T3-L2: Causal Disambiguation Benchmark

Design Document and Execution Plan

Version 2.0 (with L1/L2 Comparison Appendix)

Edward Y. Chang, Longling Geng

January 19th, 2026

Abstract

T3-L2 is a benchmark for evaluating whether Large Language Models (LLMs) can perform **causal disambiguation**—identifying what hidden information would resolve causal ambiguity in observational claims—and generate appropriate **Wise Refusals** when evidence is insufficient. This document specifies: (1) the motivation and theoretical foundation grounded in Pearl's causal hierarchy, (2) a comprehensive taxonomy of 17 trap types organized into 6 families derived from the epidemiological and econometric literature, (3) case structure with temporal disambiguation and conditional reasoning, (4) a three-layer validation framework inspired by SATBench, and (5) execution timeline and quality metrics.

Contents

1 Motivation: What Does T3-L2 Test? 3
- 1.1 The Gap Between Association and Causation . . . 3
- 1.2 Why T3-L2? The Practitioner's Skill . . . 3
- 1.3 Design Constraints: Working Within LLM Limitations . . . 3

2 Theoretical Foundation: Comprehensive Trap Taxonomy 5
- 2.1 Grounding in Established Literature . . . 5
- 2.2 The Six Families . . . 5

3 The 17 Trap Types 6
- 3.1 Family 1: Selection Effects (F1) . . . 6
- 3.2 Family 2: Statistical Artifacts (F2) . . . 7
- 3.3 Family 3: Confounding (F3) . . . 8
- 3.4 Family 4: Direction Errors (F4) . . . 9
- 3.5 Family 5: Information Bias (F5) . . . 9
- 3.6 Family 6: Mechanism Failures (F6) . . . 10
- 3.7 Quick Reference: Examples by Trap Type . . . 12

4 Case Structure 13
- 4.1 Required Components . . . 13
- 4.2 Wise Refusal Requirements . . . 13
- 4.3 Example Case . . . 14

5 Validation Framework **14**
5.1 Comparison: SATBench vs. T3-L2 . . . 14
5.2 Layer 1: Automated Structural Checks . . . 14
5.3 Layer 2: LLM-Based Consistency Checks . . . 15
5.4 Layer 3: Expert Review Protocol . . . 15

6 Target Distribution **15**
6.1 By Family and Type . . . 15
6.2 By Domain (10 Domains) . . . 16

7 Evaluation Metrics **16**
7.1 Model Output Contract . . . 16
7.2 Scoring Rubric . . . 17
7.3 Quality Targets . . . 17

8 Conclusion **17**

A Appendix: Comparison: T3-L1 vs. T3-L2 **18**
A.1 The Core Question: Fundamentally Different Tasks . . . 18
A.2 Taxonomy Comparison . . . 19
A.2.1 Type-Level Mapping . . . 19
A.2.2 Family-Level Organization . . . 19
A.3 Case Structure Comparison . . . 19
A.4 SHEEP Cases: Presence vs. Absence . . . 20
A.5 Evaluation Metrics . . . 20
A.6 Scale and Difficulty . . . 20
A.7 Summary of Key Differences . . . 21

1 Motivation: What Does T3-L2 Test?

1.1 The Gap Between Association and Causation

Pearl's Ladder of Causation [1] distinguishes three levels of causal reasoning:

Level	Name	Query	Requires
L1	Association	$P(Y\|X)$	Observational data
L2	Intervention	$P(Y\|do(X))$	Causal graph + intervention
L3	Counterfactual	$P(Y_x\|X', Y')$	Structural equations

LLMs are fundamentally **association engines**—trained on co-occurrence patterns, they excel at L1. However, the critical reasoning errors that plague real-world decision-making occur at the **L1/L2 boundary**: mistaking correlation for causation, missing confounders, reversing causal direction.

The Core Problem

Observation: LLMs readily generate causal-sounding claims from correlational evidence.
Risk: Users may trust these claims for decisions (medical, policy, business) where causal validity matters.
Question: Can LLMs recognize when a causal claim is *not* justified by the evidence?

1.2 Why T3-L2? The Practitioner's Skill

T3-L1 tests **detection**: "Is this causal claim justified?" (YES/NO)
T3-L2 tests **diagnosis**, which proceeds in stages:

1. **Classification**: What trap type is present?
2. **Pivotal Question**: What hidden information would resolve the ambiguity?
3. **Conditional Reasoning**: What is the interpretation under each possibility?
4. **Wise Refusal**: How to appropriately decline when evidence is insufficient?

