CS372 Assignment 2: T^3 Benchmark Expansion

CS372: Artificial General Intelligence for Reasoning, Planning, and Decision Making

Winter 2026

1 Overview

This assignment builds upon Assignment 1 by expanding the T^3 benchmark dataset with additional cases across all three Pearl levels. You will work in the same groups as Assignment 1 to generate cases that meet specific distribution requirements. - Each student validates a different student's file - Each student's file is validated by at least one other student

2 Target Distribution

The overall target distribution for Assignment 2 is:

- **L1 (Association)**: 500 cases
- **L2 (Intervention)**: 3,000 cases
- **L3 (Counterfactual)**: 1,500 cases
- **Total**: 5,000 cases

2.1 Group Assignment Summary

There are 10 groups (A-J), with each group responsible for generating 500 cases. The distribution across Pearl levels and domains for each group is shown in Table 1.

Assignment List: The complete group assignment list with student information is available at: Assignment Group Formulation (Google Sheets)

2.2 Cross-Validation Dataset Assignment

Each student has been assigned a dataset from another student in their group to validate. Your cross-validation assignment, including the assigned dataset file and target case counts, can be found in the Cross-Validation Assignment Form (Google Sheets).

Important: You must **ONLY download your own folder** from the assignment repository. Your folder is named `group{Letter}_{YourName}` and contains the dataset file you are assigned to validate.

Group	Domain	L1	L2	L3	Total
A	Medicine	50	300	150	500
B	Economics	50	300	150	500
C	Law & Ethics	50	300	150	500
D	Sports	50	300	150	500
E	Daily Life	50	300	150	500
F	History	50	300	150	500
G	Markets	50	300	150	500
H	Environment	50	300	150	500
I	AI & Tech	50	300	150	500
J	Social Science	50	300	150	500
Total		**500**	**3,000**	**1,500**	**5,000**

Table 1: Group Assignment Summary: Cases per Group by Pearl Level and Domain

2.2.1 Handling Case Count Differences

The assigned dataset may have a different number of cases than the target requirement. Please follow these guidelines:

- **If the assigned number is smaller than the target number:**
  - You must generate additional cases to reach the target number (170 cases total)
  - The new cases should follow the same distribution requirements: 17 L1 cases, 102 L2 cases, 51 L3 cases
  - All new cases must include proper validation fields: `initial_author`, `validator`, and `final_score`
- **If the assigned number is larger than the target number:**
  - You must validate **all cases** in the assigned dataset
  - You do not need to generate additional cases
  - Your final submission should include all validated cases from the assigned dataset

3 Deliverables

3.1 L1: Association Cases (500 cases)

For L1, you need to generate 500 cases total, distributed as follows:

- **Wolf cases**: 250 cases (cases that appear valid but contain causal reasoning traps)
- **Sheep cases**: 200 cases (cases with strong evidence for valid causal claims)
- **Ambiguous cases**: 50 cases (cases where the causal relationship is unclear or conditional)

3.1.1 WOLF Types (Selection Family)

The WOLF types focus on traps that make invalid causal claims appear valid. Table 2 shows the implementability assessment for WOLF types organized by family.

Trap Type	Tier	Status	Rationale
Selection Family (Specific → General)			
W1: Selection Bias	Core	Full	Describe sampling; LLM recognizes non-representative samples
W2: Survivorship Bias	Core	Full	Describe "only survivors observed" and missing failures
W3: Healthy User Bias	Core	Full	Describe self-selection into X and correlated lifestyle factors
W4: Regression to Mean	Adv.	Partial	Requires statistical intuition; needs careful phrasing
Ecological Family (General → Specific)			
W5: Ecological Fallacy	Core	Full	Describe aggregate correlation used to claim individual causation
W6: Base Rate Neglect	Adv.	Partial	Must provide base rates and test properties in text
Confounding Family			
W7: Confounding	Core	Full	Describe Z; LLM recognizes Z causes both X and Y
W8: Simpson's Paradox	Adv.	Partial	Must provide subgroup and aggregate numbers in text
Direction Family			
W9: Reverse Causation	Core	Full	Describe X and Y with plausible reverse direction
W10: Post Hoc Fallacy	Core	Full	Describe timing-based inference without controls or mechanism

Table 2: WOLF Types Implementability Assessment (Organized by Family)

3.1.2 SHEEP Types

The SHEEP types focus on evidence that supports valid causal claims. Table 3 shows the implementability assessment for SHEEP types.

