T3-L1 Benchmark Execution Plan

A Guide to Case Design, Template Use, and Quality Control School Project

(Internal Use) Version 2.0 (Reorganized) January 18, 2026

Abstract

This document specifies an execution plan for developing the T3-L1 benchmark, which tests whether Large Language Models (LLMs) can distinguish justified causal claims from unjustified ones using only scenario text. The plan includes: (1) an implementability assessment of all trap and evidence types under LLM constraints, (2) detailed templates for 10 WOLF (unjustified) types and 8 SHEEP (justified) types, (3) domain specifications and difficulty calibration guidelines, and (4) a team structure and timeline for producing 200 high-quality cases (100 WOLF + 100 SHEEP) across 10 domains with consistent quality control.

Contents

1 Introduction **3**
1.1 Purpose of This Document . . . 3
1.2 The Core Testing Paradigm . . . 3
1.3 Key Constraint: LLM Limitations . . . 3

2 Implementability Assessment **3**
2.1 Assessment Criteria . . . 3
2.2 WOLF Types: Implementability Summary (Organized by Family) . . . 4
2.3 SHEEP Types: Implementability Summary . . . 4

3 WOLF Templates (Unjustified Causal Claims) **5**
3.1 W1: Selection Bias (Core Tier) . . . 5
3.1.1 Definition and Structure . . . 5
3.1.2 Required Elements . . . 5
3.1.3 Scenario Template . . . 6
3.1.4 Domain Examples . . . 6
3.1.5 Difficulty Levels . . . 6
3.1.6 Validation Checklist . . . 6
3.1.7 Complete Example Case . . . 7
3.2 W2: Survivorship Bias (Core Tier) . . . 7
3.2.1 Definition and Structure . . . 7
3.2.2 Required Elements . . . 7
3.2.3 Scenario Template . . . 7
3.2.4 Domain Examples . . . 8
3.2.5 Complete Example Case . . . 8
3.3 W3: Healthy User Bias (Core Tier) . . . 8
3.3.1 Definition and Structure . . . 8
3.3.2 Required Elements . . . 9
3.3.3 Domain Examples . . . 9
3.3.4 Complete Example Case . . . 9
3.4 W5: Ecological Fallacy (Core Tier) . . . 10
3.4.1 Definition and Structure . . . 10
3.4.2 Required Elements . . . 10
3.4.3 Domain Examples . . . 10
3.4.4 Complete Example Case . . . 11
3.5 W7: Confounding (Core Tier) . . . 12
3.5.1 Definition and Structure . . . 12
3.5.2 Required Elements . . . 12
3.5.3 Scenario Template . . . 12
3.5.4 Domain Examples . . . 13
3.5.5 Difficulty Levels . . . 13
3.5.6 Validation Checklist . . . 13
3.5.7 Complete Example Case . . . 13
3.6 W9: Reverse Causation (Core Tier) . . . 14
3.6.1 Definition and Structure . . . 14
3.6.2 Required Elements . . . 14
3.6.3 Domain Examples . . . 14
3.6.4 Complete Example Case . . . 15
3.7 W10: Post Hoc Fallacy (Core Tier) . . . 15

3.7.1 Definition and Structure . . . 15
3.7.2 Required Elements . . . 15
3.7.3 Domain Examples . . . 15
3.7.4 Complete Example Case . . . 16

4 SHEEP Templates (Justified Causal Claims) 17
4.1 S1: Randomized Controlled Trial (Core Tier) . . . 17
4.1.1 Definition and Structure . . . 17
4.1.2 Required Elements . . . 17
4.1.3 Key Phrases to Include . . . 17
4.1.4 Complete Example Case . . . 17
4.2 S2: Natural Experiment (Core Tier) . . . 18
4.2.1 Definition and Structure . . . 18
4.2.2 Required Elements . . . 18
4.2.3 Complete Example Case . . . 18
4.3 S3: Lottery / Quasi-Random (Core Tier) . . . 19
4.3.1 Definition and Structure . . . 19
4.3.2 Required Elements . . . 19
4.3.3 Complete Example Case . . . 19
4.4 S4: Controlled Ablation (Core Tier) . . . 20
4.4.1 Definition and Structure . . . 20
4.4.2 Required Elements . . . 20
4.4.3 Complete Example Case . . . 20
4.5 S5: Mechanism + Dose-Response (Core Tier) . . . 21
4.5.1 Definition and Structure . . . 21
4.5.2 Required Elements . . . 21
4.5.3 Complete Example Case . . . 21

5 Advanced Cases (Reference Only) 22
5.1 W4: Regression to Mean (Advanced Tier) . . . 22
5.1.1 Definition and Structure . . . 22
5.1.2 Required Elements . . . 22
5.1.3 Key Signal Words . . . 22
5.1.4 Complete Example Case . . . 22
5.2 W6: Base Rate Neglect (Advanced Tier) . . . 23
5.2.1 Definition and Structure . . . 23
5.2.2 Required Elements . . . 23
5.2.3 Complete Example Case . . . 23
5.3 W8: Simpson's Paradox (Advanced Tier) . . . 24
5.3.1 Definition and Structure . . . 24
5.3.2 Required Elements . . . 24
5.3.3 Complete Example Case . . . 24
5.4 S6: Instrumental Variable (Advanced Tier) . . . 24
5.5 S7: Difference-in-Differences (Advanced Tier) . . . 25
5.6 S8: Regression Discontinuity (Advanced Tier) . . . 25