The T3-L2 Core Question

"What hidden information would I need to resolve the causal ambiguity?"
This is the question that:

- A **jury** asks before assigning liability or guilt
- An **economist** asks before attributing causes of inflation
- A **data scientist** asks before designing a study
- A **policy maker** asks before trusting a claim
- A **doctor** asks before recommending treatment
- A **journalist** asks before publishing a story

1.3 Design Constraints: Working Within LLM Limitations

LLMs cannot:

- Access or analyze raw statistical data

- Run experiments or interventions
- Compute $P(Y|do(X))$ from causal graphs
- Verify causal claims empirically

Therefore, T3-L2 does **not** test:

× L1 computation ($P(Y|X)$)—trivial for LLMs
× L2 computation ($P(Y|do(X))$)—impossible for LLMs

T3-L2 **does** test **causal diagnosis**—a four-stage reasoning process:

1. **Classification**: What trap type threatens causal validity?
2. **Pivotal Question**: What hidden information would resolve the ambiguity?
3. **Conditional Reasoning**: What is the causal interpretation under each possibility?
4. **Wise Refusal**: How to appropriately decline when evidence is insufficient?

Note: T3-L1 tests only **detection** ("Is this claim justified?" → YES/NO). T3-L2 subsumes detection and proceeds to diagnosis.

The Achievable Goal

T3-L2 evaluates whether LLMs can serve as **causal reasoning assistants**—not by computing causal effects, but by diagnosing causal ambiguity:

- Given a scenario, **classify** which threat to validity is present
- **Identify** the pivotal question that would resolve the ambiguity
- **Articulate** the conditional interpretations under different assumptions
- **Generate** an appropriate refusal when evidence is insufficient

This is a **pattern recognition and structured generation task** over text descriptions of evidence—precisely the kind of task LLMs can, in principle, perform.

2 Theoretical Foundation: Comprehensive Trap Taxonomy

2.1 Grounding in Established Literature

The epidemiological and econometric literature identifies three fundamental sources of bias in causal inference [2, 3]:

1. **Confounding**: Spurious association due to common causes
2. **Selection Bias**: Non-random sampling or conditioning
3. **Information Bias**: Measurement error or misclassification

Beyond these classical categories, additional threats include:

- Statistical artifacts (Simpson's paradox, regression to mean)
- Mechanism failures (Goodhart's Law, intervention mismatch)
- Dynamic effects (feedback loops, time-varying confounding)

T3-L2 organizes these into **6 families** containing **17 trap types**, each grounded in scholarly references.

2.2 The Six Families

Trap Family	Core Question	Types	Reference
F1: Selection	Who is missing from the sample?	4	Heckman (1979)
F2: Statistical	Is the pattern a mathematical artifact?	2	Robinson (1950)
F3: Confounding	What common cause is missing?	3	Pearl (2009)
F4: Direction	Which way does the arrow point?	3	Hernán (2020)
F5: Information	Is the measurement accurate?	2	Rothman (2008)
F6: Mechanism	Does intervention target the cause?	3	Goodhart (1984)

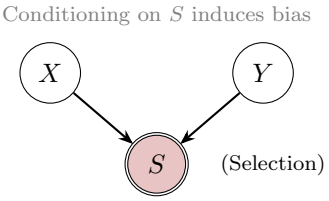
3 The 17 Trap Types

3.1 Family 1: Selection Effects (F1)

F1: Selection—"Who is missing from the sample?"

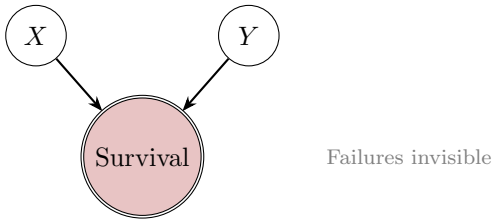
Definition: The sample is non-representative of the target population due to differential selection.
Key Reference: Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*.

T1: SELECTION Non-random sampling creates bias.



- **Hidden Question:** Who is systematically excluded?
- **Subtypes:** Healthy user effect, volunteer bias, indication bias
- **Reference:** Shrier & Platt (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*.

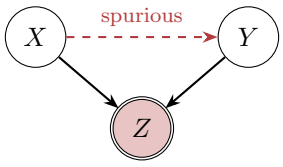
T2: SURVIVORSHIP Only survivors/successes are observed.



Only observe if Survival = 1

- **Hidden Question:** What happened to the failures?
- **Subtypes:** Business survival, publication bias, attrition
- **Reference:** Brown et al. (1992). Survivorship bias in performance studies. *Review of Financial Studies*.

T3: COLLIDER Conditioning on a common effect Z ($X \to Z \leftarrow Y$) creates spurious association.

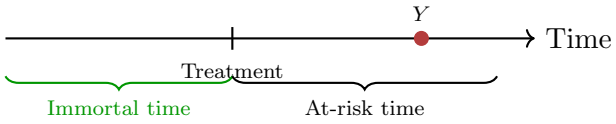


Conditioning on collider Z

- **Hidden Question:** Are we conditioning on a variable caused by both X and Y?

