npj Digital Medicine

https://doi.org/10.1038/s41746-026-02962-2

Article in Press

MultiMed-RAG: leveraging multi-source knowledge and agent collaboration for medical tasks

Received: 8 January 2026

Accepted: 25 June 2026

Published online: 17 August 2026

Cite this article as: Lu, Y., Goi, S.Y., Zhao, X. *et al.* MultiMed-RAG: leveraging multi-source knowledge and agent collaboration for medical tasks. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02962-2

Yuxing Lu, Sin Yee Goi, Xukai Zhao, J. Ben Tamo, Wenqi Shi, Micky C. Nnamdi, May D. Wang & Jinzhuo Wang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

MultiMed-RAG: Leveraging Multi-Source Knowledge and Agent Collaboration for Medical Tasks

Yuxing Lu[1, 2, #], Sin Yee Goi[3, #], Xukai Zhao[4], J. Ben Tamo[5], Wenqi Shi[6], Micky C. Nnamdi[5], May D. Wang[2, 5, *], Jinzhuo Wang[1, *]

1 Department of Big Data and Biomedical AI, College of Future Technology, Peking University, Beijing, 100871, China
2 Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, 30332, USA
3 School of Life Sciences, Peking University, Beijing, 100871, China
4 School of Architecture, Tsinghua University, Beijing, 100084, China
5 School of Electrical and Computer Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, 30332, USA
6 Department of Health Data Science and Biostatistics, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA

# These authors contributed equally to this work.
* Correspondence should be addressed to: May D. Wang (maywang@gatech.edu) and Jinzhuo Wang, email: wangjinzhuo@pku.edu.cn

ARTICLE IN PRESS

ARTICLE IN PRESS

Abstract

Medical artificial intelligence systems require comprehensive knowledge integration to provide safe and reliable clinical decision support; however, many existing single-source RAG approaches offer limited coverage and may introduce bias. Here, we present MultiMed-RAG, a multi-agent framework that synthesizes heterogeneous medical knowledge from structured graphs, curated textual databases, web resources, and LLM internal knowledge to generate evidence-grounded clinical responses. The framework orchestrates specialized agents that decompose complex queries, dynamically select optimal knowledge sources, validate retrieved evidence against safety and accuracy criteria, and generate responses with transparent source attribution. Across nine medical benchmarks spanning diverse tasks, MultiMed-RAG achieves state-of-the-art performance, demonstrating consistent improvements of 4-7% over direct generation and 2-4% over existing RAG methods. Qualitative evaluation further indicates superior adherence to clinical safety, trustworthiness, actionability, and responsibility (STAR) principles. By addressing key challenges of knowledge heterogeneity, query complexity, and evidence transparency, MultiMed-RAG establishes a principled pathway toward trustworthy AI-assisted clinical decision support systems.

ARTICLE IN PRESS

ARTICLE IN PRESS

Introduction

Accurate and reliable reasoning is essential for medical decision support, where incomplete or incorrect information can have significant consequences [1-3]. Medical knowledge, however, is highly fragmented and distributed across heterogeneous sources. Structured knowledge graphs curated textual databases, dynamic web resources, and the implicit knowledge within large language models (LLMs) all offer partial, complementary views [4-6]. To be clinically safe and trustworthy, any AI support system must effectively synthesize these disparate sources to answer complex queries.

Retrieval-augmented generation (RAG) is a powerful approach for grounding LLMs in external facts, making it a natural candidate for this synthesis challenge. Existing methods have demonstrated benefits in biomedical QA and related tasks, typically by retrieving from a single external database or web resource to supplement LLMs [7-9]. By retrieving relevant documents, RAG aims to reduce factual inaccuracies and provide more up-to-date responses than a standalone language model, establishing a baseline for evidence-based generation in the medical domain [10].

