Construction of A Causal Knowledge Graph for Research on Diabetes Comorbidities

Xueli Wu, Xuemei Yang, Longchao Wang, Yalan Huang, Xiaoli Tang
Institute of Medical Information
Chinese Academy of Medical Sciences and Peking Union Medical College
Beijing, China
e-mail: tang.xiaoli@imicams.ac.cn

Abstract—**Diabetes comorbidity is characterized by substantial mechanistic complexity and causal heterogeneity, which correlation-based approaches cannot adequately capture within deep pathological progression pathways. To resolve latent causal structures and mitigate generative hallucination in medical causal mining, this study introduces a hybrid paradigm that integrates physical anchoring, dual-channel evidence awareness, and topological reconstruction, thereby constructing the Diabetes comorbidity Causal Knowledge Graph (Diab-CKG). The framework establishes an atomized corpus indexing coordinate system to ensure traceable extraction and employs Large Language Models under strict ontology constraints to validate prior knowledge and identify novel associations, effectively suppressing generative hallucination. Experimental results demonstrate that the paradigm effectively connects unstructured text with structured reasoning, achieving robust performance in entity recognition and causal extraction, with an end-to-end Strict F1 score of 83.83% and a reduction of the Entity Hallucination Rate to 3.27%. The study delineates a comprehensive causal chain from risk exposure to clinical intervention, providing a logically coherent and computable foundation for modeling comorbidity cascades and supporting clinical decision making.**

Keywords- Diabetes mellitus; Causal knowledge graph; Comorbidity; Causal relationship extraction; Large language model

I. INTRODUCTION

Diabetes is a highly heterogeneous chronic metabolic disease whose clinical progression is frequently accompanied by complex comorbidity networks. Although epidemiological studies confirm high co-occurrence rates between diabetes and numerous comorbid conditions, most existing analyses remain confined to statistical correlations. These approaches cannot adequately capture mechanistic causal pathways, thereby limiting the development of prospective prevention and targeted intervention strategies for high-risk comorbidities. Therefore, constructing a causal knowledge graph that accurately describes disease progression directions and intermediate mechanisms has become a key task in knowledge driven healthcare.

However, building a high fidelity causal graph from unstructured biomedical literature faces multiple challenges. First, causal expressions in medical texts are highly implicit, with key pathological mechanisms often scattered across sentences, making long range logic difficult to capture for traditional methods [1]. Second, although Large Language Models (LLM) improve semantic understanding, their inherent hallucination risk may produce missing evidence or reversed logic in medical applications [2]. In addition, an internal tension exists between biological feedback loops and the Directed Acyclic Graph (DAG) structure required for computational causal reasoning. Balancing biological completeness and computational logical consistency in graph construction remains unresolved.

To address these challenges, this study proposes a hybrid computational paradigm that integrates physical anchoring, dual-channel evidence awareness, and topological reconstruction. The framework drives LLM reasoning through atomic-level evidence tracing. This design aims to resolve the precision–recall trade-off in causal mining and balance the tension between biological feedback loops and computational logic constraints. The result is an interpretable, logically consistent knowledge foundation for diabetes comorbidity research.

Finally, the remainder of this paper is organized as follows. Section 2 reviews related work, Section 3 introduces the proposed framework, Section 4 presents the experimental results and discussion, and Section 5 concludes the paper and discusses future work.

II. RELATED WORK

To contextualize the proposed framework within existing research, this section reviews advances in medical knowledge extraction, cross-sentence causal relation extraction, and hallucination control in generative models, identifying the limitations that motivate our hybrid paradigm.

A. Medical Knowledge Extraction Methods

Medical knowledge graph construction has evolved from rule-driven dictionary mapping methods such as MetaMap and SemMedDB to data-driven deep learning models including Bidirectional Encoder Representations from Transformers (BERT) and Bidirectional Long Short-Term Memory with Conditional Random Field (BiLSTM-CRF). While deep learning enhances generalization in entity and relation extraction, its reliance on annotated corpora and degraded performance on long-tail concepts limit practical deployment [3]. The rise of LLM enables few-shot and zero-shot reasoning, yet semantic drift has prompted schema-constrained controlled generation strategies [4].

