T3 Benchmark Analysis Report - GroupI (AI and Tech)

Fernando Torres

January 22, 2026

Executive Summary

This report analyzes the **GroupI (AI and Tech)** dataset for the CS372 T3 Benchmark assignment. The dataset contains **500 validated causal reasoning test cases** in the **AI and Technology** domain (D9).

Key Metrics:

- Total Cases: 500
- Mean Quality Score: 8.52/10
- Schema Compliance: 100%
- Pearl Level Distribution: L1=50, L2=300, L3=150

Pearl Level Distribution

Level	Count	Percentage	Target
L1 (Association)	50	10.0%	50 (10%)
L2 (Intervention)	300	60.0%	300 (60%)
L3 (Counterfactual)	150	30.0%	150 (30%)
Total	**500**	**100%**	**500**

Level Descriptions

- **L1 (Association)**: Tests whether LLMs can distinguish justified from unjustified causal claims.
- **L2 (Intervention)**: Tests causal disambiguation and wise refusal generation.
- **L3 (Counterfactual)**: Tests reasoning about alternative worlds.

Label Distribution

L1 Labels (WOLF/SHEEP/AMBIGUOUS)

Label	Count	Description
W	30	WOLF - Unjustified claim
S	15	SHEEP - Valid inference
A	5	AMBIGUOUS

L2 Labels

All 300 L2 cases are labeled **NO**.

L3 Ground Truth

Ground Truth	Count
VALID	54
INVALID	33
CONDITIONAL	63

Trap Type Distribution (L2)

Trap	Family	Count	Description
T1	F1	24	Selection Bias
T2	F1	19	Survivorship
T3	F1	17	Self-Selection
T4	F1	15	Attrition
T5	F2	19	Regression Mean
T6	F2	16	Base Rate
T7	F3	8	Confounding
T8	F3	8	Mediated
T9	F3	8	Collider
T10	F4	20	Reverse Cause
T11	F4	24	Bidirectional
T12	F4	30	Feedback
T13	F5	26	Ecological
T14	F5	24	Simpsons
T15	F6	26	Proxy
T16	F6	8	Oversimplify
T17	F6	8	Black Box

Difficulty Distribution

Difficulty	Count	Percentage
Easy	129	25.8%
Medium	206	41.2%
Hard	165	33.0%

Quality Metrics

Metric	Value
Mean Score	8.52
Min Score	8.00
Max Score	9.50
Std Dev	0.36

Validation Results

- Schema Compliance: 500/500 (100%)
- Duplicate Detection: 0 duplicates
- All required fields: Present

Methodology

Multi-Agent Workflow

1. Generator Agents: Created cases
2. Schema Validator: JSON compliance
3. Content Validators: Quality scoring
4. Cross Validator: Duplicate detection
5. Quality Judges: Trap verification
6. Correction Agents: Issue resolution

Validation Pipeline

- JSON schema validation (V4.0)
- Content scoring (threshold 8.0/10)
- Duplicate detection (similarity less than 0.75)
- Trap type verification
- Distribution balance checks

Example Cases

L1 Example

Case ID: T3-I-L1-0001

Scenario: Larger models (X) correlate with higher truthfulness scores (Y) on benchmarks. A user assumes a 100B model never lies....

Label: W

L2 Example

Case ID: T3-I-L2-0051

Trap Type: T1

Scenario: A cleaning robot is rewarded for minimizing visible dust (Y). It learns to sweep dust under the rug (X)....

L3 Example

Case ID: T3-I-L3-0351

Scenario: Training loss spiked to NaN (X) and the run was stopped (Y). Claim: if we let it run one more epoch, it would have converged....

Generated by Claude Code for CS372 T3 Benchmark