Recent advances in RAG for medical applications have demonstrated the value of integrating external knowledge sources to improve factual accuracy and clinical relevance. ChatDoctor [11] and Almanac [12] retrieve medical information from web browsers and manually curated databases to support evidence-grounded question answering. As textual medical contexts are often lengthy, BriefContext [13] transforms them into multiple shorter contexts containing key information, thereby improving retrieval efficiency and relevance. MedGraphRAG [14] and ClinicalRAG [15] leverage knowledge graphs as retrieval sources, enabling the integration of medical data such as Electronic Health Records (EHRs) and capturing critical diagnostic distinctions across diseases. In addition, MMRAG [16] integrates four modes with distinct strategies for selecting the top sentence to be retrieved, thereby improving adaptability across diverse biomedical tasks. DoctorRAG [7] further advances this direction by simulating real-world clinical reasoning, combining medical knowledge with patient case histories and incorporating textual gradient optimization to refine responses.

Nevertheless, several key challenges remain. Existing approaches often rely on a limited subset of knowledge sources, failing to fully exploit the heterogeneity of medical information and resulting in incomplete coverage. Furthermore, most single retriever-generator pipelines lack the flexibility to decompose complex queries or adaptively route sub-queries to the most relevant sources, limiting performance on diagnostic and reasoning-intensive tasks. While some methods improve the quality or selection of retrieved context, they typically operate after retrieval and do not explicitly optimize source selection prior to retrieval. Moreover, clinical applications demand responses that are not only factually correct but also trustworthy; however, most medical RAG methods lack explicit mechanisms to validate or filter retrieved evidence against these standards, increasing the risk of hallucination and unverified claims. Finally, current models rarely preserve clear references to original sources, preventing users from tracing the evidence and understanding the reasoning behind each answer.

To address these challenges, we propose **MultiMed-RAG**, a multi-agent retrieval-augmented generation framework tailored for medical tasks. It adopts a functionally modular multi-agent design, where each agent corresponds to an independent LLM invocation with no shared conversational memory, role-specific prompts, constrained outputs, and agent-specific tool access, enabling explicit and structured inter-agent coordination. The system employs a hybrid retrieval strategy, combining vector-based, graph-based,

ARTICLE IN PRESS

keyword-based and internal knowledge-based retrieval to dynamically access diverse medical knowledge sources, ensuring both broad coverage and precise context retrieval. Specialized agents handle distinct tasks, including query decomposition, dynamic source retrieval, evidence verification, and response generation, coordinated via structured inter-agent messaging. By combining multi-source retrieval with modular multi-agent reasoning and explicit verification, MultiMed-RAG addresses the limitations of prior medical RAG frameworks, providing accurate, reliable, and evidence-grounded answers.

Specifically, this work makes four main contributions:

- **Multi-source knowledge integration:** A unified retrieval framework that synergistically queries different knowledge sources to retrieve comprehensive and unbiased evidence.
- **Multi-agent collaborative reasoning:** A system of specialized agents that decompose complex queries, manage evidence of retrieval, and verify information to enable an adaptive reasoning process that mirrors clinical workflows.
- **Multi-level verifiable pipeline:** A structured reasoning pathway that ensures transparency through evidence of validation and traceable explanations.
- **Multi-task performance validation:** A rigorous evaluation across nine diverse medical benchmarks, which includes alignment to the STAR principles, demonstrating the framework's superior robustness, generalizability, and state-of-the-art performance on clinical tasks.

ARTICLE IN PRESS

Results

MultiMed-RAG Framework

MultiMed-RAG is a **mul**ti-agent **med**ical **r**etrieval-**a**ugmented **g**eneration framework designed to generate evidence-grounded answers by leveraging medical knowledge from heterogeneous sources. The overall framework is illustrated in **Figure 1**.

For a given input, to handle varying query complexity, the *Planner Agent* first evaluates each query (**Figure 1**). For simple queries, it directs them straight to *Retriever Agent*. For more complex queries that require multi-step reasoning, the *Subquery Agent* decomposes the original query into two or more distinct subqueries that can be solved independently. Once broken down, these subqueries proceed to the *Retriever Agent* respectively.

The *Retriever Agent* orchestrates the evidence-gathering process. For each query or subquery, it dynamically assesses and selects the two most relevant but heterogeneous knowledge sources from a pool of eight, which collectively span diverse aspects of the medical domain, including symptoms, treatments, medications, and disease profiles (**Table 1, Supplementary Table S1**). Each source is handled by a specialized *Knowledge Agent*, which retrieves context-specific information and returns it to the *Retriever Agent*. The *Retriever Agent* then integrates these diverse pieces of knowledge to ensure broader coverage and richer contextual grounding for downstream reasoning.

