When LLMs Default to Traffic: Systematic Domain Bias in Causal Mechanism Generation

Yahya Khan
University of Malakand, Chakdara, Pakistan
yahya.khan@uom.edu.pk

Huzaifa Khan
University of Malakand, Chakdara, Pakistan
huzaifa.khan@uom.edu.pk

Abstract

Large language models (LLMs) are increasingly deployed as reasoning engines in risk-sensitive domains where the quality of causal explanations carries direct operational consequences. Existing evaluation frameworks measure whether a model's answer is correct and whether its stated reasoning matches its internal computation, but neither assesses whether the explanation itself is causally valid given the facts of the case. We introduce **CausalAudit**, a post-hoc verification framework that audits LLM-generated explanations against a set of explicit, domain-grounded constraints motivated by principles of causal modeling: temporal precedence, spatial plausibility, mechanistic admissibility, non-spuriousness, and completeness.

Applying CausalAudit to six LLM families across 195 scenarios spanning six domains, we find that four of the five constraints are non-discriminative or weakly discriminative on our test distribution. C1, C2, and C5 are exactly 100%; C4 ranges from 89.7–100%, which is practically non-separating. The exception is mechanistic admissibility, which produces a clear and reproducible 29.2 percentage point gap (2.46x) between models, ranging from 20.0% (Llama 3.3 70B) to 49.2% (Qwen3 32B).

Detailed failure analysis reveals a systematic **"traffic accident default"** in our evaluation suite: models generate transportation-centered mechanisms for medical scenarios, a pattern that persists across explicit domain guidance, few-shot examples, and parameter adjustments. Human validation of the checker on 30 randomly sampled scenarios confirms substantial inter-rater agreement ($\kappa = 0.928$ for validity, 0.814 for agreement) and precision of 87.0%. Sensitivity analysis demonstrates that model ranking is stable across agreement thresholds $\theta \in [0.25, 0.45]$, with the absolute performance gap widening at stricter thresholds. We release the complete evaluation suite for reproducible research.

1 Introduction

Large language models are increasingly deployed as reasoning engines in risk-sensitive domains—healthcare diagnostics, infrastructure risk assessment, financial forecasting, and public safety analysis—where the quality of their causal explanations carries direct operational consequences. A fluent, grammatically correct explanation that asserts a physically impossible mechanism, misattributes causation across spatially disconnected components, or omits causally necessary factors is not merely inaccurate; it is **causally inadmissible**. Its deployment carries measurable risk of consequential misallocation, delayed intervention, or outright harm.

Despite this urgency, existing evaluation frameworks for LLMs remain structurally silent on a critical property: whether generated explanations satisfy a set of explicit, domain-grounded constraints on causal explanations. Benchmarks such as MMLU [1] and GSM8K [2] measure answer correctness—$P(\hat{y} = y \mid x, M)$—but do not ask whether the explanation E is causally admissible with respect to the scenario S and its domain facts $\mathcal{F}_S$. Faithfulness evaluations [4, 5] ask whether the explanation reflects the model's internal computation, but a model can

faithfully report its own flawed causal reasoning. The capability question ("can the model reason causally?") and the admissibility question ("is the explanation causally valid under domain facts?") are orthogonal; answering the former does not answer the latter.

We formalize this third property as causal admissibility: whether an explanation E satisfies a set of causal constraints C_i given the facts of a scenario S. Unlike correctness or faithfulness, which concern the answer or the reasoning process, admissibility concerns the explanation's compliance with domain-grounded causal constraints.

We built **CausalAudit** to audit this property, a post-hoc verification framework operationalizing causal admissibility through five constraints: **C1 (Temporal Precedence)**, **C2 (Spatial Plausibility)**, **C3 (Mechanistic Admissibility)**, **C4 (Non-Spuriousness)**, and **C5 (Completeness)**. The framework is model-agnostic with respect to model internals, provided the model output can be parsed into the required explanation representation.

