T3 Benchmark Analysis Report - GroupI (AI and Tech)

Fernando Torres

January 28, 2026

Executive Summary

This report analyzes the **GroupI (AI and Tech)** dataset for the CS372 T3 Benchmark Assignment 2. The dataset contains **500 validated causal reasoning test cases** in the **AI and Technology** domain (D9), migrated to the Assignment 2 Appendix B schema format.

Key Metrics:

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

1. Summary of Unvalidated vs. Validated Dataset

Metric	Before Migration	After Migration
Total Cases	500	500
Schema Version	V4.0 (Pre-remediation)	Appendix B (Assignment 2)
L1 Labels	W/S/A format	YES/NO/AMBIGUOUS format
variables.Z	Object format	Array of strings
trap field	Flat fields	Nested object
Required Fields	Partial	Complete (21 fields)

Key Improvements: - Standardized ID format: `T3-BucketLarge-I-{level}.{seq}` - Transformed L1 labels from W/S/A to YES/NO/AMBIGUOUS - Restructured variables.Z from object to array format - Created nested trap object with type, type_name, subtype, subtype_name - Added all missing required fields per Table 9

2. Pearl Level Distribution

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

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

3. Label Distribution

L1 Labels (YES/NO/AMBIGUOUS) - Per Table 10

Label	Count	Description
YES	15	Valid causal claim (SHEEP cases)
NO	30	Invalid causal claim (WOLF cases)
AMBIGUOUS	5	Unclear or conditional relationship
Total	**50**	

L2 Labels - Per Table 10

Label	Count	Description
NO	300	All L2 cases labeled NO (invalid causal claims)

L3 Labels (VALID/INVALID/CONDITIONAL) - Per Table 10

Label	Count	Percentage
VALID	54	36.0%
INVALID	33	22.0%
CONDITIONAL	63	42.0%
Total	**150**	**100%**

4. Trap Type Distribution

L1 Trap Types (W1-W10, S1-S8, A)

Category	Trap Types	Count
WOLF (W-series)	W1, W2, W3, W4, W5, W6, W7, W9, W10	30
SHEEP (S-series)	S1, S2, S3, S4, S5	15
AMBIGUOUS	A	5

WOLF Trap Type Breakdown

Type	Name	Count
W1	Selection Bias	4
W2	Survivorship Bias	3
W3	Healthy User Bias	4
W4	Regression to Mean	1
W5	Ecological Fallacy	5
W6	Base Rate Neglect	1
W7	Confounding	6
W9	Reverse Causation	3
W10	Post Hoc Fallacy	3

L2 Trap Types (T1-T17)

Trap	Family	Count	Description
T1	F1: Selection	24	Selection Bias
T2	F1: Selection	19	Survivorship
T3	F1: Selection	17	Collider Bias
T4	F1: Selection	15	Immortal Time
T5	F2: Statistical	19	Regression to Mean
T6	F2: Statistical	16	Ecological Fallacy
T7	F3: Confounding	8	Confounder
T8	F3: Confounding	8	Simpson's Paradox
T9	F3: Confounding	8	Conf-Mediation
T10	F4: Direction	20	Reverse Causation
T11	F4: Direction	24	Feedback Loop
T12	F4: Direction	30	Temporal Precedence
T13	F5: Information	26	Measurement Error
T14	F5: Information	24	Recall Bias
T15	F6: Mechanism	26	Mechanism Confusion
T16	F6: Mechanism	8	Goodhart's Law
T17	F6: Mechanism	8	Backfire Effect

L3 Trap Types (F1-F8, DomainExt)

Type	Name	Count
F1	Deterministic	21
F2	Probabilistic	16
F3	Overdetermination	15
F4	Structural	16
F5	Temporal	15
F6	Epistemic	15
F7	Attribution	20
F8	Moral/Legal	15
DomainExt	Domain Extension	17

5. Difficulty Level Distribution

Difficulty	Count	Percentage	Target Ratio
Easy	129	25.8%	~25%
Medium	206	41.2%	~50%
Hard	165	33.0%	~25%
Total	**500**	**100%**	1:2:1

Note: Distribution approximates the 1:2:1 target ratio with slight skew toward Hard cases, reflecting the complexity of AI & Technology domain scenarios.

6. Score Summary

Unvalidated Dataset Scores

Metric	Value
Mean Score	8.50
Min Score	8.00
Max Score	8.50

Validated Dataset Scores

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

Validation Impact

- Schema Compliance: 500/500 (100%)
- Duplicate Detection: 0 duplicates found
- All 21 required fields: Present in all cases
- Trap type corrections: None required for GroupI

7. Prompt Setup

LLM Configuration

Parameter	Value
Model	Claude (Anthropic)
Temperature	0.7 (generation), 0.0 (validation)
Max Tokens	4096 per case

Multi-Agent Workflow

1. **Generator Agents (10-12 parallel)**: Created cases by trap type family
2. **Schema Validator**: JSON compliance checking
3. **Content Validators**: Quality scoring using 10-point rubric
4. **Cross Validator**: Duplicate detection and distribution balance
5. **Quality Judges**: Trap type verification
6. **Correction Agents**: Issue resolution and field fixes

Validation Pipeline

- JSON schema validation (Appendix B format)
- Content scoring (threshold 8.0/10)
- Duplicate detection (similarity < 0.75)
- Trap type verification per level requirements
- Distribution balance checks

Quality Control Measures

- 95%+ pass rate threshold per batch
- Iterative correction loops until quality met
- Final validation sweep for missing fields

8. Example Case

L1 Example (Association Level)


```
{
  "id": "T3-BucketLarge-I-1.1",
  "bucket": "BucketLarge-I",
  "case_id": "0001",
  "pearl_level": "L1",
  "domain": "D9",
  "subdomain": "AI Scaling",
  "difficulty": "Easy",
  "is_ambiguous": false,
  "scenario": "Larger models (X) correlate with higher truthfulness
    scores (Y) on benchmarks. A user assumes a 100B model never lies.",
  "claim": "A 100 billion parameter model never produces false
    statements because larger models correlate with higher
    truthfulness scores.",
  "variables": {
    "X": {"name": "Parameter Count (Size)", "role": "Treatment/Factor"},
    "Y": {"name": "Truthfulness Score", "role": "Outcome"},
    "Z": ["Hallucination Rate"]
  },
  "trap": {
    "type": "W3",
    "type_name": "Healthy User Bias",
    "subtype": "Asymptotic Failure / Extrapolation",
    "subtype_name": "Extrapolation Error"
  },
  "label": "NO",
  "causal_structure": "Correlation != total elimination of defects",
  "key_insight": "Larger models can still hallucinate, sometimes
    more persuasively.",
  "hidden_timestamp": "What is the hallucination rate at 100B scale?",
  "conditional_answers": {
    "answer_if_condition_1": "If hallucination rate is zero, claim valid.",
    "answer_if_condition_2": "If hallucination persists, claim invalid."
  },
  "wise_refusal": "Parameter count correlates with benchmark scores,
    but that does not imply perfection. Larger models can still
    hallucinate; assuming the trend reaches zero defects is an
    extrapolation error.",
  "gold_rationale": "The correlation between model size and
    truthfulness does not guarantee zero hallucinations.",
  "initial_author": "Fernando Torres",
  "validator": "Fernando Torres",
  "final_score": 8.5
}
```

Report generated for CS372 Assignment 2 - T3 Benchmark Expansion Migration Date: January 28, 2026 Author: Fernando Torres