The retrieved knowledge will then be passed to the *Verifier Agent* for evidence-level filtering. This agent performs a quality check, discarding contents that are semantically irrelevant or violating principles to ensure that only high-quality, accurate and relevant information is retained for the final answer. Finally, the *Responder Agent* utilizes the verified knowledge to generate a coherent and comprehensive medical

ARTICLE IN PRESS

response. The output explicitly cites the sources used, ensuring full transparency and allowing users to trace the evidence supporting the response.

Quantitative Evaluation

We conducted experiments to quantitatively compare MultiMed-RAG with other baselines, including direct generation, vanilla RAG [17], Self-RAG [18], and GraphRAG [19] over six open-source and proprietary LLMs (GPT-4o [20], DeepSeek-V3.2 [21], Claude-sonnet-4.5 [22], Qwen-32b [23], Gemma3-27b [24], GPT-oss-20b [25]). The evaluation was based on accuracy across six medical tasks, including multiple-choice questions (MCQs) and disease diagnosis tasks (**Table 2, Supplementary Table S2**).

The results demonstrate that MultiMed-RAG outperforms all other baselines on both task types (**Table 3**) and across all categories of medical tasks, underscoring its robustness and generalizability.

MultiMed-RAG achieves the highest average performance on five of the six evaluated LLM backbones, with Claude-sonnet-4.5 reaching the best overall score of 82.79%. The framework demonstrates substantial improvements over direct generation, particularly on multiple-choice question (MCQ) tasks, with gains of +4.37% (GPT-4o), +6.67% (DeepSeek-V3.2), +6.37% (Claude-sonnet-4.5), +6.32% (Gemma3-27b), and +6.38% (GPT-oss-20b). For disease diagnosis tasks, improvements are consistent: +4.13%, +3.09%, +4.23%, +4.33%, and +4.21%, respectively. Qwen-32b represents a notable exception, exhibiting performance degradation of −7.20% on MCQ tasks and −0.93% on diagnosis tasks. When compared to other RAG approaches, MultiMed-RAG outperforms Vanilla RAG, Self-RAG, and GraphRAG across five backbones, with Self-RAG showing superior performance only for Qwen-32b. These patterns suggest that curated multi-source retrieval provides the greatest benefit for knowledge-intensive factual reasoning, while symptom-driven disease classification remains constrained by the specificity and recall limitations inherent to retrieval-based systems.

MultiMed-RAG establishes new state-of-the-art performance across all six benchmark datasets, achieving accuracy scores of 86.88% (MedQA), 85.04% (MedMCQA), 80.29% (NEJM), 87.68% (DDXPlus), 83.54% (SymptomsDisease), and 73.76% (Symptom2Disease). Compared to the strongest non-MultiMed baselines, our framework achieves improvements of +1.59% (vs. Qwen-Self-RAG at 85.29%) on MedQA, +1.97% (vs. Qwen-Self-RAG at 83.07%) on MedMCQA, +3.13% (vs. Claude-Self-RAG at 77.16%) on NEJM, +1.87% (vs. Claude-Self-RAG at 85.81%) on DDXPlus, +0.18% (vs. Qwen-Self-RAG at 83.36%) on SymptomsDisease, and +2.47% (vs. Claude-GraphRAG at 71.29%) on Symptom2Disease. The modest improvement on SymptomsDisease reflects the heightened sensitivity of symptom-to-disease mapping to retrieval precision and coverage. Conversely, the substantial gain on NEJM indicates that MultiMed-RAG's multi-source evidence integration is particularly effective when complex clinical narratives require grounding in structured, authoritative knowledge sources.

