Construction of a causal knowledge graph for research on diabetes comorbidities

Xueli Wu

Supervisor: Xiaoli Tang

Institute of Medical Information, Chinese Academy of Medical Sciences &
Peking Union Medical College

Contact email: tang.xiaoli@imicams.ac.cn

Venice, Italy, May 25, 2026

Xueli Wu wu.xueli@student.pumc.edu.cn

Xueli Wu is currently a first-year master's student majoring in Information Science at the Institute of Medical Information, Chinese Academy of Medical Sciences & Peking Union Medical College. Her research group focuses mainly on knowledge organization and data mining.

This work is supported by the Innovation Fund for Medical Sciences of the Chinese Academy of Medical Sciences (Grant No: 2021-I2M-1-033).

Outline

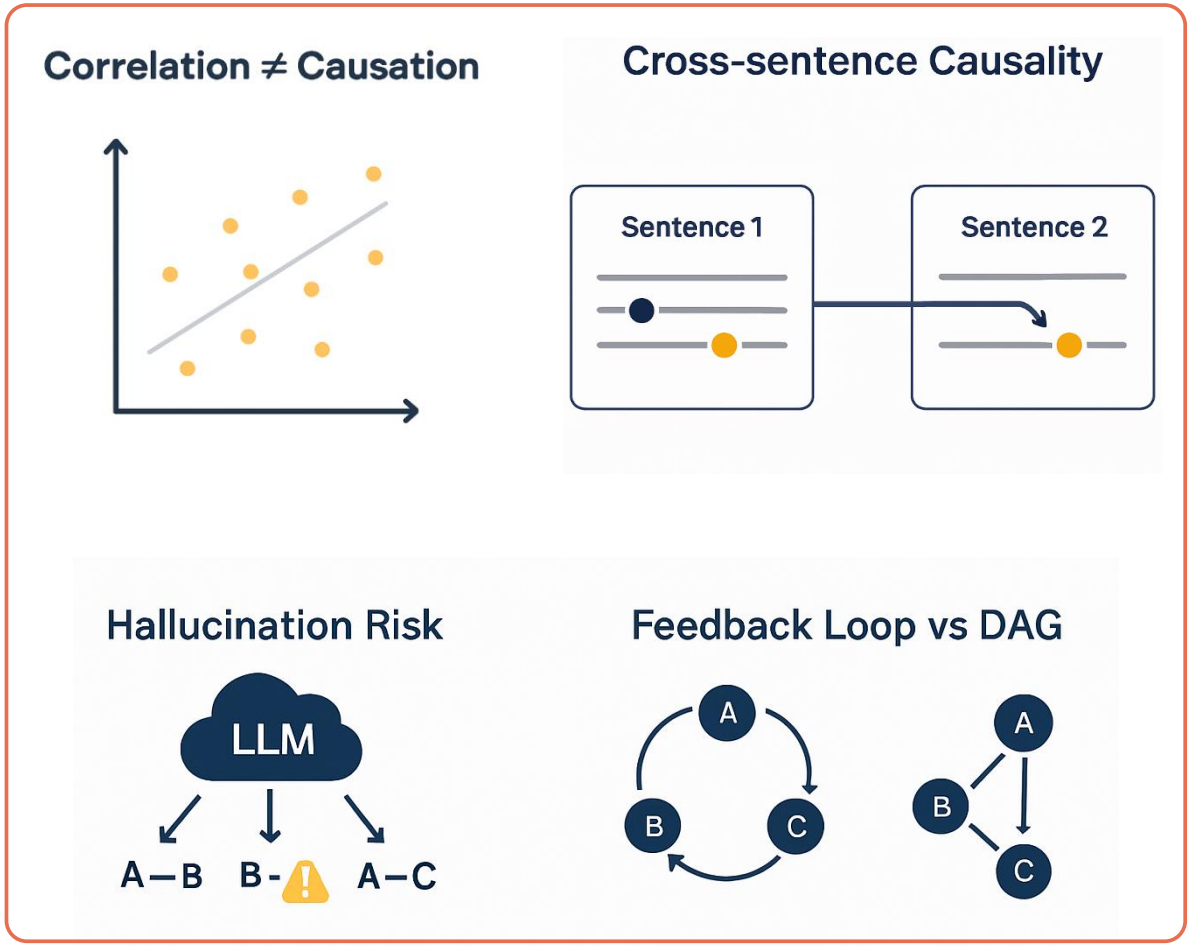
- 1. Introduction
- 2. Methodology
- 3. Case Study
- 4. Discussion

