CausalLens: Probing LLMs’ Clinical Reasoning through QA-Driven Causal Inference

Chetan Kumar Verma
chetan.kumar.verma@utexas.edu
Department of Computer Science
University of Texas at Austin

Abstract

Large Language Models (LLMs) hold significant potential for advancing clinical decision support, yet their capacity for causal reasoning in healthcare remains underexplored. We introduce CausalLens, a structured framework that generates diverse, clinically grounded causal question–answer (QA) pairs from MIMIC-IV discharge summaries. These pairs encompass medications, laboratory findings, and diagnoses, enabling a targeted evaluation of LLMs’ ability to discern cause-and-effect relationships in complex patient scenarios. By systematically probing LLM responses through controlled interventions, CausalLens exposes limitations in LLM causal reasoning and provides a reproducible pipeline for evaluating performance, enabling the development of safer, more reliable AI in clinical decision support.

Keywords

AI in healthcare, Causal reasoning, Clinical QA, LLMs

1 Introduction

Large Language Models (LLMs), such as GPT-4, Claude, and Qwen, have demonstrated remarkable proficiency in generating coherent, contextually relevant, and domain-specific text. These models leverage statistical patterns learned from extensive corpora to recall factual knowledge, follow instructions, and perform multistep reasoning to a certain extent. However, their reasoning often relies on pattern recognition rather than a grounded understanding of causal mechanisms, raising questions about their ability to handle intervention or counterfactual reasoning effectively.

In healthcare, causal reasoning is critical for informed decision-making, as it underpins the understanding of intervention effects (e.g., prescribing medications, adjusting treatment plans) and the anticipation of outcomes in hypothetical scenarios. Without robust causal reasoning, AI systems risk generating plausible but potentially unsafe recommendations in clinical settings. Evaluating LLMs’ causal reasoning capabilities is thus essential for ensuring their reliability in clinical decision support. We present **CausalLens**, a framework for generating, intervening, and evaluating causal question–answer (QA) pairs derived from clinical narratives in the MIMIC-IV dataset. This framework integrates three core components:

1. **Data Extraction and Structuring**: Extracting and organizing key clinical variables, such as diagnoses, laboratory values, and medications, from unstructured discharge summaries.
2. **Causal QA Generation**: Producing factual and counterfactual questions that target single-entity and multi-entity causal relationships.
3. **Interventional Evaluation**: Modifying discharge summaries to assess whether LLMs maintain causal consistency under altered conditions.

By systematically probing LLM responses before and after controlled interventions, CausalLens reveals specific deficiencies in LLM causal reasoning, such as cause ignorance and over-sensitivity to irrelevant context, as demonstrated in our experiments, thus informing the development of more robust clinical AI systems.

2 Related Work

Causal reasoning in clinical settings has been a focal point of research, with traditional approaches relying on structural causal models (SCMs) and Bayesian networks to infer cause-effect relationships from electronic health records (EHRs) and observational data [14, 16]. While these methods offer strong theoretical foundations, their scalability and adaptability to unstructured data are limited.

The advent of deep learning has led to the exploration of LLMs, such as GPT-4, PaLM, and domain-specific models like BioBERT and ClinicalBERT [11, 1], for question-answering tasks over clinical narratives. However, these models primarily focus on factual recall rather than causal inference, limiting their reliability for addressing “why” and “what-if” questions [5]. Recent efforts have sought to integrate causal reasoning into LLMs by augmenting prompts with structured clinical variables [18] or incorporating causal graphs into reasoning processes [6].

In the medical QA domain, prior work has explored knowledge-based QA generation from EHRs [7] and LLM-driven approaches for multi-hop reasoning over structured and unstructured data [19, 20]. Yet, there remains a significant gap in systematically evaluating LLMs’ causal reasoning limitations in controlled, domain-specific settings like MIMIC-IV discharge summaries. This gap motivates our framework, which generates causal QA pairs from clinical text to systematically evaluate LLMs’ reasoning, revealing critical limitations such as cause ignorance and spurious context sensitivity, as evidenced in our results.

