T3-L3: Counterfactual Reasoning Benchmark

Design Document and Execution Plan

Version 1.0

Edward Y. Chang, Longling Geng

January 19$^{\text{th}}$, 2026

Abstract

T3-L3 probes Large Language Model (LLM) counterfactual judgment, the top level of Pearl's causal hierarchy. While T3-L1 evaluates whether a causal claim is justified and T3-L2 evaluates what information would resolve causal ambiguity, T3-L3 evaluates judgments about alternative worlds: "what would have happened if X had been different?"

This document provides: (1) the foundation in Pearl's L3 counterfactual calculus, (2) a taxonomy of 8 families with domain-specific subtypes (98 subtypes; 100 pilot cases), (3) a case format centered on explicit invariants and counterfactual dependence (including but-for style judgments where appropriate), (4) a validation protocol, and (5) preliminary results (44–55% accuracy across five models).

Contents

1 Motivation: Why Include L3 in T3?

1.1 Why L3 Matters

Counterfactual reasoning asks "what would have happened if" and sits at the top of Pearl's hierarchy. It is qualitatively different from association and intervention because it requires evaluating an alternative world while holding specified invariants fixed. In practice, it underlies judgments such as:

- **Liability**: "But for the defendant's action, would the harm have occurred?"
- **Attribution**: "Was the intervention responsible for the outcome?"
- **Regret**: "If I had chosen differently, would things be better?"
- **Explanation**: "Why did this happen rather than something else?"

1.2 Pearl's Ladder of Causation

Pearl's Ladder of Causation [Pearl(2009)] distinguishes three levels:

Level	Name	Query	Cognitive Task
L1	Association	$P(Y \mid X)$	Seeing
L2	Intervention	$P(Y \mid do(X))$	Doing
L3	**Counterfactual**	$P(Y_x \mid X', Y')$	**Imagining**

1.3 What T3-L3 Tests

The T3-L3 Core Question

"If X had not occurred, would Y still have happened?"
A counterfactual judgment requires:

- **Invariants**: what is held fixed across worlds,
- **Mechanisms**: how changes propagate through causal structure,
- **Overdetermination awareness**: whether multiple causes suffice for Y,
- **Underdetermination detection**: when the claim cannot be resolved from the stated evidence.

The L3 Procedure: Abduction

Counterfactual evaluation is often formalized as:

1. **Abduction**: infer the relevant latent state from observed evidence,
2. **Action**: modify the antecedent (set $X \leftarrow x$),
3. **Prediction**: propagate the change under declared invariants to obtain Y_x.

1.4 Scientific Goals

T3-L3 supports three scientific goals:

Goal	Research Question
G1	**When LLMs match human judgments, what enables success?** Which cues or reasoning strategies reliably support correct counterfactual judgments?
G2	**When LLMs fail, what fails systematically?** Which failure patterns recur across families, domains, and linguistic forms?
G3	**What do these patterns suggest about model mechanisms?** How do representation, attention, and decoding interact to produce counterfactual outputs?

1.5 Expected Outcomes and Limitations

Setting Realistic Expectations

What we know:

- LLMs are trained on text via maximum likelihood and do not carry explicit causal models.
- Pilot experiments yield 44–55% accuracy across five models.

Baselines to report alongside accuracy:

- **Uniform-guess baseline** for three labels is 33%.
- **Majority-class baseline** depends on the label distribution (for the pilot, the majority label is `CONDITIONAL`).

What we aim to obtain:

- **Family profiles**: which families are easiest and hardest, and why.
- **Failure clusters**: interpretable error modes tied to invariants, mechanisms, or overdetermination.
- **A progress baseline**: a stable yardstick for future architectures and training regimes.

1.6 How T3-L3 Fits in T3

Benchmark	Core Question	Task	Output
T3-L1	Is this causal claim justified?	Binary	YES/NO
T3-L2	What info would resolve this?	Disambiguation	Trap + Refusal
T3-L3	**Is this counterfactual valid under stated invariants?**	**Ternary**	**V/I/C**

2 Theoretical Foundation: Counterfactual Semantics

2.1 Structural Causal Models (SCM)

A counterfactual Y_x ("the value Y would have taken if X had been x") is computed via:

1. **Abduction**: given observations, infer exogenous variables U,
2. **Action**: replace the equation for X with $X := x$,
3. **Prediction**: compute Y in the modified model.

