggregation Bias" vs. "Sampling Bias") to distinguish the *mechanism* of the error.

Ambiguity vs. Determinate Cases

- **L1 & L2:** These are largely **Determinate**. There is a clear statistical "Correct Answer" (e.g., in 10.3, Treatment B is mathematically superior in both subgroups).
- **L3:** These are frequently **Conditional**. For example, in Case 10.24 (Economic Counterfactual) and Case 10.26 (Fairness Audit), the "Ground Truth" is labeled as **CONDITIONAL**. This reflects the reality of Social Science: the answer depends on unobserved confounders or the validity of the chosen model (e.g., synthetic control).

Difficulty

The distribution of difficulties is:

- 5 Easy
- 21 Medium
- 19 Hard

This is different from the reported distribution in the document which is incorrect. We found these numbers with a simple Ctrl+F of the document.

2. Failure-Mode Analysis (B)

Analyzed 10 examples from the provided 45 data examples:

10.43 - Medium

ChatGPT Response

No — this reasoning is **not necessarily valid**. It commits a classic **causal inference / selection bias error**.

Why the reasoning is flawed

The statement "*charter schools significantly outperform public schools*" is an **observational comparison**, not a causal one.

Formally:

- Let **X = charter schools**
- Let **Y = public schools**

- Observed:
[
 $E[\text{Score} \mid X] > E[\text{Score} \mid Y]$
]

This does **not** imply:

$$[E[\text{Score if converted to charter}] = E[\text{Score} \mid X]]$$

because the students, families, teachers, and environments in X and Y are **not randomly assigned**.

Key issues

1. Selection effects

- Charter schools often enroll:
 - More motivated families
 - Students with higher prior achievement
 - Students whose parents actively opt in
- These factors raise scores **independently of the school type**.

2. Capacity and scaling effects

- High-performing charter schools may rely on:
 - Selective admissions (explicit or implicit)
 - High teacher workloads
 - Small class sizes
 - Exceptional leadership
- These properties may **not scale** when applied statewide.

3. Treatment vs. group effect

- What's observed is a **group difference**, not a **treatment effect**.
- The correct causal question is:
What happens to the *same students* if their school is converted to a charter?

4. Simpson's paradox (possible)

- Aggregate performance differences can reverse once you condition on:
 - Socioeconomic status
 - Prior achievement
 - English learner status
 - Special education needs

What would justify the policy?

The reasoning would be valid **only if**:

- Students were randomly assigned to charter vs. public schools, **or**

- There is strong causal evidence (e.g., lottery-based RCTs) showing:
[
\text{Charter treatment} \rightarrow \text{Higher scores}
]
and
- Evidence that the effect persists at scale.

Bottom line

Observed superiority of charter schools does not imply that converting all public schools into charters will raise statewide scores.

This is a classic example of **confusing correlation with causation** under **non-random selection**.

10.44 - Medium

ChatGPT Response

No — this conclusion is **not valid**. The critics' argument makes an **invalid causal inference**.

Why the reasoning fails

The observation is:

- **X** = gun buyback program implemented
- **Y** = gun violence rate unchanged after one year

From this, critics conclude:

"Reducing gun availability does not stop crime."

This inference does **not** logically follow.

Key problems with the argument

1. **Intervention ≠ effective treatment**
 - A buyback program does **not necessarily cause a meaningful reduction in gun availability**.
 - Buybacks often collect:
 - Old, broken, or unused guns
 - Guns not associated with criminal activity
 - If criminal access to guns is unchanged, the test of the hypothesis never actually occurred.
2. **No counterfactual**

- We do not know what gun violence would have been **without** the buyback.
- Violence could have:
 - Increased (buyback prevented a rise)
 - Decreased but was offset by other factors
- 3. **Timing and lag effects**
 - Effects of reduced availability (if any) may take longer than one year to appear.
 - Short-term measurements may miss long-term impacts.
- 4. **Single case ≠ general theory**
 - One city, one program, one outcome does not establish a general causal law about guns and crime.
- 5. **Confounding variables**
 - Gun violence is influenced by:
 - Policing strategies
 - Economic conditions
 - Gang dynamics
 - Drug markets
 - Any of these could mask or overwhelm a small effect.