Background

Challenges in Diabetes Causal Discovery

- Correlation does not imply causation
- Cross-sentence causal inference
- LLM hallucination risks
- Biological feedback loops vs DAG constraints

Motivating a constrained, evidence-aware framework



Aims and Contributions

In our paper, we aimed at:

- Improving causal mining precision & recall
- Preserving biological logic consistency
- Building an interpretable causal knowledge graph

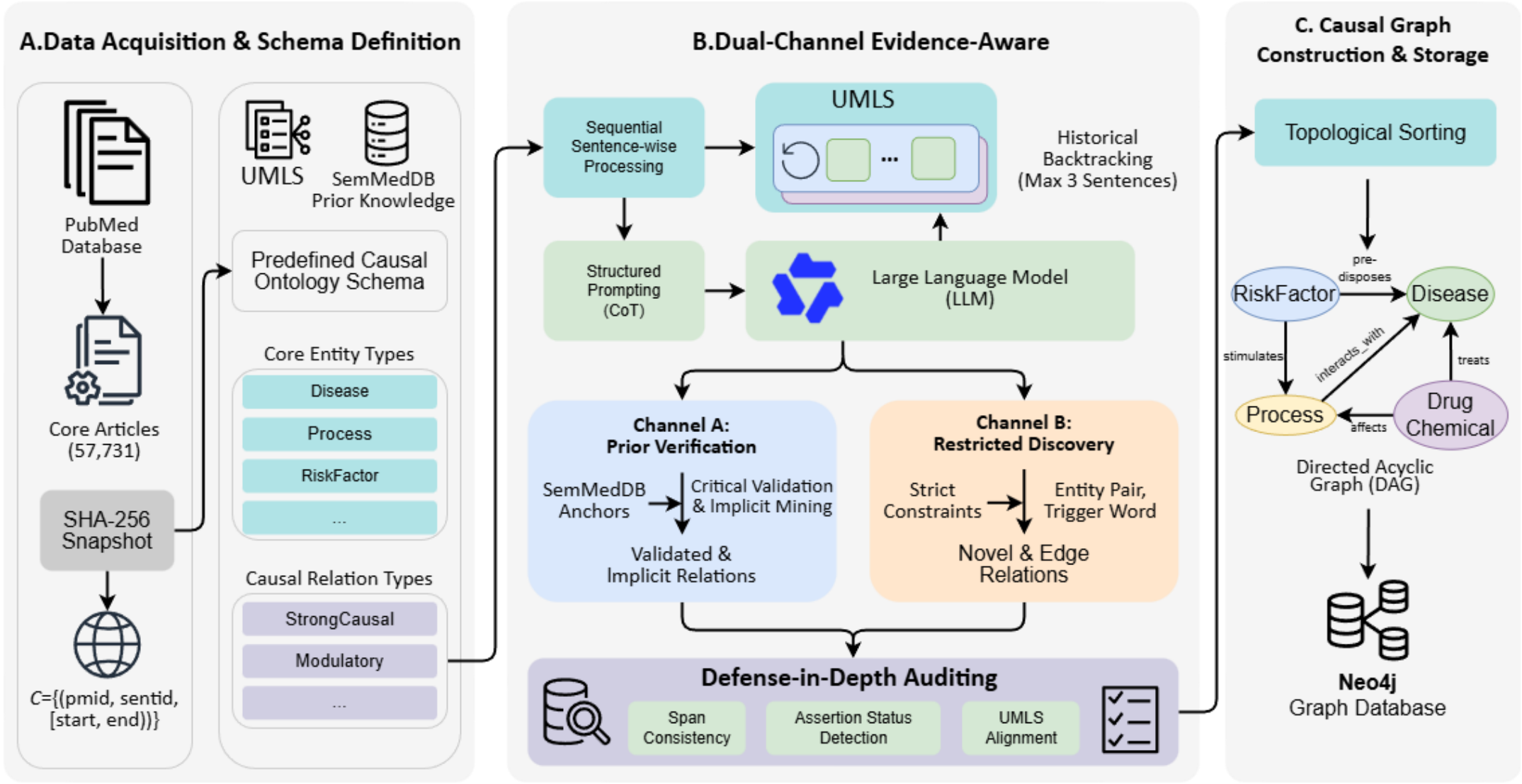
Constributions of our study:

- Proposed a hybrid computational paradigm integrating **physical anchoring, dual channel evidence awareness**, and **topological reconstruction**.
- Constructed a biologically faithful, logically consistent causal knowledge graph (**Diab-CKG)**.

Outline

- 1. Introduction
- 2. Methodology
- 3. Experimental Results
- 4. Discussion

Framework Overview



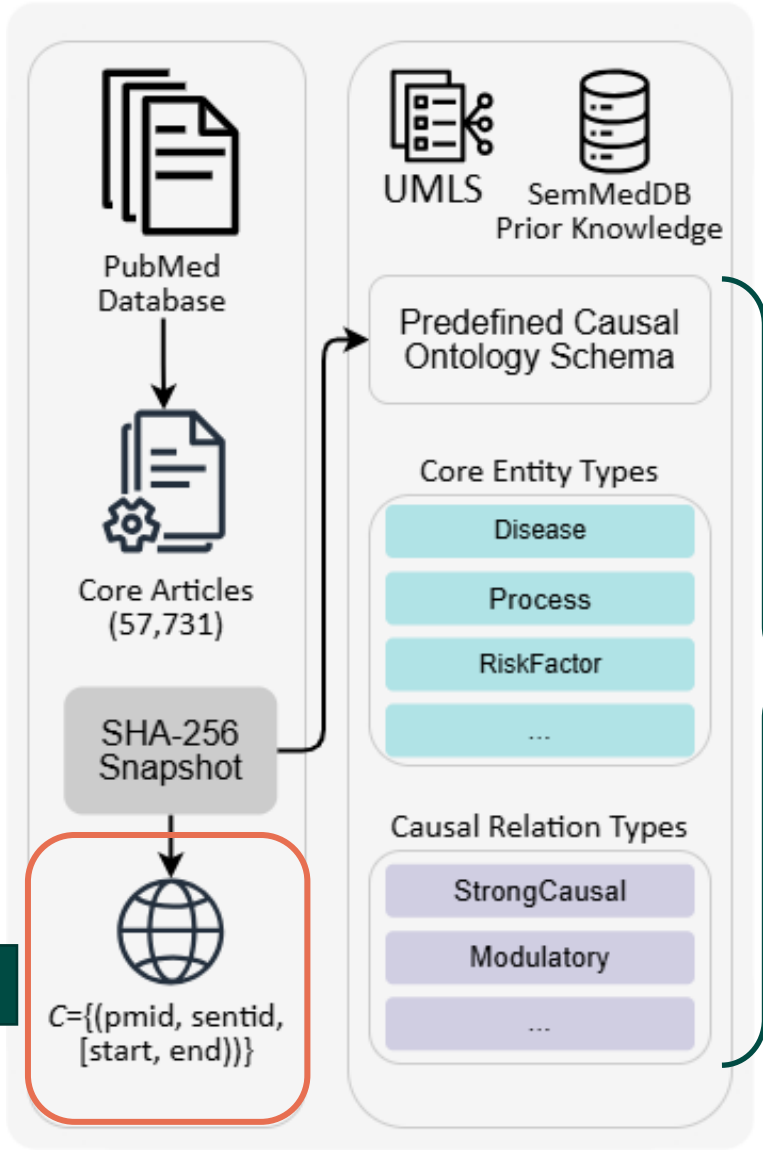
Framework — Data Acquisition & Schema

Corpus snapshot

- Systematic PubMed search with query diabetes + comorbidity