- **Subtypes:** Berkson's paradox, M-bias
- **Reference:** Berkson, J. (1946). Limitations of the application of fourfold table analysis. *Biometrics Bulletin*.

T4: IMMORTAL TIME Person-time is misclassified due to study design.



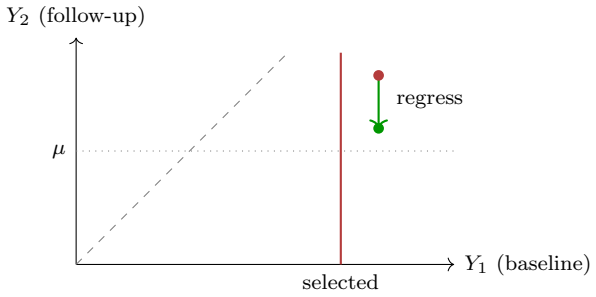
- **Hidden Question:** Was there a period when the outcome couldn't occur?
- **Subtypes:** Time-to-treatment bias, prevalent user bias
- **Reference:** Suissa, S. (2008). Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol*.

3.2 Family 2: Statistical Artifacts (F2)

F2: Statistical—"Is the pattern a mathematical artifact?"

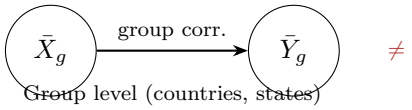
Definition: The observed pattern is an artifact of aggregation, weighting, or random variation.
Key Reference: Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *Am Sociol Rev*.

T5: REGRESSION Extreme values regress toward the mean on subsequent measurement.

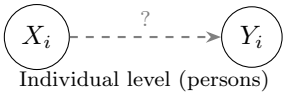


- **Hidden Question:** Were subjects selected for extreme values?
- **Subtypes:** Sports "sophomore slump," treatment of extreme cases
- **Reference:** Galton, F. (1886). Regression towards mediocrity in hereditary stature. *J Anthropological Institute*.

T6: ECOLOGICAL Group-level correlation $\neq$ individual-level correlation.



$\neq$



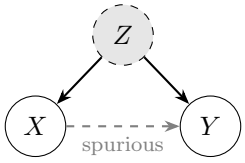
- **Hidden Question:** Does the pattern hold within each subgroup?
- **Subtypes:** Cross-level inference, compositional effects
- **Reference:** Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *Am Sociol Rev.*

3.3 Family 3: Confounding (F3)

F3: Confounding—"What common cause is missing?"

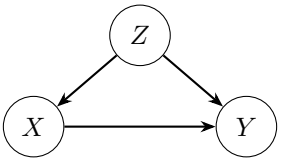
Definition: A third variable Z causes both exposure X and outcome Y, creating a spurious X-Y association. This family includes Simpson's Paradox, which is fundamentally a confounding phenomenon—stratifying by the confounder resolves the paradox.
Causal Structure: $X \leftarrow Z \rightarrow Y$
Key Reference: Pearl, J. (2009). *Causality.* Cambridge University Press.

T7: CONFOUNDER Standard confounding where Z is uncontrolled.



- **Hidden Question:** Is there an unmeasured common cause?
- **Subtypes:** Lifestyle bundle, socioeconomic confounding, genetic confounding
- **Reference:** Greenland et al. (1999). Causal diagrams for epidemiologic research. *Epidemiology.*

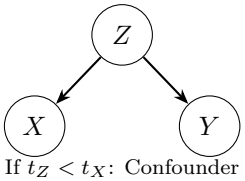
T8: SIMPSON'S Aggregate trend reverses when stratified by confounder Z.



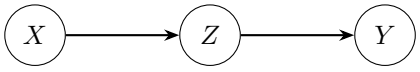
Aggregate: $X \uparrow \Rightarrow Y \uparrow$
Stratified by Z: $X \uparrow \Rightarrow Y \downarrow$

- **Hidden Question:** What happens when we stratify by the confounder Z?
- **Subtypes:** Confounded pooling, unequal base rates
- **Reference:** Bickel et al. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science.*

T9: CONF-MED (Confounder-Mediator Ambiguity) Variable Z could be confounder or mediator depending on timing.



If $t_Z < t_X$: Confounder



If $t_X < t_Z$: Mediator

- **Hidden Question:** Did Z occur before X ($t_Z < t_X$) or after ($t_X < t_Z$)?
- **Subtypes:** Environmental, pre-existing condition, lifestyle
- **Reference:** Baron & Kenny (1986). The moderator-mediator variable distinction. *J. Personality and Social Psychology.*

3.4 Family 4: Direction Errors (F4)

F4: Direction—"Which way does the causal arrow point?"

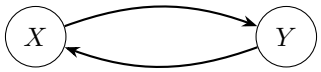
Definition: The direction of causation is misidentified or ambiguous.
Key Reference: Hernán, M.A. & Robins, J.M. (2020). *Causal Inference: What If.* Chapter 7.

