CS372 Assignment 1: T^3 Benchmark Analysis

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

Winter 2026

1 Assignment Overview

This assignment focuses on analyzing and working with the T^3 (T-cubed) Benchmark datasets. The T^3 Benchmark is designed to test reasoning capabilities across different domains and Pearl's Causality Hierarchy levels (Association, Intervention, and Counterfactual).

2 Group Structure

There will be **20 groups** for this assignment, with approximately 6 students per group (120 students total). Each of the 10 BenchmarkT3-BucketLarge files will be assigned to **2 groups** for cross-validation purposes.

2.1 Group Assignments

The following table shows the group assignments, domains, signature traps, focus areas, dataset information, and benchmark files:

Group	Domain	Signature Trap	Focus	#	Target	Benchmark File
A1, A2	Medicine	Indication Bias	Intervention	46	460	`BenchmarkT3-BucketLarge-A.pdf`
B1, B2	Economics	Equilibrium Effects	Intervention	46	460	`BenchmarkT3-BucketLarge-B.pdf`
C1, C2	Law Ethics	Attr. & Preemption	Counterfactual	46	460	`BenchmarkT3-BucketLarge-C.pdf`
D1, D2	Sports	Outcome Bias	Counterfactual	46	460	`BenchmarkT3-BucketLarge-D.pdf`
E1, E2	Daily Life	Regression to Mean	Association	45	450	`BenchmarkT3-BucketLarge-E.pdf`
F1, F2	History	Survivorship Bias	Association	45	450	`BenchmarkT3-BucketLarge-F.pdf`
G1, G2	Markets	Self-Fulfilling	Intervention	45	450	`BenchmarkT3-BucketLarge-G.pdf`
H1, H2	Environment	Feedback Loops	Intervention	45	450	`BenchmarkT3-BucketLarge-H.pdf`
I1, I2	AI & Tech	Goodhart's Law	Association	45	450	`BenchmarkT3-BucketLarge-I.pdf`
J1, J2	Social Sci.	Simpson's Paradox	Association	45	450	`BenchmarkT3-BucketLarge-J.pdf`
			Total	**454 to 4540**		

Table 1: Group assignments with domains, signature traps, focus areas, dataset sizes, and benchmark files

Note: Groups listed together (e.g., Groups A1 and A2) are cross-validation pairs that will swap datasets in Assignment 2.

2.2 Cross-Validation for Assignment 2

In Assignment 2, groups that worked on the same BenchmarkT3-BucketLarge file in Assignment 1 will **swap their datasets** for cross-validation:

- Groups A1 and A2 will exchange their datasets
- Groups B1 and B2 will exchange their datasets
- Groups C1 and C2 will exchange their datasets
- Groups D1 and D2 will exchange their datasets
- Groups E1 and E2 will exchange their datasets
- Groups F1 and F2 will exchange their datasets
- Groups G1 and G2 will exchange their datasets
- Groups H1 and H2 will exchange their datasets
- Groups I1 and I2 will exchange their datasets
- Groups J1 and J2 will exchange their datasets

This cross-validation approach ensures that each group validates the work of another group on the same dataset, providing robust evaluation and learning opportunities.

2.3 Group Formation

Students will form groups by selecting their preferred group assignment through a Google Form. The form will allow students to choose which benchmark file and category they would like to work on. Each group will have approximately 6 students.

Google Form Link: Group Selection Form

Please complete the form by the deadline specified in the course schedule. Slot is first come and first set.

3 Assignment Files

Each BenchmarkT3-BucketLarge file contains a collection of reasoning cases organized by domain and Pearl's Causality Hierarchy levels. The files are structured as follows:

3.1 File Structure

Each `BenchmarkT3-BucketLarge-*.pdf` file contains:

- **Bucket Overview:** Domain description, core themes, signature trap types, and case distribution
- **Pearl Level 1 Cases (Association):** Cases focusing on observational relationships

- **Pearl Level 2 Cases (Intervention):** Cases requiring understanding of interventions and causal effects
- **Pearl Level 3 Cases (Counterfactual):** Cases involving counterfactual reasoning

3.2 Case Format

Each case typically includes:

- **Scenario:** A description of the situation or problem
- **Variables:** Key variables involved in the causal reasoning
- **Annotations:** Additional context or background information
- **Questions:** Reasoning questions to be answered
- **Expected Analysis:** The type of reasoning required