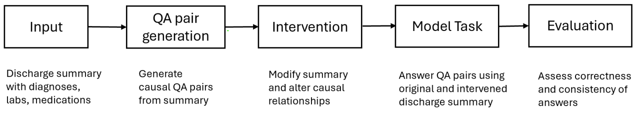
Input
QA pair generation
Intervention
Model Task
Evaluation
Discharge summary with diagnoses, labs, medications
Generate causal QA pairs from summary
Modify summary and alter causal relationships
Answer QA pairs using original and intervened discharge summary
Assess correctness and consistency of answers

Figure 1: Primary steps involved in the CausalLens framework

3 Problem Formulation

Causal reasoning in large language models (LLMs) poses a significant challenge in healthcare, where decisions must be grounded in cause-and-effect relationships rather than mere correlations. Existing LLM evaluations often focus on general-purpose reasoning or factual recall, leaving their ability to perform clinically relevant causal inference underexplored.

Discharge summaries in electronic health records (EHRs) contain a wealth of interconnected information, including diagnoses, medications, laboratory results, and clinical events, where causal relationships are often implicit. For instance, an elevated creatinine level may result from acute kidney injury, potentially triggered by specific medications or comorbidities. These causal chains are rarely annotated explicitly, posing challenges for both human and machine reasoning.

CausalLens addresses this gap by providing a structured framework for evaluating LLMs' causal reasoning abilities in clinical settings through controlled QA generation and intervention-based testing. The framework operationalizes this evaluation across the following stages (as illustrated in Figure 1):

(1) **Input**: A discharge summary containing textual descriptions of diagnoses, laboratory values, medications, and clinical events.
(2) **QA Pair Generation**: Generation of causal QA pairs based on clinical evidence within the summary, targeting cause, effect, and hypothetical reasoning.
(3) **Intervention**: Controlled modification of discharge summaries to alter causal relationships (e.g., changing laboratory values, removing diagnoses, or introducing medications) while preserving unrelated details.
(4) **Model Task**: Prompting the LLM with modified summaries and original QA pairs to generate updated answers reflecting the causal implications of interventions.
(5) **Evaluation**: Assessing the LLM's causal reasoning based on the correctness and consistency of responses before and after interventions.

CausalLens ensures a reproducible, modular approach adaptable to various datasets, medical specialties, and model architectures by providing:

- **Standardized Inputs**: Defined rules for selecting and preprocessing discharge summaries.
- **Controlled Interventions**: Reproducible modifications that isolate specific causal reasoning challenges.
- **Task-Agnostic Evaluation**: Metrics applicable across LLMs without model-specific tuning.
- **Clinical Relevance**: QA pairs grounded in real patient cases to ensure domain validity.

Key challenges include handling unstructured clinical text, detecting sparse causal signals, reasoning with domain-specific medical knowledge, and ensuring consistent responses to interventions. CausalLens bridges the gap between general causal reasoning benchmarks and clinical decision-making requirements, while pinpointing LLM failure modes such as inconsistent causal updates and over-sensitivity to irrelevant changes.

4 Methods - The CausalLens Framework

The CausalLens Framework organizes the methodology into four interconnected modules: (1) Causal QA Generation, (2) Intervention Design, (3) LLM Reasoning and Response Generation, and (4) Evaluation. Each module is designed to be independently upgradable and reusable, facilitating adaptation to diverse datasets, domains, or LLM architectures.

4.1 Causal QA Generation

This module creates clinically relevant causal question–answer (QA) pairs from MIMIC-IV discharge summaries.

- **Entity Extraction**: Identifies clinical entities (diagnoses, medications, laboratory results) using heuristic and rule-based methods.
- **Question Templates**: Instantiates domain-specific causal templates (e.g., "Could the diagnosis of [X] be related to [Y]?") with extracted entities.
- **Ground Truth Alignment**: Derives answers from the discharge summary context to serve as baseline reference outputs.

4.2 Intervention Design

This module evaluates counterfactual reasoning by introducing controlled modifications to discharge summaries.

- **Single-Variable Interventions**: Modifies one causal factor (e.g., removing a medication, altering a laboratory value) while keeping other context constant.
- **Multi-Variable Interventions**: Alters combinations of causal factors to simulate realistic clinical scenarios.
- **Consistency Check**: Ensures interventions maintain logical and clinical plausibility.