T10: REVERSE Claimed $X \to Y$, but actual direction is $Y \to X$.

Claimed: $X \to Y$ Actual: $Y \to X$

- **Hidden Question:** Did X precede Y or Y precede X?
- **Subtypes:** Protopathic bias, reactive policy
- **Reference:** Rothman (2008). Modern Epidemiology. Chapter 9.

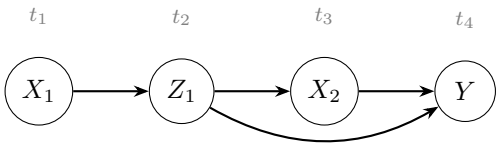
T11: FEEDBACK Bidirectional causation ($X \leftrightarrow Y$).



Reinforcing loop: $X \leftrightarrow Y$

- **Hidden Question:** Is there a reinforcing loop?
- **Subtypes:** Nocebo effect, self-fulfilling prophecy
- **Reference:** Strotz & Wold (1960). Recursive vs. nonrecursive systems. *Econometrica.*

T12: TEMPORAL Time-varying confounding or mediation.



Z is affected by past X and affects future X

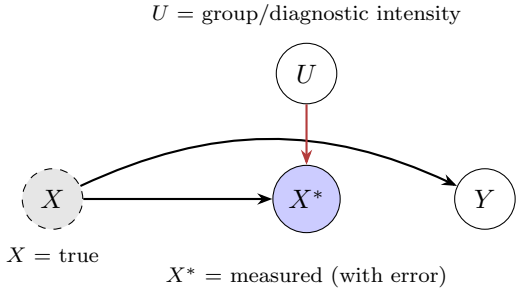
- **Hidden Question:** Does the confounding structure change over time?
- **Subtypes:** Treatment-confounder feedback
- **Reference:** Robins et al. (2000). Marginal structural models. *Epidemiology.*

3.5 Family 5: Information Bias (F5)

F5: Information—"Is the measurement accurate?"

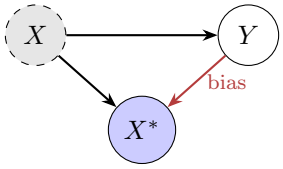
Definition: Systematic error in measuring exposure, outcome, or covariates.
Key Reference: Rothman, K.J. et al. (2008). *Modern Epidemiology*. Chapter 9.

T13: MEASUREMENT Exposure or outcome is measured with systematic error.



- **Hidden Question:** Does measurement accuracy differ between groups?
- **Subtypes:** Differential misclassification, detection bias
- **Reference:** Hernán & Cole (2009). Invited commentary: Causal diagrams and measurement bias. *Am J Epidemiol.*

T14: RECALL Participants' memory of past exposures is biased by current outcome.



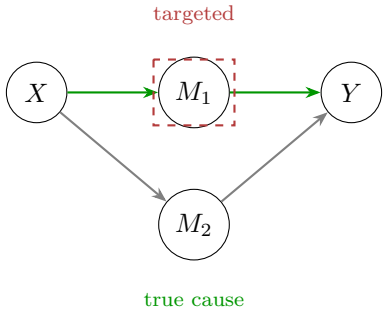
- **Hidden Question:** Do cases recall exposure differently than controls?
- **Subtypes:** Effort after meaning, rumination bias
- **Reference:** Coughlin, S. (1990). Recall bias in epidemiologic studies. *J Clinical Epidemiol.*

3.6 Family 6: Mechanism Failures (F6)

F6: Mechanism—"Does the intervention target the true cause?"

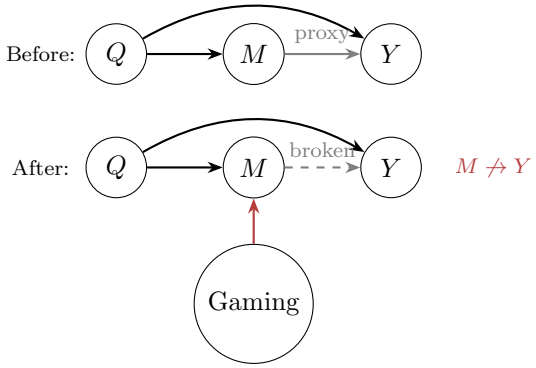
Definition: An intervention fails or backfires due to misunderstanding the causal mechanism.
Key Reference: Goodhart, C. (1984). Problems of monetary management: The U.K. experience.

T15: MECHANISM Intervention targets wrong causal path.



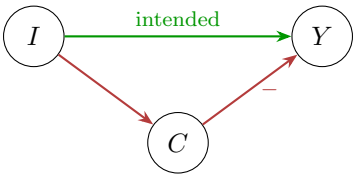
- **Hidden Question:** Did the intervention block the true mechanism?
- **Subtypes:** Wrong target population, incomplete intervention
- **Reference:** Cartwright, N. (2007). *Hunting Causes and Using Them*. Cambridge.

