ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

Reasoning over Knowledge Graphs: Enhancing LLMs for Trustworthy Medical Question Answering

Jinghui Mao[1], Menglin Cui[1,*], Yuejia Dai[1]

[1]Business School, University of Shanghai for Science and Technology, Shanghai, China
**Corresponding author*

Abstract: *Large language model (LLM)-based medical question-answering systems show promising potential for clinical consultation and health information services. However, their application in high-stakes medical scenarios is limited by issues such as hallucinated responses, unreliable reasoning, and insufficient factual grounding. To address these challenges, this paper integrates a domain-specific medical knowledge graph to constrain LLM outputs for trustworthy medical QA. In addition, a graph-validated Weighted Factuality Score is introduced to evaluate the factual reliability of generated responses by verifying atomic facts against knowledge graph evidence. Experimental results on a medical dataset show that the proposed knowledge graph-enhanced RAG framework improves the average factuality score compared with the baseline. These results demonstrate that incorporating knowledge graph constraints enhances both the factual reliability and interpretability of LLM-based medical QA systems.*

Keywords: *Knowledge graph, Large language models, Medical question answering*

1. Introduction

The rapid evolution of large language models (LLMs) has fundamentally transformed intelligent medical question-answering (QA) systems. By leveraging pre-trained knowledge and language understanding, these models can address complex queries and support clinical decision-making. However, the deployment of generative models in high-stakes healthcare environments remains impeded by critical vulnerabilities, most notably hallucinated responses, unreliable reasoning, and knowledge-reasoning decoupling [1]. Studies indicate that hallucinations in medical contexts encompass both factual fabrications and deceptive logical fallacies [2, 3], which can lead to misdiagnosis and create significant regulatory barriers for clinical adoption[4].

Current LLM-based medical QA systems often lack explicit mechanisms for factual grounding and integrative reasoning[5]. While Retrieval-Augmented Generation (RAG) effectively mitigates some hallucinations[6], conventional RAG primarily relies on unstructured text retrieval, which can still result in outputs that deviate from verified clinical evidence[7].Recent research suggests that the fusion of neuro-symbolic reasoning and knowledge graphs (KGs) offers a robust solution to these challenges[8-10]. By organizing domain knowledge into structured entities and semantic relations, KGs provide a ground truth that enhances both the interpretability and factual consistency of model outputs. Nevertheless, effectively integrating graph-structured knowledge with generative models—specifically regarding semantic query mapping and the verification of generated responses—remains a non-trivial task.

To bridge these gaps, this paper proposes KG-RAG, a Knowledge Graph-Constrained Retrieval-Augmented Generation framework. Our approach integrates a domain-specific medical knowledge graph with semantic parsing and graph reasoning to ensure that LLM-generated responses are rigorously grounded in verified medical evidence. Furthermore, to address the limitations of existing evaluation metrics, we introduce a graph-validated Weighted Factuality Score (FS), which quantitatively assesses the reliability of responses by verifying atomic facts against structured KG evidence.

Experimental results demonstrate that the proposed framework significantly improves factual accuracy and provides transparent reasoning traces, satisfying the stringent reliability requirements of clinical applications.

ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

2. Methodology

This paper proposes a KG-RAG architecture. As illustrated in Figure 1, the system processing flow consists of four major modules: constructing a high-quality MKG via the Data Layer, parsing natural language intents through the Semantic Understanding Module, retrieving structured facts using the Graph Reasoning Engine, and transforming discrete knowledge into natural language feedback via the Generation Module.