What *can* be concluded

The data only supports a much weaker statement:

This particular buyback program, as implemented in City A, did not measurably reduce gun violence within one year.

It does **not** support the broader claim that:

Reducing gun availability cannot reduce crime.

Bottom line

Failure of a specific intervention does not falsify the underlying causal hypothesis.

The critics are confusing:

- “**This policy didn’t work here**”
with
- “**The causal mechanism doesn’t exist.**”

That leap is logically unjustified.

10.5 - Medium

ChatGPT Response

Yes — **both claims can be statistically correct**. There is no contradiction.

Why this can happen

The key issue is **different units of averaging**.

- **University claim (X):**
“The average class size is 35.”
→ This is the **unweighted average over classes**.
- **Student survey (Y):**
“The average student experiences a class size of 120.”
→ This is the **average over students**, which effectively **weights large classes more**.

Intuition

Large classes contain many students, so:

- A small number of very large lectures
- Plus many small seminars

can produce:

- A small **average class size**
- But a very large **average experienced by students**

Simple example

Suppose:

- 9 classes of size 10
- 1 class of size 1,000

Average class size (university):

$$\left[\frac{9 \times 10 + 1 \times 1000}{10} = 109 \right]$$

(Adjust numbers slightly to get 35 if desired.)

Average student experience:

- Most students are in the 1,000-student class
- So the “typical” student experiences a very large class

Formally, the student average is:

$$\left[\frac{\sum (\text{class size})^2}{\sum (\text{class size})} \right]$$

which is always \geq the class average.

This phenomenon has a name

- **Size-biased sampling**
- Related to the **friendship paradox**
- Sometimes called the **class size paradox**

Bottom line

Both statements can be true simultaneously because they are averaging over different populations.

The apparent contradiction arises from confusing:

- **Average over classes**
with
- **Average over students**

10.6 - Easy

ChatGPT Response

As stated, **the comparison is not a valid basis for concluding anything about quality of care.**

What’s being claimed

- Hospital A has **higher mortality** than Hospital B
- Hospital A is a **top-tier trauma center (X)**
- Hospital B is a **community clinic (Z)**

The implicit (but invalid) inference is:

“Hospital A provides worse care than Hospital B.”

Why the reasoning is invalid

This is a classic case of **confounding / case-mix bias**.

1. **Different patient populations**

- Trauma centers receive:
 - The sickest patients
 - Severe injuries
 - Emergency transfers after complications
- Community clinics treat:
 - Less severe cases
 - Lower-risk patients

2. Higher mortality can occur **despite better care**, simply because the baseline risk is higher.

3. **Selection on severity**

- Patients are **not randomly assigned**.
- The variable “trauma center” is correlated with:
 - Injury severity
 - Comorbidities
 - Time-critical cases

4. **Simpson’s paradox (often present)**

- Within comparable severity strata (e.g., mild, moderate, severe injuries), Hospital A may have **lower mortality** in every group.
- Aggregating across groups reverses the comparison.

5. **Outcome ≠ performance**

- Raw mortality rates reflect:
 - Case mix
 - Referral patterns
 - Triage rules
- Not just quality of care.

What would make the comparison valid

To assess quality, you would need:

- Risk-adjusted mortality rates
- Comparisons within similar severity categories
- Or randomized / quasi-experimental designs

Bottom line

Higher mortality at a trauma center does not imply worse care.

The comparison ignores **who is being treated**, not **how well they are treated**, making the inference invalid.

10.7 - Hard

ChatGPT Response

No — the critics' reasoning is **not valid**. The claim that discrimination is negligible based solely on the adjusted gap ignores important nuances.

What's being observed

1. **Raw pay gap:**
[
Y = 0.80
]
(Women earn 80¢ per \$1 men earn.)
 2. **Adjusted pay gap controlling for job title and industry:**
[
Z = 0.98
]
(The gap shrinks when comparing similar roles.)
 3. **Critics' conclusion:**