T16: GOODHART Metric becomes target, ceases to be good metric.



- **Hidden Question:** Is the metric being gamed?
- **Subtypes:** Teaching to test, KPI gaming, cobra effect
- **Reference:** Goodhart, C. (1984). Monetary Theory and Practice.

T17: BACKFIRE Intervention produces opposite of intended effect.



C = Compensatory behavior

Net effect may be opposite of intended

- **Hidden Question:** Could the intervention trigger compensatory behavior?
- **Subtypes:** Reactance, Streisand effect, moral licensing
- **Reference:** Schultz et al. (2007). The constructive, destructive, and reconstructive power of social norms. *Psych Science*.

3.7 Quick Reference: Examples by Trap Type

Family	Type	Hidden Question Pattern	Primary Reference
F1: Selection	T1: SELECTION	Who is excluded?	Heckman (1979)
	T2: SURVIVORSHIP	What about failures?	Brown et al. (1992)
	T3: COLLIDER	Conditioning on effect?	Berkson (1946)
	T4: IMMORTAL TIME	Protected time period?	Suissa (2008)
F2: Statistical	T5: REGRESSION	Extreme value selection?	Galton (1886)
	T6: ECOLOGICAL	Within-group pattern?	Robinson (1950)
F3: Confounding	T7: CONFOUNDER	Unmeasured common cause?	Pearl (2009)
	T8: SIMPSON'S	Reversal after stratification?	Bickel et al. (1975)
	T9: CONF-MED	$t_Z < t_X$ or $t_X < t_Z$?	Baron & Kenny (1986)
F4: Direction	T10: REVERSE	$t_X < t_Y$ or $t_Y < t_X$?	Rothman (2008)
	T11: FEEDBACK	Bidirectional loop?	Strotz & Wold (1960)
	T12: TEMPORAL	Time-varying structure?	Robins et al. (2000)
F5: Information	T13: MEASUREMENT	Differential accuracy?	Hernán & Cole (2009)
	T14: RECALL	Differential memory?	Coughlin (1990)
F6: Mechanism	T15: MECHANISM	Right causal path?	Cartwright (2007)
	T16: GOODHART	Metric gaming?	Goodhart (1984)
	T17: BACKFIRE	Compensatory response?	Schultz et al. (2007)

Type	Example
T1: SELECTION	Surveying gym members about exercise habits overstates population fitness (non-exercisers excluded).
T2: SURVIVORSHIP	Studying traits of successful startups misses that failed startups had the same traits.
T3: COLLIDER	Among the hospitalized, diabetes appears protective against flu (both independently cause hospitalization).
T4: IMMORTAL TIME	Oscar winners appear to live longer—but they had to survive long enough to win.
T5: REGRESSION	The "Sports Illustrated cover jinx"—athletes featured after peak performances naturally regress.
T6: ECOLOGICAL	Countries with higher chocolate consumption have more Nobel laureates (aggregation artifact).
T7: CONFOUNDER	Coffee drinkers have higher lung cancer rates—but smoking causes both coffee drinking and cancer.
T8: SIMPSON'S	UC Berkeley appeared to discriminate against women overall, but favored women within each department.
T9: CONF-MED	Exercise is associated with lower cholesterol and less heart disease—but did low cholesterol precede or follow exercise adoption?
T10: REVERSE	Depressed people watch more TV—but does TV cause depression, or do depressed people seek passive entertainment?
T11: FEEDBACK	Self-esteem and academic performance reinforce each other in a bidirectional loop.
T12: TEMPORAL	Blood pressure medication affects kidney function, which affects future medication dosing—time-varying confounding.
T13: MEASUREMENT	Patients with diagnosed disease are examined more thoroughly, detecting more comorbidities (surveillance bias).
T14: RECALL	Mothers of children with birth defects recall pregnancy exposures more thoroughly than mothers of healthy children.
T15: MECHANISM	Teaching "critical thinking" to improve test scores fails because the test actually measures memorization.
T16: GOODHART	Hospitals penalized for high mortality stop admitting terminal patients—mortality drops but care doesn't improve.
T17: BACKFIRE	Anti-drug campaigns in schools increase teen curiosity and experimentation (reactance effect).