6 References 26
6.1 Selection Family (W1–W4) . . . 26
6.2 Ecological Family (W5–W6) . . . 26
6.3 Confounding Family (W7–W8) . . . 26
6.4 Direction Family (W9–W10) . . . 27

1 Introduction

1.1 Purpose of This Document

The T3-L1 benchmark tests a fundamental capability: given a textual description of evidence, can an LLM correctly judge whether a causal claim is justified? This document provides the execution plan for developing T3-L1, including:

1. **Implementability Assessment**: Which types can be tested reliably given LLM constraints?
2. **Template Specifications**: Templates for each WOLF (trap) type and SHEEP (evidence) type, including required elements, domain examples, and validation checklists.
3. **Execution Plan**: Team structure, timeline, and quality control procedures.

1.2 The Core Testing Paradigm

T3-L1 operates at the boundary between Pearl's Level 1 (Association) and Level 2 (Intervention). The core question is:

T3-L1 Core Question

Given this described evidence, is the causal claim justified?

- **WOLF cases (Answer: NO)**: The evidence describes correlation or flawed reasoning; the causal claim is not justified.
- **SHEEP cases (Answer: YES)**: The evidence describes a valid causal design (RCT, natural experiment, etc.); the causal claim is justified.

1.3 Key Constraint: LLM Limitations

LLMs read text descriptions; they cannot:

- Collect or analyze raw statistical data
- Compute conditional probabilities from datasets
- Verify numerical claims independently
- Access external databases or run experiments

Therefore, T3-L1 tests reasoning about *described evidence*, not statistical computation. All necessary information must be contained in the scenario text.

2 Implementability Assessment

This section evaluates which types can be effectively tested on LLMs, given their limitations.

2.1 Assessment Criteria

For each type, we assess:

1. **Text-describable**: Can the flaw or validity be conveyed in natural language?
2. **Recognition-testable**: Can an LLM recognize the pattern without computation?
3. **Ground-truth determinable**: Is the correct answer unambiguous from the text?

2.2 WOLF Types: Implementability Summary (Organized by Family)

The 10 WOLF trap types are organized into four families based on the nature of the inferential error:

Table 1: WOLF Types Implementability Assessment (Organized by Family)

Trap Type	Tier	Status	Rationale
Selection Family (Specific → General): W1–W4			
W1: Selection Bias	Core	Full	Describe sampling; LLM recognizes non-representative samples
W2: Survivorship Bias	Core	Full	Describe "only survivors observed" and missing failures
W3: Healthy User Bias	Core	Full	Describe self-selection into X and correlated lifestyle factors
W4: Regression to Mean	Adv.	Partial	Requires statistical intuition; needs careful phrasing
Ecological Family (General → Specific): W5–W6			
W5: Ecological Fallacy	Core	Full	Describe aggregate correlation used to claim individual causation
W6: Base Rate Neglect	Adv.	Partial	Must provide base rates and test properties in text
Confounding Family: W7–W8			
W7: Confounding	Core	Full	Describe Z; LLM recognizes Z causes both X and Y
W8: Simpson's Paradox	Adv.	Partial	Must provide subgroup and aggregate numbers in text
Direction Family: W9–W10			
W9: Reverse Causation	Core	Full	Describe X and Y with plausible reverse direction
W10: Post Hoc Fallacy	Core	Full	Describe timing-based inference without controls or mechanism

Summary: 7 WOLF types are fully implementable (Core tier). 3 types require numerical or methodological support in the scenario text (Advanced tier). Assignment 2 does not require implement the Advance tier W4, W6, and W8.

Table 2 provides the references for WOLF cases.

Table 2: WOLF Types with Primary Scholarly References

Type	Primary Reference	Additional References
Selection Family		
W1: Selection Bias	Heckman (1979) *Econometrica*	Hernán et al. (2004); Berkson (1946)
W2: Survivorship	Wald (1943) WWII aircraft	Brown et al. (1992); Elton et al. (1996)
W3: Healthy User	Shrank et al. (2011) *JGIM*	Brookhart et al. (2007); Petitti (2002)
W4: Regression	Galton (1886)	Barnett et al. (2005); Bland & Altman (1994)
Ecological Family		
W5: Ecological	Robinson (1950) *ASR*	Freedman (1999); Piantadosi et al. (1988)
W6: Base Rate	Kahneman & Tversky (1973)	Tversky & Kahneman (1974); Bar-Hillel (1980)
Confounding Family		
W7: Confounding	Pearl (2009) *Causality*	Greenland et al. (1999); Hernán & Robins (2020)
W8: Simpson's	Simpson (1951); Bickel (1975)	Pearl (2014)
Direction Family		
W9: Reverse	Rothman et al. (2008) Ch. 9	Davey Smith & Hemani (2014)
W10: Post Hoc	Kahneman (2011) Ch. 19	Gilovich (1991); Hume (1748)

2.3 SHEEP Types: Implementability Summary

Summary: 5 SHEEP types are fully implementable (Core tier). 3 require methodological understanding (Advanced tier).