2.2 The Three Answer Types

T3-L3 cases have three possible ground truth answers:

Answer	Meaning
VALID	The claim is supported under the stated invariants: changing X would change Y, or would materially change the probability of Y.
INVALID	The claim is not supported: changing X would not change Y under the stated invariants (for example due to overdetermination, spurious linkage, or causal independence).
CONDITIONAL	The scenario underdetermines the answer: reasonable completions of missing invariants lead to different conclusions.

Policy for stochastic scenarios. If the counterfactual is worded deterministically ("would"), then a purely stochastic link typically forces `CONDITIONAL` unless the scenario pins down a near-deterministic mechanism or an explicit probability threshold. If the claim is probabilistic ("more likely," "reduces risk"), then `VALID` can be used when a material probability shift follows from stated mechanisms.

2.3 Key Counterfactual Concepts

But-for causation. X is a but-for cause of Y if Y would not have occurred but for X. Formally: $Y_{x=0} = 0$ when $Y_{x=1} = 1$.

Overdetermination. When multiple causes each suffice for Y, removing one does not prevent Y. Example: two assassins act independently; removing either still results in death.

Preemption. An early cause brings about Y and blocks a backup cause. The early cause is the actual cause even though the backup would have sufficed.

Invariants. Variables held fixed across worlds. Different invariant choices can yield different counterfactual conclusions, so invariants must be stated explicitly in each case.

3 Taxonomy of Counterfactual Trap Types

T3-L3 groups counterfactual challenges into **8 theory-grounded families** (F1–F8) plus a separate bucket for **domain extensions**. Across the pilot, the taxonomy contains **98 subtypes** instantiated into **100 cases**.

3.1 Coverage Summary (F1–F8 + Domain Extensions)

Family	Core challenge	Subtypes	Cases	Reference
F1: Deterministic	Mechanistic or rule-based necessity	17	19	[Lewis(1973)]
F2: Probabilistic	Uncertainty and stochastic outcomes	5	5	[Pearl(2009)]
F3: Overdetermination	Multiple sufficient causes, preemption	6	6	[Mackie(1974)]
F4: Structural	Trigger vs. background structure	9	9	[Halpern(2016)]
F5: Temporal	Timing, sequencing, path dependence	8	8	[Woodward(2003)]
F6: Epistemic	Underdetermination and unknowability	13	13	[Dawid(2000)]
F7: Attribution	Credit assignment and contribution	16	16	[Shpitser(2018)]
F8: Moral/Legal	Responsibility under norms and standards	4	4	[Hart and Honoré(1959)]
Domain extensions	Application-specific counterfactuals	20	20	—
Total		**98**	**100**	

Domain extensions policy. These cases require substantial domain knowledge (for example, AI systems, markets, military strategy) and do not map cleanly onto a single theoretical family. They are retained for ecological validity, but excluded from family-level error analysis to keep analysis interpretable.

Promotion rule. A domain extension may be promoted into F1–F8 if it can be rewritten so that (i) the core causal mechanism is expressible without specialized theory and (ii) the case's main difficulty matches an existing family's guiding question.

3.2 Family 1: Deterministic Counterfactuals

F1: Deterministic (19 cases). Guiding question: "Would the mechanism still operate?"

Counterfactuals governed by physical, logical, or rule-based necessity, where a correct judgment hinges on identifying an invariant mechanism rather than extrapolating from surface association.

Representative subtype clusters:

- **Mechanistic necessity**: removing an essential component breaks the outcome.
- **Rule-based determinism**: outcomes fixed by explicit rules (formal constraints, protocols).
- **Necessary condition**: the outcome cannot occur without X, given stated invariants.
- **Valid state comparison**: the counterfactual is resolved by comparing known states under the same rules.
- **Spurious linkage**: the scenario explicitly lacks a causal connection (superstition).