Applying CausalAudit to six LLM families across 195 scenarios spanning six domains, we observe a striking and systematic pattern. Four of the five constraints are non-discriminative or weakly discriminative on our test distribution. C1, C2, and C5 are exactly 100%; C4 ranges from 89.7–100%, which is practically non-separating. We interpret this as evidence that these constraints are not sufficiently challenged by our scenario set. We flag this limitation explicitly and focus our analysis on the fifth constraint.

The fifth constraint, mechanistic admissibility, is where all the signal is. We operationalize mechanistic admissibility through two subcomponents: C_{3a} (mechanistic validity) and C_{3b} (reference-mechanism agreement). Pass rates range from 20.0% for Llama 3.3 70B to 49.2% for Qwen3 32B, a 29.2 percentage point gap (2.46x), and the gap is not explained by model scale (a 70B model performs worse than a 32B model from a different family). Detailed failure analysis reveals the central discovery of this work: a systematic **domain bias** we term the "traffic accident default." We observe in our evaluation suite that models generate traffic accident mechanisms for medical scenarios. In a case of premature hospital discharge due to bed pressure, the model generates: *"Vehicle malfunction → Brake failure → Loss of control → Collision..."* In a scenario about delayed sepsis recognition, the model produces: *"Driver failure to yield → Vehicle collision → Emergency Department response..."* This pattern persists across explicit domain guidance, few-shot examples, prohibited terminology lists, temperature adjustments, and token budget increases—suggesting that the observed behavior is not readily eliminated by the prompting interventions tested.

The observed traffic default is consistent with a distributional-prior explanation, although our experiments do not directly identify its origin in model training. The failure is one of domain adaptation: models appear unable to shift their causal reasoning from traffic accidents to healthcare when the scenario distribution changes.

1.1 Contributions

This paper makes four contributions:

1. **CausalAudit**: A constraint-based auditing framework with auditable checkers that runs post hoc on any API-accessible model.

2. **Cross-Model Evaluation**: A systematic evaluation across six model families and 195 scenarios spanning six domains that establishes a controlled cross-domain diagnostic baseline for causal explanation quality and documents the mechanistic admissibility gap.

3. **Discovery**: A description and taxonomy of the traffic accident default, a systematic domain-inappropriate mechanism substitution pattern that resists standard prompting fixes.

4. **Human Validation**: A checker validation study with 30 human-annotated scenarios confirms substantial inter-rater agreement ($\kappa = 0.928$ for validity, 0.814 for agreement) and precision of 87.0%.

1.2 Paper Structure

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the CausalAudit framework. Section 4 describes the evaluation methodology. Section 5 presents the empirical results. Section 6 analyzes the failure patterns. Section 7 validates the checker with human annotations. Section 8 discusses implications. Section 9 concludes with limitations and future work.

2 Related Work

2.1 LLM Evaluation and Faithfulness

Standard NLP benchmarks such as MMLU [1], GSM8K [2], and BIG-Bench [3] score whether a model's final answer is correct. A model can do this well while producing an explanation that does not actually hold up on inspection. Faithfulness research [4, 5] addresses a different question: whether a stated explanation corresponds to the model's own internal computation. Recent work [12] surveys the broader landscape of reasoning evaluation, including faithfulness as one dimension. However, these frameworks do not ask whether the explanation is causally valid under domain facts.

2.2 Causal Reasoning in LLMs

A separate line of work investigates whether LLMs possess genuine causal reasoning capabilities. Zečević et al. [6] describe LLMs as "causal parrots" that recite causal associations from training data without an underlying causal model. Benchmarks in this space include CLadder [7], which tests reasoning across Pearl's three rungs of causation [14] using synthetic causal graphs; CausalBench; InterveneBench; and EconCausal, where frontier models achieve only 9.5% accuracy on null-effect recognition. Prompting strategies such as chain-of-thought [8] and self-consistency decoding [9] improve scores somewhat.

These are capability diagnostics run on synthetic or semi-synthetic inputs. They tell us whether a model can execute causal reasoning under controlled conditions, not how its explanations fail when deployed on realistic, domain-varied scenarios. They provide no automated verification mechanism for causal admissibility in real-world scenarios. He et al. [13] propose symbolic verification for causal correctness in LLM reasoning, representing a related but distinct direction.