"Discrimination is negligible."
-

Why this reasoning is flawed

1. ****Controlling for job and industry may mask discrimination**
 - Women may be **underrepresented in high-paying industries or senior roles** due to:
 - Bias in hiring or promotion
 - Historical discrimination
 - Workplace culture
 - By adjusting for job and industry, you **remove the portion of the gap that could itself be caused by discrimination** ("explained vs. unexplained gap" issue).
2. **Indirect discrimination**
 - Even if pay within the same role is equal, structural barriers can prevent women from reaching those roles.
 - The \$0.98 figure **does not capture this upstream discrimination**.
3. **Other factors**
 - Factors like:
 - Hours worked, overtime access
 - Negotiation outcomes

- Leadership opportunities
 - May also reflect systemic biases and are often omitted from simple adjustments.
 - 4. **Causal interpretation is tricky**
 - The **adjusted comparison** only tells us:
Among women and men in the same job and industry, pay is closer to equal.
 - It **does not prove discrimination is negligible overall**.
-

Bottom line

Shrinking the pay gap after controlling for certain factors **does not prove discrimination is negligible**. It only shows that **within those controlled categories, the raw pay difference is smaller**, but systemic or structural discrimination can still exist upstream.

This is a classic case of **misinterpreting adjusted statistics as causal proof**.

If you want, I can make a **tiny diagram showing the difference between raw gap, controlled gap, and structural discrimination** — it makes this concept much clearer visually. Do you want me to?

10.22 - Hard

ChatGPT Response

No — this reasoning is **not valid**. The claim that environment determines success based on this single twin pair is **far too strong**.

Why it's invalid

1. Sample size = 1
 - The “study” is **one pair of twins**.
 - You cannot generalize from a single example; this is anecdotal evidence, not statistical proof.
2. Genetics and other factors ignored
 - Twins share genetics (even if adopted, they may differ in inherited traits such as intelligence, temperament, health predispositions).

- Differences in success could be influenced by both **nature** and **nurture**, not just environment.
 - 3. Selection and adoption effects
 - Adoption itself may involve selection bias: the type of household able to adopt may differ systematically.
 - Individual personality, motivation, social networks, luck, and chance events also matter.
 - 4. Correlation \neq causation
 - Twin A's career and Twin B's career correlate with their environments, **but that alone does not prove causation**.
 - 5. Ignored confounders
 - Cultural expectations, neighborhood opportunities, schools, peers, health, mentorship, or random life events could all contribute.
-

Correct conclusion

From this anecdote, the only thing you can reasonably say is:

“In this one case, the twins ended up in different socioeconomic outcomes. This is **consistent with environmental influence**, but it does not prove that environment alone determines success.”

✓ **Key point:**

One data point cannot establish causality. To claim that environment determines success, you would need **large, controlled studies**, ideally **randomized or natural experiments**, or statistically rigorous **twin/adoption studies**.

If you want, I can **make a simple “why anecdotes mislead” diagram** using this twin example — it really clarifies the fallacy visually. Do you want me to do that?

10.31 - Hard

ChatGPT Response

Yes — this is **statistically valid** and is a classic example of **Simpson's Paradox**.

What's happening

- **Observed fact:**
(\text{Average tax rate in Country A} > \text{Average tax rate in Country B})
- **Conditional fact:**
Within every income bracket,
(\text{Tax rate in Country A} < \text{Tax rate in Country B})

At first glance, this seems contradictory, but it's possible.

Why it happens

1. **Different income distributions**
 - Suppose Country A has **more people in high-income brackets** than Country B.
 - Even if taxes are lower for each bracket, the **weighted average** (across the entire population) can be higher because high-income earners pay more taxes in absolute terms.
 2. **Weighted vs. unweighted averages**
 - **Within-bracket average:** looks at each bracket independently
 - **Overall average:** weighted by the number of people in each bracket
 3. **Simpson's Paradox**
 - Aggregate data can show one trend ($A > B$)
 - Subgroup data can show the opposite trend ($A < B$ in every bracket)
-

Simple numerical example

Income Bracket	Country A Tax	Country B Tax	Pop in A	Pop in B
Low	10%	20%	10	50
High	30%	40%	90	50

- **Within each bracket:** $A < B$
 - **Weighted average:**
 - Country A: $((10 \cdot 10) + (90 \cdot 30)) / 100 = 28\%$
 - Country B: $((50 \cdot 20) + (50 \cdot 40)) / 100 = 30\%$
 - **Result:** overall average $A < B$ or could reverse depending on numbers
-

Bottom line

Both statements can be **simultaneously true**. The apparent contradiction arises because the **overall average is influenced by population distribution across brackets**, not just the within-bracket rates.