3.3 Family 2: Probabilistic Counterfactuals

F2: Probabilistic (5 cases). Guiding question: "How does uncertainty change what can be concluded?"

Counterfactuals with stochastic outcomes, where the task is to distinguish changes in probability from deterministic claims, and to respect background risk.
Representative subtype clusters:

- **Sufficiency-style queries under uncertainty**: individual-level dependence with stochasticity.
- **Probabilistic exposure**: causal links that are real but non-deterministic.
- **Background risk**: outcomes that can occur without X at non-trivial rates.
- **Sensitivity/chaos**: small changes can yield divergent trajectories, limiting determinacy.
- **Chance vs. necessity**: separating "could have happened anyway" from mechanistic dependence.

3.4 Family 3: Overdetermination

F3: Overdetermination (6 cases). Guiding question: "Would another cause have sufficed?"

Cases where more than one cause is sufficient for the outcome. The central difficulty is distinguishing necessity, sufficiency, and preemption under a but-for style query.
Representative subtype clusters:

- **Symmetric overdetermination**: multiple sufficient causes occur together.
- **Preemption**: an early cause brings about Y and blocks a backup cause.
- **Simultaneous lethal actions**: "double-assassin" style structures.
- **Threshold effects**: several factors jointly push the system past a threshold.

3.5 Family 4: Structural vs. Contingent Causes

F4: Structural (9 cases). Guiding question: "Was this the trigger or the root cause?"

Distinguishes proximate triggers from background enabling conditions. The trap is attributing the outcome to the most salient event when structural forces dominate.
Representative subtype clusters:

- **Trigger vs. structure**: spark vs. fuel.
- **Agent vs. system**: individual action vs. underlying constraints.
- **Technological and institutional framing**: invention/policy vs. structural necessity.
- **Strategy vs. resources**: contingent tactics vs. structural capacity.

3.6 Family 5: Temporal and Path-Dependent

F5: Temporal (8 cases). Guiding question: "Does timing or path matter?"

Counterfactuals where sequencing, delays, windows, or accumulated history changes what can be held fixed. The trap is treating the world as memoryless.
Representative subtype clusters:

- **Path dependence**: early choices constrain later options.
- **Timing windows**: the same action can succeed or fail depending on when it occurs.
- **Chain framing**: proximate vs. distal links in a causal chain.
- **Downstream propagation**: separating warranted propagation from speculation.

3.7 Family 6: Epistemic Limits

F6: Epistemic (13 cases). Guiding question: "Is the counterfactual resolvable from what is stated?"

Cases where the scenario underdetermines the answer, either because key mechanisms are unknown, measurements are intrusive, or the counterfactual changes identity conditions.
Representative subtype clusters:

- **Unverifiable counterfactuals**: no feasible test or identifying evidence.
- **Mechanism dependence**: the answer hinges on an unstated mechanism.
- **Observer effects**: measuring or intervening changes the system.
- **Non-identity**: the alternative world implies a different individual or reference class.

3.8 Family 7: Causal Attribution

F7: Attribution (16 cases). Guiding question: "How much credit does X deserve?"

Attribution cases quantify or compare causal contribution rather than asking only for a binary but-for judgment. The trap is collapsing contribution into necessity or into moral blame.
Representative subtype clusters:

- **Attributable fraction**: population-level contribution.
- **Sufficiency-style individual attribution**: contribution given an observed outcome.
- **Path-specific effects**: distinguishing direct vs. mediated contribution.
- **Principal strata / complier logic**: contribution defined for specific latent groups.
- **Additionality**: "would it have happened anyway?"

3.9 Family 8: Moral and Legal Causation

F8: Moral/Legal (4 cases). Guiding question: "Who is responsible under a standard?"

Counterfactuals embedded in normative standards (legal proof thresholds, omissions, intent), where correctness depends on matching the stated standard rather than importing outside norms.

Representative subtype clusters:

- **But-for under uncertainty**: causal standards with probabilistic evidence.
- **Moral luck**: identical actions with divergent outcomes.
- **Action vs. omission**: doing harm vs. allowing harm.
- **Process effects**: selection into legal outcomes (for example, pleas) altering observed histories.