4 Case Structure

4.1 Required Components

Every T3-L2 case contains:

Component	Description
`Scenario`	Narrative describing observed correlation between X and Y, with Z present
`Variables`	Explicit labeling: X (exposure), Y (outcome), Z (ambiguous third variable)
`Annotations`	Case ID, Domain, Trap Type, Subtype, Difficulty, Causal Structure
`Hidden Question`	The temporal or structural question that would resolve ambiguity
`Answer if A`	Causal interpretation if condition A holds
`Answer if B`	Causal interpretation if condition B holds
`Wise Refusal`	What to say when the information is unavailable

4.2 Wise Refusal Requirements

Wise Refusal Template

A proper Wise Refusal must:

1. **Identify** the specific causal ambiguity
2. **State** what information is missing
3. **Present** both conditional interpretations
4. **Decline** to endorse the causal claim

Template:

> "The [claim] is ambiguous due to [trap type]. We cannot determine whether [A] or [B] without knowing [hidden information]. If [A], then [interpretation A]. If [B], then [interpretation B]. Without this information, the causal claim is not justified."

4.3 Example Case

Case 1.11: The Pet Ownership Effect
Scenario. Families who adopted a dog (X) report walking 30% more miles (Y). These families also moved to the suburbs (Z) recently.
Variables.

- X = Dog Adoption
- Y = Walking Miles
- Z = Move to Suburbs (Ambiguous)

Annotations.

- **Trap Type:** T2 (CONF-MED)
- **Difficulty:** Medium
- **Causal Structure:** Either $Z \to X, Y$ or $X \to Z \to Y$

Hidden Question. Did walking increase (Y) start after the move (Z) but *before* the dog arrived?
Answer if $t_Z < t_X$ (Suburbs are Confounder). The move (Z) provided sidewalks/nature, encouraging walking (Y) and providing space for a dog (X). The environment is the driver.
Answer if $t_X < t_Z$ (Dog is Cause). They wouldn't walk without the dog (X), even in the suburbs.
Wise Refusal. "Did the environment change behavior, or did the pet? If walking increased immediately after the move (Z) prior to the adoption, the dog is not the primary cause of the exercise."

5 Validation Framework

Inspired by SATBench [4], which achieves consistency through solver-verified ground truth and multi-stage validation, T3-L2 employs a three-layer validation pipeline.

5.1 Comparison: SATBench vs. T3-L2

Aspect	SATBench	T3-L2
Ground truth	SAT solver (provable)	Expert consensus (verifiable)
Case generation	Fully automated (formula → story)	Semi-automated (seed → expansion)
Structural check	Solver validates assignment	Automated schema checks
Consistency check	LLM validates story-formula match	LLM validates trap type
Human review	Spot-check (>90% pass)	100% two-pass review

5.2 Layer 1: Automated Structural Checks

Deterministic Validation (All Cases Must Pass)

1. **Variable Completeness**: X, Y, Z defined
2. **Hidden Question Present**: Temporal or structural question stated
3. **Conditional Duality**: Both "if A" and "if B" answers present

4. **Wise Refusal Present**: Refusal statement included
5. **Trap Type Valid**: Type $\in \{T1, \ldots, T17\}$
6. **Structure Notation**: Causal graph uses arrows ($\rightarrow$, $\leftarrow$, $\leftrightarrow$)

5.3 Layer 2: LLM-Based Consistency Checks

Three validation prompts ensure internal consistency:

Prompt 1: Trap Type Verification

```
Given this case, what trap type does it demonstrate?
Does your answer match the labeled type?
```

Prompt 2: Conditional Answer Validity

```
Do the conditional answers logically follow from
their respective conditions?
Are they mutually exclusive and exhaustive?
```

Prompt 3: Wise Refusal Quality

```
Does the refusal (1) identify ambiguity, (2) state
missing info, (3) present both interpretations,
(4) decline to conclude?
```

5.4 Layer 3: Expert Review Protocol

Pass 1: Trap Type and Hidden Question (2 reviewers)

- Is the trap type correct?
- Is the hidden question correctly formulated?
- Agreement $\rightarrow$ accept; Disagreement $\rightarrow$ third reviewer or PI

Pass 2: Quality Scoring (1 reviewer)

Criterion	Points	Description
Scenario clarity	2	X, Y, Z clearly defined
Hidden question quality	2	Identifies key ambiguity
Conditional answer A	1.5	Logically follows from A
Conditional answer B	1.5	Logically follows from B
Wise refusal quality	2	Follows template
Difficulty calibration	1	Label matches complexity
Total	**10**	≥ 8 accept; 6–7 revise; < 6 reject

6 Target Distribution

6.1 By Family and Type

Target: **3,000 cases** with difficulty distribution: 28% Easy, 44% Medium, 28% Hard.

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** (27.8%)	**1,330** (44.3%)	**835** (27.8%)	**3,000** (100%)

6.2 By Domain (10 Domains)

Target: **300 cases per domain** (±10%). Each trap type should have cases from at least 5 domains.

