Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Medical HyperRAG: A Hypergraph-Enhanced Retrieval-Augmented Generation Framework for Clinical Question Answering

Fangfang Xu
Wuhan University of Science and Technology
Yifan Kuai
Wuhan University of Science and Technology
Chao Gao
Wuhan University of Science and Technology
Dan Luo
Lehigh University
Jinguang Gu
simon@ontoweb.wust.edu.cn

Wuhan University of Science and Technology

Article

Keywords:

Posted Date: April 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8497459/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Medical HyperRAG: A Hypergraph-Enhanced Retrieval-Augmented Generation Framework for Clinical Question Answering

Fangfang Xu[1,2,+], **Yifan Kuai**[1,2,+], **Chao Gao**[1,3,+], **Dan Luo**[4,+], **and Jinguang Gu**[1,2,*]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China
[3]Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam 1081 HV, The Netherlands
[4]Lehigh University, Bethlehem 18016, PA, USA
[*]simon@ontoweb.wust.edu.cn
[+]They played the same role in the process of writing the thesis and were co-first authors.

ABSTRACT

Clinical question answering plays a crucial role in intelligent healthcare, where reliable reasoning over patient-specific conditions and up-to-date medical evidence is essential for clinical decision support. However, large language models often suffer from hallucinations, outdated internal knowledge, and weak interpretability when facing multi-entity and multi-condition reasoning tasks. To address these limitations, this paper proposes **Medical HyperRAG**, a hypergraph-enhanced Retrieval-Augmented Generation framework that unifies structured and unstructured medical knowledge into a patient-centered reasoning paradigm. Specifically, heterogeneous information from electronic health records, clinical cases, and medical guidelines is organized through a ClinBridge HyperGraph, in which hyperedges capture high-order relations among symptoms, tests, diagnoses, and treatments. This design preserves logical constraints and temporal dependencies inherent to real clinical workflows. During inference, the framework performs structure-aware retrieval and context construction by verbalizing hyperedges into natural-language representations, enabling the LLM to generate answers grounded in both individualized and guideline-based knowledge. Experiments on real-world clinical datasets demonstrate that Medical HyperRAG substantially improves diagnostic and treatment accuracy compared with existing RAG baselines, achieving notable gains in factual precision, coherence, and semantic relevance. Ablation studies further confirm that the fusion of personalized case data and authoritative guidelines, along with phased retrieval and natural-language hyperedge representation, provides complementary benefits for reasoning interpretability and robustness. Overall, Medical HyperRAG establishes a unified, interpretable pathway for integrating high-order medical knowledge into large language models, thereby advancing trustworthy AI-assisted clinical question answering and decision support.

Introduction

Answering clinical questions is a critical challenge in modern healthcare, where accurate and timely medical decisions directly affect patient outcomes. Large Language Models (LLMs) have demonstrated remarkable progress in general-domain question answering tasks due to their strong semantic understanding and generation capabilities[1]. However, when applied to clinical question answering systems[2], these models face unique challenges that significantly limit their practical utility in medical settings.

Consider a typical complex clinical scenario: a 65-year-old diabetic patient with chronic kidney disease presents with chest pain and elevated troponin levels, while currently taking metformin and warfarin. The treating physician must determine the optimal diagnostic approach, accounting for drug interactions, imaging study contraindications related to kidney function, and the patient's cardiovascular risk profile. This scenario exemplifies multi-entity semantics, in which multiple interacting entities (patient demographics, comorbidities, medications, symptoms, and test results) must be considered simultaneously. Additionally, it involves complex logical reasoning with multiple constraints: the diagnostic approach must account for (1) contrast-induced nephropathy risk due to chronic kidney disease, (2) potential drug interactions between metformin and contrast agents, (3) bleeding risk assessment due to warfarin use, and (4) the need for urgent cardiac evaluation despite these constraints. Existing approaches struggle to capture these interconnected relationships and preserve the logical constraints during reasoning.

LLMs are prone to generating 'hallucinated' content when handling professional medical questions[3], such as fabricating causes of diseases, diagnostic methods, or drug usage information that lack clinical evidence. Furthermore, latency in training

data prevents them from capturing the latest drug approvals, clinical guidelines, and research progress in real time, leading to outdated internal knowledge. These limitations seriously hinder the practical application of these models in clinical scenarios, where high reliability and strong interpretability are required[4].

In the medical domain, Retrieval-Augmented Generation (RAG) has been applied to mitigate the limitations of large language models, including hallucinations and outdated knowledge. Conventional text-based RAG grounds responses in biomedical literature, clinical guidelines, or electronic health records, thereby improving factual precision but struggling with multi-entity reasoning and complex diagnostic rules, often leading to fragmented context[5]. Knowledge graph-based RAG (KG-RAG) addresses these challenges by integrating structured entity–relation knowledge, enabling more specific and interpretable clinical reasoning. Recent systems, such as MedRAG, demonstrate notable improvements by combining diagnostic knowledge graphs with EHR retrieval[6]. Nonetheless, existing knowledge graphs often struggle to incorporate newly emerging biomedical knowledge[7]. Moreover, simple triple representations remain insufficient for capturing multi-entity semantics and complex logical constraints, underscoring the need for more expressive modeling approaches in future medical KG-RAG systems.

Hypergraphs address these limitations by modeling n-ary relations via hyperedges, allowing simultaneous representation of multiple entities and attributes (e.g., $\{S1, S2, T > V, D\}$[8]. Unlike traditional graphs, in which edges connect exactly two nodes, hypergraphs use hyperedges that can connect any number of nodes, thereby enabling the representation of complex multi-entity relationships. In medical contexts, a hyperedge may simultaneously connect a patient node to symptom, medication, diagnostic test, and outcome nodes, thereby preserving the joint semantic constraints among these entities. This structure enhances knowledge completeness and provides stronger logical support for RAG systems in complex medical question answering.

Despite their expressiveness, hypergraphs face practical challenges. Patient records (individualized, real-time) and guidelines/cases (population-level) often reside in separate repositories, hindering integration. Conventional RAG frameworks struggle to leverage multi-dimensional hyperedge semantics during retrieval and generation, making it challenging to match patient-specific conditions or preserve multi-condition constraints.

We propose **Medical HyperRAG**, a patient-centered hypergraph-augmented retrieval framework. Heterogeneous knowledge fragments (records, guidelines, drugs, examinations) are organized into n-ary hyperedges anchored by patient nodes, enabling deep integration of population-level and individual knowledge. Structure-aware subgraph retrieval identifies hyperedges that closely match patient conditions, whereas generation explicitly encodes hyperedge constraints, preserving multi-condition logic during Large Language Model reasoning. This design enhances reliability, personalization, and interpretability of complex medical question answering.

The main contributions of this work are as follows:

1. The concept of an individual-centered medical hypergraph is proposed. Clinical cases, guidelines, drug information, and examination results are unified through n-ary hyperedges, providing structured support for personalized, multi-source integration and high-order semantic reasoning. Furthermore, a two-stage enhancement mechanism is introduced that improves the model's generation quality by leveraging collaborative strategies across different knowledge sources.

2. The Medical HyperRAG framework is designed to perform hyper-subgraph retrieval and evidence aggregation anchored on patient nodes. This approach achieves deep integration of population-level normative knowledge and individual cases. At the same time, multi-condition constraints are explicitly preserved during both retrieval and generation, thereby improving the accuracy and interpretability of complex clinical question answering.

3. The proposed method is constructed and evaluated on a clinical question-answering task. Results indicate that Hypergraph RAG outperforms existing RAG baselines in terms of answer accuracy, consistency, and individualization.