Table 3: SHEEP Types Implementability Assessment

Evidence Type	Tier	Status	Rationale
S1: RCT	Core	Full	Describe random assignment and control group
S2: Natural Experiment	Core	Full	Describe exogenous event and comparison group
S3: Lottery/Quasi-Random	Core	Full	Describe random allocation among applicants
S4: Controlled Ablation	Core	Full	Describe removal of X while holding other factors constant
S5: Mechanism + Dose	Core	Full	Describe known pathway plus dose-response gradient
S6: Instrumental Variable	Adv.	Partial	Requires IV logic to be described cleanly
S7: Diff-in-Diff	Adv.	Partial	Requires time and control group with parallel pre-trends
S8: Regression Discont.	Adv.	Partial	Requires cutoff assignment and local comparison

3 WOLF Templates (Unjustified Causal Claims)

WOLF cases present scenarios where a causal claim is made but the evidence does not justify it. The LLM should answer **NO** for all WOLF cases. This section provides detailed templates for all 10 WOLF types, organized by family.

Selection Family (W1–W3)

Specific → General Inference Errors: Sample ≠ Population

3.1 W1: Selection Bias (Core Tier)

3.1.1 Definition and Structure

W1: Selection Bias

Definition: The sample studied is not representative of the population to which the causal claim is generalized. Results from the biased sample may not apply to the broader population. **Key Issue**: Non-random sampling creates systematic differences between the sample and the target population. **Primary Reference**: Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161. [Nobel Prize work]

3.1.2 Required Elements

#	Required Element
1	Sampling method is described (non-random)
2	Sample differs systematically from target population
3	Causal claim generalizes beyond the sample
4	No mention of correction for selection or weighting

Table 4: W1 Required Elements Checklist

3.1.3 Scenario Template

Template: "Researchers surveyed [biased sample] and found that [X relates to Y]. They concluded that [X causes Y] in the general population." *Key flaw*: The sample is not representative; those observed differ from the population.

3.1.4 Domain Examples

#	Domain	Example (Biased Sample)
1	Health	Hospital patients (sicker than general population)
2	Education	Survey respondents (more engaged than non-respondents)
3	Business	Successful companies (ignoring failed companies)
4	Technology	App users who did not uninstall (satisfied users only)
5	Finance	Investors who stayed in market (risk-tolerant survivors)
6	Social Media	Viral posts (not representative of all posts)
7	Medicine	Clinical trial volunteers (healthier, more motivated)
8	Workplace	Employees who stayed (survivors, not those who left)
9	Research	Published studies (publication bias)
10	Consumer	Online reviewers (extreme opinions overrepresented)

Table 5: W1 Domain Examples

3.1.5 Difficulty Levels

Level	Description
Easy	Selection mechanism explicitly stated
Medium	Selection implied by sample description
Hard	Selection subtle; requires inference about who is included

Table 6: W1 Difficulty Calibration

3.1.6 Validation Checklist

☐ Is the sample clearly non-representative?

☐ Does the claim generalize beyond the sample?

☐ Would a random sample likely show a different result?

☐ Is ground truth unambiguously NO?

3.1.7 Complete Example Case

Example: W1-TECH-001 (Easy Difficulty)

Scenario: A software company surveyed users who had been using their product for over 2 years and found 95% satisfaction. They concluded that their software causes high user satisfaction and used this finding in their marketing materials. **Claim**: "The software causes high user satisfaction." **Ground Truth**: NO **Selection Bias**: Only long-term users were surveyed. Dissatisfied users likely stopped using the software and were not included in the survey. **Why Flawed**: The 95% satisfaction rate reflects survivor bias in the sample, not the software's effect on all users who tried it.

3.2 W2: Survivorship Bias (Core Tier)

3.2.1 Definition and Structure

W2: Survivorship Bias

Definition: Only "survivors" (successful cases) are observed, while failures are invisible. This can create a false impression that some attribute X caused success Y. **Causal Structure**: This aligns with a collider pattern: conditioning on being observed (success) can induce misleading associations. **Primary Reference**: Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. Statistical Research Group, Columbia University. [Declassified 1980]

3.2.2 Required Elements

#	Required Element
1	Only successful or surviving cases are observed
2	Failures are invisible or excluded
3	X is attributed as cause of success Y
4	Failed cases with similar X plausibly exist but are unobserved

Table 7: W2 Required Elements Checklist

3.2.3 Scenario Template

Template: "Looking at [successful cases], researchers found they all [had X]. They concluded that [X leads to success/Y]." *Key flaw*: Failed cases with X are missing from the analysis.