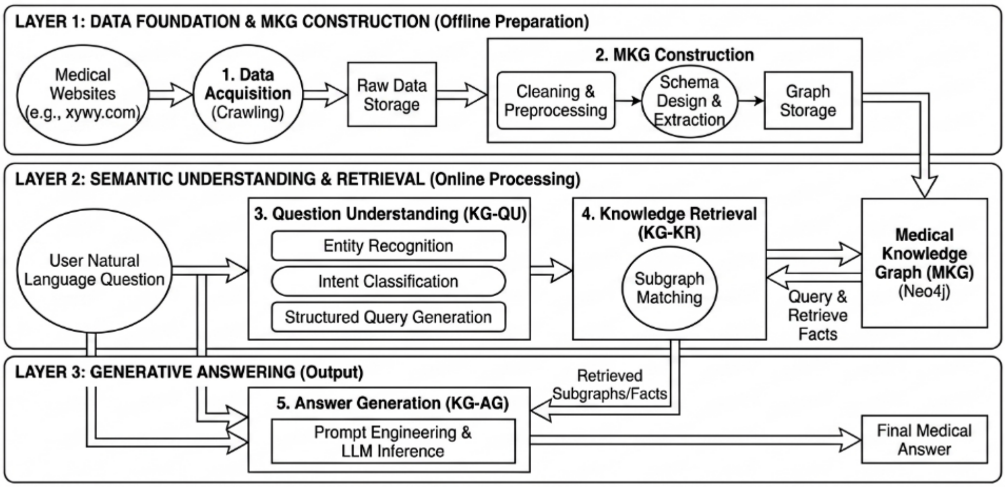

Figure 1: Architecture of the Medical KG-RAG Framework

2.1. Dataset

To build a trustworthy MKG, this paper selects the authoritative medical platform "xywy.com" as the primary data source, capturing medical knowledge across seven dimensions: disease attributes, pathology, clinical symptoms, auxiliary checks, treatment plans, dietary nursing, and medication recommendations. Given the multi-source heterogeneity of medical data, we establish a "Bottom-Up" data governance paradigm. The data preprocessing function is defined as: $f_{ETL}: D_{raw} \rightarrow E_{norm}$ .This function maps unstructured raw web data D_{raw} into a set of standardized entities E_{norm} that conform to the graph ontology. During this process, the system implements schema alignment for non-standard tags, utilizes heuristic rules to discretize long-text fields (e.g., decomposing complex dietary advice into independent relations such as *do_eat*, *not_eat*, and *recommand_eat*), and applies regularization filters for noise reduction and entity normalization.

In order to evaluate the system’s performance, we also construct a dedicated test dataset of 600 samples using a reverse-generation strategy, where each entry contains complete ground truth including entities, triples, and executable query paths. The system performs natural language reverse-mapping of structured facts across 20 core medical intents via the MKG. To validate safety and robustness, 25% adversarial samples targeting knowledge–reasoning decoupling (e.g., erroneous disease association queries) are incorporated. This dual-track approach ensures that the underlying data and evaluation metrics are intrinsically consistent with validated medical logic.

2.2. Medical Knowledge Graph Construction

To support complex clinical reasoning, the medical knowledge graph is formally represented as a directed labeled graph:

$$G = \langle E, R, T \rangle \tag{1}$$

where the components are defined as follows:

- **Entity set (*E*):** $E = \{e_1, e_2, \ldots, e_n\}$ represents discrete knowledge nodes. Entities are mapped to a type set $C = \{\text{Disease,Symptom,Drug,Food,Check,Department,Producer}\}$ such that $\forall e \in E, Type_e \in C$.
- **Relation set (*R*):** $R = \{r_1, r_2, \ldots, r_m\}$ defines the semantic connections, covering the logic from pathological mechanisms (e.g., has_symptom) to diagnosis and treatment plans (e.g., need_check).

ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

- **Triple set (*T*):** $T \subseteq E \times R \times E$ constitutes the basic factual units. Each triple $t = \langle h, r, t \rangle \in T$ describes a determinate medical fact, where h is the head entity, t is the tail entity, and r is the semantic edge.

In the visualization, node colors are used to distinguish semantic entity categories. Specifically, disease, symptom, drug, examination, and food entities are encoded using yellow, blue, orange, green, and pink colors. As shown in Figure 2, the resulting graph exhibits a disease-centric star-shaped topology, where disease entities function as semantic anchors linking heterogeneous medical knowledge.