Related Work

The combination of Large Language Models and RAG has been widely adopted to enhance the factual consistency of medical question answering. ChatDoctor[9] augmented a general-purpose LLM with a medical knowledge base retrieval mechanism; MetaSepsisKnowHub[10] further introduced a sepsis phenotype-literature matching mechanism to provide personalized management recommendations. Ye et al.[11] and Xu et al.[5] utilized Electronic Medical Records (EMRs) as the retrieval source, while Ou et al. (2025)[12] integrated Electronic Health Records (EHRs) into a RAG framework to enhance discharge question answering performance. These studies collectively demonstrate that when retrieval granularity is refined to the individual patient course, the model not only yields more targeted diagnostic hypotheses but also generates treatment plans that better match the patient, thereby significantly improving clinical interpretability and practical value.

Despite the notable success of RAG, which shows potential in clinical decision support[13,14,15] and question-answering systems[16,17], traditional RAG still faces bottlenecks such as high semantic redundancy and weak entity correlation[18] when dealing with queries that require cross-document reasoning or complex semantic associations (e.g., multivariate medical relationships). Its mechanism, based on semantic similarity text retrieval, is insufficient for handling complex queries that require cross-document inference or involve details not explicitly present in the dataset.

To mitigate the semantic drift inherent in traditional RAG, researchers have embedded Knowledge Graphs into medical question-answering systems, aiming to explicitly model relationships among medical entities and strengthen reasoning[19,20,21]. A review by Genesis points out that KGs can serve as a "semantic bridge" to connect heterogeneous data and alleviate entity ambiguity arising from embedding matching[19]. KG-Med[20] directly injected triplets into the dialogue model to enhance entity tracking and reply consistency; its subsequent work[22] further utilized an enhanced graph structure to depict contextual evolution in multi-turn consultations. Soman et al.[21] converted KG triplets into structured prompts, significantly reinforcing the factual accuracy of LLMs. However, these methods remain constrained by the representational bottleneck of triplets: complex diagnostic rules are often decomposed into isolated fragments, lacking explicit "combinational condition" nodes. This forces the LLM to independently complete the logical chain across multiple triplets, usually leading to reasoning chain breaks. Furthermore, the completeness of entity and relation extraction directly determines the quality of the graph, and the absence of critical nodes or edges further exacerbates the context insufficiency caused by missing combinational conditions[23].

To more naturally represent complex relationships, the Hypergraph, a generalized graph structure supporting n-ary connections[8], has emerged. It repackages multivariate conditions that would otherwise be decomposed into self-contained reasoning units, allowing the LLM to acquire complete context in a single retrieval operation. By providing a more complete reasoning context than KGs, hypergraphs have demonstrated significant advantages in complex information modeling tasks across multiple domains in recent years. Luo et al. proposed HyperGraphRAG[24], which, for the first time, introduced hypergraphs to open-domain question answering, enhancing multi-hop reasoning capabilities. Existing studies have applied hypergraphs to model clinical pathways[25] and to analyze drug-drug interactions[26], thereby validating their superiority for representing complex clinical interactions. More recent developments such as Cog-RAG[27], Cross-Granularity Hypergraph RAG (HGRAG)[28], EBM-HyperRAG[29], and PRoH[30] have further extended this paradigm toward dual-hypergraph modeling, cross-granularity reasoning, evidence-based medical retrieval, and dynamic hypergraph planning, respectively—collectively highlighting a shift from static graph-based RAG toward cognitively inspired, multi-granular, and dynamically adaptive hypergraph retrieval frameworks.

Methods

Ethics statement

This study did not involve human participants, human experiments, or the use of human tissue samples, nor did it involve the direct collection of any human-related data; the dataset used herein was derived from doctors' reference to real patient cases in clinical practice and rewritten as virtual patient clinical data, without involving any doctor–patient private information.

Framework Overview

We propose a unified RAG framework that integrates both structured and unstructured medical knowledge to enhance the factual grounding and reasoning capability of LLMs for medical question answering and clinical decision support.

As illustrated in Fig. 1, the framework bridges two complementary knowledge sources—Personal Case data and Medical Guideline texts—by constructing two corresponding knowledge graphs and integrating them through a unified ClinBridge HyperGraph. This hierarchical design aligns instance-level knowledge from clinical cases with domain-level standards from medical guidelines.

Let the original medical dataset be denoted as:

$$\mathscr{D} = \{d_i\}_{i=1}^{N}$$

Where each sample d_i contains the following fields: *patient query* (q_i), *physician response* (a_i), *diagnosis label* (δ_i), *treatment recommendation* (α_i), and *contextual background* (c_i).

To distinguish between case-level and guideline-level knowledge, we partition $\mathscr{D}$ into two subsets:

- **Case Corpus** $\mathscr{D}_{\text{case}}$: consisting of real-world patient trajectories and doctor–patient interactions.
- **Guideline Corpus** $\mathscr{D}_{\text{guide}}$: consisting of authoritative medical guidelines and clinical protocols.

From these two sources, we construct two complementary knowledge graphs:

- **Case-centered Knowledge Graph (Case-centered KG)**:

  $$G_{\text{case}} = (V_{\text{case}}, R_{\text{case}})$$

  where V_{case} denotes clinical entities such as *symptoms*, *tests*, *diagnoses*, and *treatments*, and R_{case} denotes relations such as "leads to," "diagnosed as," or "treated by."

- **Guideline-based Knowledge Graph (Guideline-based KG)**:

$$G_{\text{guide}} = (V_{\text{guide}}, R_{\text{guide}})$$

where V_{guide} represents standardized medical concepts (e.g., *disease definitions*, *diagnostic criteria*, *treatment protocols*), and R_{guide} represents semantic or ontological relations such as "is a subtype of," "diagnosable by," or "recommended for."

To align the two graphs, we introduce a **semantic anchoring function**:

$$\phi : V_{\text{case}} \to V_{\text{guide}}$$

This mapping normalizes case-specific entities (e.g., "chest tightness") to standardized guideline terminology (e.g., "angina pectoris"), enabling coherent linkage between empirical and normative knowledge.

ClinBridge HyperGraph Construction

Medical data often exhibits firm semantic heterogeneity and non-uniform knowledge distribution across different sources, such as personal cases and standardized medical guidelines. To integrate these heterogeneous knowledge structures, we build on the two base graphs to construct a unified ClinBridge HyperGraph that bridges case-level and guideline-level knowledge.

Formally, the hypergraph is denoted as:

$$\mathscr{H} = (\mathscr{V}, \mathscr{E})$$

where $\mathscr{V} = V_{\text{case}} \cup V_{\text{guide}}$ represents the complete set of medical entities, and each hyperedge $e_i \in \mathscr{E}$ corresponds to one patient case, connecting all related entities from both the **Case-centered KG** and the **Guideline-based KG**:

$$e_i = \{v_1, v_2, \ldots, v_n\}, \quad v_j \in \mathscr{V}$$

Case Representation. Each patient case is represented as a temporally ordered set of medical triplets:

$$P_i = \{(s_1, r_1, o_1), (s_2, r_2, o_2), \ldots, (s_T, r_T, o_T)\}$$

where each triplet (s, r, o) denotes a relation among clinical entities such as *symptom*, *examination*, *diagnosis*, and *treatment*. This structure explicitly models the sequential reasoning process of real clinical decision-making.

Knowledge Alignment. To connect the two knowledge layers, we introduce a disease-centered semantic anchoring function:

$$\phi : V_{\text{case}} \to V_{\text{guide}}$$

This aligns free-form case entities with standardized guideline terminology. For example,

$$\phi(\text{"chest tightness"}) = \text{"angina pectoris"}$$

Anchors a case-specific entity to its corresponding guideline-defined concept, ensuring consistent medical semantics.

Higher-order Structure. Unlike conventional knowledge graphs that rely solely on binary relations (v_i, r, v_j), The ClinBridge HyperGraph introduces higher-order connections among multiple entities, thereby overcoming the expressive limitation of traditional pairwise graphs. Each hyperedge e_i thus represents a complete patient-centered reasoning unit that integrates dynamic case data with static guideline knowledge.