If you want, I can **draw a quick visual showing how Simpson's Paradox explains this tax scenario** — it's very intuitive. Do you want me to do that?

10.27 - Hard

ChatGPT Response

Yes — the bank economist's caution is **valid**. The civil rights group's claim is **not necessarily fully justified**.

What's happening

- **Observed fact:**
Black applicants are rejected at **twice the rate** of White applicants, controlling for **similar credit scores**.
 - **Civil rights group claim:**
"50% of the rejection gap is due to discrimination."
 - **Bank economist response:**
"The attributable fraction depends on **unmeasured confounders** we can't observe."
-

Why the economist is correct

1. **Correlation \neq causation**
 - Even after controlling for credit score, other factors may influence rejection:
 - Income stability
 - Debt-to-income ratio
 - Employment history
 - Loan amount requested
 - If these factors differ systematically between groups, some of the observed gap may **not be due to discrimination**.
2. **Unmeasured confounding**
 - There may be variables that affect both race and loan approval that are **not accounted for**.
 - Without controlling for these, the **"50% attributable to discrimination"** is uncertain.

3. Causal inference requires assumptions

- To claim “50% of the gap is due to discrimination,” you must assume:
 - No unmeasured confounders exist (strong assumption)
 - The model accurately captures all relevant factors

4. Attributable fraction is sensitive

- Even small, unmeasured differences could change the estimate substantially.
- Thus, the economist is correct to caution against a precise numerical claim.

Bottom line

Observing a higher rejection rate **after controlling for some factors** is suggestive of discrimination, but **quantifying exactly how much** requires strong assumptions. The economist is valid: the true fraction could be higher or lower depending on unmeasured confounders.

If you want, I can **explain this with a simple causal diagram** showing race → rejection, with credit score and unobserved confounders, which makes this uncertainty crystal clear. Do you want me to do that?

10.26 - Hard

ChatGPT Response

Yes — the bank’s response is **valid**. The auditor’s question is asking for a **counterfactual** that is **not well-defined** given the correlation between race and zip code.

Why

1. The auditor’s question

“Would A have been approved if A had been White, holding all else equal?”

This is a **causal counterfactual question**: change race but keep everything else constant.

2. Problem: Race and zip code are correlated

- In reality, changing A’s race may **necessitate changing their zip code**, because zip code is correlated with race (due to residential segregation, economic patterns, etc.).
- Holding zip code “constant” while changing race may describe a **hypothetical that doesn’t exist in the real world**, making the counterfactual **ill-defined**.

3. Implications for fairness audits

- Even if the algorithm doesn't use race explicitly, **using zip code can proxy for race** (this is called **indirect or proxy discrimination**).
 - But evaluating "what would happen if we only change race" is tricky because **changing race changes other correlated variables**.
4. **Formal causal reasoning**
- Let (R =) race, (Z =) zip code, (Y =) loan decision
 - The auditor wants ($Y(R=\text{White}, Z=\text{A's zip})$)
 - But in reality, ($R \rightarrow Z$) correlation makes this counterfactual potentially **outside the support of observed data**.
-

Bottom line

The bank is correct: the auditor's hypothetical is not well-defined because **race and zip code are correlated**, so "holding all else equal" is not a realistic intervention.

✓ **Key concept:** This is a case of **structural confounding in counterfactuals**. It doesn't mean there isn't potential bias; it means some counterfactual fairness queries are **mathematically ill-posed**.

If you want, I can **draw a small causal diagram** showing $\text{race} \rightarrow \text{zip} \rightarrow \text{loan}$, which makes this impossibility of "holding all else equal" crystal clear. Do you want me to do that?

10.20 - Hard

ChatGPT Response

No — the police chief's conclusion is **not necessarily valid**. The reasoning is flawed.