3.2.4 Domain Examples

#	Domain	Example (Survivors vs. Invisible Failures)
1	Business	Successful startups (failed startups invisible)
2	Music	Famous musicians and practice habits
3	Military	Returning aircraft damage patterns
4	Architecture	Ancient buildings that survived
5	Investing	Successful investors' strategies
6	Academia	Published researchers (file drawer effect)
7	Sports	Professional athletes' training
8	Restaurants	Long-standing restaurants only
9	Books	Bestselling authors' habits
10	Medicine	Recovered patients only

Table 8: W2 Domain Examples

3.2.5 Complete Example Case

Example: W2-BUS-001 (Medium Difficulty)

Scenario: A business magazine analyzed 50 highly successful entrepreneurs and found that 80% of them had dropped out of college. The article concluded that dropping out of college increases your chances of entrepreneurial success and encouraged young entrepreneurs to consider leaving school. **Claim**: "Dropping out of college causes increased entrepreneurial success." **Ground Truth**: NO **Survivorship Bias**: The analysis only included successful entrepreneurs. Many college dropouts who attempted entrepreneurship and failed are not visible in the dataset. **Why Flawed**: Without knowing outcomes among all who attempted entrepreneurship (including failures), we cannot infer that dropping out helps.

3.3 W3: Healthy User Bias (Core Tier)

3.3.1 Definition and Structure

W3: Healthy User Bias

Definition: People who voluntarily engage in health behavior X differ systematically from those who do not, in ways that independently affect health outcome Y. This is a special case of confounding where self-selection is central. **Key Issue**: Voluntary health behaviors are often markers of overall health consciousness, not necessarily causes of outcomes. **Primary Reference**: Shrank, W. H., Patrick, A. R., & Brookhart, M. A. (2011). Healthy user and related biases in observational studies of preventive interventions. *Journal of General Internal Medicine*, 26(5), 546–550.

3.3.2 Required Elements

#	Required Element
1	X is a voluntary health behavior or choice
2	People who choose X differ systematically from those who do not
3	Y could be driven by overall lifestyle, not X specifically
4	Evidence is observational (not randomized)

Table 9: W3 Required Elements Checklist

3.3.3 Domain Examples

#	Domain	Example (X chosen by health-conscious people)
1	Supplements	Vitamin takers (health-conscious in general)
2	Exercise	Gym members (already fit, health-aware)
3	Diet	Organic food buyers (wealthy, health-aware)
4	Screening	Health checkup attendees (proactive about health)
5	Wellness	Meditation practitioners
6	Vaccination	Early adopters (correlated behaviors)
7	Preventive care	Regular dental visits
8	Lifestyle	Non-smokers (bundle of behaviors)
9	Fitness	Sports participants
10	Medicine	Preventive medication adherence

Table 10: W3 Domain Examples

3.3.4 Complete Example Case

Example: W3-SUPP-001 (Medium Difficulty)

Scenario: A 10-year study tracked 20,000 adults and found that those who took daily multivitamins had 20% lower mortality rates than those who did not. The researchers concluded that taking multivitamins extends lifespan and recommended daily supplementation. **Claim**: "Taking multivitamins causes longer lifespan." **Ground Truth**: NO **Healthy**

User Bias: People who take daily vitamins tend to be more health-conscious overall and differ in diet, exercise, smoking, care-seeking, and socioeconomic status. **Why Flawed**:

Without randomization, we cannot separate the effect of vitamins from the effect of being the type of person who takes vitamins.

Ecological Family (W5)

General → Specific Inference Errors: Aggregate ≠ Individual

3.4 W5: Ecological Fallacy (Core Tier)

3.4.1 Definition and Structure

W5: Ecological Fallacy

Definition: Inferring individual-level causation from group-level (aggregate) data. Group correlations do not directly imply individual causal effects. **Key Issue**: Aggregate patterns can arise from composition and confounding. **Primary Reference**: Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. [Foundational paper]

3.4.2 Required Elements

#	Required Element
1	Data is at group or aggregate level (countries, states, schools)
2	Claim is about individual-level causation
3	Within-group variation is ignored
4	Individual-level relationship could differ

Table 11: W5 Required Elements Checklist

3.4.3 Domain Examples

#	Domain	Example (Group correlation, individual claim)
1	Health	Countries: chocolate consumption → Nobel prizes
2	Education	States: class size → test scores
3	Economics	Countries: X → GDP per capita
4	Crime	Cities: police presence → crime rates
5	Environment	Countries: car ownership → lifespan
6	Politics	States: turnout → outcomes
7	Social	Countries: religiosity → happiness
8	Diet	Countries: fat intake → heart disease
9	Technology	Countries: phone use → productivity
10	Labor	Countries: immigration → wages

Table 12: W5 Domain Examples

3.4.4 Complete Example Case

Example: W5-HEALTH-001 (Easy Difficulty)

Scenario: A study found that countries with higher per-capita chocolate consumption have more Nobel Prize winners per capita. The researchers suggested that eating chocolate improves cognitive function and increases your chances of winning a Nobel Prize. **Claim**:

"Eating chocolate causes improved cognitive function (individual level)." **Ground Truth**:

NO **Ecological Fallacy**: The data is country-level, but the claim is about individuals.

The analysis does not show that Nobel Prize winners personally eat more chocolate. **Why**

Flawed: Aggregate correlation cannot establish individual-level causation.