Temporal Preservation. While the hyperedges model multi-entity relations as unordered sets, We preserve temporal information by maintaining timestamped triplets within each Case-centered KG. During the subsequent verbalization step, these temporal dependencies are reconstructed through ordered natural language descriptions, allowing the LLM to recover the original diagnostic flow.

Example. For instance, a case path such as:

"cough and fever" $\to$ "chest X-ray" $\to$ "pneumonia" $\to$ "antibiotic treatment"

Is semantically anchored to the guideline entities:

"respiratory symptom", "diagnostic imaging", "infection management"

through the mapping ϕ.

Template-based Verbalization. To make the ClinBridge HyperGraph compatible with LLM input, each hyperedge e_i is verbalized into a natural language sentence using a template-based linearization function:

$$\text{Linearize}(e_i) = \text{“Patients with disease } D \text{ typically present with symptoms } S\text{, undergo } E\text{, and receive } T\text{.”}$$

These verbalized sentences constitute the structured textual corpus $\mathscr{C}_{\text{hyper}}$. Together with the original guideline corpus $\mathscr{C}_{\text{guide}}$, They form the unified natural language knowledge base:

$$\mathscr{C} = \mathscr{C}_{\text{hyper}} \cup \mathscr{C}_{\text{guide}}$$

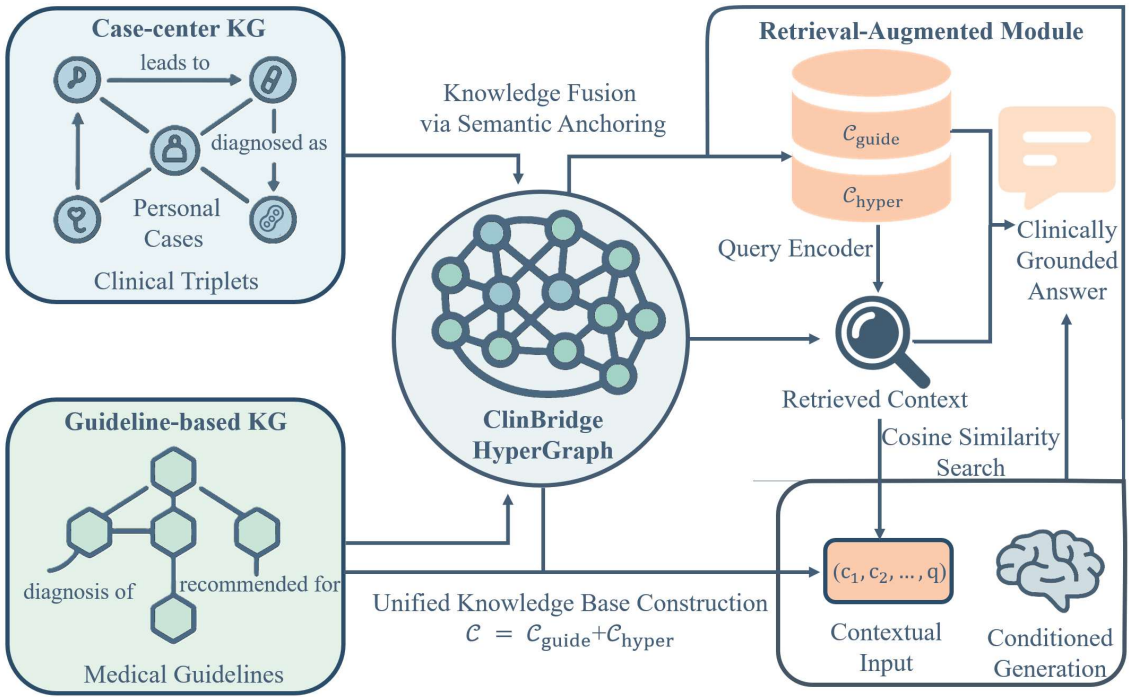


Figure 1. Overview of the proposed RAG framework. Personal Cases and Medical Guidelines are modeled as Case-centered KG and Guideline-based KG, respectively, and fused via the ClinBridge HyperGraph to enable unified retrieval and generation.

Retrieval-Augmented Module

Based on the unified knowledge base constructed in the previous section, we design a **Retrieval-Augmented Module** to enhance the reasoning and factual grounding of the language model. This module allows both structured knowledge from the **ClinBridge HyperGraph** and unstructured textual information from the **Medical Guideline** to be accessed within a single retrieval interface.

Query Encoding. Given a user query q (e.g., a new patient description or clinical question), We first encode it into a dense semantic vector using a pretrained biomedical encoder $f_\theta(\cdot)$:

$$\mathbf{q} = f_\theta(q)$$

Knowledge Retrieval. Retrieval is then performed **over** the unified natural language knowledge base:

$$\mathscr{C} = \mathscr{C}_{\text{guide}} \cup \mathscr{C}_{\text{hyper}}$$

where $\mathscr{C}_{\text{guide}}$ contains guideline paragraphs and $\mathscr{C}_{\text{hyper}}$ consists of **verbalized hyperedges** produced by the ClinBridge HyperGraph. Each element $c_i \in \mathscr{C}$ is embedded into the same semantic space:

$$\mathbf{c}_i = f_\theta(c_i)$$

And the top-k relevant entries are retrieved according to cosine similarity:

$$\text{Top-}k(q, \mathscr{C}) = \underset{c_i \in \mathscr{C}}{\arg\text{topk}}(\cos(\mathbf{q}, \mathbf{c}_i))$$

Definition of Entries. Here, the term "entries" explicitly refers to both (1) the verbalized case-level descriptions derived from the ClinBridge HyperGraph, and (2) the guideline-based paragraphs from authoritative sources. This ensures that the retrieval space jointly encodes empirical and normative medical knowledge.

Context Construction. The retrieved passages are concatenated with the user query to form the contextual input for the LLM:

$$\text{Input}_{\text{LLM}} = [c_1; c_2; \ldots; c_k; q]$$

This unified prompt integrates normative guidelines with case-anchored reasoning, providing the LLM with both factual reliability and contextual relevance.

Model Generation. Conditioned on the augmented input, the LLM generates an output sequence:

$$\hat{y} = \text{LLM}(\text{Input}_{\text{LLM}})$$

The generation process thus benefits from external medical grounding while retaining. The LLM's natural language fluency.

Advantages. The Retrieval-Augmented Module achieves:

- **Representation Fusion**: bridges structured (graph-derived) and unstructured (textual) knowledge;
- **Unified Indexing**: embeds all medical knowledge into a shared dense vector space;
- **Shared Retrieval Interface**: enables a single query to access heterogeneous medical data sources.

This design allows the system to leverage both patient-centered reasoning and standardized medical guidance, improving factual accuracy, adaptability, and clinical interpretability of the generated answers.

Methodological Comparison with Existing Hypergraph-RAG Frameworks To further clarify the conceptual distinction and methodological advantages of our proposed framework, we compare Medical HyperRAG with several representative hypergraph-based Retrieval-Augmented Generation methods, including HyperGraphRAG[24], Hyper-RAG[31], Cog-RAG[27], and Cross-Granularity HGRAG[29]. Table1 summarizes the conceptual characteristics across these methods in terms of their knowledge structure, integration strategy, retrieval mechanism, and primary innovation focus.