2.3 Neuro-Symbolic Verification

Closer to our approach, several recent systems pair an LLM with a symbolic checker. Explanation-Refiner [10] combines LLMs with theorem provers to generate and formalize explanatory sentences. Logic-Explainer [11] uses a backward-chaining solver to refine natural language explanations. PEIRCE generalizes this into an iterative generate-and-critique loop. All three verify logical, propositional validity. CausalAudit checks a different property—causal admissibility grounded in domain facts—and to our knowledge no prior system combines a constraint-based auditing approach with post-hoc black-box compatibility across multiple application domains in one framework.

2.4 Positioning of CausalAudit

CausalAudit occupies a distinct position. Table 1 summarizes the positioning. No prior framework combines: (1) a constraint-based auditing approach motivated by causal principles, (2) post-hoc black-box compatibility, (3) multi-domain coverage, and (4) an explicit violation taxonomy.

Table 1: Positioning of CausalAudit relative to prior work.

Dimension	**Prior Work**	**CausalAudit**
Target	Accuracy, faithfulness, logical validity	Causal admissibility (domain-grounded)
Approach	LLM evaluation, prompting, fine-tuning	Constraint-based auditing
Constraints	Logical (propositional)	Causal principles
Operation	Generation-time or evaluation	Post-hoc verification
Compatibility	White-box or limited	Model-agnostic (output-parsable)
Coverage	Single domain or synthetic	6 application domains

3 The CausalAudit Framework

3.1 Formal Setup

A scenario S consists of a natural language description, a domain label $d \in \mathcal{D}$, and a set of domain facts $\mathcal{F}_S$, of which $F_{\min} \subseteq \mathcal{F}_S$ is the minimal sufficient set of causal factors. Given an explanation E, we parse it into causal triples $T(E) = \{(c_k, r_k, e_k)\}_{k=1}^{n}$, where c_k is the asserted cause, e_k the asserted effect, and r_k the connective.

For each constraint i, a checker $V_i(E, S) \in \{0, 1\}$ returns 1 if satisfied. E is admissible iff:

$$A(E, S) = \prod_{i=1}^{5} V_i(E, S)$$

3.2 The Five Constraints

C1: Temporal Precedence. For every causal triple, the asserted cause must precede the asserted effect in time.

C2: Spatial Plausibility. The location of the cause must be plausibly connected to the location of the effect.

C3: Mechanistic Admissibility. The asserted causal chain must be physically possible and use domain-appropriate terminology. We operationalize this through two subcomponents:

C_{3a}: Mechanistic Validity. The generated mechanism is assessed against domain-specific physical rules and prohibited terminology.

C_{3b}: Reference-Mechanism Agreement. The generated mechanism is compared against a ground-truth mechanism using:

$$\text{Score}_{C3b} = 0.4 \cdot \text{Similarity} + 0.6 \cdot \text{StepMatch}$$

An explanation passes when $\text{Score}_{C3b} \geq \theta$, with $\theta = 0.3$ as default; Section 5.2 examines sensitivity to this choice. For the cross-model comparison, we use C_{3b} as the discriminative operational measure. Human validation in Section 7 establishes the relationship between C_{3a} and C_{3b}.

C4: Non-Spuriousness. The explanation must not present correlated variables as causal.

C5: Completeness. All $F_{\min}$ factors must be present.

3.3 Implementation

CausalAudit runs in four stages: claim extraction, parallel constraint checking, conjunction verdict, and audit logging. On standard hardware (Intel i7-1165G7 CPU, 16GB RAM, Python 3.10) the pipeline runs with median latency of 74ms. Figure 1 illustrates the pipeline.