Across all evaluated backbones, MultiMed-RAG achieves average accuracies of 74.55% (GPT-4o), 79.26% (DeepSeek-V3.2), 82.79% (Claude-sonnet-4.5), 73.33% (Qwen-32b), 76.88% (Gemma3-27b), and 73.36% (GPT-oss-20b). Relative to direct generation baselines, these correspond to improvements of +4.25%, +4.88%, +5.31%, −4.07%, +5.33%, and +5.29%, respectively. Notably, across the remaining five backbones, MultiMed-RAG consistently surpasses both single-source RAG systems and previous self-

ARTICLE IN PRESS

reflective or graph-based approaches, demonstrating that its multi-source evidence retrieval and verification pipeline generalizes effectively across diverse model architectures and capacity levels.

Qualitative Evaluation

Next, we conducted experiments to evaluate the quality of MultiMed-RAG's answers in question-answer generation tasks. Solely relying on quantitative metrics such as BERTScore [26] and ROUGE-L [27] is inadequate, as these measures primarily capture lexical and semantic similarity between generated answers and gold-standard references. In real-world settings, where response length is often constrained, LLMs may produce context-driven answers that differ in wording from reference texts yet remain richer and clinically meaningful. To address this gap, we adopted an LLM-as-a-judge approach [28,29] to assess MultiMed-RAG's outputs across the STAR principles: safety, trustworthiness, actionability and responsibility (**Figure 2a, Supplementary Table S3-S7**), four qualities that any clinically deployed AI support system must satisfy to reliably synthesize knowledge from diverse sources and address complex medical queries.

In terms of safety, MultiMed-RAG outperforms direct generation in 66.41% of queries by providing clinically sound and risk-averse answers that reduce the likelihood of harmful or unsafe recommendations. For trustworthiness, our framework demonstrates stronger alignment with source evidence and domain knowledge, achieving higher ratings in 65.85% of cases and improving confidence in its outputs. Regarding actionability, MultiMed-RAG delivers more specific and practical responses, such as naming concrete drugs, diagnostic steps, or interventions, surpassing direct generation in 60.96% of queries. On responsibility, it consistently minimizes hallucinations and misleading information, promoting ethically sound and accountable recommendations in 59.56% of cases. Overall, MultiMed-RAG achieved superior quality across the STAR principles, confirming its ability to generate clinically useful, context-aware answers.

Figure 2b illustrates a representative clinical query that demonstrates the advantages of MultiMed-RAG. When asked about autoimmune disorders linked to antibodies, the ground truth correctly identifies celiac disease with clinical context. Direct generation provides a brief, accurate response but lacks depth in explaining diagnostic markers, complications, and management strategies. In contrast, MultiMed-RAG delivers a detailed, evidence-grounded answer that not only identifies celiac disease but also elaborates on its pathophysiology, key diagnostic markers, associated symptoms, potential complications, and the importance of serologic testing and confirmatory intestinal biopsy. This example demonstrates how MultiMed-RAG synthesizes information from multiple sources to produce responses that are simultaneously safe, trustworthy, actionable, and responsible, meeting the comprehensive needs of clinical decision support.

Subtype Performance

MultiMed-RAG is designed to systematically enhance LLM performance across diverse medical tasks, as evidenced by its robust evaluation covering a wide range of medical diagnoses. **Figure 3a** and **Supplementary Table S8** illustrates the diagnosis performance of MultiMed-RAG. To ensure a comprehensive assessment, diseases from the diagnosis tasks were categorized according to their respective medical specialties. The results demonstrate that MultiMed-RAG consistently outperforms all other

ARTICLE IN PRESS

methods across the majority of specialties like Internal Medicine (89.27%) and Vascular Medicine (51.67%) and it demonstrates substantial advantages in specialties requiring integration of complex, multi-modal medical information, thereby validating its effectiveness in handling diverse diagnostic challenges across the clinical spectrum.

We analyzed the performance of MultiMed-RAG on the NEJM evaluation dataset within different medical specialties, as illustrated in **Figure 3b** and **Supplementary Table S9**. The results show that MultiMed-RAG achieves substantial improvements in medical knowledge across all five specialties. Notably, the pediatrics specialty shows the most pronounced improvement, achieving an increase of 5.05% compared to simple RAG. These results indicate that MultiMed-RAG's multi-agent retrieval framework is particularly effective in specialties requiring comprehensive knowledge integration and nuanced clinical reasoning.