Evidence Type	Tier	Status	Rationale
S1: RCT	Core	Full	Describe random assignment and control group
S2: Natural Experiment	Core	Full	Describe exogenous event and comparison group
S3: Lottery/Quasi-Random	Core	Full	Describe random allocation among applicants
S4: Controlled Ablation	Core	Full	Describe removal of X while holding other factors constant
S5: Mechanism + Dose	Core	Full	Describe known pathway plus dose-response gradient
S6: Instrumental Variable	Adv.	Partial	Requires IV logic to be described cleanly
S7: Diff-in-Diff	Adv.	Partial	Requires time and control group with parallel pre-trends
S8: Regression Discont.	Adv.	Partial	Requires cutoff assignment and local comparison

Table 3: SHEEP Types Implementability Assessment

3.2 L2: Intervention Cases (3,000 cases)

For L2, you need to generate 3,000 cases organized by Family Type and Trap Type. The distribution is shown in Table 4.

3.3 L3: Counterfactual Cases (1,500 cases)

For L3, you need to generate 1,500 cases with two distribution requirements:

3.3.1 By Domain (10 Domains)

Target: 150 cases per domain (±10%). Each trap type should have cases from at least 5 domains.

3.3.2 By Family (8 Families)

The L3 cases are also organized by family type as shown in Table 6.

Note: The family distribution totals 1,000 cases, with the remaining 500 cases distributed across domains to meet the 1,500 total requirement.

4 Group Assignments

You will work in the same groups as Assignment 1. While you are organized into groups for coordination and validation purposes, **all submissions must be done individually**. Each student is responsible for generating their own cases according to the distribution requirements above, with specific focus on your assigned domain and trap types from Assignment 1.

5 Submission Guidelines

5.1 Submission Requirements

All submissions must be done individually. Each student must submit their own work.

For individual submission, you are required to generate:

- **170 cases total minimum** (more cases will receive bonus points)
- Distribution: 17 L1 cases, 102 L2 cases, 51 L3 cases

5.2 What to Submit

You must submit the following files for Assignment 2:

5.2.1 Dataset Files

All dataset files should be named using the format: `group{Letter}_{StudentName}_{Type}.json`, where {`Type`} is one of `schema`, `score`, or `dataset`.

1. **Schema File:** `group{Letter}_{StudentName}_schema.json`
    - A summarized schema of your dataset structure
    - Should include field definitions, types, and examples

- Documents the structure of all cases in your dataset

2. **Score File:** `group{Letter}_{StudentName}_score.json`
   - Contains quality scores for the previous dataset (from Assignment 1) that you validated
   - Each case should include the following scoring fields:
     - **Scenario clarity** (2 points): X, Y, Z clearly defined
     - **Hidden question quality** (2 points): Identifies key ambiguity
     - **Conditional answer A** (1.5 points): Logically follows from condition A
     - **Conditional answer B** (1.5 points): Logically follows from condition B
     - **Wise refusal quality** (2 points): Follows template
     - **Difficulty calibration** (1 point): Label matches complexity
     - **Total** (10 points): ≥ 8 accept; 6–7 revise; < 6 reject
3. **Final Dataset File:** `group{Letter}_{StudentName}_dataset.json`
   - Your final validated dataset with **170 cases minimum** (more cases will receive bonus points)
   - Each case must include the following fields:
     - **initial_author**: The student who originally created the case
     - **validator**: The student who validated this case
     - **final_score**: The quality score assigned during validation
   - All standard case fields: Scenario, Variables, Annotations, Hidden Timestamp, Conditional Answers, Wise Refusal
4. **Your coding pipeline (if used)**