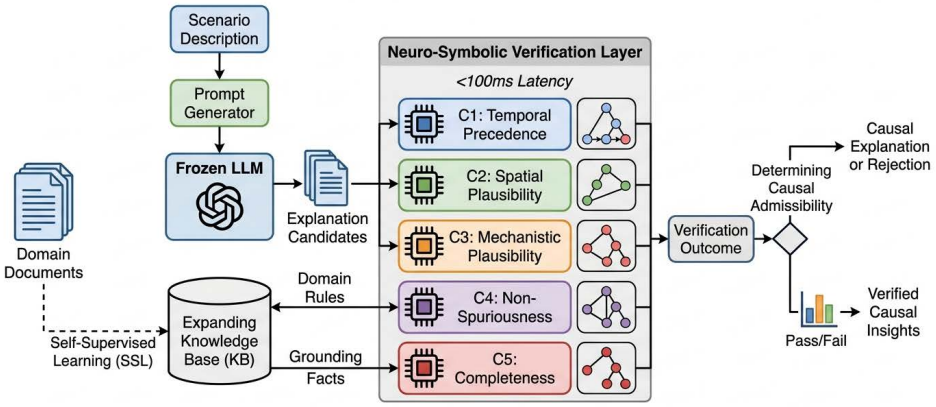

Figure 1: The CausalAudit pipeline architecture. A frozen LLM produces explanation candidates, which are checked against a knowledge base by five parallel constraint checkers; the conjunction of their verdicts determines admissibility. The pipeline executes in parallel with sub-100ms latency and is compatible with any model whose output can be parsed into the required explanation representation.

4 Evaluation Methodology

4.1 Scenarios

We constructed 195 scenarios across six application domains: Healthcare (42), Traffic Accident (30), Weather (33), Finance (30), Public Event (30), and Road Maintenance (30). Two annotators independently validated the scenarios; disagreements were resolved through adjudication. Cohen's $\kappa \geq 0.81$ for all domains.

4.2 Models

We evaluated six LLM families (Table 2): Llama 3.3 70B, Llama 3.1 8B, GPT-OSS 120B, GPT-OSS 20B, DeepSeek R1 70B, and Qwen3 32B.

Table 2: Models evaluated in this study.

Model	Size	Family	Provider	Type
Llama 3.3 70B	70B	Llama	Meta	General-purpose baseline
Llama 3.1 8B	8B	Llama	Meta	Lightweight comparison
GPT-OSS 120B	120B	GPT-OSS	OpenAI	Large open-source
GPT-OSS 20B	20B	GPT-OSS	OpenAI	Compact open-source
DeepSeek R1 70B	70B	DeepSeek	DeepSeek	Reasoning-oriented
Qwen3 32B	32B	Qwen	Alibaba	Strong open-source

4.3 Protocol

All evaluations used zero-shot prompting with domain context. Temperature = 0.3, max_tokens = 1500, top-p = 0.9, 3 retries.

4.4 Analysis

Results are reported as pass rates per constraint. Sensitivity analysis evaluates C_{3b} stability across $\theta \in \{0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45\}$.

5 Results

An early pilot on 41 transportation scenarios (Llama 3.3 70B) showed 52.5% of explanations violated at least one constraint. A controlled temperature experiment confirmed verification failure: models accepted physically impossible black ice formation at 2°C and 5°C despite correctly identifying it at -2°C.

5.1 Main Evaluation

Table 3 summarizes performance across all constraints.

Table 3: Per-constraint pass rates (%) across six LLM families.

Model	C1	C2	C3	C4	C5	Avg
Llama 3.3 70B	100.0	100.0	20.0	89.7	100.0	81.9
Llama 3.1 8B	100.0	100.0	31.8	94.9	100.0	85.3
GPT-OSS 120B	100.0	100.0	27.2	97.9	100.0	85.0
GPT-OSS 20B	100.0	100.0	35.1	97.9	100.0	86.6
DeepSeek R1 70B	100.0	100.0	48.7	100.0	100.0	89.7
Qwen3 32B	100.0	100.0	49.2	100.0	100.0	89.8

C1, C2, C5 are exactly 100%. C4 ranges 89.7–100%, practically non-separating. C3 is the discriminative constraint, with a 29.2 percentage point gap (2.46x). Figure 2 shows this gap. The cross-model comparison does not show a monotonic relationship between parameter count and C3 performance.