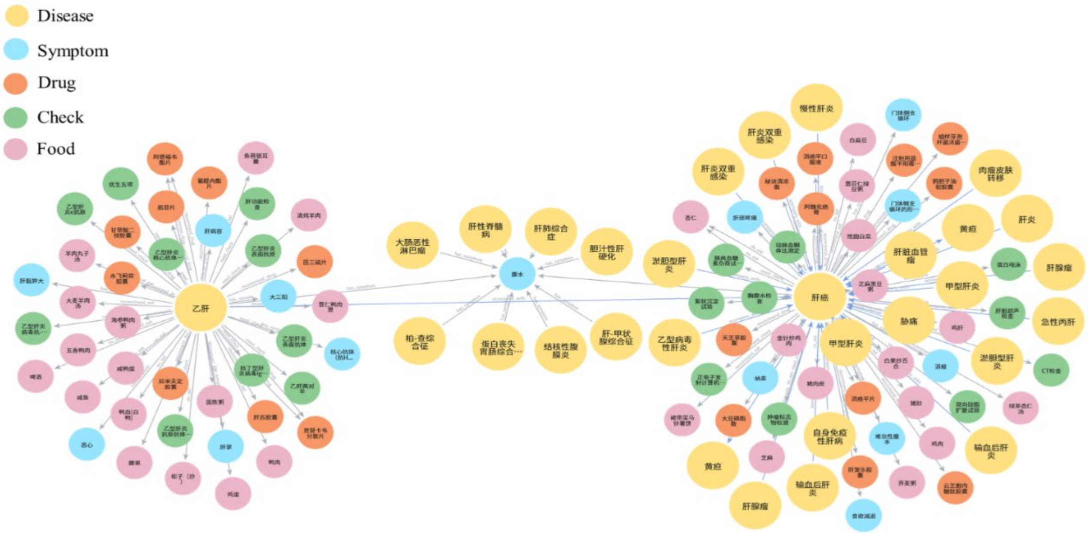

Figure 2: Example of Local Knowledge Subgraph Retrieval Based on Disease-Centric Schema.

Based on the above definitions, the medical knowledge graph constructed in this paper contains 44,111 entity nodes and 294,148 semantic relation edges. Detailed statistics on the distribution of graph entities and relations are shown in Table 1.

Table 1: Statistics of nodes and relations in the medical knowledge graph.

Type	Label/Relation	Count	Semantic Description
Nodes	Disease	8,807	Core entities covering common and intractable diseases
	Symptom	5,998	Clinical manifestations of diseases
	Drug	3,828	Western and traditional Chinese medicines
	Food	4,870	Ingredients for dietary recommendations
	Check	3,353	Medical examination items for auxiliary diagnosis
	Producer	17,201	Pharmaceutical manufacturers
	Department	54	Hospital clinical departments
	Total	**44,111**	—
Edges	recommand_drug / common_drug	74,115	Medication guidance: recommended and common drugs
	recommand_eat / do_eat / no_eat	84,706	Dietary recommendations and restrictions
	need_check	39,423	Required diagnostic examinations for diseases
	has_symptom	5,998	Associations between diseases and symptoms
	acompany_with	12,029	Comorbidity or concurrent disease relationships
	belongs_to	20,562	Organizational hierarchy (e.g., department affiliation)
	Total	**294,148**	Total number of semantic relations

2.3. Semantic Understanding and Graph Reasoning

Semantic understanding and graph reasoning module utilizes a cascaded architecture to transform unstructured queries Q into machine-executable semantic representations $\langle I, E \rangle$.In the entity linking phase, to resolve nested ambiguity, the system uses the AC Automaton algorithm to extract a candidate entity set S and applies the Maximal Span Principle. The filtering function is defined as:

$$E = \{ \mathrm{e} \in S \mid \nexists e' \in S, \mathrm{s.t.}\ e \subset e' \} \quad (2)$$

ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

This mechanism ensures the retention of the most specific semantic units with the highest information entropy. For intent inference, a deterministic engine based on first-order logic activates intent slots when a specific entity type $Type_e$ and a linguistic feature word $w \in F$ co-occur within the query window. To ensure clinical safety, a contextual negation detection mechanism scans for negation operators to trigger intent flipping logic.