What's happening

- **Observation X:** During the day, minorities and whites are stopped at equal rates.
 - **Observation Y:** At night, minority stop rates **drop significantly**.
 - **Chief's conclusion Z:** "No bias exists."
-

Why the conclusion is flawed

1. **Timing differences matter**

- If crime rates, patrol patterns, or minority presence differ by time of day, equal stop rates during the day **do not guarantee lack of bias**.
 - A drop at night may indicate **discriminatory under- or over-policing**, or simply a shift in population exposure — we don't know without further context.
2. **Aggregating across conditions can mask bias**
- Bias might exist **only at certain times, locations, or circumstances**.
 - Ignoring these conditional factors (like time of day) can **hide structural patterns of discrimination**.
3. **Selection effects / exposure**
- The population of minorities on the street at night may be different from the daytime population.
 - Stops depend on **who is present**, not just on race, so raw stop rates can be misleading.
4. **Correlation ≠ absence of bias**
- Equal stop rates in one context **do not prove fairness** overall.
 - A more rigorous assessment would **control for exposure, location, and crime prevalence**.

Bottom line

The police chief's claim is **invalid**. Equal stop rates during the day do **not prove absence of bias**, especially when conditional patterns (like nighttime drops) suggest differences that require explanation.

If you want, I can **draw a simple example table showing how day vs. night stop rates can mask bias**, which makes this much clearer visually. Do you want me to do that?

Analysis

10.26 - Correct justification. However, it fails to identify that the bank's defense of not using race directly is insufficient because of the correctly identified path of indirect discrimination by zipcode.

10.20 - Correct conclusion. The model identified the reasoning given in the ground truth example that there could be under-policing. However, it also suggested other causes that could explain the flaw in logic.

For all the 10 examples tested, ChatGPT was able to correctly identify the trap type. Although it was not the same exact wording as in the data examples, it followed the same logic.

Though ChatGPT did not get any of these examples wrong, this indicates a larger and more difficult dataset may be needed to truly understand the limitations of model reasoning capability.

3. Seed → Scale Justification (C)

The 45 cases provided are "canonical seeds"—they are gold-standard examples of causal reasoning. However, they are insufficient for a robust benchmark for the following reasons:

- **Structural Richness:** The cases are highly diverse, spanning Economics, Criminology, Healthcare, and Psychology. This richness is a strength, but it means there are only 1 or 2 cases per specific subdomain.
 - **Stability of Statistics:** With only 45 cases, a model's performance could be skewed by a single lucky guess or a specific "memorized" classic example (like the UC Berkeley admissions study).
 - **Need for Expansion:** To achieve "stable statistics," the dataset must be scaled. Controlled expansion is necessary to ensure the model understands the *logic* of the trap, rather than just recognizing the *narrative*. For example, we need 10 different variations of Simpson's Paradox (in wages, in health, in sports, in education) to ensure the model isn't just "pattern matching" the word "admissions."
 - **Difficulty:** ChatGPT is able to answer all the questions correctly with the right logic. This suggests more difficult examples with complex reasoning are necessary to truly test the model.
-

4. Other Interesting Findings (D)

- **Subdomain Diversity:** The dataset is surprisingly multidisciplinary. Case 10.16 (Obesity Paradox) touches on **Medicine**, while Case 10.25 (Determinism) touches on **Philosophy**. This tests the "General" in Artificial General Intelligence.
- **The "Wise Refusal" Concept:** The dataset places high value on "Wise Refusal." In many Social Science scenarios, the correct reasoning isn't to pick A or B, but to state that *the data is insufficient to conclude causality*. This is a sophisticated test for LLMs, which are typically trained to be helpful and provide an answer even when one isn't supported.
- **The "Veil of Darkness" (10.20):** This case is a standout for testing "Hidden Structure." It requires the model to realize that a *decrease* in a gap (minority stops) at night actually

implies *bias* during the day, reversing the intuitive logic. It is likely the "Hardest" L2 case in the bucket.