- After exclusions, 57,731 heterogeneous core publications were retained.
- SHA-256 hashed PMID list for snapshot identification

Coordinate system

C = (pmid, sentid, [start, end))



Semantic normalization based on UMLS

Ontology schema

Entity types

Disease, Process, RiskFactor, ClinicalManifestation, DrugChemical

Relation types

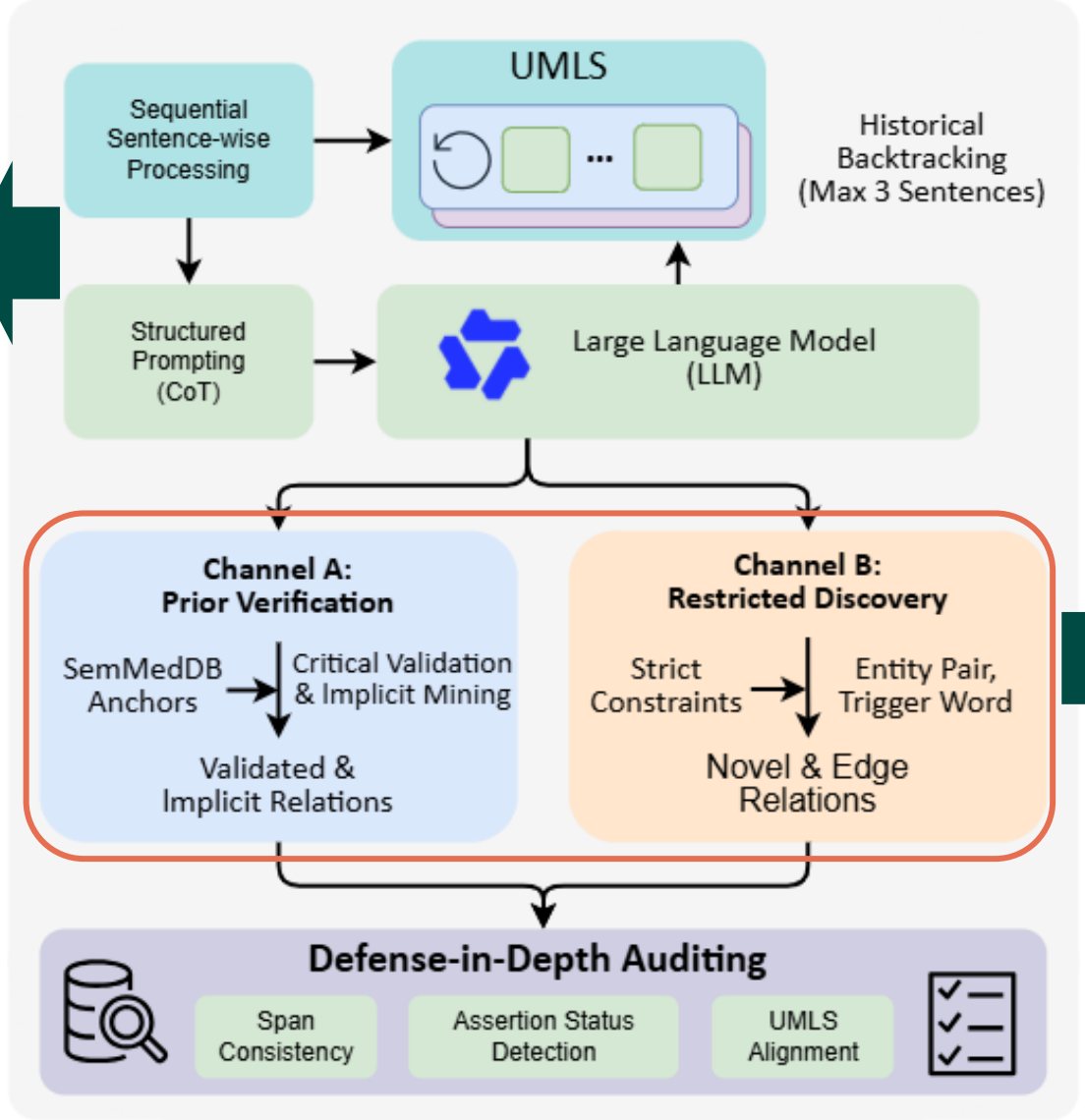
StrongCausal, RiskCausal, Modulatory, Mechanistic, WeakAssociation, Intervention