3.10 Domain Extensions (20 cases)

Domain extensions (20 cases). Guiding question: "Is the counterfactual answerable without specialized domain theory?"

Application-driven counterfactuals that are realistic but not cleanly attributable to a single family.

Examples:

- **AI/Technology**: base-model capability, emergent behaviors, scaling constraints.
- **Finance/Markets**: valuation mechanisms, liquidity vs. structure, market impact.
- **Military/Strategy**: defense efficacy, asymmetry, deterrence logic.
- **Business**: strategy choice, business-model viability, organizational constraints.
- **Science**: consensus dynamics, evolutionary and natural history counterfactuals.

Policy for analysis: keep these in the benchmark release, but exclude them from family-level theoretical plots unless a clear mapping is later established.

4 Case Structure

4.1 Case Schema (Required Fields)

Each T3-L3 case is a self-contained text instance with a fixed set of fields. All information required to judge the counterfactual must be present in the `Scenario`.

Field	Specification
`CaseID`	Unique identifier (e.g., `F1-D1-001`).
`Domain`	One of the predefined domains (D1–D10).
`Family/Subtype`	Trap family (F1–F8, or `DomainExt`) and subtype label.
`Difficulty`	`Easy`, `Medium`, or `Hard`.
`Scenario`	World A narrative describing facts, events, and any relevant mechanisms. No external data required.
`CounterfactualClaim`	The explicit claim: "If X had been different, then Y".
`Variables`	Identify X (antecedent), Y (consequent), and Z (context/mechanism variables).
`Invariants`	1–3 bullets specifying what is held fixed across worlds; if unknown, state explicitly.
`GroundTruth`	One of `VALID`, `INVALID`, `CONDITIONAL`.
`Justification`	Short explanation grounded only in the scenario and listed invariants.
`WiseResponse`	Expected high-quality reasoning style for a strong model (brief, structured).

4.2 Field Constraints and Quality Checks

- **Self-contained**: the scenario includes all facts needed; no external knowledge is required to establish the intended label.
- **Single antecedent and consequent**: X and Y are identifiable and not conflated with multiple unrelated changes.
- **Invariant clarity**: invariants are explicit; if the verdict is `CONDITIONAL`, the missing invariants are explicitly listed.
- **Determinacy**: the label is defensible under a normal reading and stable under reasonable paraphrases of the claim.
- **Annotation correctness**: the chosen family/subtype matches the core reasoning challenge in the case.

4.3 Labeling Policy: Determinacy and "CONDITIONAL"

A case is labeled `CONDITIONAL` when the scenario and invariants do not determine a unique answer, and at least two reasonable completions of missing invariants lead to different labels. Annotators must list the missing invariants and provide two short completions showing why the label depends on them.

A case is labeled `VALID` or `INVALID` only when the scenario and invariants jointly pin down the relevant mechanism or independence relation without importing outside facts.

5 Target Scale and Assignment

5.1 Scaling to 1,000 Cases

The current 100-case pilot will be expanded to 1,000 cases, with priority given to under-represented families:

Family	**Description**	**Current**	**Target**	**Priority**
F1	Deterministic	19	150	Normal
F2	Probabilistic	5	120	**High**
F3	Overdetermination	6	100	**High**
F4	Structural	9	120	Normal
F5	Temporal	8	100	Normal
F6	Epistemic	13	120	Normal
F7	Attribution	16	140	Normal
F8	Moral/Legal	4	100	**High**
Subtotal (Theoretical)		*80*	*950*	
DomainExt	Domain extensions	20	50	Low
Total		**100**	**1,000**	

5.2 Step-by-Step Procedure

Day 1: Study existing cases.

1. Read all 100 pilot cases.
2. Study the cases in your assigned family carefully.
3. Identify the core reasoning pattern that defines your family.
4. Note the distribution of `VALID`/`INVALID`/`CONDITIONAL`.

Day 2: Identify subtypes and gaps.