Method	Domain	Knowledge Structure	Knowledge Source	Key Characteristics
ClinBridge (Ours)	Medical QA / Clinical Decision Support	Dual Knowledge Graphs + Unified ClinBridge HyperGraph	Case-centered KG + Medical Guidelines	Integrates empirical case data with normative guidelines; semantic anchoring for entity alignment; preserves temporal reasoning chains.
HyperGraphRAG[24]	Multi-domain QA	Hypergraph (n-ary relation representation)	Text corpora / KBs	Enhances knowledge completeness by modeling multi-entity relations.
Hyper-RAG[31]	Biomedical QA	Hypergraph (low- and high-order associations)	Textual medical data	Mitigates hallucinations by capturing complex relational dependencies.
Cog-RAG[27]	General QA	Dual Hypergraphs (Theme-level + Entity-level)	Textual corpora	Cognitive-inspired two-stage retrieval with theme–entity alignment.
HGRAG[29]	Multi-hop QA	Cross-granularity Hypergraph (Entity–Paragraph)	Paragraph-level texts	Combines fine-grained entities with coarse-grained contextual links.

Table 1. Conceptual comparison between ClinBridge and existing Hypergraph-RAG frameworks.

As shown in Table 1, prior Hypergraph-RAG studies have primarily focused on enhancing multi-entity relation modeling and semantic retrieval by introducing high-order graph structures[24,27,29,31]. While these methods improve relational expressiveness and retrieval accuracy, they generally operate on homogeneous textual data sources and do not explicitly align heterogeneous domain knowledge.

In contrast, **ClinBridge** introduces a hierarchical knowledge integration paradigm that unifies *case-level empirical knowledge* with *guideline-level normative knowledge* through the ClinBridge HyperGraph. The semantic anchoring function $\phi : V_{\text{case}} \rightarrow V_{\text{guide}}$ normalizes free-form clinical expressions to standardized medical terminology, enabling consistent cross-source semantics. Moreover, temporal preservation within the case graph allows the framework to reconstruct the clinical reasoning process, which is then verbalized via template-based linearization for LLM consumption. This design enables a single retrieval interface for both structured and unstructured medical knowledge, thereby enhancing factual grounding, interpretability, and transparency in reasoning in medical question answering and clinical decision support.

Experimental Design and Analysis

Evaluation Metrics

We evaluate **Medical HyperRAG** on medical question-answering (QA) tasks to assess its ability to generate clinically accurate and reliable responses. The evaluation focuses on the alignment between model-generated outputs and expert physician answers derived from real clinical cases and guideline references.

Four primary dimensions are considered: (1) diagnostic accuracy, (2) treatment recommendation accuracy, (3) textual consistency, and (4) semantic similarity. Specifically, we adopt $\text{F1}_{\text{disease}}$, $\text{F1}_{\text{advice}}$, Average F1 Score, ROUGE-L, and BERTScore as core metrics to comprehensively assess factual precision, linguistic fluency, and semantic fidelity. Formal definitions and metric descriptions are provided in **Appendix A**.

Combining these multi-dimensional evaluation criteria provides a comprehensive view of model performance across various medical QA scenarios. Particular emphasis is placed on $\text{F1}_{\text{disease}}$ and $\text{F1}_{\text{advice}}$, which directly reflect the model's capability to produce clinically grounded diagnostic and therapeutic reasoning. The results collectively demonstrate the effectiveness of the proposed **Medical HyperRAG** framework, in which the ClinBridge HyperGraph supports structure-aware retrieval and multi-condition–constrained generation, thereby enhancing the reliability and interpretability of reasoning.

Experimental Setup

Experiments are conducted using the RJUA-QA dataset[32,33], which includes over 2,000 expert-authored question–answer pairs covering real-world urological clinical scenarios. Each instance consists of physician-authored questions, corresponding expert responses, diagnoses, treatment recommendations, and context passages extracted from authoritative guidelines.

This dataset is publicly available on OpenKG.CN via the official project page: https://github.com/alipay/RJU_Ant_QA. For our experiments, we adopted Version V1, which comprises the training, validation, and test sets as defined by the dataset constructors.

We compare five representative lightweight large language models—LLaMA3.2[34], LLaVA[35], Mistral[36], Neural-Chat[37], and Qwen2.5-7B[38]—under a unified RAG pipeline. All models use consistent prompts, vector-embedding backbones, and retrieval databases to ensure fairness. Details of the implementation setup, including embedding models, prompt templates, and hyperparameter configurations, are described in **Appendix B**.

Comprehensive evaluation results of all baseline LLMs are provided in **Appendix C**. The comparison focuses on the fundamental reasoning and generation capabilities of each model under a unified RAG framework. Among the evaluated models, **Qwen2.5-7B** demonstrates the strongest overall performance in diagnostic and treatment reasoning metrics. Therefore, all subsequent experiments and ablation studies in this work are conducted using Qwen2.5-7B as the backbone generator.

Comparison Experiments

Comparative Analysis of Different RAG Methodologies To systematically evaluate the performance of different RAG strategies for medical question answering, we fix **Qwen2.5-7B** as the backbone Large Language Model and conduct comparative experiments across six representative RAG paradigms, each employing distinct retrieval or reasoning mechanisms:

1. **OpenRAG**[39]: an end-to-end optimized RAG framework where the retriever and generator are jointly trained, enabling in-context retrieval learning and direct optimization for generation quality.

2. **Graph-CoT**[40]: a graph-based chain-of-thought reasoning framework that integrates retrieval and multi-hop reasoning to enhance factual inference.

3. **TRACE**[41]: a knowledge-grounded reasoning framework that constructs triplet-based retrieval chains, where extracted (*question*, *answer*, *context*) triplets serve as the fundamental retrieval units for downstream generation.

4. **HyperGraphRAG**[24]: an open-domain hypergraph retrieval framework that constructs high-order relational structures but lacks domain-specific medical specialization.

5. **Baseline (Qwen2.5-7B)**: the plain generation model without retrieval augmentation.

6. **Medical HyperRAG (Ours)**: our framework that instantiates each patient's historical diagnostic information—including *symptoms*, *examinations*, *diagnoses*, and *treatments*—as a hyperedge in the **ClinBridge HyperGraph**, retrieving semantically similar hyperedges to form structured, context-aware input.

We evaluate all paradigms using five metrics: $F1_{disease}$, $F1_{advice}$,Average F1 Score, ROUGE-L, and BERTScore. The experimental results on the RJUA-QA dataset are summarized in Table 2.

Method	$F1_{disease}$	$F1_{advice}$	Average F1 Score	ROUGE-L	BERTScore
Qwen2.5-7B (Baseline)	62.68%	29.78%	51.71%	13.14%	32.14%
OpenRAG[39]	68.51%	40.54%	59.18%	11.67%	32.54%
Graph-CoT[40]	68.56%	46.91%	61.35%	16.64%	31.50%
TRACE[41]	66.71%	41.57%	58.33%	14.04%	35.23%
HyperGraphRAG[24]	64.59%	35.52%	54.90%	13.46%	34.10%
Medical HyperRAG (Ours)	**73.19%**	**47.97%**	**64.79%**	**24.29%**	**37.19%**

Table 2. Comparison of different RAG paradigms on the RJUA-QA dataset using Qwen2.5-7B as the base model.

All RAG paradigms significantly outperform the baseline Qwen2.5-7B model, confirming the effectiveness of retrieval augmentation for medical question answering. The proposed Medical HyperRAG achieves the highest overall performance, with $F1_{disease} = 73.19\%$ and $F1_{advice} = 47.97\%$, outperforming OpenRAG by $+4.68\%$ and $+7.43\%$, respectively, and surpassing TRACE by $+6.48\%$ and $+6.40\%$.

This advantage arises from its explicit modeling of *high-order clinical dependencies* through patient-instantiated hyperedges, which capture complex interrelations between symptoms, diagnoses, tests, and treatments. For example, in an RJUA-QA query describing "a 58-year-old female presenting with left flank pain and fever (Tmax 38.8 °C), CT showing left ureteral upper-segment stone with hydronephrosis and perinephric exudation, blood tests revealing WBC 20.6 × 10^9/L, CRP 310 mg/L, and PCT 26.4 ng/mL", Medical HyperRAG accurately retrieves and aligns hyperedges linking left ureteral calculus, acute pyelonephritis, PCT > 10 (indicating urosepsis), and ureteral stent insertion with broad-spectrum antibiotic therapy. In contrast, TRACE retrieves fragmented information—such as "hydronephrosis" and "antibiotic treatment"—without capturing the temporal-causal dependency between obstruction and infection, leading to less accurate treatment recommendations. In contrast, TRACE retrieved semantically relevant but fragmented snippets (e.g., cough–fever pairings) without coherent temporal or causal structure, resulting in lower accuracy in treatment recommendations.