Once the semantic representation $\{I, E\}$ is established, the mapping function Ψ utilizes a predefined Cypher template library to convert these intents into Graph Database Query Language. To address composite medical needs, we propose a Multi-path Aggregation Mechanism. Based on the parsed intent I, the system dynamically generates parallel sub-queries for different relation types, such as *do_eat* and *recommand_eat*. These sub-queries are executed to retrieve discrete subgraphs, which are subsequently merged through logical aggregation into a unified knowledge context. This multi-path retrieval strategy significantly enhances information coverage and prevents the omission of critical clinical evidence.

2.4. Knowledge-Constrained Generation

To mitigate hallucinations in closed-domain QA, the framework employs Qwen-Plus as its generative core, integrated within a retrieval-augmented constrained generation mechanism. First, a serialization function $Serialize(\cdot)$ maps the retrieved subgraph G_{sub} into a context sequence C in the natural language prompt space:

$$C = Serialize(G_{sub}) = [FACTS] \oplus \sum_{<h,r,t> \in G_{sub}} Relation(h, r, t) \oplus [END] \tag{3}$$

where $\oplus$ represents string concatenation, and $Relation(\cdot)$ converts structured triples into declarative sentences using natural language templates.

To further ensure the trustworthiness of clinical advice, we introduce a Negative Constraint Mechanism to regulate decoding boundaries. The LLM is permitted to generate a response if and only if the retrieval module provides valid evidence $\boldsymbol{C}$. If the evidence set is empty ($\boldsymbol{C} = \emptyset$), the system is forced to output a predefined rejection token (e.g., "No records in the database"). This hard constraint, combined with a low-temperature sampling strategy ($\boldsymbol{\tau} \rightarrow \boldsymbol{0}$), blocks the model from fabricating facts based on pre-trained parameters, ensuring the consistency and rigor of the medical advice.

3. Experimental Setup and Model Evaluation

To achieve an objective evaluation of the proposed medical question-answering system, this paper establishes a complete experimental setup including dataset construction, system inference, and factuality assessment. The overall evaluation architecture is illustrated in Figure 3.

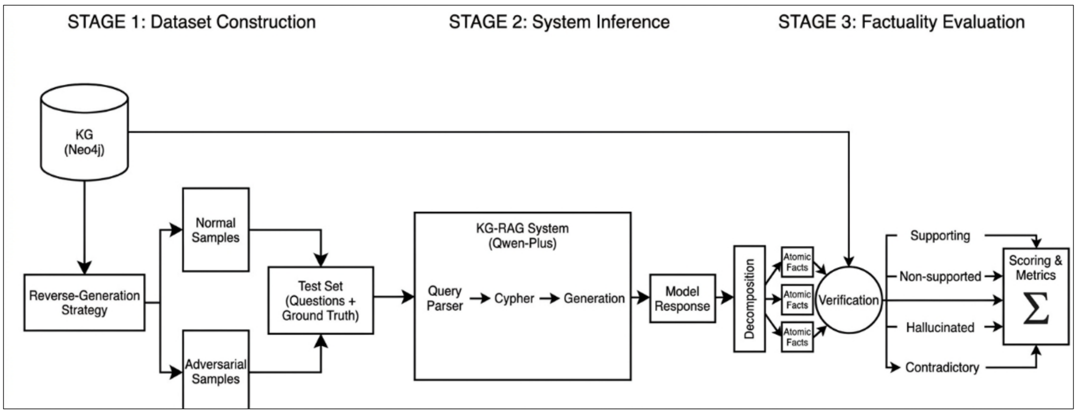

Figure 3: Overall Evaluation Architecture Diagram

3.1. Overall Factuality Evaluation

Following the trustworthy medical QA evaluation framework proposed by Wang et al. (2025)[11], this paper designed a graph-validated Weighted Factuality Score (FS). This metric adopts the atomic fact

ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

decomposition paradigm of FactScore[12] and incorporates the knowledge graph path validation logic proposed by FACE-KEG[13], achieving a fine-grained evaluation that is more rigorously aligned with medical logic than shallow metrics like BLEU or ROUGE. The methodological core is "Decomposition-Verification-Scoring": first, the response is decomposed into atomic factual units; then, an EvidenceVerifier module maps each unit back to the KG for triple matching. Based on the matching results, each unit is assigned a specific verification label and score (s_i): **Supporting (+1.0)**, where explicit triple evidence for the entity and relation exists; **Non-supported (+0.5)**, where the entity exists but no direct relational path is retrieved; **Hallucinated (0.0)**, which involves unknown entities not present in the graph; and **Contradictory (-1.0)**, which conflicts with mutually exclusive relations (e.g., contraindications) and represents a severe medical logic error.

To quantitatively aggregate the response quality, the *FactualityScorer* module calculates the final FS using a weighted average method. The formula is as follows:

$$FS = \frac{\sum_{i=1}^{N} w_i \times s_i}{\sum_{i=1}^{N} w_i} \quad (4)$$

where N is the total number of factual units, w_i represents the **category weight** (Core=0.5, Advice=0.3, Supplementary=0.2), and s_i is the verification score. Responses are classified into three credibility tiers: **High-Factuality** ($FS \geq 0.75$), **Partially Factual** ($0.4 \leq FS < 0.75$), and **Low-Factuality** ($FS < 0.4$).

To objectively assess the performance of the proposed **KG-RAG** architecture, a **Standard RAG** model was introduced as a baseline. This baseline utilizes a non-structured text retrieval paradigm based on dense vector matching. The comparative results across the full test set (600 samples) are shown in Table 2.

Table 2: Comparative Factuality Evaluation: Standard RAG vs. KG-RAG System

Model Architecture	Average FS	High-Factuality (FS≥0.75)	Partially Factual (0.4≤FS<0.75)	Low-Factuality (FS<0.4)
Standard RAG (Baseline)	0.7752	385 (64.17%)	145 (24.17%)	70 (11.67%)
KG-RAG (Proposed)	**0.8167**	**452 (75.33%)**	**72 (12.00%)**	**76 (12.67%)**

The KG-RAG system's average score (0.8167) significantly outperforms the baseline (0.7752), with the proportion of High-Factuality responses increasing by approximately 5.3%, demonstrating the effectiveness of the proposed KG-constrained mechanism in improving factual reliability. Standard RAG showed a heavy accumulation of samples in the Partially Factual range (24.17%), reflecting semantic drift or vague phrasing; KG-RAG compressed this to 12.00% through "hard constraints" of structured facts. The similar proportions in the Low-Factuality range are primarily due to the 25% adversarial samples: for KG-RAG, these represent successful compliant rejections (e.g., "No records in the database") which, by providing no positive factual units, naturally fall into this scoring interval.

3.2. Fine-grained Knowledge Dimensions and Interpretability Trade-offs

To explore performance across different medical knowledge types, factual units were categorized by semantic importance. As the table 3 showed, average scores for specific knowledge dimensions were calculated using "Session-level Aggregation" to reflect the overall reliability of global responses when specific dimensions are triggered.

Table 3: Score Distribution and Feature Analysis by Knowledge Dimension

Knowledge Dimension	Average Score	Sample Proportion	Feature Description
Supplementary	0.9219	89.78%	Highly standardized; extremely high accuracy.
Advice	0.5876	2.38%	Moderate confidence; occasional KG coverage gaps.
Core	0.4732	7.84%	Distinctive low-score profile; requires deep attribution.

