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

Code & Paper: https://github.com/ckvermaAI/CausalLens

II: Introduction and Motivation

Introduction and Motivation

Why Causal Reasoning in Healthcare?

- Large Language Models (LLMs) excel in generating coherent, domain-specific text but often rely on pattern recognition rather than causal understanding.
- In healthcare, causal reasoning is critical for:
  - Understanding intervention effects (e.g., prescribing medications).
  - Anticipating outcomes in hypothetical scenarios.
- Without robust causal reasoning, LLMs risk producing unsafe recommendations in clinical settings.

Research Gap

- Existing LLM evaluations focus on factual recall or general reasoning, not clinical causal inference.
- Limited systematic approaches to assess LLMs' ability to handle cause-and-effect in complex medical narratives.

Our Contribution

- We present **CausalLens**, a framework to evaluate LLMs' causal reasoning using clinically grounded question-answer (QA) pairs.
- Reveal specific LLM limitations, such as cause ignorance and over-sensitivity, to inform safer clinical AI development.

II: CausalLens Framework

CausalLens Framework

Overview

- **CausalLens**: A structured, reproducible pipeline for generating and evaluating causal QA pairs from MIMIC-IV discharge summaries.
- Targets medications, lab results, and diagnoses to probe LLM causal reasoning.

Key Components

- **Data Extraction and Structuring**:
  - Extracts clinical entities (diagnoses, medications, lab values) from unstructured discharge summaries.
- **Causal QA Generation**:
  - Produces factual and counterfactual QA pairs using domain-specific templates.
  - Ensures clinical relevance by aligning with real patient data.
- **Interventional Evaluation**:
  - Modifies summaries (e.g., altering lab values or diagnoses) to test causal consistency.
- **Evaluation Metrics**:
  - Measures **causal consistency** (agreement when no change is expected).
  - Assesses **causal sensitivity** (correct updates for altered factors).
  - Evaluates **overall reasoning accuracy** against human annotations.

Objective

Expose LLM limitations in clinical causal reasoning through controlled interventions.

III: Experiments

Experiments (1)

- **Dataset**
  - Utilized **MIMIC-IV** database (discharge summaries, lab results, diagnoses).
  - Selected 50 patients (aged 18-89) with identifiable causal links.
  - Preprocessed data by normalizing units, terminology, and removing identifiable information.

- **Causal QA Pair Generation**
  - Generated 10 QA pairs (single and multi entity) using domain-specific templates:
    - **Factual**: “What caused [X]?”
    - **Counterfactual**: “If [Y] had not occurred, would [X] still be likely?”
  - Evaluated on original and intervened summaries.

- **Intervention Design**
  - **Single-variable interventions**: Altered one causal factor (e.g., modifying creatinine values).
  - **Multi-variable interventions**: Changed multiple factors for realistic scenarios.
  - Ensured clinical plausibility of interventions.

Experiments (2)

- **Model Setup**
  - Used **Qwen3-8B** LLM for all experiments.
  - Prompted with discharge summary, question, and reasoning instruction.
  - Ran experiments on A100-40GB GPU via Google Colab.

- **Evaluation Protocol**
  - **Human Evaluation**: Manual evaluation assessed the correctness of causal reasoning for original and intervened summaries.
  - **Cosine Similarity Analysis**: Measured response changes post-intervention.
  - Metrics: Causal consistency, sensitivity, and overall accuracy.

IV: Results & Conclusion

Example 1 - Cause Ignorance

- **Scenario**: Single-entity intervention removing osteoporosis from patient history.
- **Question Types**:
  - **Factual**: “What caused the bilateral tibial plateau fractures?”
  - **Counterfactual**: “What would have happened if the patient did not have osteoporosis?”
- **Expected Behavior**:
  - Factual: Exclude osteoporosis from explanation.
  - Counterfactual: Reflect reduced fracture risk without osteoporosis.
- **Observed Behavior**:
  - Factual: Incorrectly retained osteoporosis as cause.
  - Counterfactual: Correctly adapted to absence of osteoporosis.
- **Interpretation**: Model exhibited **cause ignorance**, failing to register removal of a key causal factor in factual reasoning.

Example 2 - Downstream Persistence

- **Scenario**: Multi-entity intervention changing unsuccessful RFA to successful RFA.

- **Question Types**:
  - **Factual**: “What caused the patient’s need for a chest tube due to RFA?”
  - **Counterfactual**: “What if the RFA was successful and no pneumothorax occurred?”

- **Expected Behavior**:
  - Both answers reflect no pneumothorax or chest tube need.

- **Observed Behavior**:
  - Factual: Retained downstream effects of unsuccessful RFA.
  - Counterfactual: Correctly updated to reflect successful RFA.

- **Interpretation**: Model showed **downstream persistence**, failing to propagate causal changes in factual reasoning.

Example 3: Over-sensitivity

- **Scenario**: Irrelevant/random intervention adding “Follow-up care included physiotherapy.”

- **Question Types**:
  - **Factual**: “Why were oxycodone and docusate sodium prescribed?”
  - **Counterfactual**: “What if the patient did not take docusate sodium?”

- **Expected Behavior**:
  - No change in answers, as intervention is unrelated.

- **Observed Behavior**:
  - Both factual and counterfactual answers changed unnecessarily.
  - Corrected previously incorrect counterfactual reasoning.

- **Interpretation**: Model displayed **over-sensitivity**, altering reasoning due to non-causal context

Key Findings and Conclusion

Key Findings

- **CausalLens revealed significant LLM limitations**:
  - **Cause Ignorance**: Failed to exclude removed causal factors (e.g., osteoporosis) in factual reasoning.
  - **Downstream Persistence**: Retained outdated effects after upstream changes (e.g., RFA success).
  - **Over-sensitivity**: Altered responses due to irrelevant context (e.g., physiotherapy notes).
- Demonstrated through three curated examples, highlighting distinct failure modes.

Conclusion

- **CausalLens** is a framework that systematically exposes LLMs' limited capability in clinical causal reasoning.
- By applying targeted interventions, it uncovers critical failure modes, enhancing the understanding of LLM weaknesses.
- Serves as a diagnostic tool for improving the reliability of medical QA systems in high-stakes clinical settings.

Future Directions

- SCALE TO DIVERSE CLINICAL DOMAINS (E.G., CARDIOLOGY, NEUROLOGY).
- INTEGRATE WITH MODEL FINE-TUNING FOR CAUSAL ROBUSTNESS.
- AUTOMATE INTERVENTION GENERATION USING MEDICAL ONTOLOGIES.
- BENCHMARK ACROSS MULTIPLE LLMS TO IDENTIFY ARCHITECTURE-SPECIFIC WEAKNESSES.
- DEVELOP TARGETED STRATEGIES TO MITIGATE IDENTIFIED FAILURE MODES.

Thank You!

Questions?

Mail to: Chetan.kumar.verma@utexas.edu