T3 Annotation Cheat Sheet

Labeling Rules for Ground Truth, Pearl Levels, Trap Types, and Subtype

Table 1: Definition of ground-truth labels, illustrative examples, and their relationship to trap types.

Label	Definition	Example	Trap Type Allowed
YES	The claim is supported *as stated* by the given scenario under the appropriate Pearl level.	"Higher consumer sentiment is associated with higher stock prices," based on an observational study reporting a positive correlation.	No (`trap = NONE`)
NO	The claim is invalid *as stated* due to a violated causal or statistical assumption.	"Raising the minimum wage increases employment," inferred from observational data affected by policy endogeneity.	Yes (exactly one)
AMBIGUOUS	The claim cannot be definitively evaluated given the available information.	"The policy caused economic growth," without specifying timing, controls, or comparison group.	No (`trap = NONE`)

Table 2: Summary of trap subtypes by Pearl level, with representative examples. Each instance has exactly one trap type and at most one subtype.

Pearl Level	Trap Type	Trap Subtype	Example
L1: Association	Confounding	Confounding by Indication	"Sicker patients receive Drug A and also have higher mortality."
		Omitted Variable	"Ice cream sales correlate with drowning due to temperature."
		Socioeconomic	"Wealthier students attend tutoring and score higher."
	Reverse Causation	Outcome-driven Selection	"Firms invest more because profits are already rising."
		Policy Endogeneity	"Minimum wage is raised when the economy is improving."
	Selection Bias	Sampling-on-the-Outcome	"Only successful startups are analyzed."
		Attrition Bias	"Lower-performing students drop out of the study."
	Collider	Conditioning on Participation	"Among admitted students, test scores and essays appear correlated."
		Case-Control Sampling	"Cases and controls are selected based on disease status."
	Simpson's Paradox	Aggregation Bias	"Treatment looks harmful overall but helpful in every age group."
		Imbalanced Group Composition	"Hospital A treats sicker patients overall."
	Regression to the Mean	Extreme-Group Selection	"Lowest scorers improve on the next test."
		Noise-Induced Extremes	"Outliers regress after random measurement error."
	Survivorship Bias	Selective Observation	"Only companies that survived a recession are studied."
		Historical Filtering	"We observe only technologies that remained popular."
	Base-rate Neglect	Prior Ignorance	"A positive test is assumed to imply disease despite rarity."
		Conditional Fallacy	"P(Disease—Positive) confused with P(Positive—Disease)."
	Goodhart's Law	Static Metric Gaming	"Teaching to the test improves scores but not learning."
		Proxy Drift	"Click-through rate stops reflecting user satisfaction."
L2: Intervention	Confounding	Unblocked Backdoor	"Intervention fails to block socioeconomic influence."
		Time-varying Confounding	"Past outcomes affect future treatment assignment."
	Reverse Causation	Reactive Intervention	"Policy enacted because outcomes were already worsening."
	Selection Bias	Post-intervention Selection	"Only compliant patients are analyzed."
	Collider	Conditioning on Compliance	"Among those who followed treatment, outcomes differ."
	Confounder–Mediator Error	Mediator Adjustment Error	"Controlling for a variable caused by the intervention."
	Simpson's Paradox	Stratified Intervention Reversal	"Policy helps overall but harms every subgroup."
	Goodhart's Law	Policy Target Gaming	"Hospitals optimize wait-time metrics, not care quality."
	Feedback Loops	Policy–Response Loop	"Drivers reroute in response to congestion pricing."
L3: Counterfactual	Preemption	Early Preemption	"Fire is extinguished before the sprinkler activates."
		Late Preemption	"Backup generator would have powered the system later."
	Confounding	Cross-world Confounder	"Motivation differs between actual and hypothetical worlds."
	Reverse Causation	Outcome-dependent Worlds	"Knowing the outcome constrains the counterfactual."
	Confounder–Mediator Error	Mediator Fixing Error	"Holding recovery constant while changing treatment."
	Feedback Loops	Dynamic World Divergence	"Small change alters long-term system evolution."
	Selection Bias	Counterfactual Conditioning	"Conditioning on survival when asking what would have happened."

Step 1: Identify the Pearl Level

- **L1 (Association)**: Observational correlation only.
- **L2 (Intervention)**: Explicit or implicit `do(X)`.
- **L3 (Counterfactual)**: "What would have happened if X had not occurred?"

Step 2: Decide the Label

- **YES**: Claim follows from stated information.