5.2 Threshold Sensitivity

Table 4 shows C3 pass rates across thresholds. The ranking is stable for $\theta \geq 0.25$.

Table 4: C_3 pass rates (%) across agreement thresholds.

Model	0.15	0.20	0.25	0.30	0.35	0.40	0.45
Llama 3.3 70B	85.1	70.8	32.8	20.0	6.7	3.6	3.6
Llama 3.1 8B	84.6	75.4	40.5	31.8	24.6	22.1	21.5
GPT-OSS 120B	86.2	73.8	33.3	27.2	25.6	25.6	25.6
GPT-OSS 20B	90.2	84.0	44.8	35.1	31.4	30.9	30.9
DeepSeek R1 70B	99.5	99.5	48.7	48.7	48.7	48.7	48.7
Qwen3 32B	100.0	100.0	49.2	49.2	49.2	49.2	49.2

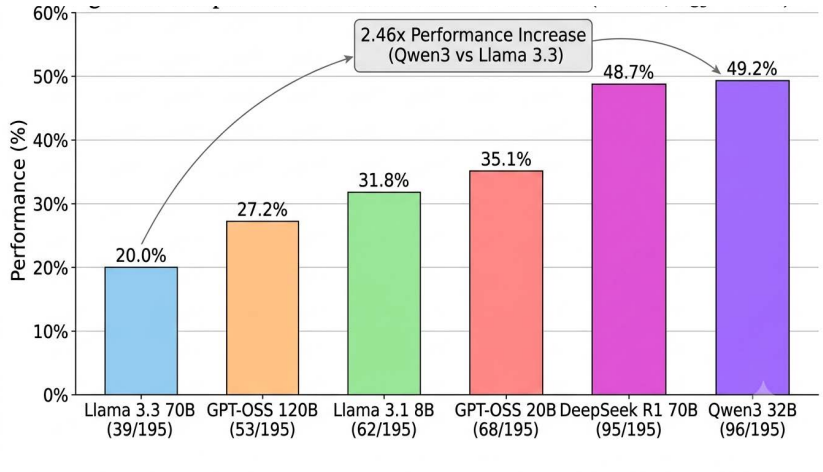

Figure 2: C_3 (mechanistic admissibility) pass rates by model at $\theta = 0.30$. The 29.2 percentage point gap (2.46x) from 20.0% to 49.2% does not track parameter count. Llama 3.3 70B (70B) is the worst performer, while Qwen3 32B (32B) is the best.

6 The Traffic Accident Default

The pattern behind most C3 failures is a specific, repeatable substitution in our evaluation suite. Figure 3 breaks failures into four types. The dominant type—traffic accident default—accounts for 62.5% of healthcare failures.

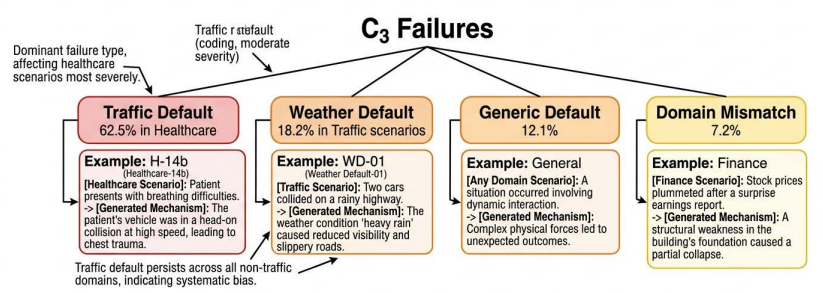

Figure 3: C_3 failure taxonomy. The traffic accident default—domain-inappropriate mechanism substitution—accounts for most healthcare failures (62.5%). The taxonomy identifies four distinct failure types: Traffic Default, Weather Default, Generic Default, and Domain Mismatch.

Table 5 shows the traffic default's prevalence by domain.