Prior knowledge

75,185 qualified prior triples used as admission standards for reasoning

Framework — Dual-Channel Evidence

- Cross-sentence causal reasoning in medical text
- Dynamic context for sentence-wise reasoning
- Structured prompting with role, task, schema, and CoT



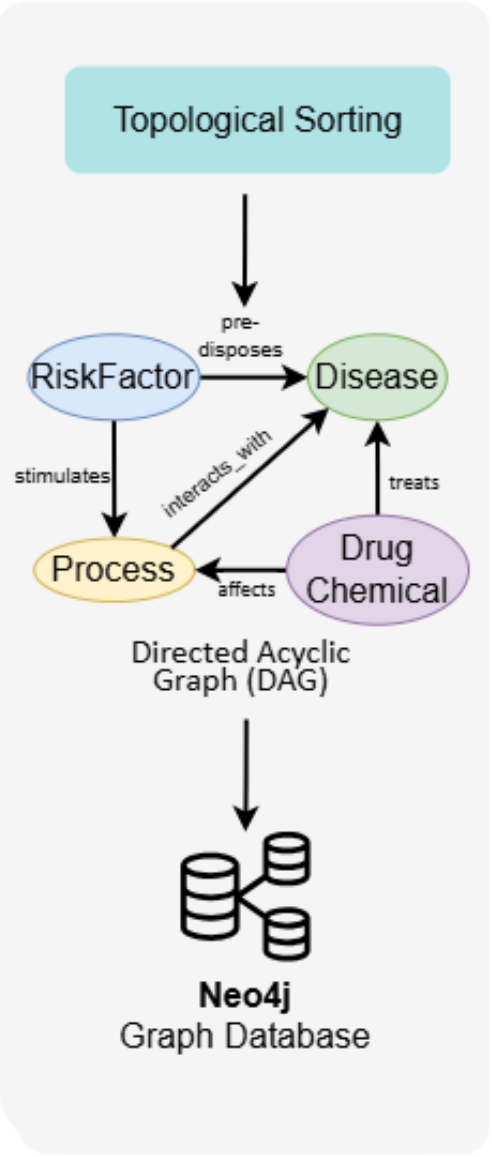
Channel A (prior verification)

With SemMedDB anchor window:
→ LLM verifies existing relations
→ Top-20 neighbors for discovering new associations

Channel B (constrained discovery)