B. Cross Sentence Causal Relation Extraction

Unlike conventional semantic association extraction, causal mining demands stronger logical discrimination. Existing approaches incorporate causal connectives or syntactic dependency trees, but largely focus on explicit intra-sentence relations. Capturing cascade mechanisms across paragraphs remains challenging, particularly in diabetes research [5]. Current document-level models also lack robustness in coreference resolution and multiple negation handling common in biomedical texts, hindering strict causal logic enforcement [6].

C. Hallucination Control and Evidence Traceability in Generative Models

Retrieval-augmented generation reduces knowledge errors by introducing external bases, yet noisy retrieval can still induce misleading outputs. Verification strategies such as self-consistency checking and multi-agent debate improve accuracy at higher reasoning cost. However, they emphasize overall text quality and rarely provide atomic-level evidence auditing for knowledge graph construction [7]. Without precise linkage between generated triples and original text offsets, existing graphs struggle to satisfy evidence-based medicine requirements.

D. Research Focus

Overall, current paradigms often increase semantic depth at the expense of traceability and logical consistency. This study therefore injects structured prior knowledge into LLM reasoning and establishes a strict physical coordinate mapping system to develop an extraction paradigm that unifies semantic generalization with rule-level rigor, supporting diabetes comorbidity research.

III. METHODS

To address the common problems of hallucination generation and causal logic disconnection in medical text mining, this study establishes a hybrid computational paradigm that integrates physical anchoring, dual channel evidence awareness, and topological reconstruction.

A. Data Acquisition and Schema Definition

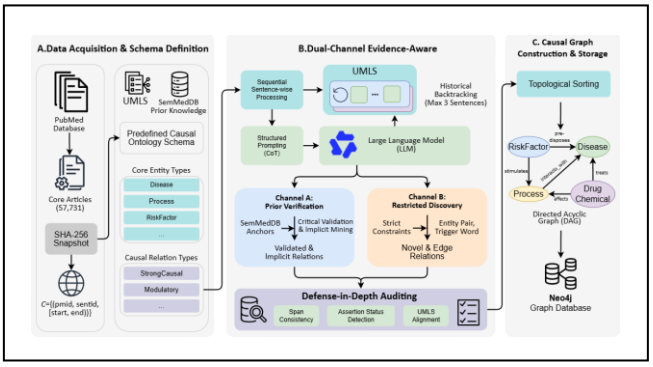

Figure 1. Dual-Channel Evidence-Aware Causal Knowledge Extraction Framework.

TABLE I. CAUSAL KNOWLEDGE GRAPH ENTITY TYPES

Entity Types	*Main UMLS content*	*Relation types*	*Main content*
Disease	dsyn, neop, mobd, cgab	StrongCausal	CAUSES, PRODUCES, COMPLICATES
Process	patf, phsf, phpr, menp, biof, genf, orgf, moft, ortf, celf	RiskCausal	PREDISPOSES
RiskFactor	bhvr, inbe, socb, acty, food, ocac; humn	Modulatory	AFFECTS, AUGMENTS, STIMULATES
Clinical Manifestation	sosy, fndg, lbtr, clna, inpo, anst	Mechanistic	PROCESS_OF, MANIFESTATION_OF
DrugChemical	phsu, clnd, horm, imft, antb, bacs; chvs, aapp, nnon, lipd, carb, enzy, nsba	WeakAssociation	ASSOCIATED_WITH, INTERACTS_WITH
		Intervention	TREATS

Building a high-fidelity medical causal graph requires a traceable data foundation and strict semantic specification. We conducted a systematic PubMed search using the query ((Diabetes Mellitus OR diabetes) AND comorbidit*). After excluding animal experiments, duplicates, and reviews, 57,731heterogeneous core publications were retained. To ensure reproducibility and physical traceability, the Secure Hash Algorithm 256-bit (SHA-256) hash of the PubMed Identifier (PMID) list was computed to generate a unique corpus snapshot identifier, Snapshot ID. Abstracts were mapped into a tamper-resistant global three-dimensional coordinate systems

$$C=\{(\text{pmid}, \text{sentid}, [\text{start}, \text{end}))\}. \quad (1)$$

Here, [start, end) denotes zero-based Unicode character offsets.