1. List all subtypes currently in your family (Section 3).
2. Identify subtypes with only 1 case; these need expansion.
3. Propose 3–5 new subtypes not currently covered.
4. Document each proposed subtype with a short pointer to relevant causal reasoning concepts or literature.

Days 3–4: Draft new cases. For each new case, provide all fields from the case schema (Section 4.1):

```
CaseID:             [Family]-[Domain]-[Number]
Domain:             D1--D10
Family/Subtype:     F[1-8] / [Subtype name]   (or DomainExt / [Subtype])
Difficulty:         Easy / Medium / Hard

Scenario:           [2-5 sentences: what happened in World A]

CounterfactualClaim: "If [X had been different], then [Y]."

Variables:
  X = [Antecedent]
  Y = [Consequent]
  Z = [Mechanism/Context]

Invariants:
  - [1-3 bullets; what is held fixed across worlds]
  - [If unknown, state: "Not specified: ..."]

GroundTruth:          VALID / INVALID / CONDITIONAL

Justification:        [2-4 sentences grounded in Scenario + Invariants]

WiseResponse:         [Brief structured reasoning template]
```

Day 5: Balance the distribution.

1. Ground truth: aim for about 35% `VALID`, 25% `INVALID`, 40% `CONDITIONAL`.
2. Difficulty: aim for about 25% Easy, 45% Medium, 30% Hard.
3. Cover multiple domains; avoid clustering most cases in one domain.
4. Verify each subtype has at least 3 cases.

Day 6: Peer review.

1. Exchange cases with another group for review.
2. Check label correctness and whether the justification relies only on the scenario and invariants.
3. Check family fit and whether the main difficulty matches the guiding question.
4. Revise based on feedback.

Day 7: LLM validation (debugging and difficulty estimation).

1. Run 3 LLMs on your new cases (for example, GPT-4, Claude, Gemini).
2. Use model agreement patterns to estimate difficulty and to detect ambiguous wording.
3. Do not use LLM agreement to decide ground truth; ground truth is defined by Scenario + Invariants.
4. Cases with mixed results are often the most diagnostic; keep them if the label remains defensible.

Day 8: Final submission.

1. Format all cases in LaTeX (template provided).
2. Submit case file, validation report, and peer review notes.
3. Group 10 integrates submissions into T3-L3 v2.0.

5.3 Quality Criteria Checklist

Each case must satisfy:

Criterion	Requirement
Self-contained	Scenario includes all facts needed; no external knowledge required
Clarity	X, Y, Z, and invariants are unambiguous and well-defined
Correctness	Label is defensible; justification is sound under stated invariants
Family fit	Case clearly tests the assigned counterfactual pattern
Novelty	Not a trivial variant of an existing case
Realism	Scenario is plausible (real-world or realistic hypothetical)

5.4 Deliverables per Group

1. **Case file**: 90 new cases in LaTeX format
2. **Subtype documentation**: list of subtypes with definitions and brief rationale
3. **Validation report**: model outputs used for difficulty estimation and ambiguity checks
4. **Peer review report**: feedback given and received

5.5 Timeline Summary

Day	Milestone
1	Study existing cases; understand assigned family
2	Identify subtypes and gaps; propose new subtypes
3–4	Draft 90 new cases
5	Balance distribution; internal quality check
6	Peer review with partner group
7	LLM validation for debugging and difficulty estimation; revisions
8	Final submission; Group 10 integrates

6 Conclusion

T3-L3 addresses the highest level of Pearl's causal hierarchy: counterfactual imagination. The benchmark is deliberately conservative and text-grounded.

What T3-L3 is:

- A diagnostic probe into LLM counterfactual judgment under explicit invariants
- A tool for studying when and why LLMs match (or fail to match) human judgments
- A baseline for measuring progress as architectures and training regimes evolve

What T3-L3 is not:

- A claim that LLMs perform full SCM-based counterfactual inference
- A capability benchmark for causal modeling in the traditional sense

Key findings from the 100-case pilot:

1. **Accuracy is modest**: 44–55% accuracy suggests a substantial gap, especially relative to majority-class baselines implied by the label distribution.
2. **Scaling is not monotone**: larger models do not reliably improve, consistent with heuristic pattern matching rather than stable counterfactual competence.
3. **Systematic biases emerge**: models differ in how often they predict `VALID` vs. `CONDITIONAL`, indicating distinct strategies for handling underdetermination.