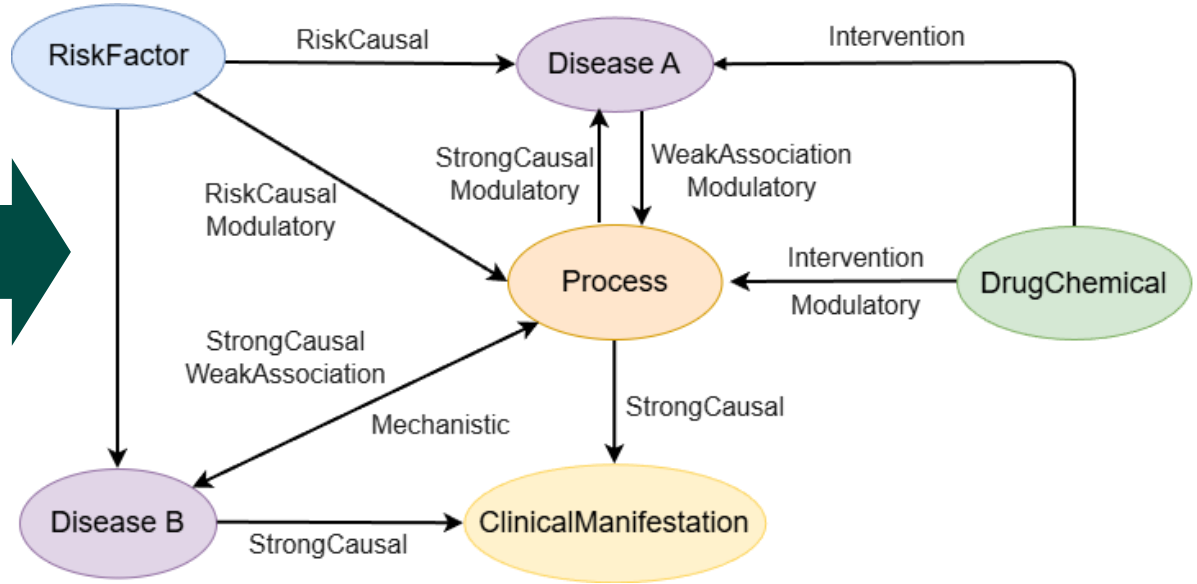
No-anchor blind zone discovery:
→ Entity type constraints
→ Mandatory causal triggers
→ Top-10 cutoff to suppress open-ended hallucinations