To reduce semantic heterogeneity and establish reasoning boundaries, a predefined causal ontology schema was constructed on the physical corpus as the blueprint for automated reasoning (Table I). At the entity level, Unified Medical Language System (UMLS) served as the normalization backbone to cluster five core categories. At the relation level, to compensate for the lack of strength differentiation in SemMedDB predicates, a six-category causal semantic system was predefined and used as the admission standard, yielding 75,185 qualified prior triples.

B. Dual-Channel Evidence-Aware Extraction

1) Dynamic Context and Structured Prompting

Given the prevalence of cross-sentence dependencies in medical texts, we adopted a sequential sentence-wise processing strategy rather than isolated sentence analysis. For each core sentence, a dynamic context window was iteratively constructed to support reasoning, forming a local semantic unit with clearly defined boundaries and complete contextual information. A structured prompting template was embedded into the LLM, integrating role settings, task specifications, schema hard constraints enforcing JSON output, and chain-

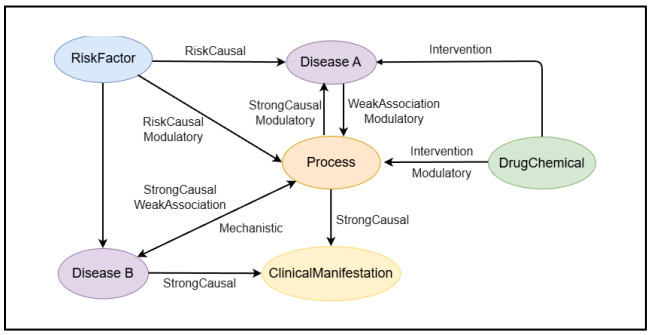

Figure 2. Predefined Topology Framework.

of-thought guidance. Under this design, dual-channel evidence-aware extraction w-
as performed based on contextual information density, as shown in Figure 1.

2) Dual-Channel Routing Mechanism

To optimize precision and recall, the task flow is routed by prior knowledge richness. For windows containing valid SemMedDB anchors, Channel A (prior verification mode) was activated. The LLM functioned as discriminator and generator, verifying prior relations while mining hidden associations among entity pairs satisfying the allowed type list and Top 20 proximity threshold. For anchor-absent knowledge blind areas, Channel B (constrained discovery mode) was applied, introducing three restrictions: limited entity type pairs, mandatory causal triggers, and a stricter Top 10 truncation threshold. Together, these constraints suppressed open-ended hallucinations while supplementing marginal knowledge.

3) Defense-in-Depth Auditing

All candidate triples underwent multidimensional defense-in-depth auditing. Span consistency checking first enforces exact character-level mapping within coordinate system C via inverse backtracking, physically blocking ungrounded hallucinations. Subsequently, assertion-state detection applies a deep coupling strategy to identify and remove negated and speculative expressions, ensuring atomic-level knowledge certainty.

4) Standardization and UMLS Alignment

To unify the semantic space and eliminate gaps between LLM outputs and prior knowledge, UMLS alignment was enforced, producing a semantically consistent high-fidelity causal graph.

C. Causal Graph Construction and Topological Storage

Given the directionality and temporal irreversibility of medical causality, a DAG was adopted to represent dependency pathways among diseases, risk factors, and interventions. Standardized triples aligned via UMLS were instantiated as nodes and directed edges in the initial network. A global topological sorting algorithm was then applied to detect and block cyclic dependencies violating medical temporal order, ensuring monotonic causal flow. The final graph follows the predefined topology in Figure 2, allowbreak

TABLE II. LAYERED PERFORMANCE EVALUATION OF ENTITY AND RELATION EXTRACTIONS

Evaluation tier	*Precision (%)*	*Recall (%)*	*Strict F1 (%)*
Entity Level (NER)	96.73	89.41	92.93
Relation Level (RE)	82.41	93.52	87.61
Strict Triple	79.11	89.15	83.83

constructing a logically self-consistent and unambiguous computable causal knowledge foundation.

IV. RESULTS AND DISCUSSION

To validate the effectiveness of the proposed hybrid paradigm and assess the quality of the constructed Diab-CKG, this section reports the evaluation methodology, extraction performance, hallucination audit results, and topological characteristics of the causal graph.