The Core dimension obtained a lower score of 0.4732, a phenomenon resulting from a necessary trade-off for white-box explainability. To resolve the "black box" issue, the system is designed to forcefully output the [Graph Path] for interpretability. While highly valuable to clinicians, this structured process data is difficult for automated triple-verifiers to parse and is often conservatively judged as "no direct evidence (+0.5)". This statistical low score is a positive signal, indicating the system preserves reasoning traces instead of end-to-end probabilistic generation.

ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

3.3. Safety and Boundary Defense Evaluation

The system enforces fine-grained deterministic classification to strictly constrain generation boundaries via two core defense mechanisms: a Mutually Exclusive Relation Dictionary that imposes severe "Contradictory" penalties on semantic relations violating graph contraindications to intercept high-risk medical logic errors, and a strict Truth-Alignment Penalty that reclassifies any out-of-scope generated entities as "Hallucinated" to curb factual fabrication.

Table 4: Safety and Risk Interception Assessment

Fact Label	Detection Count	Overall Proportion	Safety Implication
Supporting	2,648	92.68%	High knowledge precision.
Non-supported	16	0.56%	Good knowledge completeness.
Hallucinated	154	5.39%	Rejection capability: few hallucinations under 150 malicious attacks.
Contradictory	39	1.37%	Risk interception: extremely low rate of severe medical logic errors.

As shown in Table 4, under high-pressure adversarial testing, these mechanisms successfully suppressed the "Hallucinated" label rate to 5.39%, effectively triggering the compliant rejection mechanism. Meanwhile, the detection rate of critical "Contradictory" errors remained under 1.5%. This verifies that hard graph constraints form a robust defense against fatal clinical risks.

4. Case Study and Qualitative Analysis

To evaluate the operational mechanisms of the system in practical clinical scenarios, this section selects two representative cases from the real-world logs for qualitative analysis. These cases focus on two distinct intent categories: "high-precision clinical facts" and "complex dietary guidance".

4.1. Precise Mapping and Structural Constraints of Diagnostic Plans

For medical examination-related queries with stringent factual accuracy requirements, the proposed system achieves precise responses via a KG-constrained retrieval mechanism. Taking the query *"What examinations are required for vulvar malignant schwannoma?"* as an example, the Natural Language Understanding (NLU) module accurately identifies this long-tail disease entity and accomplishes entity type anchoring. It then constructs a targeted Cypher query statement to lock onto the *need_check* semantic relation, directly retrieving specialized examination items (e.g., neurological immunological testing) from the MKG. Meanwhile, the complete reasoning path from the disease entity to specific examination items is explicitly presented, with the detailed running logs shown in Table 5.

Table 5: Diagnostic Query Log for Vulvar Malignant Schwannoma

User Input: What examinations are required for vulvar malignant schwannoma?
Dictionary Matching (Dicts): final_dict : {'Schwannoma': ['disease']} diseases_dict : {'Schwannoma': ['disease']} medical_dict : {'Schwannoma': ['disease']}
Graph Query Statement (Cypher): ["MATCH (m:Disease)-[r:need_check]->(n:Check) where m.name = 'Schwannoma' return m.name, r.name, n.name"]
Knowledge Graph Path: Disease (Schwannoma) → [need_check relation] → Examination Items (Neurological Immunological Testing, Neurological Cytological Examination, etc.)
Answer: In response to your question "What examinations are required for vulvar malignant schwannoma?", we have professionally summarized the diagnostic and examination information of schwannoma based on the medical knowledge graph as follows: 1. Neurological Examination: As a basic assessment method, it is used to initially determine whether the tumor involves the surrounding nerve structures and their functional status. 2. Neurological Cytological Examination: Morphological analysis of cells from relevant tissue or body fluid samples helps to clarify the nature of the tumor. 3. Neurological Immunological Testing: Detection of specific antigen expression (e.g., S-100 protein, SOX10) via immunohistochemical techniques serves as a critical basis for the diagnosis and differentiation of benign and malignant schwannoma. 4. Neurological Nuclear Medicine Examination: Techniques such as PET-CT assist in evaluating tumor metabolic