Confounding Family (W7)

Third Variable Explains X–Y Correlation

3.5 W7: Confounding (Core Tier)

3.5.1 Definition and Structure

W7: Confounding

Definition: A third variable Z causes both the claimed cause X and the claimed effect Y, making the X–Y correlation spurious. **Causal Structure**:

$$Z \to X \quad \text{and} \quad Z \to Y$$

The observed association between X and Y does not establish that X causes Y. **Primary Reference**: Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press. Chapter 3.

3.5.2 Required Elements

#	Required Element
1	X–Y correlation is described (observational evidence)
2	Z is present in scenario (explicit or strongly implied)
3	Z → X relationship is plausible
4	Z → Y relationship is plausible
5	Causal claim "X causes Y" is explicitly made
6	No mention of controlling for Z or randomization

Table 13: W7 Required Elements Checklist

3.5.3 Scenario Template

Template: "Studies show that people who [do X] tend to [have Y]. Researchers concluded that [X causes Y]." *Hidden in scenario*: Z (for example, socioeconomic status, age, or lifestyle) explains both X and Y.

3.5.4 Domain Examples

#	Domain	Example (X → Y, confounded by Z)
1	Health	Exercise → Longevity (Z = SES, genetics)
2	Education	Private school → College admission (Z = family wealth)
3	Business	Training program → Productivity (Z = motivated employees)
4	Technology	Early adoption → Success (Z = tech-savvy personality)
5	Finance	Financial advisor → Wealth (Z = already wealthy)
6	Environment	Organic food → Health (Z = health-conscious lifestyle)
7	Social	Marriage → Happiness (Z = personality traits)
8	Medicine	Vitamin supplements → Health (Z = health awareness)
9	Sports	Expensive equipment → Performance (Z = dedication)
10	Policy	Neighborhood programs → Safety (Z = community engagement)

Table 14: W7 Domain Examples

3.5.5 Difficulty Levels

Level	Description
Easy	Z is explicitly mentioned in the scenario
Medium	Z is implied but not named
Hard	Z is not mentioned; must be inferred from context

Table 15: W7 Difficulty Calibration

3.5.6 Validation Checklist

☐ Is Z a plausible common cause of both X and Y?

☐ Would controlling for Z reduce or eliminate the X–Y association?

☐ Is the scenario realistic and domain-appropriate?

☐ Is the causal claim clearly stated?

☐ Is ground truth unambiguously NO?

3.5.7 Complete Example Case

Example: W7-HEALTH-001 (Medium Difficulty)

Scenario: A large observational study followed 50,000 adults over 10 years. The researchers found that people who drink wine regularly have 25% lower rates of heart disease compared to non-drinkers. The study concluded that moderate wine consumption protects against heart disease. **Claim**: "Wine consumption causes reduced heart disease risk." **Ground Truth**: NO **Confounder (Z)**: Socioeconomic status and overall lifestyle factors. Wine drinkers tend to have higher income, better diet, more exercise, and better access to healthcare. **Why Flawed**: The study is observational. People who choose to drink wine moderately differ systematically from non-drinkers in ways that independently affect heart disease risk.

Direction Family (W9–W10)

Causal Arrow Direction or Timing Errors

3.6 W9: Reverse Causation (Core Tier)

3.6.1 Definition and Structure

W9: Reverse Causation

Definition: The claimed causal direction is wrong. Y actually causes X, not X causes Y. The association exists, but the arrow points in the opposite direction. **Key Issue**: Direction is not identified by correlation alone; timing or design is required. **Primary Reference**: Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* (3rd ed.). Lippincott Williams & Wilkins. Chapter 9.

3.6.2 Required Elements

#	Required Element
1	X–Y association is described
2	Claim states X causes Y
3	Y causing X is plausible (often more plausible)
4	No evidence that rules out reverse direction

Table 16: W9 Required Elements Checklist

3.6.3 Domain Examples

#	Domain	Example (Claimed X→Y, likely Y→X)
1	Health	Inactivity → Depression (reverse plausible)
2	Education	Library visits → Reading ability
3	Business	Confidence → Success
4	Social	Social media → Loneliness
5	Economics	Poor health → Poverty
6	Psychology	Avoidance → Anxiety
7	Medicine	Treatment-seeking → Illness severity
8	Technology	Tool adoption → Skill
9	Finance	Financial literacy → Wealth
10	Relationships	Communication issues → Conflict

Table 17: W9 Domain Examples

3.6.4 Complete Example Case

Example: W9-SOC-001 (Medium Difficulty)

Scenario: A study found that teenagers who spend more time on social media report higher levels of loneliness. The researchers concluded that social media use causes loneliness in teenagers and recommended that parents limit their children's screen time. **Claim**: "Social media use causes loneliness." **Ground Truth**: NO **Reverse Causation**: Lonely teenagers may turn to social media seeking connection. Loneliness can drive social media use. **Why Flawed**: Without a design that identifies direction (for example, longitudinal timing or random assignment), the causal arrow is not determined.

3.7 W10: Post Hoc Fallacy (Core Tier)

3.7.1 Definition and Structure

W10: Post Hoc Fallacy

Definition: The assumption that because Y followed X in time, X must have caused Y (post hoc ergo propter hoc). **Key Issue**: Temporal sequence alone does not establish causation. **Primary Reference**: Kahneman, D. (2011). *Thinking, Fast and Slow.* Farrar, Straus and Giroux. Chapter 19: The Illusion of Understanding.