A. Evaluation Methodology

To overcome the limitations of static benchmarks and fragmented expert consensus in biomedical knowledge graph evaluation, this study adopts a dynamic auditing system based on the LLM-as-a-Judge framework, replacing conventional single-metric assessment. The flagship Qwen-Max model serves as the core auditor within a schema-constrained layered evaluation design. Assessment spans two dimensions: entity atomicity and causal topological logic. To quantify physical anchoring and noise resistance, the Entity Hallucination Rate (EHR) is defined as

$$EHR = 1 - Precision_{Entity}, \quad (2)$$

measuring the proportion of hallucinations untraceable to source text spans. At the logical level, a causal topology verification mechanism leverages LLM reasoning to detect causal inversion and ensure that the constructed DAG conforms to medical pathology. Performance is finally assessed under strict matching, where true positives require simultaneous verification of entity boundaries, semantic types, relation categories, and causal direction. The resulting Strict F1 serves as the decisive indicator of clinical usability.

B. Performance Analysis and Hallucination Audit

1) Overall Performance and Mechanism Validation

Deep auditing by Qwen-Max (Table II) shows that the dual-channel evidence-aware framework maintains high robustness from atomic anchoring to causal reasoning. Entity recognition achieves 96.73% precision and an F1 score of 92.93%, surpassing standard baselines. This improvement is attributed to atomic indexing, which precisely anchors medical term boundaries and reduces the Entity Hallucination Rate to 3.27%. In complex semantic reasoning, the model attains 87.61% F1 and 93.52% recall, reflecting strong sensitivity to implicit associations. The end-to-end strict F1 reaches 83.83%, confirming substantial clinical utility for automated high-fidelity graph construction.

2) Fine-grained Robustness

Fine-grained auditing (Table III) reveals differential adaptability across semantic spaces. Atomic indexing yields

TABLE III. FINE-GRAINED PERFORMANCE EVALUATION ACROSS PREDEFINED ENTITY AND RELATION TYPES

Schema Type	*Category*	*Precision*	*Recall*	*Strict F1*
Entity Types	Disease	98.25	89.84	93.85
	Process	98.46	88.89	93.43
	RiskFactor	95.56	75.44	84.31
	ClinicalManifestation	93.85	96.06	94.94
	DrugChemical	97.44	88.37	92.68
	Macro	96.71	87.72	91.84
Relation Types	StrongCausal	74.07	91.95	82.05
	RiskCausal	86.16	87.26	86.71
	Modulatory	78.69	88.89	83.48
	Mechanistic	92.31	85.71	88.89
	WeakAssociation	76.65	90.37	82.95
	Intervention	76.47	81.25	78.79
	Average (Macro)	80.73	87.57	83.81

98.46% precision for the Process category, whereas the 84.31% F1 for RiskFactor reflects boundary ambiguity in non-restrictive expressions. At the relation level, Mechanistic relation achieve 92.31% precision, ensuring accurate disease mechanism representation, while StrongCausal relations reach 91.95% recall, indicating a calibrated trade-off that captures major causal links with limited noise.

C. Topological Characteristics of Diab CKG

The complete Diab-CKG, derived from 57,731 diabetes publications, contains 15,573 entities and 272,140 directed causal edges, with an average degree of 17.5, exhibiting a den-

TABLE IV. STATISTICS AND DISTRIBUTION OF ENTITIES AND RELATIONS IN THE CONSTRUCTED DIAB-CKG

Category	*Type (Schema)*	*Count*
Nodes	DrugChemical	4326
	Process	1808
	Disease	4209
	ClinicalManifestation	4307
	RiskFactor	923
	Total Nodes（unique node）	15573
Edges	StrongCausal	17548
	Mechanistic	162558
	Modulatory	21851
	RiskCausal	27376
	Intervention	13264
	WeakAssociation	29543
	Total Edges	272140

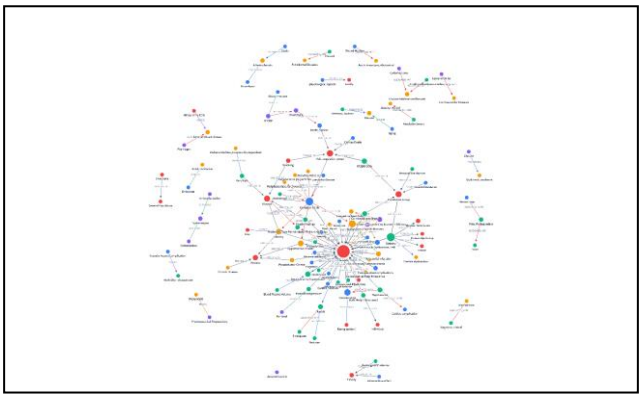

Figure 3. Example of Diab-CKG Local Causal Sub-graph.