4.3 LLM Reasoning and Response Generation

This module prompts the LLM (Qwen) with original and intervened discharge summaries to generate answers to causal QA pairs.

- **Original Context Responses**: Establishes baseline model performance.
- **Post-Intervention Responses**: Measures reasoning shifts in response to altered contexts.
- **Prompt Format**: Uses structured prompts containing context, question, and answer request for reproducibility.

4.4 Evaluation

This module measures the LLM's causal reasoning performance, specifically targeting weaknesses such as cause ignorance, downstream persistence, and spurious context sensitivity, as observed in our experiments.

- **Causal Consistency**: Assesses agreement between original and intervened responses when no causal change is expected.
- **Causal Sensitivity**: Evaluates appropriate response changes when causal factors are modified.
- **Overall Reasoning Accuracy**: Measures the proportion of QA pairs answered correctly according to expected causal logic.

5 Experiments

5.1 Dataset and Preprocessing

Experiments[1] utilized the MIMIC-IV database, focusing on discharge summaries, laboratory results, and diagnoses. A subset of 50 randomly selected patients, aged 18 to 89 years, was extracted, each with free-text discharge summaries and associated structured data.

Preprocessing involved:

- Removing personally identifiable information (already de-identified in MIMIC-IV).
- Normalizing measurement units and medical terminology.
- Filtering summaries to retain cases with at least one identifiable causal link (e.g., a laboratory abnormality linked to a diagnosis or medication).

5.2 Causal QA Pair Generation

Using the CausalLens Framework, ten causal QA pairs were generated from discharge summaries with domain-specific templates, covering:

- **Factual Causal Reasoning**: Questions like "What is the likely cause of [X]?"
- **Counterfactual Reasoning**: Questions like "If [Y] had not occurred, would [X] still be likely?"

QA pairs were evaluated on both original and intervened discharge summaries.

5.3 Intervention Design

Interventions tested the model's ability to handle single-variable causal changes:

- For explicit causal patterns (e.g., "elevated creatinine → acute kidney injury"), targeted single-variable interventions were applied (e.g., modifying creatinine values).
- For summaries without explicit causal patterns, random but clinically plausible interventions were introduced.

Interventions preserved unrelated details while altering the causal landscape relevant to the QA pair.

5.4 Model Setup

The Qwen3-8B large language model was used for all experiments.

- **Prompting Strategy**: Queries included the discharge summary, causal question, and an instruction to provide reasoning.
- **Evaluation Scope**: Focused solely on Qwen3-8B without additional baselines.
- **Hardware**: Experiments were conducted on an A100-40GB GPU via Google Colab.

5.5 Evaluation Protocol

Evaluation followed a two-part process:

1. **Human Evaluation**: Manual evaluation (combined with LLMs to check domain knowledge) assessed the correctness of causal reasoning for original and intervened summaries.
2. **Cosine Similarity Analysis**: Computed similarity between embeddings of Qwen's responses for original and intervened summaries to detect reasoning changes. High similarity (cosine similarity score > 0.85) was expected for unrelated interventions, and low similarity for targeted causal changes.

5.6 Limitations of Experiment Scope

This study focused on Qwen3-8B without comparing other LLMs. The limited number of QA pairs (n=10) facilitated detailed manual evaluation. Future work will expand the dataset, intervention diversity, and baseline comparisons.

6 Results and Error Analysis

The CausalLens intervention–evaluation framework allowed us to uncover significant limitations in LLM causal reasoning, including cause ignorance, downstream persistence, and over-sensitivity to irrelevant context, as demonstrated by model responses to targeted and non-targeted changes in discharge summaries. Across three curated examples (more details are available Table-1 and Table-2), distinct causal reasoning failure modes were identified.

6.1 Example 1 – Cause Ignorance (Single-entity causal link)

In this scenario, the targeted intervention removed a relevant causal factor (osteoporosis) from the patient history.

- **Expected Behavior**: The factual answer should exclude osteoporosis from its explanation, and the counterfactual should reflect its absence.
- **Observed Behavior**: The factual answer incorrectly retained osteoporosis as a cause, while the counterfactual appropriately adapted.
- **Interpretation**: The model exhibited cause ignorance in factual reasoning, failing to recognize the removal of a key causal entity.