Although the ROUGE-L (24.29%) and BERTScore (37.19%) of Medical HyperRAG are only marginally higher than those of TRACE (35.23%), the difference is below 1%, demonstrating that structural retrieval enhances factual accuracy without sacrificing language fluency.

In comparison, HyperGraphRAG, which applies an open-domain hypergraph strategy without domain adaptation, introduces redundant entities during retrieval, resulting in diluted critical information and reduced clinical interpretability ($F1_{disease} = 64.59\%$, $F1_{advice} = 35.52\%$). Meanwhile, Graph-CoT achieves moderate improvements in reasoning precision but lacks the domain grounding necessary for accurate medical treatment prediction.

Overall, these results validate that grounding LLM reasoning in the ClinBridge HyperGraph—which unifies structured and unstructured medical knowledge—provides superior interpretability and reliability in clinical decision support.

Effectiveness Verification To verify the generalizability and effectiveness of the proposed Hypergraph-RAG (HyperRAG) framework across different LLMs, we conducted comparative experiments on several popular open-source models, including LLaMA3.2, Mistral, Neural-Chat, LLaVA, Qwen2.5, and GLM4-9B. Figure 2 presents the performance comparison results of these models with and without the Medical HyperRAG enhancement across five core evaluation metrics: $F1_{disease}$, $F1_{advice}$, Average F1 Score, ROUGE-L, and BERTScore.

Ablation Experiments

To further investigate the contribution of the hypergraph structure in RAG and the impact of different knowledge components, a series of ablation experiments was designed to evaluate how various combinations of knowledge components affect model performance

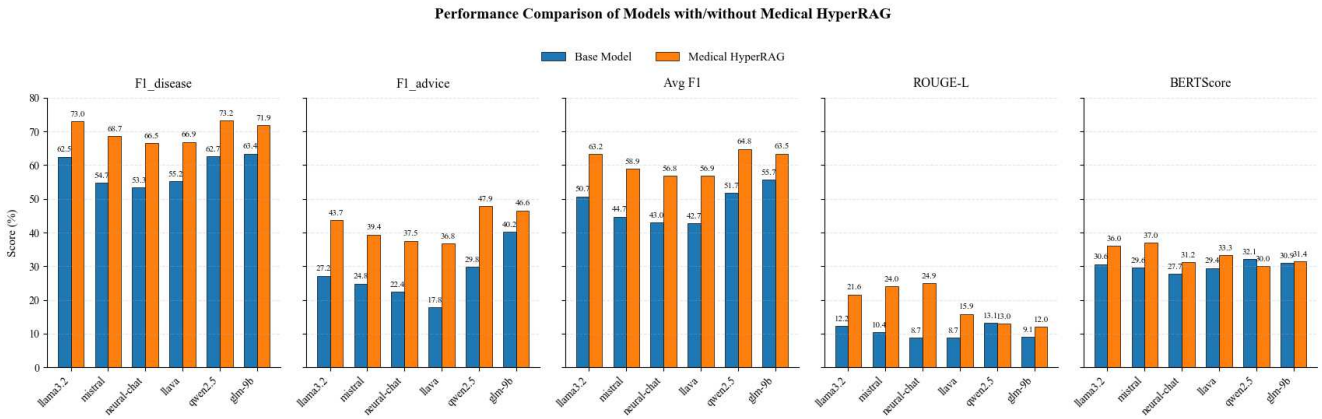


Figure 2. Performance comparison of different open-source LLMs with and without Medical HyperRAG enhancement across five core metrics ($F1_{\text{disease}}$, $F1_{\text{advice}}$, Average F1 Score, ROUGE-L, BERTScore) on the RJUA-QA dataset

Ablation on Different Knowledge Component Configurations We design three knowledge-enhanced configurations and conduct ablation studies using the Qwen 2.5 model. The specific configurations are as follows:

- **Scheme I (With Personal Case):** This configuration uses only the patient's historical diagnostic information as knowledge input to construct a personalized QA context. This study aims to evaluate the independent contribution of individual case data to answer generation.
- **Scheme II (With Medical Guideline):** This configuration relies solely on medical guideline texts as knowledge input, introducing external expert knowledge to support QA. It examines the effectiveness of general medical knowledge in the absence of individual case context.
- **Scheme III (With Personal Case & Medical Guideline):** This configuration jointly incorporates patient-specific information and medical guideline texts. By integrating personalized context with general medical knowledge, it aims to enhance the coverage and accuracy of model-generated responses.

The experimental results of these configurations on the medical QA task are summarized in Table 3.

Component	$\mathbf{F1_{disease}}$	$\mathbf{F1_{advice}}$	Average F1 Score	ROUGE-L	BERTScore
Scheme I	69.08%	45.36%	61.17%	30.72%	40.21%
Scheme II	66.31%	41.51%	58.04%	11.88%	33.40%
Scheme III	73.19%	47.97%	64.79%	24.29%	37.19%

Table 3. Ablation Study on Knowledge Component Configurations

Based on the experimental results in Table 3, the ablation study on different knowledge component configurations demonstrates the distinct and complementary roles of personalized and general medical knowledge in enhancing medical question answering (QA) performance.

Scheme II, which relies solely on general medical guidelines, yields the weakest results across all metrics, with $\mathbf{F1_{disease}}$ of 66.31% and $\mathbf{F1_{advice}}$ of 41.51%. The low ROUGE-L (11.88%) and BERTScore (33.40%) indicate that the absence of patient-specific context substantially limits semantic coherence and factual alignment, leading to less context-aware and less precise recommendations. These findings suggest that while general guidelines provide authoritative domain knowledge, they lack the personalization needed for case-specific reasoning.

In contrast, **Scheme I**, which uses only individual patient case information, achieves marked improvements in both disease recognition (**69.08**%) and treatment recommendation (**45.36**%), as well as stronger linguistic consistency (ROUGE-L 30.72%, BERTScore 40.21%). This demonstrates that patient-specific diagnostic history provides crucial contextual grounding, enabling more accurate and tailored medical advice generation even without external expert knowledge.

Most notably, **Scheme III**—which integrates both personal case data and medical guideline texts—achieves the best overall performance, with $\mathbf{F1_{disease}}$ reaching 73.19%, $\mathbf{F1_{advice}}$ 47.97%, and the highest Average F1 of 64.79%. This integrated

configuration effectively combines the contextual depth of personalized information with the systematic reliability of expert medical knowledge, yielding more complete and consistent reasoning outcomes. The moderate yet meaningful gains in all evaluation metrics indicate a synergistic interaction between the two knowledge sources: while case data improves contextual relevance, medical guidelines enhance factual correctness and generalizability.

In summary, the results confirm that relying on either personalized or general medical knowledge alone is suboptimal. The **fusion of patient-specific and guideline-based knowledge** achieves the most balanced and robust performance, providing a solid foundation for high-accuracy, interpretable, and clinically reliable QA systems.

Ablation Study on Retrieval Strategies We investigate how different collaborative strategies between medical **Guidelines (G)** and clinical **Case Records (C)** within RAG influence medical question-answering performance. This ablation study compares two distinct knowledge retrieval strategies.

1. **Scheme I (C+G Joint Retrieval):** The model performs knowledge retrieval simultaneously based on both clinical case records and medical guidelines, utilizing the combined retrieval results for generating disease identification, treatment recommendations, and the final answer. This scheme provides the model with comprehensive context, integrating authoritative medical knowledge with individualized patient information.