Domain	Description	Target
D1	Daily Life	300
D2	Health/Medicine	300
D3	Education	300
D4	Business/Economics	300
D5	Technology	300
D6	Environment	300
D7	Policy/Government	300
D8	Finance	300
D9	Sports	300
D10	Social Science	300
Total		**3,000**

7 Evaluation Metrics

7.1 Model Output Contract

```
{
  "trap_type": "T1|T2|...|T17",
  "trap_family": "F1|F2|F3|F4|F5|F6",
```

```
  "hidden_question": "string",
  "conditional_answers": {
    "if_A": "interpretation under A",
    "if_B": "interpretation under B"
  },
  "wise_refusal": "2-4 sentence refusal",
  "confidence": "low|medium|high"
}
```

7.2 Scoring Rubric

Metric	Weight	Evaluation Method
Trap Type Accuracy	20%	Exact match (automatic)
Trap Family Accuracy	10%	Exact match (automatic)
Hidden Question Quality	20%	LLM-as-judge (rubric)
Conditional Answer Quality	25%	LLM-as-judge (rubric)
Wise Refusal Quality	25%	LLM-as-judge (rubric)

7.3 Quality Targets

Metric	Target	Minimum
Inter-rater trap type agreement	>85%	80%
Inter-rater family agreement	>95%	90%
Cohen's κ (trap type)	>0.75	0.70
Average quality score	>8.0/10	7.0/10
LLM validation agreement	>90%	85%

8 Conclusion

T3-L2 fills a critical gap in LLM evaluation: testing whether models can identify **what they don't know** about causal claims. By focusing on disambiguation rather than computation, T3-L2:

1. **Works within LLM limitations**: Tests pattern recognition over text, not causal computation
2. **Has practical value**: The questions it tests are exactly what practitioners need
3. **Is grounded in scholarship**: All 17 trap types trace to foundational references
4. **Ensures quality**: Three-layer validation adapted from SATBench

The benchmark's core insight—that recognizing the **pivotal question** is more valuable than applying the correct **label**—makes T3-L2 both achievable and useful.

References

[1] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

[2] Hernán, M.A. & Robins, J.M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

[3] Rothman, K.J., Greenland, S., & Lash, T.L. (2008). *Modern Epidemiology* (3rd ed.). Lippincott Williams & Wilkins.

[4] Wei, A., et al. (2025). SATBench: Benchmarking LLMs' logical reasoning via Boolean satisfiability. *arXiv:2505.14615*.

[5] Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.

[6] Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3), 47–53.

[7] Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357.

[8] Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction. *J. Personality and Social Psychology*, 51(6), 1173–1182.

[9] Goodhart, C. (1984). Problems of monetary management: The U.K. experience. In *Monetary Theory and Practice*. Macmillan.

[10] Bickel, P.J., Hammel, E.A., & O'Connell, J.W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404.

[11] Greenland, S., Pearl, J., & Robins, J.M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1), 37–48.

[12] Suissa, S. (2008). Immortal time bias in pharmaco-epidemiology. *American Journal of Epidemiology*, 167(4), 492–499.

[13] Coughlin, S. (1990). Recall bias in epidemiologic studies. *J Clinical Epidemiology*, 43(1), 87–91.

[14] Robins, J.M., Hernán, M.A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

[15] Hernán, M.A. & Cole, S.R. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology*, 170(8), 959–962.

A Appendix: Comparison: T3-L1 vs. T3-L2

This appendix provides a systematic comparison between T3-L1 and T3-L2, clarifying their complementary roles in evaluating LLM causal reasoning capabilities.

A.1 The Core Question: Fundamentally Different Tasks

Aspect	**T3-L1**	**T3-L2**
Core Question	"Is this causal claim justified?"	"What hidden information would resolve the ambiguity?"
Task Type	Binary Classification (YES/NO)	Disambiguation + Generation
Skill Tested	Detection	Explanation & Conditional Reasoning
Output	Simple answer	Trap type, hidden question, conditional answers, wise refusal

T3-L1 tests whether the model can *recognize* a problem.

T3-L2 tests whether the model can *articulate* what's missing and provide conditional interpretations.

A.2 Taxonomy Comparison

A.2.1 Type-Level Mapping

T3-L1 WOLF Types (10)	T3-L2 Trap Types (17)	Notes
W1: Selection Bias	T1: SELECTION	Same concept
W2: Survivorship Bias	T2: SURVIVORSHIP	Same concept
W3: Healthy User Bias	T1: SELECTION (subtype)	Merged into Selection
W4: Regression to Mean	T5: REGRESSION	Same concept
W5: Ecological Fallacy	T6: ECOLOGICAL	Same concept
W6: Base Rate Neglect	—	Dropped in L2
W7: Confounding	T7: CONFOUNDER	Same concept
W8: Simpson's Paradox	T8: SIMPSON'S	Same concept
W9: Reverse Causation	T10: REVERSE	Same concept
W10: Post Hoc Fallacy	—	Subsumed under Direction
—	T3: COLLIDER	**New**: Berkson's paradox
—	T4: IMMORTAL TIME	**New**: Time-to-event bias
—	T9: CONF-MED	**New**: Confounder-Mediator
—	T11: FEEDBACK	**New**: Bidirectional loops
—	T12: TEMPORAL	**New**: Time-varying confounding
—	T13: MEASUREMENT	**New**: Information bias
—	T14: RECALL	**New**: Recall bias
—	T15: MECHANISM	**New**: Wrong intervention target
—	T16: GOODHART	**New**: Metric gaming
—	T17: BACKFIRE	**New**: Intervention reversal