6.2 Example 2 – Downstream Persistence (Multi-entity causal chain)

Here, the intervention altered an upstream causal event (RFA unsuccessful → RFA successful).

- **Expected Behavior**: Both factual and counterfactual answers should reflect the absence of pneumothorax and chest tube requirements.
- **Observed Behavior**: The factual answer retained downstream effects of the unsuccessful RFA, while the counterfactual correctly updated.
- **Interpretation**: The model demonstrated downstream persistence, failing to fully propagate causal changes in factual reasoning.

6.3 Example 3 – Over-sensitivity / Spurious Causal Influence (Multi-entity, random intervention)

An irrelevant detail ("Follow-up care included physiotherapy sessions twice a week") was appended to the discharge summary.

---

[1] Code and steps for reproducing the result https://github.com/ckvermaAI/CausalLens

Table 1: Results and Error Analysis for Targeted Interventions

Example	Intervention Type	Changed Factual	Changed Counterfactual	Observed Model Behavior / Error Analysis
1 (Osteoporosis)	Single-entity causal intervention (osteoporosis → no history of osteoporosis)	No	Yes	Model maintained factual reasoning (fall as cause) but correctly adapted counterfactual to reduced fracture risk. Indicates causal sensitivity in counterfactual but stability in factual.
2 (RFA / Pneumothorax)	Multi-entity causal intervention (unsuccessful RFA → successful RFA)	Yes	Yes	Model updated both factual and counterfactual, but factual reasoning was slightly oversimplified in intervention version. Shows causal update propagation but minor loss of detail.
3 (Oxycodone / Docusate)	Irrelevant/random intervention (added follow-up physiotherapy note)	Yes	Yes	Model unnecessarily altered both factual and counterfactual answers despite intervention being unrelated. Reveals vulnerability to spurious context and low causal stability.

- **Expected Behavior**: Factual and counterfactual answers should remain unchanged, as the intervention is causally unrelated.
- **Observed Behavior**: Both answers changed, including a correction of previously incorrect counterfactual reasoning.
- **Interpretation**: The model displayed over-sensitivity, altering reasoning due to non-causal context, indicating susceptibility to spurious influences.

7 Conclusion

We introduced CausalLens, a novel framework that systematically exposes the limited capability of large language models in performing causal reasoning tasks within clinical contexts. By applying targeted interventions to discharge summaries, CausalLens reveals critical failure modes, such as cause ignorance, downstream persistence, and over-sensitivity to irrelevant context, as evidenced in our results. These findings underscore the framework's value as a diagnostic tool for enhancing the reliability of medical QA systems, particularly in high-stakes clinical environments where consistent causal reasoning is paramount.

8 Future Directions

To extend this work, we propose:

1. **Scaling to Broader Clinical Domains**: Apply CausalLens to specialties like cardiology and neurology [17, 12] and diverse clinical note types [4].
2. **Integration with Model Training Loops**: Incorporate framework outputs into fine-tuning [15] to optimize for causal robustness [10].
3. **Automated Intervention Generation**: Develop automated intervention pipelines using medical ontologies [3] and causal graphs [8].
4. **Cross-Model Benchmarking**: Compare reasoning patterns across LLM architectures [13, 2] to identify architecture-specific weaknesses.
5. **Clinical Validation Layer**: Combine automated scoring with clinician review [9] to ensure detected errors are clinically relevant.
6. **Targeted Mitigation Strategies**: Develop techniques to address specific LLM failure modes, such as cause ignorance and over-sensitivity, identified by CausalLens, to enhance causal reasoning robustness.

References

[1] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. DOI: 10.18653/v1/W19-1909.

[2] Rohan Anil et al. 2023. Palm 2 technical report. https://arxiv.org/abs/2305.10403. (2023).

[3] Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32, suppl_1, D267–D270.

[4] William Campbell et al. 2013. Comparison of structured and narrative electronic health record data for identifying patients with chronic diseases. *BMJ Open*, 3, 8, e003354.