Framework — Graph Construction & Storage



- Causal triples → nodes and directed edges
- Topological sorting removes cycles

DAGs preserve directional causal flow



Neo4j supports risk prediction and comorbidity analysis

Outline

- 1. Introduction
- 2. Methodology
- 3. Experimental Results
- 4. Discussion

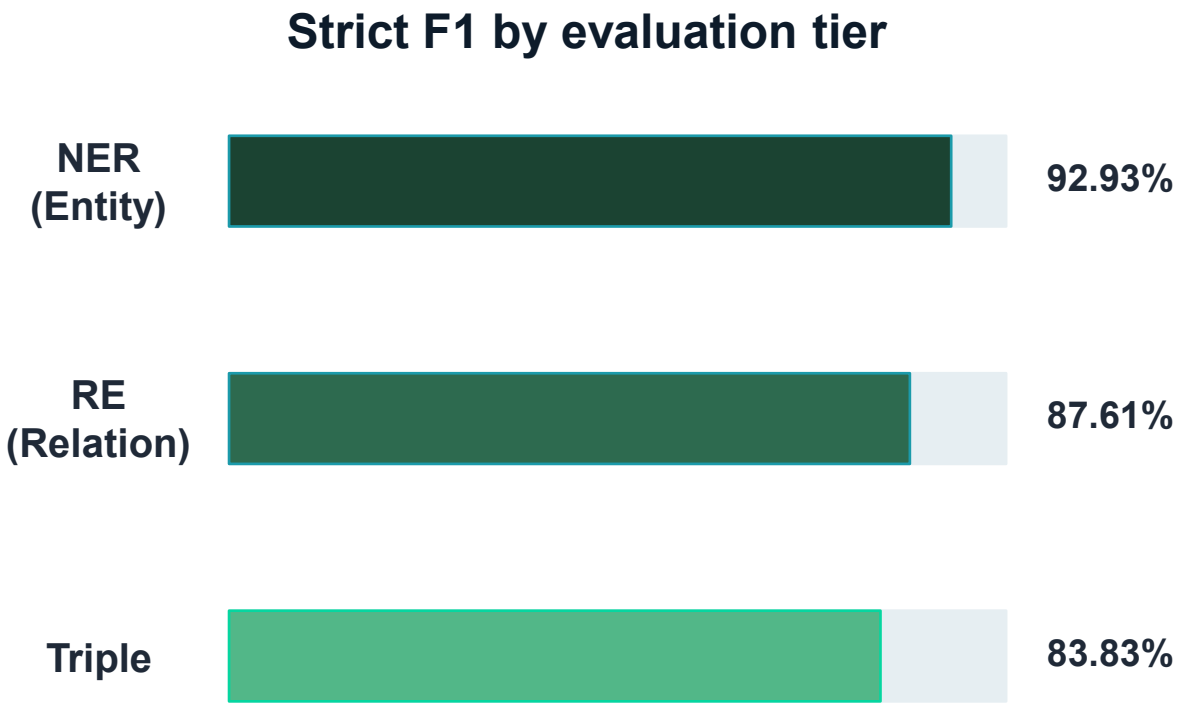
Evaluation Methodology

LLM-as-a-Judge · Qwen-Max

Layer	Name	Key Metric	Mechanism	Description
01	Schema Constraint	Schema Conformance Rate (%)	Ontology-Driven Admission	5 entity types + 6 relation types as admission whitelist
02	Entity & Relation Atomicity	EHR = 1 − Precision_Entity; Relation P / R / F1	Atomic Anchoring + Dual-Channel	Locks entity boundaries against hallucinations; ensures relation completeness
03	Causal Topological Logic	Causal Inversion Detection Strict F1 (Final Criterion)	LLM Reasoning + Strict Matching	Detects causal inversions in DAG; all four criteria must pass

Layer 1 Structural Compliance | Layer 2 Factual Validity & Hallucination Control | Layer 3 Causal Logic & Clinical Usability

Overall Performance



Entity hallucination rate 3.27%

Entity precision 96.73%

Relation recall 93.52%

Atomic Anchoring

Physical anchoring reduces hallucinations

Deep Causal Reasoning

Dual-channel improves causal recall

Clinical Applicability

High strict-F1 supports clinical applicability

Fine-grained Robustness Analysis

Entity Level

- Process→Precision 98.46%
- RiskFactor→F1 84.31%

Relatino Level

- Mechanistic→Precision 92.31%
- StrongCausal→Recall 91.95%
- Intervention→F1 78.79%

- Best performance on well-defined entities and mechanistic relations.
- Intervention remains challenging.