4. `CONDITIONAL` **is diagnostic**: cases that require explicit invariants test whether models can recognize epistemic limits instead of forcing a determinate answer.

Future directions:

- Expand to 1,000 cases with priority for under-represented families (F2, F3, F8)
- Interpretability analysis: connect failures to attention patterns and training signal hypotheses
- Evaluate neuro-symbolic and causal-model-augmented systems
- Use T3-L3 cases as targeted training data for improving counterfactual approximation

Together, T3-L1/L2/L3 provide a diagnostic of LLM causal cognition, measuring how far pattern-based systems can approximate causal reasoning, and where they break.

References

[Dawid(2000)] A. P. Dawid. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000.

[Halpern(2016)] J. Y. Halpern. *Actual Causality*. MIT Press, Cambridge, MA, 2016.

[Hart and Honoré(1959)] H. L. A. Hart and T. Honoré. *Causation in the Law*. Oxford University Press, Oxford, UK, 1959.

[Lewis(1973)] D. Lewis. Counterfactuals. *Journal of Philosophy*, 70(17):556–567, 1973.

[Mackie(1974)] J. L. Mackie. *The Cement of the Universe: A Study of Causation*. Oxford University Press, Oxford, UK, 1974.

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

[Shpitser(2018)] I. Shpitser. A causal query framework for statistical analysis of counterfactual fairness. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 187–196, 2018.

[Woodward(2003)] J. Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, UK, 2003.

APPENDICES

A Suggested Evaluation Protocol (RCA)

A.1 RCA Steps (Recommended Rubric)

Models may be prompted to follow the Regulated Causal Anchoring (RCA) rubric. This is recommended for analysis consistency, but is not required to define a case.

1. **LABEL**: state the verdict (`VALID/INVALID/CONDITIONAL`)
2. **GRAPH**: describe the causal structure in words (or a simple arrow sketch)
3. **INVARIANTS**: state what is held fixed across worlds
4. **JUSTIFICATION**: explain how changing X affects Y
5. **EVIDENCE**: cite specific facts from the scenario text
6. **MISSING**: if `CONDITIONAL`, state what information would resolve it

B Illustrative Examples (Non-normative)

B.1 Example A: Deterministic

Case: The Missed Flight
Scenario. Alice missed her flight (X) by 5 minutes. The plane later crashed (Y). She claims: "If I had arrived on time, I would have died."
Variables. X = Arrival Time; Y = Survival; Z = Boarding rules.
Invariants.

- Boarding rules and crash outcome are fixed.
- Arriving on time implies boarding.

Ground Truth. `VALID`
Justification. Under the scenario, arriving on time implies boarding. Since the crash was fatal, boarding implies death in the alternative world.

B.2 Example B: Conditional on an Unobserved Mediator

Case: The Bitcoin Investment
Scenario. You did not buy Bitcoin in 2010 ($X = 0$). You claim: "If I had bought $100 of Bitcoin, I would be a millionaire today."
Variables. X = Purchase; Y = Wealth; Z = Selling decision (unobserved mediator).
Invariants.

- Not specified: whether the agent holds or sells during volatility.

Ground Truth. `CONDITIONAL`
Justification. The outcome depends on whether the agent holds through later volatility. The scenario does not fix that invariant, so the claim is not determinable as stated.

B.3 Example C: No Causal Link

Case: The Lucky Shirt
Scenario. A fan wore a red shirt (X) and their team won (Y). They claim: "If I had not worn this shirt, they would have lost."
Variables. X = Shirt color; Y = Game result; Z = Player performance.
Invariants.

- Game outcome is governed by player performance and gameplay factors.
- Fan clothing does not affect player performance.

Ground Truth. `INVALID`
Justification. The outcome is causally independent of the fan's clothing choice under the stated invariants for sports outcomes.