[5] Qian Chen and et al. 2023. Limitations of large language models in causal reasoning. *arXiv preprint arXiv:2301.00000.*

[6] Omar Choudhury and et al. 2024. Incorporating causal graphs into llm reasoning for healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[7] Dina Demner-Fushman and et al. 2009. A knowledge-based approach to medical question answering. In *AMIA Annual Symposium Proceedings*, 171–175.

[8] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10, 524.

[9] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17, 1, 195.

[10] Kun Kuang et al. 2020. Causalbert: injecting causal knowledge into pre-trained language models with minimal supervision. *arXiv preprint arXiv:2009.12700.*

[11] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36, 4, 1234–1240. DOI: 10.1093/bioinformatics/btz682.

[12] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. 2018. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19, 6, 1236–1246.

[13] OpenAI. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774. (2023).

[14] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference.* (2nd ed.). Cambridge University Press. ISBN: 9780521895606.

[15] Colin Raffel et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research.*

[16] Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search.* (2nd ed.). MIT Press. ISBN: 9780262194402.

[17] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25, 1, 44–56.

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting

Table 2: Example QA pairs and Qwen responses (original vs. after intervention).

Example	Question	Qwen Response (Original)	Qwen Response (After Intervention)
Example-1 (Factual)	What caused the bilateral tibial plateau fractures?	Yes, the bilateral tibial plateau fractures were caused by a fall, as indicated in the discharge summary under the "Discharge Diagnosis" section which lists "Mechanical Fall" as one of the diagnoses. The patient reported an unwitnessed fall with knee trauma, leading to the fractures.	Yes, the bilateral tibial plateau fractures were caused by a fall.
Example-1 (Counterfactual)	What would have happened if the patient did not have osteoporosis?	Yes, the patient would have had a less severe fracture. Osteoporosis contributes to increased bone fragility, making fractures more likely and potentially more severe. Without it, the fractures might have been less severe or even non-existent.	Yes, because the patient's bilateral tibial plateau fractures would have been less likely to occur without osteoporosis. The discharge summary indicates that the patient has a history of osteoporosis, which increases the risk of fractures from falls. Without osteoporosis, the bones would be stronger, making such fractures less probable.
Example-2 (Factual)	What caused the patient's need for a chest tube and how did it relate to the unsuccessful radiofrequency ablation (RFA) procedure?	Yes, the unsuccessful radiofrequency ablation (RFA) procedure resulted in a pneumothorax, which required the placement of a chest tube to treat the air collection between the lung tissue and chest wall.	Yes, the patient required a chest tube due to a pneumothorax that occurred during the unsuccessful RFA procedure, as the procedure inadvertently traversed the right lung, leading to air leakage into the pleural space.
Example-2 (Counterfactual)	What would have happened if the RFA procedure was successful and did not result in a pneumothorax?	Yes, the patient would have had a successful RFA without complications, and would have been discharged with the same medications and follow-up instructions as listed.	Yes, the patient would have had a successful treatment for his hepatocellular carcinoma (HCC) without complications, avoiding the need for a chest tube and pneumothorax.
Example-3 (Factual)	Why were both oxycodone and docusate sodium prescribed to the patient?	Yes, because oxycodone is a pain medication to manage post-procedural pain, and docusate sodium is a stool softener to prevent constipation, which can exacerbate hemorrhoid symptoms.	Yes, because the patient had a hemorrhoidectomy and required pain management with oxycodone, while docusate sodium was prescribed to prevent constipation, a common side effect of opioid use.
Example-3 (Counterfactual)	What would happen if the patient did not take docusate sodium while on oxycodone?	No, the patient would not experience increased pain or other complications from not taking docusate sodium while on oxycodone.	No, the patient would not experience severe complications, but may have constipation due to the opioid effect. The docusate sodium helps prevent constipation, which is a common side effect of opioids like oxycodone. Without it, the patient might need more aggressive management for constipation.

elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*. Vol. 35, 24824–24837.

[19] Vikas Yadav and et al. 2023. Multi-hop reasoning in clinical question answering using llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

[20] Shuo Zhang and et al. 2024. Hybrid reasoning over structured and unstructured clinical data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.