3.7.2 Required Elements

#	Required Element
1	Clear temporal sequence (X happened before Y)
2	Causal claim based primarily on timing
3	No mechanistic or controlled evidence
4	Alternative explanations plausible (coincidence, natural progression)

Table 18: W10 Required Elements Checklist

3.7.3 Domain Examples

#	Domain	Example (X before Y, not causal)
1	Medicine	Took remedy, cold resolved (natural recovery)
2	Sports	Changed routine, then won (coincidence)
3	Business	New CEO, profits rose (market conditions)
4	Weather	Rain dance, then rain (coincidence)
5	Technology	Software update, speed improved (other changes)
6	Policy	New law, crime dropped (trend already occurring)
7	Personal	Lucky socks, passed exam (chance)
8	Agriculture	Planting ritual, good harvest (weather)
9	Health	Started supplement, felt better (placebo)
10	Economics	Tax cut, economy grew (global factors)

Table 19: W10 Domain Examples

3.7.4 Complete Example Case

Example: W10-MED-001 (Easy Difficulty)

Scenario: A patient suffering from a cold took an herbal remedy recommended by a friend. Three days later, the cold symptoms were completely gone. The patient concluded that the herbal remedy cured the cold and recommended it to others. **Claim**: "The herbal remedy cured the cold." **Ground Truth**: NO **Post Hoc Fallacy**: The only evidence is that the remedy preceded recovery. Common colds often resolve in 3–7 days without treatment. **Why Flawed**: Without a control group, we cannot attribute the recovery to the remedy rather than natural healing.

4 SHEEP Templates (Justified Causal Claims)

SHEEP cases S1–S5 present scenarios where a causal claim IS justified by the evidence. The LLM should answer **YES** for all SHEEP cases.

4.1 S1: Randomized Controlled Trial (Core Tier)

4.1.1 Definition and Structure

S1: Randomized Controlled Trial

Definition: Participants are randomly assigned to treatment or control groups. Randomization eliminates confounding by ensuring groups are comparable.

Why Valid: Random assignment ensures any difference in outcomes can be attributed to the treatment.

4.1.2 Required Elements

#	Required Element
1	Random assignment explicitly stated
2	Treatment group and control group identified
3	Outcome measured in both groups
4	Difference attributed to treatment

Table 20: S1 Required Elements Checklist

4.1.3 Key Phrases to Include

- "randomly assigned," "randomized," "random allocation"
- "treatment group vs. control group"
- "placebo-controlled" (strengthens validity)
- "double-blind" (strengthens validity)

4.1.4 Complete Example Case

Example: S1-EDU-001

Scenario: A school district wanted to evaluate a new math curriculum. They randomly assigned 500 students to either the new curriculum or the standard curriculum, ensuring that classrooms were balanced on prior math ability, socioeconomic status, and school. After one year, students in the new curriculum scored 12 points higher on standardized tests than the control group ($p < 0.001$).

Claim: "The new curriculum causes improved math performance."

Ground Truth: YES

Why Valid: Random assignment ensures that the treatment and control groups are comparable. The 12-point difference can be attributed to the curriculum, not to confounders.

4.2 S2: Natural Experiment (Core Tier)

4.2.1 Definition and Structure

S2: Natural Experiment

Definition: An external event (policy change, natural disaster, arbitrary rule) creates variation in treatment that is "as-if" random.

Why Valid: The external event assigns treatment independent of potential outcomes.

4.2.2 Required Elements

#	Required Element
1	External event/policy creates variation in X
2	Event is plausibly exogenous (not caused by Y)
3	Comparison between affected and unaffected groups
4	Groups were similar before the event

Table 21: S2 Required Elements Checklist

4.2.3 Complete Example Case

Example: S2-ECON-001

Scenario: In 2015, State A suddenly raised its minimum wage from \$7.25 to \$10.00 per hour, while neighboring State B kept its minimum wage at \$7.25. Researchers compared employment in counties along the state border before and after the change. Before 2015, employment trends in border counties were nearly identical. After the increase, employment in State A's border counties declined by 3% relative to State B's.

Claim: "The minimum wage increase caused reduced employment."

Ground Truth: YES

Why Valid: The policy change was exogenous. Border counties are highly comparable. Workers didn't choose which state to be in based on future minimum wage changes.

4.3 S3: Lottery / Quasi-Random (Core Tier)

4.3.1 Definition and Structure

S3: Lottery / Quasi-Random

Definition: Treatment is assigned by a lottery or random process among those who want the treatment.

Why Valid: All applicants wanted the treatment; the lottery randomly determines who gets it.

4.3.2 Required Elements

#	Required Element
1	Lottery or random assignment mechanism described
2	Winners vs. non-winners compared
3	Both groups wanted the treatment (no self-selection)
4	Outcome measured for both groups

Table 22: S3 Required Elements Checklist

4.3.3 Complete Example Case

Example: S3-HEALTH-001

Scenario: In 2008, Oregon expanded Medicaid but had limited spots. They used a lottery to randomly select 30,000 winners from 90,000 low-income applicants. Researchers compared lottery winners (who received Medicaid) to non-winners (who remained uninsured). Winners had 40% higher rates of outpatient care and significantly lower rates of depression.