2. **Scheme II (Phased Retrieval):** The model first uses joint retrieval based on case records and guidelines (C+G) to generate the disease and recommendation components. Subsequently, during the final answer generation phase, the model performs a secondary retrieval and enhanced generation using *only* the clinical case records (C). This strategy aims to leverage guideline knowledge to improve the accuracy of structured judgments (disease and recommendations), while focusing on patient-specific information during answer generation to enhance the relevance and semantic consistency of the generated text.

Table 4 presents the comparative results of the two retrieval strategies.

Component	$F1_{disease}$	$F1_{advice}$	Average F1 Score	ROUGE-L	BERTScore
Scheme I (C+G Joint)	73.25%	47.89%	64.79%	13.01%	30.00%
Scheme II (Phased)	73.19%	47.97%	64.79%	24.29%	37.19%

Table 4. Ablation Study on Retrieval Strategies

The results from the retrieval strategy ablation experiment indicate that the configuration of knowledge sources during the retrieval phase significantly affects the model's overall performance. Regarding disease and recommendation identification, the Average F1 scores for Scheme I and Scheme II are nearly identical (73.25% vs. 73.19% and 47.89% vs. 47.97%), demonstrating that both schemes effectively integrate medical guidelines and clinical records to achieve high accuracy in diagnosis and advice. However, in terms of generation quality, Scheme II shows a notable improvement over Scheme I in both ROUGE-L and BERTScore. This demonstrates that the phased retrieval strategy, while maintaining sufficient knowledge utilization, enhances the semantic coherence and textual relevance of the generated answers.

From a methodological perspective, this finding reveals a crucial complementary mechanism between joint and phased retrieval. Direct joint retrieval (Scheme I) provides a more comprehensive knowledge base, aiding accurate disease identification and recommendations. However, during the free-text generation stage, the inclusion of authoritative guideline information may introduce abstract or overly general descriptions, causing the answer to drift from the specific clinical context. Scheme II mitigates this by relying only on clinical records for secondary retrieval during the final generation stage, enabling the content to be more focused on patient characteristics and the therapeutic process. This improves the specificity and naturalness of the generated text without sacrificing accuracy. In conclusion, phased retrieval (Scheme II) achieves a vital balance between knowledge breadth and semantic precision in medical QA. This strategy retains the structured guidance of medical guidelines while preventing their interference with individualized expression during answer generation, offering a generalizable knowledge fusion pathway for future medical RAG system designs.

Ablation experiments on representation formats To further explore the optimal representation of structured medical knowledge within LLMs, this study designs an ablation study focusing on representation formats. It compares the practical effects of two hypergraph input strategies on the medical QA task. **Scheme I** takes the direct input of structured hypergraph information, where question-answer pairs serve as hyperedges and diagnostic/treatment information as nodes, with the hypergraph constructed and supplied to the model in its structured form. In contrast, **Scheme II** converts the hypergraph into a natural-language expression while retaining the underlying structural information, which is then used as input to the model for QA generation.

Component	$F1_{disease}$	$F1_{advice}$	Average F1 Score	ROUGE-L	BERTScore
Scheme I	62.99%	37.94%	54.64%	13.03%	20.62%
Scheme II	73.19%	47.97%	64.79%	24.29%	37.19%

Table 5. Ablation Study on Representation Formats

Table 5 displays the comparative results of the representation format ablation experiment.

Based on the results in Table 5, the ablation study on knowledge representation formats clearly demonstrates that the input form of knowledge exerts a substantial influence on model performance. **Scheme II**, which converts the hypergraph into natural-language descriptions, consistently outperforms **Scheme I**, which directly uses the structured hypergraph, across all evaluation metrics. Specifically, $F1_{disease}$ and $F1_{advice}$ increase from 62.99% and 37.94% in Scheme I to 73.19% and 47.97% in Scheme II, respectively, while the **Average F1** improves by more than 10 points (from 54.64% to 64.79%). This indicates that even when the underlying medical knowledge remains identical, expressing it in natural language enables the model to integrate clinical information more effectively and reason over it, yielding more accurate disease recognition and treatment recommendation outcomes.

From the perspective of generation quality, Scheme II also achieves higher ROUGE-L (24.29% vs. 13.03%) and BERTScore (37.19% vs. 20.62%), reflecting superior lexical alignment, semantic consistency, and contextual fluency in generated responses. These results suggest that natural-language representation mitigates information loss caused by the model's limited ability to directly parse structured inputs, thereby enhancing the overall expressiveness and coherence of the generated answers.

Methodologically, this comparison reveals a clear gap in representation between structured medical knowledge and the pre-training objectives of large language models. While structured hypergraphs explicitly encode relational semantics, current LLMs are not inherently optimized for processing complex graph structures. By contrast, natural-language representations preserve the essential semantics of medical entities and relations while aligning with the model's linguistic reasoning mechanisms. Consequently, this alignment substantially improves the efficiency of knowledge utilization and downstream QA performance.

In summary, the experiment underscores that the **form of knowledge representation is a decisive factor** in effective knowledge integration. Given the current limitations of language models for structured reasoning, converting structured medical knowledge into coherent natural-language descriptions is a practical and effective strategy. This finding provides empirical evidence supporting the design of adaptive, linguistically aligned knowledge representations for high-stakes applications, such as medical question answering.

Conclusion and Future Work

This study presents Medical HyperRAG, a unified Retrieval-Augmented Generation framework that integrates structured and unstructured medical knowledge through a novel ClinBridge HyperGraph. By bridging real-world clinical cases with standardized medical guidelines, the framework enables language models to perform more accurate, interpretable, and evidence-grounded reasoning in medical question answering and decision support.

The proposed ClinBridge HyperGraph aligns case-level entities with guideline-level terminology via a semantic anchoring function, capturing higher-order and temporally coherent relations among symptoms, diagnoses, and treatments. This structure transforms heterogeneous medical knowledge into a unified retrieval space, enhancing factual grounding while preserving the reasoning flow of clinical practice.

Experiments on the RJUA-QA dataset show that Medical HyperRAG outperforms existing RAG paradigms—such as OpenRAG, TRACE, and HyperGraphRAG—achieving the highest diagnostic and treatment accuracies. Ablation studies further demonstrate that (1) jointly incorporating case and guideline knowledge yields the best overall performance; (2) a phased retrieval strategy improves coherence during generation; and (3) verbalized hypergraph representation aligns best with LLM reasoning.

In summary, Medical HyperRAG provides a practical approach to integrating heterogeneous clinical knowledge into large language models, thereby improving the reliability and interpretability of medical reasoning. Future work will extend the framework to multi-modal data, dynamically evolving hypergraphs, and uncertainty-aware reasoning to further advance trustworthy AI-assisted clinical decision support.

Data Availability

The dataset used in this study, RJUA-QA, is an open-source medical question-answering dataset jointly developed by Renji Hospital and Ant Group. The dataset is officially described and maintained on its GitHub project page: https://github.com/alipay/RJU_Ant_QA. The data files are hosted on the OpenKG.CN platform (http://data.openkg.

cn/dataset/rjua-qadatasets). For direct download, please use the following link: Download RJUA-QA V1. Please note that this server is located in mainland China and may not be accessible from some regions due to network restrictions. As an alternative download source, the dataset is also available from the HyperAI platform (https://hyper.ai/cn/datasets/28970), where a copy of the same dataset can be found. To ensure accessibility for all readers, the data supporting the findings of this study are also available from the corresponding author upon reasonable request. For our experiments, we used Version V1, including the training, validation, and test sets as defined by the dataset constructors. When citing or using this dataset, please refer to the official citations: Lyu, S. et al. (2023). RJUA-QA: A comprehensive QA dataset for urology. arXiv:2312.09785. RJU, AntGroup. (2023). RJUA-QA Datasets: QA reasoning dataset constructed from specialized urological medical knowledge. The dataset was derived from doctors' reference to real patient cases in clinical practice and rewritten as virtual patient clinical data, without involving any doctor–patient private information.

