T3 Benchmark Expansion: Methodology Report

CS372: Artificial General Intelligence for Reasoning, Planning, and Decision Making Stanford University - Winter 2026

Authors: Fernando Torres, Alessandro Balzi **Group:** I1 **Date:** January 12, 2026

---

1. Introduction

This report describes our methodology for expanding the T3 Benchmark dataset from 49 original cases to 454 validated cases. The T3 Benchmark evaluates AI systems' causal reasoning capabilities by testing their ability to identify "reasoning traps" - common fallacies in causal inference.

Our benchmark cases are structured around Pearl's Ladder of Causation, covering three levels of causal reasoning: Association (L1), Intervention (L2), and Counterfactual (L3).

---

2. Theoretical Framework

2.1 Pearl's Ladder of Causation

Each case is classified by its required level of causal reasoning:

- **L1 (Association):** Observational queries of the form P(Y|X). Tests whether AI can distinguish correlation from causation.
- **L2 (Intervention):** Interventional queries P(Y|do(X)). Tests whether AI correctly applies do-calculus and identifies backdoor paths.
- **L3 (Counterfactual):** What-if queries requiring structural causal models. Tests AI's ability to reason about alternative outcomes.

2.2 CRIT Scoring

We evaluated case quality using the CRIT (Causal Reasoning Integrity Test) framework, scoring each case on five dimensions: scenario clarity, variable definition, trap mechanism, reasoning chain, and wise refusal. Cases required a mean CRIT score of 7.0 or higher for acceptance.

---

3. Methodology

3.1 Hybrid Approach

We employed a five-phase hybrid strategy combining automated generation with agent-based gap filling:

Phase 1: Bug Fixes We identified and fixed critical issues in the existing pipeline: - Duplicate case ID generation in the finalization step - Missing template variables causing schema errors - Placeholder case detection for quality filtering

Phase 2: Threshold Adjustment We calibrated the similarity threshold to 0.75, balancing diversity requirements against case quantity goals.

Phase 3: Automated Pipeline Run We executed the orchestrator pipeline with 8 parallel generators, each specializing in a trap type (e.g., Goodhart's Law, Confounding, Instrumental Convergence). This phase produced 281 validated cases.

Phase 4: Agent-Based Gap Filling To reach our 454-case target, we used 4 parallel AI agents to generate 173 additional cases. Each agent received: - Specific trap type and Pearl level assignments - Anti-similarity context (existing scenarios to avoid) - High-quality examples as templates - Unique subdomain assignments for diversity

Phase 5: Final Validation All cases underwent three-stage validation: 1. **DAG Validation:** Checking acyclicity and backdoor criterion compliance 2. **Content Validation:** CRIT rubric scoring 3. **Cross Validation:** Duplicate detection and distribution verification

3.2 Quality Assurance

Cases failing validation entered a revision cycle (maximum 3 iterations). Issues were categorized by severity (Critical, High, Medium, Low), with Critical issues requiring immediate resolution.

---

4. Results

4.1 Dataset Statistics

Metric	Value
Total Cases	454
Original Cases	49
Generated Cases	405
Mean CRIT Score	8.54/10
DAG Validity Rate	96.9%
Unique IDs	100%

4.2 Distribution Achieved

Pearl Level Distribution: - L1 (Association): 52 cases (11.5%) - L2 (Intervention): 277 cases (61.0%) - L3 (Counterfactual): 125 cases (27.5%)

Difficulty Distribution: - Easy: 91 cases (20.0%) - Medium: 197 cases (43.4%) - Hard: 166 cases (36.6%)

Top Trap Types: - Goodhart's Law: 93 cases (20.5%) - Counterfactual Fallacies: 91 cases (20.0%) - Selection/Spurious Correlation: 47 cases (10.4%) - Specification Gaming: 42 cases (9.3%)

---

5. Key Insights

1. **Template Limitations:** Purely automated generation hit a ceiling around 280 cases due to template similarity constraints. Agent-based writing was essential for reaching our target.
2. **Validation Importance:** The three-stage validation pipeline caught issues that simple schema checks would miss, particularly in causal structure consistency.

3. **Pearl Level Balance:** L2 (Intervention) cases dominated our distribution, reflecting the practical importance of do-calculus reasoning in AI safety scenarios.
4. **Trap Diversity:** Goodhart's Law and Counterfactual fallacies were the most represented trap types, aligning with their prevalence in real-world AI safety concerns.

---

6. Conclusion

We successfully expanded the T3 Benchmark from 49 to 454 cases using a hybrid automated-plus-agent approach. The resulting dataset provides comprehensive coverage of causal reasoning traps across Pearl's three levels, with validated quality metrics meeting or exceeding all course requirements.

---

7. Deliverables

- **GroupI1_datasetV2.0.json:** 454 validated cases with metadata
- **GroupI1_methodology.pdf:** This methodology report

Group I1 - Stanford CS372 Winter 2026