ISSN 2616-5775 Vol. 9, Issue 5: 20-27, DOI: 10.25236/AJCIS.2026.090503

activity, localizing lesions and screening for distant metastases, which is of great value especially for malignant or recurrent cases. Note: The knowledge graph does not separately define specific examination items for the "vulvar" or "malignant" subtypes. The above examinations are universal and key diagnostic methods for schwannoma. In clinical practice, a comprehensive judgment should be made combining imaging examinations (e.g., MRI) and pathological biopsy. It is recommended to complete an individualized examination plan under the guidance of a specialist.

This case illustrates the effectiveness of KG-based structural constraints in limiting the generation space of the LLM. Because responses must be supported by retrieved triples, the model is prevented from producing generalized or unrelated examinations. Consequently, the generated answer maintains high factual precision while preserving transparent reasoning paths derived from the knowledge graph.

4.2. Multi-source Knowledge Aggregation under Complex Intents

For queries with broad semantics or implicit requirements, the system deconstructs intents and aggregates multi-source knowledge. As shown in Table 6, when faced with the unstructured question *"Dietary suggestions for minimal change nephrosis?"*, the system decomposes it into two sub-requirements: "suitable ingredients" and "recommended recipes".

The logs reveal that the system executes parallel Cypher queries: one retrieving basic ingredients like "black beans" via the *do_eat* relation, and another retrieving complex dishes like "minced meat with spinach" via the *recommand_eat* relation. The generation module then performs a logical fusion of these multi-dimensional retrieval results.

Table 6: Multi-source Knowledge Aggregation in Complex Dietary Scenarios

User Input: What dietary suggestions are there for minimal change nephrosis?
Dictionary Matching: final_dict : {'Minimal Change Nephrosis': ['disease']} diseases_dict : {'Minimal Change Nephrosis': ['disease']} medical_dict : {'Minimal Change Nephrosis': ['disease']}
Graph Query Statement (Cypher): disease_do_food ["MATCH (m:Disease)-[r:do_eat]->(n:Food) where m.name = 'Minimal Change Nephrosis' return m.name, r.name, n.name", "MATCH (m:Disease)-[r:recommand_eat]->(n:Food) where m.name = 'Minimal Change Nephrosis' return m.name, r.name, n.name"]
Knowledge Graph Path: Disease (Minimal Change Nephrosis) → [do_eat / recommand_eat relation] → Food (Spinach and Celery Congee, Black Beans, etc.)
Answer: Hello, rational dietary adjustment plays a positive role in the disease control and recovery of minimal change nephrosis. Based on current clinical recommendations, we have compiled the following dietary guidelines for you: 1. Recommended Foods: Appropriate intake of high-quality protein and nutrient-rich foods such as black beans, eggs, mung beans and sunflower seed kernels is advised, as these foods help supplement the nutritional needs of the body with relatively low metabolic burden. 2. Recommended Recipes: Choose light, easily digestible and nutritionally balanced dishes, such as stewed black carp (rich in high-quality protein and unsaturated fatty acids), Spinach and Celery Congee, Spinach and Milk Congee (helpful for iron and calcium supplementation), and Carrot and Tomato Soup (rich in antioxidants), which are conducive to maintaining a good physical state. 3. Avoided Foods: Please try to avoid high-salt and irritating foods such as pickles and chili peppers to prevent increasing renal burden or inducing discomfort such as edema.

By decomposing broad "dietary advice" into parallel query paths and achieving cross-relational aggregation in the final response, the system effectively maps vague inquiries to structured storage. This multi-path retrieval strategy significantly enhances information coverage and validates the logical integrity of the architecture when handling long-tail lifestyle guidance.