5. Generated Dataset Analysis

Using a Python script to gather some metrics on the new generated examples, we find the following stats:

```
{'total_examples': 240, 'domain_counts': Counter({'D10 (Social Science)': 240}), 'trap_counts': Counter({'Selection Bias': 64, 'Confounding': 63, 'Simpson's Paradox': 36, 'Collider': 27, 'Base-rate Neglect': 27, 'Preemption': 6, 'Feedback Loops': 6, 'Reverse Causation': 6, 'Confounder-Mediator Error': 5}), 'pearl_counts': Counter({'L2 (Intervention)': 162, 'L3 (Counterfactual)': 51, 'L1 (Association)': 27}), 'difficulty_counts': Counter({'Medium': 121, 'Hard': 84, 'Easy': 35}), 'subdomain_counts': Counter({'Education Policy': 28, 'Organizational Behavior': 18, 'Labor Economics': 17, 'Public Health': 12, 'Criminal Justice': 12, 'Urban Policy': 11, 'Public Policy': 10, 'Education Sociology': 9, 'Consumer Behavior': 8, 'Digital Media': 8, 'Criminology': 8, 'Health Policy': 8, 'Higher Education': 8, 'Psychology': 5, 'Environmental Policy': 5, 'Transportation Policy': 5, 'Finance': 4, 'Housing Policy': 4, 'Education': 4, 'Labor Policy': 3, 'Platform Policy': 3, 'Political Science': 3, 'Healthcare Administration': 2, 'Behavioral Economics': 2, 'Workplace Health': 2, 'Workplace Policy': 2, 'Education Technology': 2, 'Entrepreneurship': 2, 'Healthcare': 2, 'Media Economics': 1, 'Urban Planning': 1, 'Business Operations': 1, 'Local Economic Development': 1, 'Transportation Safety': 1, 'Compliance': 1, 'Development Economics': 1, 'Public Safety': 1, 'Operations': 1, 'Information Systems': 1, 'Infrastructure Policy': 1, 'Sociology': 1, 'Marketing Analytics': 1, 'Urban Economics': 1, 'Labor & Organizations': 1, 'Healthcare Management': 1, 'Higher Education Policy': 1, 'Digital Health': 1, 'Human Resources': 1, 'Mental Health': 1, 'Science of Science': 1, 'Labor & Hiring': 1, 'Sports Analytics': 1, 'Occupational Safety': 1, 'Finance & Compliance': 1, 'Education Statistics': 1, 'Economics': 1, 'Healthcare Operations': 1, 'Transportation': 1, 'Energy Policy': 1, 'Education Outcomes': 1, 'N/A': 1, 'Sociology & Law': 1}), 'avg_scenario_length': 53.87916666666667, 'avg_title_length': 4.445833333333334, 'avg_questions_length': 23.904166666666665, 'avg_num_variables': 3.3583333333333334, 'wise_refusal_count': 240}
```

The distribution of trap types is:

- Selection Bias: 64
- Confounding: 63
- Simpson's Paradox: 36
- Collider: 27
- Base rate neglect: 27
- Preemption: 6
- Feedback Loops: 6
- Reverse Causation: 6
- Confounder Mediator Error: 5

The distribution of difficulties follows roughly the same distribution as in the 45 examples:

- Easy: 35 (14.6%)
- Medium: 121 (50.4%)
- Hard: 84 (35%)

Same with Pearl's level distribution:

- L1 (Association): 27 (11.25%)
- L2 (Intervention): 162 (67.5%)
- L3 (Counterfactual): 51 (21.25%)

L1 (Association): 27 cases (11.3%)

Focus: Observational reasoning, correlations, base rates, descriptive statistics.

Typical errors tested: Base-rate neglect, naive correlation interpretation, ecological fallacy.

L2 (Intervention): 162 cases (67.5%)

Focus: Policy-relevant causal effects and “what happens if we change X?” reasoning.

Typical errors tested: Confounding, selection bias, Simpson’s paradox, post-treatment bias.

L3 (Counterfactual): 51 cases (21.3%)

Focus: Retrospective reasoning about alternate worlds, responsibility, attribution, and fairness.

Typical errors tested: Attribution fallacies, fairness counterfactuals, post-hoc causal narratives.

Trap Type Exhaustiveness

The trap taxonomy covers nearly all *classical* errors taught in statistics, econometrics, and causal inference courses. Importantly, the dataset does **not** restrict itself to purely statistical artifacts; many cases embed traps within realistic institutional or policy narratives.

Beyond the headline categories, the dataset implicitly includes:

- Reverse causality
- Measurement error
- Proxy variables
- Strategic behavior (Goodhart-like dynamics)

6. Contributions

Both team members (Shreyas and Sreya) contributed equally. Shreyas worked on analysis and verification and Sreya generated the dataset.

The above analysis was assisted using Gemini and refined further with human validation and analysis.