References

1. Zhao, W. X. e. a. A survey of large language models. *Prepr. at arXiv* (2023). ArXiv:2303.18223.
2. Yan, Y., Guo, X.-R., Wang, T.-J., Rao, Q. & Wang, K.-J. Survey on question answering system research. *Comput. Syst. Appl.* **32**, 1–18 (2023).
3. Guo, H. e. a. Research on a massively large artificial intelligence model and its application in medicine. *Sci. China Life Sci.* **54**, 482–486 (2024).
4. Bunnell, D. J., Bondy, M. J., Fromtling, L. M., Ludeman, E. & Gourab, K. Bridging ai and healthcare: a scoping review of retrieval-augmented generation—ethics, bias, transparency, improvements, and applications. *Prepr. at medRxiv* (2025).
5. Lewis, P. e. a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Adv. Neural Inf. Process. Syst.*, vol. 33, 9459–9474 (2020).
6. Zhang, D. e. a. Cmedragbot: a chinese medical chatbot based on graph rag and large language models. *Interdiscip. Sci. Comput. Life Sci.* **17**, 844–859 (2025).
7. Zhou, D. e. a. Evaluating knowledge graph based retrieval augmented generation methods under knowledge incompleteness. *Prepr. at alphaxiv* (2025).
8. Wei, S. e. a. Graph learning and its advancements on large language models: a holistic survey. *Prepr. at arXiv* (2022). ArXiv:2212.08966.
9. Li, Y. *et al.* Chatdoctor: a medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* **15**, e40895 (2023).
10. Zhang, C. e. a. A knowledge-enhanced platform (metasepsisknowhub) for retrieval augmented generation–based sepsis heterogeneity and personalized management: development study. *J. Med. Internet Res.* **27**, e67201 (2025).
11. Ye, C. Exploring a learning-to-rank approach to enhance the retrieval augmented generation (rag)-based electronic medical records search engines. *Informatics Heal.* (2024).
12. Ou, J. e. a. Experience retrieval-augmentation with electronic health records enables accurate discharge qa. *Prepr. at alphaxiv* (2025).
13. Zakka, C. e. a. Almanac: retrieval-augmented language models for clinical medicine. *Prepr. at arXiv* (2024).
14. Shi, W. e. a. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proc. ACM Conf. Bioinform. Comput. Biol. Health Inform. (BCB 2023)*, 1–10 (2023).
15. Thompson, W. E. e. a. Large language models with retrieval-augmented generation for zero-shot disease phenotyping. *Prepr. at arXiv* (2023). ArXiv:2312.06457.
16. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking retrieval-augmented generation for medicine. In *Proc. Annu. Meet. Assoc. Comput. Linguist.* (2024).
17. Jeong, M., Sohn, J., Sung, M. & Kang, J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics* **40**, i119–i129 (2024).
18. Li, H. e. a. A survey on retrieval-augmented text generation. *Prepr. at arXiv* (2022). ArXiv:2202.01110.
19. Genesis, J. & Keane, F. Integrating knowledge retrieval with generation: a comprehensive survey of rag models in nlp. *Preprints* (2025). Doi:10.20944/preprints202504.0351.v1.
20. Varshney, D., Zafar, A., Behera, N. K. & Ekbal, A. Knowledge graph assisted end-to-end medical dialog generation. *Artif. Intell. Med.* **139** (2023).

21. Soman, K. e. a. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics* **40** (2024).

22. Varshney, D. e. a. Knowledge grounded medical dialogue generation using augmented graphs. *Sci. Rep.* **13** (2023).

23. Verma, G. e. a. Hybridcontextqa: a hybrid approach for complex question answering using knowledge graph construction and context retrieval with llms. *Work. on Knowl. Base Constr. from Pre-Trained Lang. Model.* (2024).

24. Luo, H. e. a. Hypergraphrag: retrieval-augmented generation via hypergraph-structured knowledge representation. In *Proc. Conf. Neural Inf. Process. Syst.* (2025).

25. Li, M. e. a. Hyperbolic hypergraph neural networks for multi-relational knowledge hypergraph representation. *Prepr. at arXiv* (2024). ArXiv:2412.12158.

26. Chen, X., Ye, X., Mohanty, M. & Manoharan, S. Detecting offensive posts on social media. In *Proc. Int. Conf. Electr. Comput. Energy Technol. (ICECET)*, 1–6, DOI: 10.1109/ICECET58911.2023.10389225 (2023).

27. Hu, H. e. a. Cog-rag: cognitive-inspired dual-hypergraph with theme alignment retrieval-augmented generation. *Prepr. at arXiv* (2025). ArXiv:2511.13201.

28. Wang, C. e. a. Cross-granularity hypergraph retrieval-augmented generation for multi-hop question answering. *Prepr. at arXiv* (2025). ArXiv:2508.11247.

29. Dou, C. e. a. Enhancing llm generation with knowledge hypergraph for evidence-based medicine. *Prepr. at alphaxiv* (2025).

30. Zai, X. e. a. Proh: dynamic planning and reasoning over knowledge hypergraphs for retrieval-augmented generation. *Prepr. at arXiv* (2025). ArXiv:2510.12434.

31. Feng, Y. e. a. Hyper-rag: combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *Prepr. at alphaxiv* (2025).

32. Lyu, S. e. a. Rjua-qa: a comprehensive qa dataset for urology (2023). ArXiv:2312.09785.

33. RJU, AntGroup. Rjua-qa datasets: Qa reasoning dataset constructed from specialized urological medical knowledge. Available at http://data.openkg.cn/dataset/rjua-qadatasets (2023).

34. Touvron, H. e. a. Llama: open and efficient foundation language models. *Prepr. at arXiv* (2023). ArXiv:2302.13971.

35. Liu, H., Li, C., Li, Y. & Lee, Y. J. Visual instruction tuning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 26705–26715 (2024).

36. Jiang, A. e. a. Mistral 7b. *Prepr. at arXiv* (2023). ArXiv:2310.06825.

37. Intel. neural-chat-7b-v3-1. Available at https://huggingface.co/Intel/neural-chat-7b-v3-1 (2023). Accessed 2025-12-28.

38. Bai, J. e. a. Qwen2 technical report. *Prepr. at arXiv* (2024). ArXiv:2407.10671.

39. Zhou, J. & Chen, L. Openrag: optimizing rag end-to-end via in-context retrieval learning. *Prepr. at arXiv* (2025). ArXiv:2503.08398.

40. Jin, B. e. a. Graph chain-of-thought: augmenting large language models by reasoning on graphs. *Prepr. at arXiv* (2024). ArXiv:2404.07103.

41. Fang, J., Meng, Z. & MacDonald, C. Trace the evidence: constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings Assoc. Comput. Linguist.: EMNLP 2024*, 8472–8494 (2024).

42. Luo, Y. e. a. Oneke: a dockerized schema-guided llm agent-based knowledge extraction system. *World Wide Web* (2025).

Funding

This work was supported by the following projects: the National Natural Science Foundation of China (Grant No. 62576254), the Key Research and Development Program of Wuhan (Grant No. 2022012202015070), and the Natural Science Foundation of Hubei Province Educational Committee(ID: B2019009).

Author contributions statement

Y.K. and F.X. conceived the core framework of Medical HyperRAG and designed the overall experimental plan. C.G. led the development of the ClinBridge HyperGraph architecture, while F.X. implemented the retrieval and alignment modules and performed the experiments and data analysis. Y.K. contributed to model optimization, experimental validation, and assisted in

data interpretation. D.L. supported the implementation and conducted supplementary analyses. J.G. supervised the project, provided methodological guidance, and contributed to the interpretation of the results. All authors contributed to the writing, revision, and final approval of the manuscript. F.X., Y.K., C.G., and D.L. contributed equally to this work and share first authorship.