Figure 4 shows examples from Llama 3.3 70B. In the medical case, the model generates "Vehicle malfunction → Brake failure" for a premature hospital discharge scenario, scoring 15% while reporting 95% confidence. In the weather case, it produces a traffic sequence for a scenario never about vehicles, scoring 18% with 88% confidence.

6.1 Exploratory Prompting Interventions

In exploratory experiments, explicit domain guidance, few-shot examples, prohibited terminology, temperature adjustment, and token budget increases did not produce consistent reduction. Only model switching (Llama → Qwen) produced the 29.2 percentage point improvement.

6.2 Relationship Between Model Scale and C3 Performance

The cross-model comparison does not show a monotonic relationship between parameter count and C3 performance (Table 6).

Table 5: Traffic default prevalence by domain.

Domain	% Transportation	% Domain-Appropriate	% Other
Healthcare	62.5	25.0	12.5
Finance	45.8	33.3	20.9
Public Event	38.9	41.7	19.4
Weather	22.2	55.6	22.2
Traffic Accident	8.3	83.3	8.4
Road Maintenance	16.7	66.7	16.6

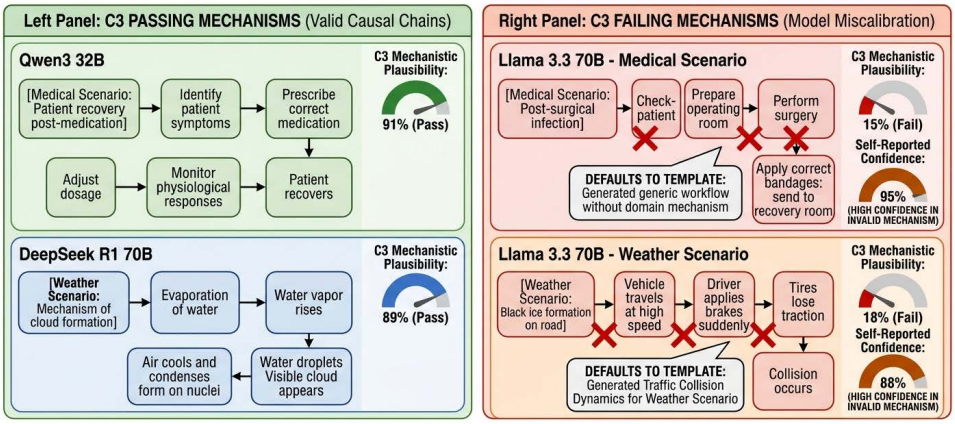

Figure 4: Passing and failing mechanisms side by side. Failing mechanisms default to a transportation template regardless of domain, and are reported with high self-stated confidence. In the medical case, the model generates "Vehicle malfunction → Brake failure" for a premature hospital discharge scenario, scoring 15% while reporting 95% confidence. In the weather case, it produces a full traffic sequence for a scenario never about vehicles, scoring 18% with 88% confidence.

7 Human Validation of the Checker

We validated C3 checker outputs with a human annotation study on 30 random scenarios. Two independent annotators labeled each scenario on validity (C3a) and agreement (C3b).

7.1 Inter-Annotator Agreement

Inter-annotator agreement was substantial: 96.7% for validity ($\kappa = 0.928$) and 90.0% for agreement ($\kappa = 0.814$).

7.2 Checker Performance

For binary evaluation, human validity labels were converted as: Valid = positive, Invalid/Partial = negative. C3b output was positive when $\text{Score}_{C3b} \geq \theta$. C3 achieved precision of 87.0%, recall of 74.1%, and F1 of 79.8% (Table 7). The confusion matrix (Table 8) shows TP=20, FP=3, FN=7, TN=0.

7.3 C3a vs. C3b

Table 9 shows the relationship between validity and agreement. 20/21 (95%) valid mechanisms achieved agreement; 0/6 invalid mechanisms achieved agreement.

Table 6: C_3 pass rate by model scale at $\theta = 0.30$.