5.2.2 Analysis Report

Submit a PDF report (maximum 10 pages) that includes the following sections:

1. **Summary of Unvalidated vs. Validated Dataset**
   - Comparison of dataset characteristics before and after validation
   - Key improvements and changes made during validation
2. **Pearl Level Distribution**
   - Distribution of cases across L1 (Association), L2 (Intervention), and L3 (Counterfactual)
   - Comparison between unvalidated and validated datasets
3. **Label Distribution**
   - **L1**: Yes/No/Ambiguous label distribution
   - **L2**: All cases should be labeled as "No" (invalid causal claims)
   - **L3**: Valid/Invalid/Conditional label distribution
   - Comparison between unvalidated and validated datasets

4. **Trap Type Distribution**
    - **L1**: Distribution across 10 wolf cases, 8 sheep cases, and 2 ambiguous cases
    - **L2**: Distribution across 17 trap types (T1–T17)
    - **L3**: Distribution across 8 families (F1–F8) and their subtypes
    - Comparison between unvalidated and validated datasets
5. **Difficulty Level Distribution**
    - **L1**: Roughly 1:2:1 ratio (Easy:Medium:Hard)
    - **L2**: Roughly 1:2:1 ratio (Easy:Medium:Hard)
    - **L3**: Roughly 1:2:1 ratio (Easy:Medium:Hard)
    - Comparison between unvalidated and validated datasets
6. **Score Summary**
    - Summary of quality scores for unvalidated dataset
    - Summary of quality scores for validated dataset
    - Analysis of score improvements and validation impact
7. **Prompt Setup**
    - Description of the prompt engineering approach used
    - LLM configuration and parameters
    - Generation methodology and quality control measures
8. **Example Case**
    - Include at least one complete example case from your validated dataset
    - Should demonstrate all required fields and proper structure

5.3 Deadline

Submission Deadline: January 28, 2026, 11:59 PM PST

All submissions must be uploaded to Gradescope by the deadline. Late submissions will be subject to the course late policy.

6 Contact

- **Instructor:** Prof. Edward Y. Chang
    - Email: chang@stanford.edu
- **Course Assistant:** Longling Gloria Geng
    - Email: gll2027@stanford.edu

Good luck with your assignment2!

A Scoring Method

All cases across L1 (Association), L2 (Intervention), and L3 (Counterfactual) levels should be evaluated using a quality scoring system. The scoring method ensures consistency and quality across all generated cases.

A.1 10-Point Quality Scoring Rubric

The detailed scoring rubric is used for evaluating cases across all three Pearl levels (L1, L2, L3), with each case receiving a score out of 10 points based on the following criteria:

A.2 Level-Specific Scoring Considerations

In addition to the 10-point rubric above, the following level-specific elements must be considered when scoring cases:

A.3 Scoring Guidelines

- **Acceptance Threshold:** Cases with a total score of ≥ 8 points should be accepted
- **Revision Threshold:** Cases with a total score of 6–7 points should be revised
- **Rejection Threshold:** Cases with a total score of < 6 points should be rejected, regenerated, and re-scored

A.4 Application Across Pearl Levels

The 10-point quality scoring rubric specified in Table 7 applies to all three Pearl levels. Additionally, level-specific considerations from Table 8 must be evaluated:

- **L1 (Association):** Evaluate using the 10-point rubric plus verify final label (W/S/A) matches trap type (W1–W10 or S1–S8) and that the trap is correctly implemented
- **L2 (Intervention):** Evaluate using the 10-point rubric plus verify all cases are labeled "No" and trap type (T1–T17) correctly matches the family and difficulty level
- **L3 (Counterfactual):** Evaluate using the 10-point rubric plus verify final label (V/I/C) matches counterfactual validity and trap type (F1–F8) is correctly classified

All validated cases in your final dataset must include the `final_score` field, which should reflect the quality score assigned during the validation process.