5. Conclusions and Future Works

Based on the aforementioned analysis and discussion, this study verifies that utilizing a medical knowledge graph as an external constraint source significantly mitigates the hallucination risks of Large Language Models. By employing graph database query languages to directly anchor user intents to verified structured triples, the system achieves a paradigm shift from probabilistic generation to symbolic retrieval. This ensures that outputs possess a clear medical basis, which not only elevates the overall factuality score but also demonstrates robust compliant rejection capabilities during adversarial sample testing. Furthermore, the system's scoring performance in the core factual dimension reflects a necessary trade-off made for interpretability. The system is designed to explicitly output graph evidence; although such structured information is difficult for automated verification tools to fully capture, the transparent reasoning traces it provides substantially enhance output interpretability, offering added clinical reference value for medical practitioners.

Despite its excellent performance, the system still faces limitations, including a singular source of knowledge, insufficient adaptability of intent recognition to highly colloquial queries, and a lack of multi-modal data processing capabilities. Therefore, future work will focus on constructing agent-driven dynamic knowledge update mechanisms, developing a hybrid intent recognition architecture that merges pre-trained semantic models with rule engines, and introducing multi-modal capabilities to parse medical images and laboratory reports. While continuously expanding the boundaries of clinical applications, future research will also further strengthen medical ethics and safety alignment mechanisms to ensure that intelligent services strictly adhere to medical norms and legal boundaries.

Acknowledgements

This work is supported by Humanities and Social Science Fund of Ministry of Education of China (Grant No. 24YJCZH035).

References

[1] Bedi S, Jiang Y, Chung P, et al. Fidelity of medical reasoning in large language models[J]. JAMA Network Open, 2025, 8(8): e2526021.
[2] Roy S, Khatua A, Ghoochani F, et al. Beyond accuracy: Investigating error types in GPT-4 responses to USMLE questions[C]//Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval. 2024: 1073-1082.
[3] Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation[J]. NPJ digital medicine, 2025, 8(1): 274.
[4] Freyer O, Wiest I C, Kather J N, et al. A future role for health applications of large language models depends on regulators enforcing safety standards[J]. The Lancet Digital Health, 2024, 6(9): e662-e672.
[5] de Hond A, Leeuwenberg T, Bartels R, et al. From text to treatment: the crucial role of validation for generative large language models in health care[J]. The Lancet Digital Health, 2024, 6(7): e441-e443.
[6] Amugongo L M, Mascheroni P, Brooks S, et al. Retrieval augmented generation for large language models in healthcare: A systematic review[J]. PLOS Digital Health, 2025, 4(6): e0000877.
[7] Li X, Cui M, Li J, et al. A hybrid medical text classification framework: Integrating attentive rule construction and neural network[J]. Neurocomputing, 2021, 443: 345-355.
[8] Prenosil G A, Weitzel T K, Bello S C, et al. Neuro-symbolic AI for auditable cognitive information extraction from medical reports[J]. Communications Medicine, 2025, 5(1): 491.
[9] Sheth A, Khandelwal V, Roy K, et al. NeuroSymbolic Knowledge-Grounded Planning and Reasoning in Artificial Intelligence Systems[J]. IEEE Intelligent Systems, 2025, 40(2): 27-34.
[10] Cui M, Li X, Qin P. Explainable Knowledge-Based Learning for Online Medical Question Answering[C]. International Conference on Knowledge Science, Engineering and Management. Singapore: Springer Nature Singapore, 2024: 294-304.
[11] Wang Y, Wang B, Mercer R, et al. Trustworthy medical question answering: An evaluation-centric survey[C]. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. 2025: 27477-27490.
[12] Min S, Krishna K, Lyu X, et al. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 12076-12100.
[13] Vedula N, Parthasarathy S. Face-keg: Fact checking explained using knowledge graphs[C]. Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 526-534