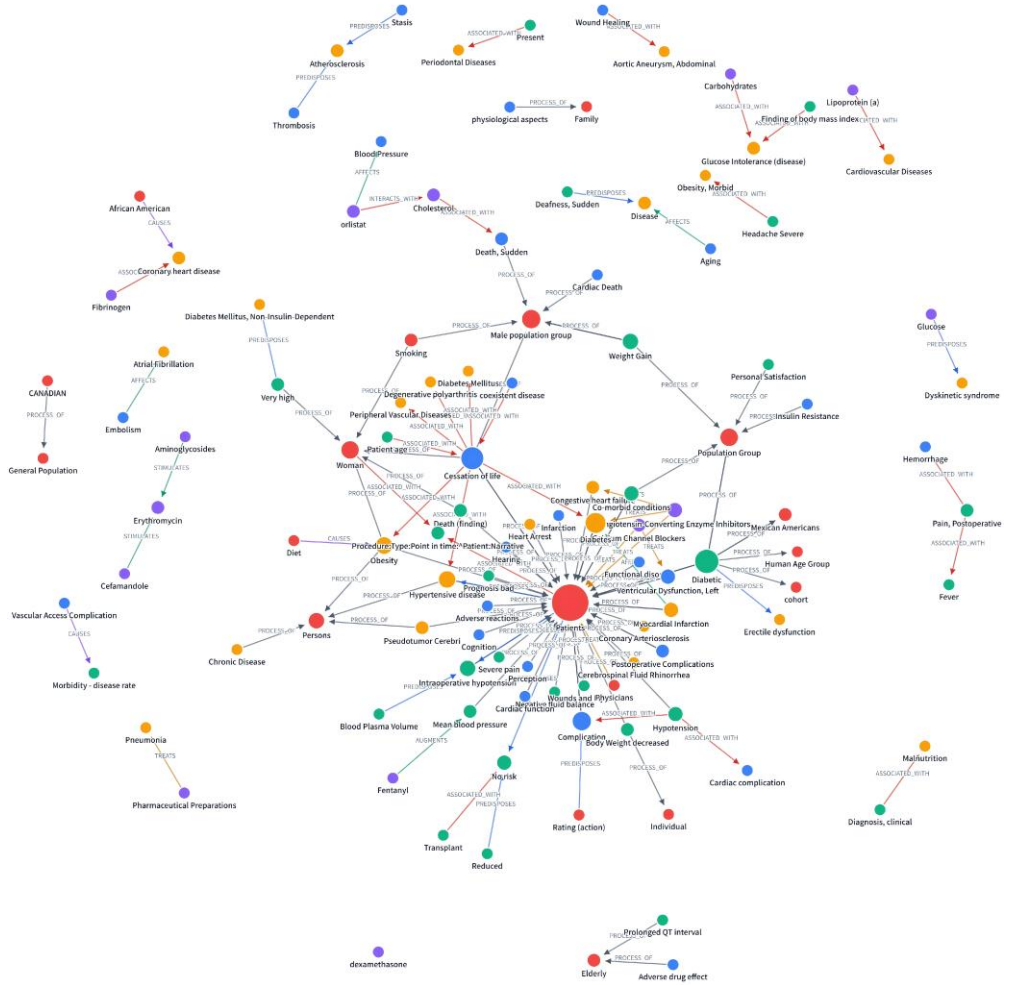
Topological Characteristics of Diab-CKG

Diab-CKG Scale：

- 15,573 nodes
- 272,140 causal edges
- Average degree: 17.5

Structural Features：

- StrongCausal + Mechanistic = 66.2%
- DrugChemical 27.8%



Example of Diab-CKG Local Causal Sub-graph

Example Causal Pathway in Diab-CKG

Disease pathway：

Obesity(RiskFactor)

→ Type 2 diabetes(RiskCausal)

→ Hyperglycemia(Mechanistic)

→ Diabetic nephropathy(StrongCausal)

→ Proteinuria(ClinicalManifestation)

Mechanism：

Chronic inflammation(Process)

→ **Modulatory** → Insulin resistance

Intervention:

Metformin(DrugChemical)

→ **Intervention** → Hyperglycemia ↓

An interpretable pathway from exposure to complication and intervention.

Outline

- 1. Introduction
- 2. Methodology
- 3. Experimental Results
- 4. Discussion

Conclusion and Future Work

Conclusion:

- Built **Diab-CKG** via physical anchoring, dual-channel evidence, and topological reconstruction.
- Achieved end-to-end **strict F1 83.83%** and **entity hallucination rate of 3.27%.**
- Revealed complete causal chains from risk exposure to intervention, providing a computable foundation for diabetic cascade analysis.

Future work:

- Extend to **full-text articles** and **EHRs** for broader comorbidity coverage.
- Combine **GNNs with causal DAGs** for early warning and drug repurposing.
- Enable **dynamic clinical decision support** in precision medicine.

Thank you！

Venice, Italy, May 25, 2026