Family Type	Easy	Med	Hard	Total
F1: Selection				
T1: SELECTION	55	90	55	200
T2: SURVIVORSHIP	50	80	50	180
T3: COLLIDER	45	70	45	160
T4: IMMORTAL TIME	40	60	40	140
Subtotal	*190*	*300*	*190*	*680*
F2: Statistical				
T5: REGRESSION	50	80	50	180
T6: ECOLOGICAL	45	70	45	160
Subtotal	*95*	*150*	*95*	*340*
F3: Confounding				
T7: CONFOUNDER	60	100	60	220
T8: SIMPSON'S	50	80	50	180
T9: CONF-MED	55	90	55	200
Subtotal	*165*	*270*	*165*	*600*
F4: Direction				
T10: REVERSE	55	90	55	200
T11: FEEDBACK	45	70	45	160
T12: TEMPORAL	40	60	40	140
Subtotal	*140*	*220*	*140*	*500*
F5: Information				
T13: MEASUREMENT	50	80	50	180
T14: RECALL	45	70	45	160
Subtotal	*95*	*150*	*95*	*340*
F6: Mechanism				
T15: MECHANISM	50	80	50	180
T16: GOODHART	50	80	50	180
T17: BACKFIRE	50	80	50	180
Subtotal	*150*	*240*	*150*	*540*
Total	**835**	**1,330**	**835**	**3,000**
Percentage	**(27.8%)**	**(44.3%)**	**(27.8%)**	**(100%)**

Table 4: L2 Distribution by Family Type and Trap Type

Domain	Description	Target
D1	Daily Life	150
D2	Health/Medicine	150
D3	Economics	150
D4	Law & Ethics	150
D5	Sports	150
D6	History	150
D7	Markets	150
D8	Environment	150
D9	AI & Tech	150
D10	Social Science	150
Total		**1,500 (150 per domain)**

Table 5: L3 Distribution by Domain

Family	Description	Current	Target	Priority
F1	Deterministic	19	150	Normal
F2	Probabilistic	5	120	High
F3	Overdetermination	6	100	High
F4	Structural	9	120	Normal
F5	Temporal	8	100	Normal
F6	Epistemic	13	120	Normal
F7	Attribution	16	140	Normal
F8	Moral/Legal	4	100	High
Subtotal (Theoretical)		*80*	*950*	
DomainExt	Domain extensions	20	50	Low
Total		**100**	**1,000**	

Table 6: L3 Distribution by Family

Criterion	Description	Points
Scenario clarity	X, Y, Z clearly defined	1.0
Hidden question quality	Identifies key ambiguity	1.0
Conditional answer A	Logically follows from condition A	1.5
Conditional answer B	Logically follows from condition B	1.5
Wise refusal quality	Follows template	2.0
Difficulty calibration	Label matches complexity	1.0
Final label	Correct label for level (W/S/A for L1, None for L2, V/I/C for L3)	1.0
Trap type	Correct trap type classification (see Table 8)	1.0
Total		**10.0**

Table 7: 10-Point Quality Scoring Rubric

Level	Final Label	Trap Type	Additional Considerations
L1 (Association)	W/S/A (Wolf/Sheep/Ambiguous)	W1–W10 (Wolf types) or S1–S8 (Sheep types) or A(Ambiguous)	Label must correctly identify trap type; WOLF cases should appear valid but contain traps; SHEEP cases should have strong evidence
L2 (Intervention)	None (all cases labeled "No")	T1–T17 (17 trap types across 6 families)	All L2 cases must be invalid causal claims; trap type must match family (F1–F6) and difficulty level
L3 (Counterfactual)	V/I/C (Valid/Invalid/Conditional)	F1–F8 (8 families with subtypes)	Label must match counterfactual validity; trap type must align with family classification

Table 8: Level-Specific Scoring Considerations for L1, L2, and L3