se scale-free structure. Topological statistics (Table IV) show DrugChemical entities account for 27.8%, aligning with the clinical reliance on pharmacological intervention. StrongCausal and Mechanistic relations form the backbone, exceeding 66.2%, supporting inference of complication pathways, while RiskCausal and Modulatory relations enable modeling of complex biological feedback loops.

se scale-free structure. Topological statistics (Table IV) show DrugChemical entities account for 27.8%, aligning with the clinical reliance on pharmacological intervention. StrongCausal and Mechanistic relations form the backbone, exceeding 66.2%, supporting inference of complication pathways, while RiskCausal and Modulatory relations enable modeling of complex biological feedback loops.

Figure 3 illustrates a local comorbidity subgraph, visualizing cross-level causal cascades around core diseases and confirming that the graph captures both explicit associations and latent pathological evolution, providing a structured foundation for drug repositioning and comorbidity risk mitigation.

V. CONCLUSION AND FUTURE WORK

We address the challenge of converting dispersed biomedical narratives into computable causal evidence for diabetes comorbidity research. We introduce a hybrid framework combining physical anchoring, dual-channel evidence awareness, and topological reconstruction, constraining large language model reasoning within traceable evidence coordinates and a predefined causal ontology.

This evidence-bounded approach enables construction of a biologically faithful, logically consistent Diab-CKG, reducing hallucinations while supporting reliable end-to-end performance. Beyond extraction accuracy, Diab-CKG captures multi-stage causal cascades linking risk factors, pathological mechanisms, and therapeutic interventions, providing a computable foundation for mechanistic exploration and clinical knowledge discovery.

Limitations include reliance on abstracts, restricting temporal depth and phenotype granularity, and DAG enforcement, which may oversimplify feedback loops. Future work will integrate full-text literature and de-identified EHRs, and explore neuro-symbolic models combining graph neural networks with causal priors, advancing Diab-CKG toward

predictive and decision-support applications in precision medicine.

ACKNOWLEDGMENT

This research was supported by the National Key Research and Development Program of China (Project No.: 2023YFC2308803).

REFERENCES

[1] S. Wang, H.Y. Wang, and J. Du, "Discovering causal paths to diabetic nephropathy by combining computable biomedical knowledge with graph mining algorithms," AMIA Annu. Symp. Proc., pp. 1118-1124, Apr. 29, 2023.

[2] L. Huang, et al, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Transactions on Information Systems, pp. 1-55, Jan. 24, 2025.

[3] S. Sun, Q. Hu, F. Xu, F. Hu, Y. Wu and B. Wang, "Medical named entity recognition based on domain knowledge and position encoding," BMC Med. Inform. Decis. Mak., p. 235, Jul. 1, 2025, doi:10.1186/s12911-025-03037-0.

[4] R. Wagner, E. Kitzelmann and I. Boersch, "Mitigating Hallucination by Integrating Knowledge Graphs into LLM Inference – a Systematic Literature Review," Association for Computational Linguistics, pp. 795-805, Jul. 2025, doi: 10.18653/v1/2025.acl-srw.53.

[5] X. Jin, X. Wang, X. Luo, S. Huang and S. Gu, "Inter-sentence and Implicit Causality Extraction from Chinese Corpus, Advances in Knowledge Discovery and Data Mining, pp. 739–751.May 6 2020 doi: 10.1007/978-3-030-47426-3_57.

[6] J. Wang, et al, "Document-level biomedical relation extraction using graph convolutional network and multihead attention: Algorithm development and validation," JMIR Med. Inform., p. 17638 Jul. 31, 2020, doi: 10.2196/17638.

[7] Choi S, et al. Knowledge Graph Construction: Extraction, Learning, and Evaluation (2022–2024 Survey). Appl Sci. 2025;15(7):37.