SHEEP Types (Justified Claims): T3-L1 includes 8 SHEEP types (S1: RCT, S2: Natural Experiment, S3: Lottery, S4: Controlled Ablation, S5: Mechanism + Dose-Response, S6: Instrumental Variable, S7: Difference-in-Differences, S8: Regression Discontinuity) to test whether models can recognize *valid* causal designs. T3-L2 does not include SHEEP equivalents because its task is **disambiguation**, not classification. When evidence is sufficient to justify a causal claim, there is no ambiguity to resolve—no hidden question to identify, no conditional interpretations to articulate, and no refusal to generate.

A.2.2 Family-Level Organization

T3-L1 Families (4)	T3-L2 Families (6)
Selection (W1–W4)	F1: Selection (T1–T4)
Ecological (W5–W6)	F2: Statistical (T5–T6)
Confounding (W7–W8)	F3: Confounding (T7–T9)
Direction (W9–W10)	F4: Direction (T10–T12)
—	**F5: Information (T13–T14)** — NEW
—	**F6: Mechanism (T15–T17)** — NEW

A.3 Case Structure Comparison

T3-L1 Case Structure:

```
Scenario: [Narrative]
Claim: "[X causes Y]"
```

```
Ground Truth: YES or NO
[Trap/Evidence Type]: [Label]
Why Flawed/Valid: [Explanation]
```

T3-L2 Case Structure:

```
Scenario: [Narrative with X, Y, Z labeled]
Variables: X (exposure), Y (outcome), Z (ambiguous)
Annotations: Trap Type, Difficulty, Causal Structure
Hidden Question: [Temporal/structural question]
Answer if A: [Interpretation under condition A]
Answer if B: [Interpretation under condition B]
Wise Refusal: [What to say when info unavailable]
```

Key Difference: T3-L2 requires explicit **conditional reasoning** (if A vs. if B) and a **Wise Refusal** template.

A.4 SHEEP Cases: Presence vs. Absence

	T3-L1	**T3-L2**
Positive cases (justified claims)	8 SHEEP types	None
Purpose	Distinguish valid from invalid	Focus on ambiguous cases

T3-L1 includes justified causal claims (SHEEP: RCT, Natural Experiment, Lottery, Controlled Ablation, Mechanism + Dose-Response, etc.) so the model must distinguish both valid and invalid evidence. T3-L2 focuses exclusively on cases requiring disambiguation—testing depth of understanding rather than breadth of classification.

A.5 Evaluation Metrics

T3-L1	**T3-L2**
Binary accuracy (correct YES/NO)	Multi-component scoring: • Trap Type Accuracy (20%) • Trap Family Accuracy (10%) • Hidden Question Quality (20%) • Conditional Answer Quality (25%) • Wise Refusal Quality (25%)

T3-L2 uses **LLM-as-judge** for qualitative components, whereas T3-L1 scoring is purely objective.

A.6 Scale and Difficulty

	T3-L1	**T3-L2**
Total Cases	200 (100 WOLF + 100 SHEEP)	3,000
Domains	10	10
Difficulty Levels	Easy / Medium / Hard	Easy / Medium / Hard
Trap Types	7 Core + 3 Advanced	17 (all testable)

A.7 Summary of Key Differences

1. **Different Cognitive Levels:** T3-L1 tests *recognition* (can you spot the problem?), while T3-L2 tests *explanation* (can you articulate what's missing and reason conditionally?). This follows a natural pedagogical progression.

2. **Taxonomy Evolution:** T3-L2 expands the taxonomy significantly:

   - Adds 2 new families (Information, Mechanism)
   - Adds 10 new trap types
   - Drops Base Rate Neglect
   - Merges Healthy User Bias into Selection

3. **Complementary Design:** The two benchmarks work together:

   - L1: Can the model classify? (Simpler, objective scoring)
   - L2: Can the model explain what's missing? (Harder, requires conditional reasoning)

4. **Practical Value:** T3-L2's "Wise Refusal" requirement directly addresses sycophancy—the model must explicitly decline to endorse a claim while explaining why. This aligns with Regulated Causal Anchoring (RCA) research on AI safety control.

5. **Intentional Overlap:** The overlap in trap types ensures L1 and L2 test the same conceptual space but at different depths. A model that passes L1 should find L2's classification component easier, but the disambiguation and wise refusal components raise the bar significantly.