Table 3: Pearl levels, their causal semantics, and representative examples.

Pearl Level	Definition	Example
L1: Association	Observational relationships of the form $P(Y \mid X)$ without intervention.	"People who exercise more have lower blood pressure."
L2: Intervention	Interventional claims involving do(X) and causal effects of actions.	"If we increase the minimum wage, employment will rise."
L3: Counterfactual	Counterfactual reasoning across hypothetical worlds.	"Had the policy not passed, unemployment would have increased."

Table 4: Trap types as primary causal failure modes, with examples.

Trap Type	Core Definition	Example
Confounding	A common cause affects both exposure and outcome.	"Sicker patients receive Drug X and also have higher mortality."
Reverse Causation	Outcome (or its causes) influences the exposure.	"Cities raise minimum wage because the economy is already improving."
Selection Bias	Conditioning on a non-random subset distorts inference.	"Only successful startups are analyzed."
Collider Bias	Conditioning on a common effect induces spurious association.	"Among admitted students, test scores and essays appear negatively correlated."
Simpson's Paradox	Aggregated trends reverse within subgroups.	"Treatment A looks worse overall but better in every age group."
Regression to the Mean	Extreme observations regress toward the average.	"Top scorers drop and low scorers improve next year."
Survivorship Bias	Failures are systematically unobserved.	"Studying only companies that survived a recession."
Goodhart's Law	Optimizing a proxy breaks its correlation with the target.	"Teaching to the test improves scores but not learning."
Base-rate Neglect	Ignoring prior probabilities.	"A positive test is assumed to imply disease despite rarity."
Feedback Loops	Bidirectional or adaptive causation.	"Traffic policy changes driving behavior that alters congestion."
Preemption	An alternative cause prevents another from acting.	"A backup system would have failed if the primary one had not."

- **NO**: Claim is invalid due to a causal trap.
- **AMBIGUOUS**: Missing information is *critical* to causal validity.

Step 3: Assign Exactly One Trap Type (NO cases only)

Use this decision order strictly:

1. Is there a missing common cause? → **CONFOUNDING**
2. Does outcome (or its causes) influence exposure? → **REVERSE**
3. Are we conditioning on a selected or filtered sample? → **SELECTION / COLLIDER**
4. Does aggregation reverse subgroup trends? → **SIMPSONS**
5. Are extremes selected and naturally reverting? → **REGRESSION**
6. Is a proxy optimized instead of the target? → **GOODHART**
7. Is causation bidirectional or adaptive? → **FEEDBACK**
8. (L3 only) Is an alternative cause preempting the hypothesized one? → **PREEMPTION**

Subtype Rule

- Assign a subtype **only after** trap type is fixed.
- Use a subtype only if it captures a recurring mechanism.
- If unsure, leave subtype empty.

Disambiguation Rules (Most Common Confusions)

- **Confounding vs Reverse**:
  $Z \to X$ and $Z \to Y \Rightarrow$ Confounding;
  Y (or its causes) $\to X \Rightarrow$ Reverse.
- **Regression vs Confounding**:
  No causal variable needed $\Rightarrow$ Regression;
  Latent variable needed $\Rightarrow$ Confounding.
- **Simpson's vs Selection**:
  Aggregation reversal $\Rightarrow$ Simpson's;
  Who enters dataset matters $\Rightarrow$ Selection.

Adjudication Principle

1. If annotators disagree, select the trap that explains the error with the **minimal causal graph (fewest nodes/edges)**. 2. One instance $\to$ One Pearl level $\to$ One trap type $\to$ Optional subtype.

Listing 1: Example JSON instance (Bucket B)

```
{
  "id": "T3-BucketB-0041",
  "bucket": "BucketLarge-B",
  "pearl_level": "L1",
  "domain": "Economics",
  "scenario": "States that raised minimum wage saw employment increase the
      following year.",
  "claim": "Raising the minimum wage increases employment.",
  "label": "NO",
  "is_ambiguous": false,
  "trap": {
    "type": "REVERSE",
    "subtype": "Policy_Endogeneity"
  },
  "variables": {
    "X": "Minimum wage increase",
    "Y": "Employment level",
    "Z": ["Economic growth"]
  },
  "gold_rationale": "Minimum wage increases are often enacted in response
     to improving economic conditions, which independently raise
     employment. The observed association does not establish causation.",
  "annotation": {
    "author": ABC,
    "num_annotators": 3,
    "adjudicated": true
  }
}
```