Additional information

Competing interests: The authors declare no competing interests.

Appendix A. Evaluation Metrics Details

This appendix provides the formal definitions and characteristics of the evaluation metrics introduced in Section . Five key metrics are adopted to comprehensively evaluate diagnostic accuracy, treatment recommendation quality, textual consistency, and semantic similarity.

A.1 F1 Scores for Disease and Advice

The F1 score evaluates the balance between precision (P) and recall (R):

$$F1 = \frac{2 \times P \times R}{P + R} \tag{1}$$

We compute two task-specific metrics, $\mathbf{F1}_{\mathbf{disease}}$ and $\mathbf{F1}_{\mathbf{advice}}$, which assess the model's diagnostic and therapeutic reasoning accuracy, respectively. Higher values indicate better performance in identifying correct diseases and generating appropriate treatment suggestions.

A.2 Average F1 Score

To provide an overall performance indicator across multiple task objectives, we define a weighted average:

$$F1_{\text{avg}} = \alpha \times F1_{\text{disease}} + (1 - \alpha) \times F1_{\text{treatment}} \tag{2}$$

where α controls the relative importance of diagnostic versus treatment accuracy.

A.3 ROUGE-L

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence) measures global textual overlap:

$$ROUGE\text{-}L = \frac{(1 + \beta^2) \times P \times R}{R + \beta^2 \times P} \tag{3}$$

Here, P and R are precision and recall derived from the longest common subsequence (LCS) between generated text S_1 and reference text S_2. This metric captures structural alignment and fluency in long-form generation tasks.

A.4 BERTScore

BERTScore leverages contextual embeddings from a pretrained BERT model to measure semantic similarity between generated and reference texts. It effectively captures deeper semantic alignment that surface-level matching metrics (e.g., ROUGE) may miss, making it suitable for complex medical QA contexts.

A.5 Summary Table

For clarity, we compile all discussed metrics into a single summary table. Table 6 compares their analytical focus, distinctive characteristics, and practical use cases, serving as a quick reference for evaluating different aspects of model performance.

Appendix B. Experimental Setup

B.1 Hardware and Environment

All experiments were conducted on a Windows-based host equipped with an NVIDIA GeForce RTX 4060 GPU (8 GB VRAM) and an AMD Ryzen 9 7945HX CPU.

Metric	Focus	Characteristics	Use Cases
$\mathbf{F1}_{\mathbf{disease}}$	Diagnostic accuracy	Combines precision and recall	Medical QA, diagnostic reasoning
$\mathbf{F1}_{\mathbf{treatment}}$	Treatment accuracy	Combines precision and recall	Recommendation generation
Average F1	Overall task performance	Unified multi-task evaluation	Joint QA objectives
ROUGE-L	Textual consistency	Based on longest common subsequence	Summarization, QA generation
BERTScore	Semantic similarity	Contextual embedding comparison	Complex QA, long-text generation

Table 6. Descriptions and characteristics of evaluation metrics.

B.2 Dataset Schema and Preprocessing

The experiments in this study utilize the **RJUA-QA Dataset**, which contains 2,132 high-quality clinician-authored QA pairs. Each pair consists of a *Question*, an expert-verified *Answer*, a diagnosed *Disease*, a clinical *Advice*, and a *Context* paragraph manually extracted from medical guideline documents. The dataset covers approximately 97.6% of real urological disease cases collected between 2019 and 2023, spanning outpatient, emergency, inpatient, and perioperative scenarios.

Each record follows a six-field schema: *Id*, *Question*, *Context*, *Answer*, *Disease*, and *Advice*. The data are divided into training, validation, and test subsets for model development and evaluation. After denoising and entity normalization, the average text length of the *Question* field is 186 characters, and that of the *Context* field is 1,744 characters.

Knowledge Extraction. Entity and relation extraction are conducted using the **OneKE** framework, which employs **LLaMA2-13B** as the backbone language model and integrates Group-Query Attention with LoRA fine-tuning. The framework identifies entities and relationships in urological QA texts, producing structured triplets that capture medical semantics, such as *symptom–examination–diagnosis–treatment*. Extraction is performed in two stages: (1) initial automated extraction followed by manual correction of sampled errors, and (2) incremental fine-tuning with contrastive learning until extraction accuracy exceeds 95%.

Hypergraph Construction. Triplets extracted from both *RJUA-QA* and *RJUA-Context* subsets are merged, cleaned, and deduplicated to form the **RJUA-HKG** (Hypergraph-based Knowledge Graph). Entities include diseases, symptoms, examinations, treatments, and etiologies, while relations capture causal, temporal, and semantic dependencies. After alignment, 16,049 triplets were obtained from RJUA-QA and 18,189 from RJUA-Context, with 1,038 unique patient-centered hyperedges (Table 7). These hyperedges link guideline-derived concepts with individual patient records, providing a unified semantic representation framework for complex clinical reasoning.

Data Description	Number of Triplets
RJUA-QA	16,049
RJUA-Context	18,189
Hyperedges	**Count**
Individual Case Paths	1,038

Table 7. Results of Triplet Extraction

B.3 Prompt Template

```
Question: {question}
Context: {context}
Return including:
'answer': '<detailed response to patient question>'
'disease': '<disease name>'
'advice': '<treatment or recommendation>'
```

B.4 Implementation Details

The entity and relation extraction stage is implemented using the **OneKE**[42] model with **LLaMA2-13B** as the backbone large language model. This phase is responsible for structured knowledge extraction and hypergraph construction but is not involved in downstream QA generation.

For all **RAG experiments**, lightweight open-source large language models are used as the backbone generators, including **LLaMA3.2**, **LLaVA**, **Mistral**, **Neutral-Chat**, and **Qwen2.5-7B**. Each model is integrated via the LangChain framework for modular retrieval and reasoning. Text embeddings are generated using the **nomic-embed-text** encoder and stored in a **Chroma** vector database to support efficient low-latency retrieval.

Hyperparameters are kept consistent across all LLMs to ensure fair comparison: temperature = 0.3, top_k = 5, top_p = 0.6.

Appendix C. Additional Experimental Results

This appendix presents the detailed results of baseline large language models evaluated for their fundamental reasoning and generation capabilities. Five representative lightweight models—LLaMA3.2, LLaVA, Mistral, Neural-Chat, and Qwen2.5-7B—are compared across multiple metrics, including $F1_{disease}$, $F1_{advice}$, Average F1 Score, ROUGE-L, and BERTScore. All results are averaged over three independent runs on the RJUA-QA test set.

Model	**$F1_{disease}$**	**$F1_{advice}$**	Average F1 Score	ROUGE-L	BERTScore
LLaMA3.2	62.47%	27.18%	50.70%	12.20%	30.62%
LLaVA	55.20%	17.80%	42.73%	8.74%	29.43%
Mistral	54.69%	24.82%	44.73%	10.37%	29.57%
Neural-Chat	53.30%	22.36%	42.99%	8.74%	27.67%
Qwen2.5-7B	62.68%	29.78%	51.71%	13.14%	32.14%
Medical HyperRAG (Ours)	**73.19%**	**47.97%**	**64.79%**	**24.29%**	**37.19%**

Table 8. Model performance comparison on the RJUA-QA test set.

The results consistently demonstrate that the proposed **Medical HyperRAG** achieves higher diagnostic and treatment accuracy, as well as improved textual and semantic consistency. These findings confirm the advantages of the **ClinBridge HyperGraph** for enabling structure-aware retrieval and multi-condition-constrained generation, thereby enhancing the reliability and interpretability of clinical question-answering systems.