Claim: "Medicaid coverage causes increased healthcare utilization."

Ground Truth: YES

Why Valid: All 90,000 applicants wanted Medicaid coverage. The lottery randomly determined who got coverage. Any difference can be attributed to having insurance.

4.4 S4: Controlled Ablation (Core Tier)

4.4.1 Definition and Structure

S4: Controlled Ablation

Definition: X is removed or disabled in a controlled manner, and Y disappears or changes as a result.

Why Valid: If removing X eliminates Y (while everything else stays constant), X must be causing Y.

4.4.2 Required Elements

#	Required Element
1	X was present and Y was observed
2	X is removed in a controlled manner
3	Y disappears or significantly changes
4	Nothing else changed during the ablation

Table 23: S4 Required Elements Checklist

4.4.3 Complete Example Case

Example: S4-BIO-001

Scenario: To test whether gene ABC is necessary for tumor growth, researchers created genetically identical mice with gene ABC knocked out. When exposed to a carcinogen under identical conditions, normal mice developed tumors within 8 weeks, but ABC-knockout mice did not develop any tumors even after 16 weeks.

Claim: "Gene ABC causes (is necessary for) tumor growth."

Ground Truth: YES

Why Valid: The only difference between the two groups is the presence or absence of gene ABC. All other factors were identical. The disappearance of tumors when ABC is removed demonstrates ABC's causal role.

4.5 S5: Mechanism + Dose-Response (Core Tier)

4.5.1 Definition and Structure

S5: Mechanism + Dose-Response

Definition: A known mechanistic pathway connects X to Y, AND there is a dose-response relationship (more X leads to more/less Y).

Why Valid: The combination of mechanism + dose-response gradient provides strong evidence for causation.

4.5.2 Required Elements

#	Required Element
1	Known mechanism connecting X to Y is described
2	Dose-response relationship (gradient effect) is shown
3	Higher exposure leads to stronger effect
4	Mechanism is scientifically established

Table 24: S5 Required Elements Checklist

4.5.3 Complete Example Case

Example: S5-TOX-001

Scenario: Lead is known to interfere with neurotransmitter function by blocking calcium channels in neurons. Studies of children across multiple countries show a clear dose-response relationship: blood lead levels of 5, 10, and 20 μg/dL are associated with IQ reductions of approximately 2, 4, and 8 points respectively.

Claim: "Lead exposure causes cognitive impairment."

Ground Truth: YES

Why Valid: (1) Known biological mechanism. (2) Clear dose-response gradient. The combination provides strong causal evidence.

5 Advanced Cases (Reference Only)

5.1 W4: Regression to Mean (Advanced Tier)

5.1.1 Definition and Structure

W4: Regression to Mean (Advanced)

Definition: When cases are selected based on extreme values, subsequent measurements naturally tend to move toward the average, regardless of intervention. This is often mistaken for a treatment effect. **Key Issue**: Extreme values partly reflect noise; on retest, noise tends to average out. **Primary Reference**: Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263. [Original discovery]

5.1.2 Required Elements

#	Required Element
1	Selection based on extreme Y values (highest or lowest)
2	Intervention X applied after selection
3	Retest shows movement toward average
4	Claim attributes change to intervention, not statistical artifact
5	No control group of similarly extreme cases without intervention

Table 25: W4 Required Elements Checklist

5.1.3 Key Signal Words

- "worst performers," "lowest scores," "most severe cases"
- "top performers," "highest scores," "best cases"
- followed by "improvement" or "decline" on retest

5.1.4 Complete Example Case

Example: W4-EDU-001 (Medium Difficulty)

Scenario: A school identified the 50 students with the lowest scores on the fall math exam (bottom 10%) and enrolled them in an intensive after-school tutoring program. On the spring exam, these students' scores improved by an average of 15 points. The school board concluded that the tutoring program was highly effective. **Claim**: "The tutoring program caused the score improvement." **Ground Truth**: NO **Regression to Mean**: Students were selected because they scored extremely low. Some of that low performance reflects temporary factors. On retest, scores can improve even without tutoring. **Why Flawed**: Without a control group of equally low-scoring students who did not receive tutoring, the tutoring effect cannot be separated from regression to the mean.

5.2 W6: Base Rate Neglect (Advanced Tier)

5.2.1 Definition and Structure

W6: Base Rate Neglect (Advanced)

Definition: Ignoring the prior probability (base rate) of a condition when interpreting diagnostic evidence. This often involves confusing $P(A \mid B)$ with $P(B \mid A)$. **Key Issue**: A highly sensitive test does not imply a high probability of the condition after a positive result when the base rate is low. **Primary Reference**: Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.