Scale	Model	C_3 Pass Rate
8B	Llama 3.1 8B	31.8%
20B	GPT-OSS 20B	35.1%
32B	Qwen3 32B	**49.2%**
70B	Llama 3.3 70B	20.0%
70B	DeepSeek R1 70B	**48.7%**
120B	GPT-OSS 120B	27.2%

Table 7: C3 checker performance against human labels.

Metric	Value
Precision	87.0%
Recall	74.1%
F1 Score	79.8%
Inter-Annotator Agreement (Validity)	96.7% ($\kappa = 0.928$)
Inter-Annotator Agreement (Agreement)	90.0% ($\kappa = 0.814$)
Sample Size	30

8 Discussion

8.1 Non-Discriminative Constraints

C1, C2, C5 are exactly 100%. C4 ranges 89.7–100%. These constraints are non-discriminative on our test distribution. This does not imply LLMs are robust; it means our scenarios are not adversarial enough.

8.2 Why Prompting Alone May Be Insufficient

The fact that C3 is the only discriminative constraint, combined with the traffic default's resistance to prompting, suggests the need for a verification layer. A prompt can tell a model what topic to write about; it cannot easily override which mechanism template the model reaches for by default.

8.3 The Neuro-Symbolic Path Forward

A promising direction separates knowledge acquisition from generation: frozen LLM, SSL-expanding knowledge base, and neuro-symbolic verification.

9 Conclusion

We introduced CausalAudit, a constraint-based auditing framework for LLM-generated causal explanations. Four constraints were non-discriminative or weakly discriminative. The fifth, mechanistic admissibility, produced a clear 29.2 percentage point gap between models. Failure analysis revealed a systematic traffic accident default in our evaluation suite. Human validation confirmed checker precision of 87.0%.

Trustworthy AI requires verification systems that can support validation before outputs reach decision-makers. CausalAudit provides a step in this direction.

Table 8: Confusion matrix: C3 checker vs. human validity.

	Human Label		
	Valid	**Invalid**	**Total**
Checker Pass	20	3	23
Checker Fail	7	0	7
Total	27	3	30

Table 9: C3a (validity) vs. C3b (agreement) breakdown.

	C3b: Yes	**C3b: Partial**	**C3b: No**	**Total**
C3a: Yes	20	1	0	21
C3a: Partial	0	0	3	3
C3a: No	0	1	5	6
Total	20	2	8	30

Limitations

Our scenario set is a single fixed collection. C3 depends on chosen thresholds (sensitivity analysis provided). We evaluated open-weight models only. Our knowledge bases are not independently audited. The human validation sample is modest (n=30).

References

[1] Hendrycks, D., et al. (2021). Measuring Massive Multitask Language Understanding. *ICLR*.

[2] Cobbe, K., et al. (2021). Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*.

[3] Srivastava, A., et al. (2022). Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv:2206.04615*.

[4] Jacovi, A., & Goldberg, Y. (2020). Towards Faithfully Interpretable NLP Systems. *ACL*.

[5] Turpin, M., et al. (2024). Language Models Don't Always Say What They Think. *NeurIPS*.

[6] Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. *TMLR*. arXiv:2308.13067.

[7] Zečević, M., Willig, M., Dhami, D. S., & Kersting, K. (2023). CLadder: Assessing Causal Reasoning in Language Models. *NeurIPS*.

[8] Wei, J., et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS*.

[9] Wang, X., et al. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ICLR*.

[10] Quan, X., Valentino, M., Dennis, L. A., & Freitas, A. (2024). Verification and Refinement of Natural Language Explanations through LLM-Symbolic Theorem Proving. *EMNLP*.

[11] Quan, X., Valentino, M., Dennis, L. A., & Freitas, A. (2024). Enhancing Ethical Explanations of Large Language Models through Iterative Symbolic Refinement. *EACL*.

[12] Mondorf, P., & Plank, B. (2024). Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models. *arXiv:2404.01869*.

[13] He, P., Huang, Y., Sachan, M., & Jin, Z. (2025). Uncovering Hidden Correctness in LLM Causal Reasoning via Symbolic Verification. *arXiv:2601.21210*.

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