4 Assignment Instructions

4.1 Objectives

1. Analyze the assigned BenchmarkT3-BucketLarge file
2. Understand the causal reasoning challenges presented in each case
3. Identify the types of reasoning required (Association, Intervention, Counterfactual)
4. Apply the $\mathbf{T^3}$ architecture principles learned in class
5. Document your analysis and findings

4.2 Key Concepts to Apply

When working on your assigned file, consider:

- **Pearl's Causality Hierarchy:** Association $\rightarrow$ Intervention $\rightarrow$ Counterfactual
- **$\mathbf{T^s}$ Architecture:** Sycophancy and Skepticism mechanisms
- **Causal Reasoning:** Understanding cause-effect relationships
- **Confounding Variables:** Identifying and handling confounders
- **Selection Bias:** Recognizing and addressing selection issues
- **Collider Bias:** Understanding collider structures
- **Instrumental Variables:** Using instruments for causal inference

4.3 Deliverables

Each group should prepare a pdf and file, including:

- A comprehensive analysis of their assigned benchmark file (450 or 460 instances)
- Identification and classification of reasoning types in the cases
- Discussion of the causal reasoning challenges
- Application of T^3 architecture concepts
- A summary report of findings
- Each individual's participation

5 Resources

5.1 Course Materials

Refer to the following course materials:

- Lecture slides on T^3 Architecture for Sycophancy and Skepticism
- AGI Book, Volume #2, Chapters 6 and 7
- Lecture on Pearl's Causality Hierarchy
- Assignment #1 Specification (from Lecture 3)

5.2 Additional Reading

- Multi-LLM Collaborative Intelligence (MACI), The Path to AGI
- Course readings on SocraSynth.com

6 Submission Guidelines

6.1 Dataset Submission Format

For NLP reasoning and evaluation purposes, your expanded dataset must be submitted in **machine-readable text formats** (NOT PDF). This is essential because NLP systems require text-based data that can be programmatically processed.

6.1.1 Required Dataset Formats

Your expanded dataset must be submitted in one of the following **text-based formats**:

- **JSON (.json): Required format** for NLP reasoning. Structured JSON with all required fields preserved. This format is most compatible with automated evaluation systems and NLP pipelines.

Important: PDF files are NOT acceptable for dataset submissions. PDFs cannot be easily processed by NLP reasoning systems and automated evaluation tools. Only text-based, machine-readable formats will be accepted.

6.1.2 Required Dataset Structure

Your submitted cases must follow the same structure as the original benchmark cases, including:

- **Scenario:** A clear description of the situation or problem
- **Variables:** Key variables with their roles (Treatment, Outcome, Confounder, etc.)
- **Annotations:** Structured metadata including:
  - Case ID
  - Pearl Level (L1: Association, L2: Intervention, L3: Counterfactual)
  - Domain
  - Trap Type
  - Trap Subtype (if applicable)
  - Difficulty level
  - Subdomain
  - Causal Structure
  - Key Insight
- **Hidden Timestamp:** A question that reveals temporal/causal ordering
- **Conditional Answers:** "Answer if..." sections for different scenarios
- **Wise Refusal:** A response that identifies missing information or potential biases

6.1.3 Dataset Format Requirements

- All cases must include all required fields listed above
- Variable notation should be consistent (e.g., X for treatment, Y for outcome, Z for confounders)
- Case IDs should follow the same numbering scheme as the original dataset
- Pearl Level classifications must be accurate and consistent
- All metadata fields must be populated for each case
- The file must be machine-readable (no PDF, no scanned documents, no images)

6.2 Analysis Report Submission Format

Your analysis report (separate from the dataset) may be submitted in:

- **PDF (.pdf):** For the written analysis and discussion

6.3 General Submission Requirements

- Submit **two separate files:**
  - Your expanded dataset (in JSON - **NOT PDF**)
  - Your analysis report (PDF)
- Clearly name your files (e.g., `GroupA1_dataset.json` and `GroupA1_report.pdf`)

7 Important Dates

- **Assignment #1 Out:** January 7, 2026 (Lecture 2)
- **Assignment Group Formulation Due:** January 8, 2026
- **Assignment #1 Due:** January 14, 2026 (Lecture 4)

8 Contact

For questions about this assignment, please contact:

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

Good luck with your assignment1!