5.2.2 Required Elements

#	Required Element
1	Low base rate is stated or implied
2	Test properties are provided (sensitivity and specificity, or equivalent)
3	Interpretation confuses conditional probabilities
4	Claim overstates P(Condition | Positive)

Table 26: W6 Required Elements Checklist

5.2.3 Complete Example Case

Example: W6-MED-001 (Medium Difficulty)

Scenario: A screening test for a rare disease (affecting 1 in 1,000 people) has 99% sensitivity and 95% specificity. A patient tests positive. The doctor says they almost certainly have the disease because the test is "99% accurate." **Claim**: "A positive test means you almost certainly have the disease." **Ground Truth**: NO **Base Rate Neglect**: The doctor confused P(Positive | Disease) with P(Disease | Positive). **Illustration** (per 100,000 people):

- 100 have disease → 99 test positive (true positives)
- 99,900 do not have disease → 4,995 test positive (false positives)
- P(Disease | Positive) $= 99/(99 + 4{,}995) \approx 2\%$

5.3 W8: Simpson's Paradox (Advanced Tier)

5.3.1 Definition and Structure

W8: Simpson's Paradox (Advanced)

Definition: A trend that appears in aggregate data reverses or disappears when data is stratified by a variable Z. **Key Issue**: Pooling heterogeneous subgroups can produce misleading aggregate comparisons; numbers must be included. **Primary Reference**: Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404.

5.3.2 Required Elements

#	Required Element
1	Aggregate data shows X → Y pattern
2	Stratified data shows an opposite pattern or invalidates the conclusion
3	Z is related to both X and Y
4	Numerical data provided in scenario

Table 27: W8 Required Elements Checklist

5.3.3 Complete Example Case

Example: W8-MED-001 (Medium Difficulty)

Scenario: Hospital A has an overall surgery success rate of 90%, while Hospital B has 85%. A health board recommended Hospital A. However, the detailed data shows:

- **Low-risk patients**: Hospital A: 95%; Hospital B: 98%
- **High-risk patients**: Hospital A: 65%; Hospital B: 70%

Hospital A treats 80% high-risk patients; Hospital B treats only 20%. **Claim**: "Hospital A provides better surgical care overall." **Ground Truth**: NO **Simpson's Paradox**: Hospital B has better outcomes within both risk groups. Hospital A's higher overall rate reflects a different patient mix, not better care.

5.4 S6: Instrumental Variable (Advanced Tier)

S6: Instrumental Variable (Advanced)

Definition: An instrument Z affects Y *only through* X, allowing estimation of causal effect even with confounding.

Structure: $Z \to X \to Y$ (Z affects Y only through X)

Example: Using "distance to college" as instrument for education's effect on earnings.

5.5 S7: Difference-in-Differences (Advanced Tier)

S7: Difference-in-Differences (Advanced)

Definition: Compare change in outcomes over time between treatment and control groups.

Key Assumption: Parallel trends—absent treatment, both groups would follow similar trajectories.

Example: Comparing smoking rates in states that raised cigarette taxes vs. states that didn't.

5.6 S8: Regression Discontinuity (Advanced Tier)

S8: Regression Discontinuity (Advanced)

Definition: Treatment assigned based on whether a continuous variable crosses a cutoff. Units just above and below are nearly identical except for treatment.

Why Valid: Near the cutoff, assignment is as-if random.

Example: Scholarship awarded to students scoring 70+; compare students scoring 69 vs. 70.

6 References

6.1 Selection Family (W1–W4)

W1: Selection Bias

- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.
- Hernán, M. A., Hernández-Díaz, S., & Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15(5), 615–625.
- Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3), 47–53.

W2: Survivorship Bias

- Wald, A. (1943). A method of estimating plane vulnerability based on damage of survivors. Statistical Research Group, Columbia University.
- Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. *Review of Financial Studies*, 5(4), 553–580.

W3: Healthy User Bias

- Shrank, W. H., Patrick, A. R., & Brookhart, M. A. (2011). Healthy user and related biases in observational studies of preventive interventions. *Journal of General Internal Medicine*, 26(5), 546–550.
- Petitti, D. B. (2002). Hormone replacement therapy and heart disease prevention: Experimentation trumps observation. *JAMA*, 288(1), 99–101.

W4: Regression to Mean

- Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.
- Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34(1), 215–220.

6.2 Ecological Family (W5–W6)

W5: Ecological Fallacy

- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357.
- Freedman, D. A. (1999). Ecological inference and the ecological fallacy. In *International Encyclopedia of the Social & Behavioral Sciences* (Vol. 6, pp. 4027–4030). Elsevier.

W6: Base Rate Neglect

- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233.

6.3 Confounding Family (W7–W8)

W7: Confounding

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

- Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1), 29–46.
- Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

W8: Simpson's Paradox

- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B*, 13(2), 238–241.
- Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175), 398–404.
- Pearl, J. (2014). Comment: Understanding Simpson's paradox. *The American Statistician*, 68(1), 8–13.

6.4 Direction Family (W9–W10)

W9: Reverse Causation

- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* (3rd ed.). Lippincott Williams & Wilkins.
- Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1), R89–R98.

W10: Post Hoc Fallacy

- Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
- Gilovich, T. (1991). *How We Know What Isn't So: The Fallibility of Human Reason in Everyday Life*. Free Press.
- Hume, D. (1748). *An Enquiry Concerning Human Understanding*. Section VII.