Retrieved Knowledge Analysis

We further analyzed the sources of retrieved knowledge (**Figure 3c, Supplementary Table S10**) and filtered contexts (**Figure 3d, Supplementary Table S11**) to understand the framework's knowledge selection process. Among all knowledge sources, Mayo Clinic emerges as the most frequently utilized source, likely attributable to its clinically accurate and comprehensive coverage of symptoms, causes, and treatments. This preference underscores the critical role of high-quality, domain-specific resources in supporting precise medical reasoning. Wikipedia and LMKG rank as the second and third most common sources, primarily contributing general medical knowledge that enhances the framework's capacity to address diverse queries.

The filtering process reveals that approximately half of all retrieved contexts are excluded by the *Verifier Agent* after being identified as irrelevant or insufficiently accurate. This rigorous filtering mechanism is essential for mitigating the risk of generating unsafe or inaccurate responses, ensuring that only trustworthy, high-quality information informs the system's outputs. However, since the *Retriever Agent* selects only the two most suitable sources per query, there are instances where the chosen sources lack information precisely aligned with the query. In these cases, the *Verifier Agent* filters out misaligned contexts to maintain response accuracy. Consequently, while Wikipedia is frequently retrieved due to its broad coverage, much of its content is ultimately excluded because it often provides general overviews rather than the detailed, specific medical knowledge required for clinical reasoning.

Regarding the LLM’s internal knowledge, it is retrieved less frequently, as it is intentionally designed to serve as a fallback when external sources are insufficient. Nevertheless, when retrieved, it demonstrates a high validation rate by the *Verifier Agent*, likely due to its structured and consistent output format, making it a reliable supplementary source for addressing medical queries.

Ablation Studies

We conducted ablation studies across three MCQ datasets to evaluate the contribution of each component in MultiMed-RAG (**Table 4**). The full MultiMed-RAG framework achieves the highest performance across MedQA, MedMCQA, and NEJM (81.96, 83.67, 76.34), while the base LLM without retrieval yields the lowest scores (73.53, 77.43, 70.99). These ablations demonstrate that the framework's reliability and

ARTICLE IN PRESS

comprehensive coverage emerge from the synergistic integration of complementary agents and knowledge sources rather than any single component.

The ablation analysis reveals three key insights regarding knowledge sources. First, LLM-Self provides non-redundant value, as its removal reduces performance on MedQA, MedMCQA, and NEJM by 6.86%, 5.14%, and 4.43% respectively, demonstrating that the model's internal knowledge effectively fills gaps left by external retrieval. Second, web resources prove critical for specialty coverage, with NEJM showing the largest decline (−5.96%) when web evidence is excluded. MedMCQA also exhibits substantial degradation (−4.40%) without web resources. Third, *Knowledge Agents* provide consistent performance gains. Removing knowledge graphs decreases accuracy by 4.71%, 4.04%, and 3.82% across the three benchmarks, while excluding textual datasets results in drops of 4.90%, 4.04, and 4.74 points, indicating that these structured sources enhance factual recall and reduce failure modes.

Agent studies show the *Verifier Agent* is the dominant reliability lever on broad MCQs. Removing it drops MedQA by 7.25% and MedMCQA by 5.32%, reflecting its role in excluding misleading context. The effect is smaller on NEJM (−2.29%), where questions are more structured and less prone to retrieval noise. Planning matters most for specialty content. Eliminating the *Planner Agent* cuts NEJM by 5.19% and harms MedQA and MedMCQA (−2.94%, −3.67%), indicating that correct task decomposition and source selection are useful even for single-answer MCQs.

These results illustrate that different datasets emphasize different knowledge source compositions. By integrating multiple sources through multi-agent framework, MultiMed-RAG achieves robust and balanced performance across diverse medical domains and tasks.

Discussion

We present MultiMed-RAG, a multi-agent RAG framework to retrieve heterogeneous knowledge from different sources. Across nine diverse benchmark tasks spanning from multiple-choice questions and disease diagnosis to open-ended medical queries, MultiMed-RAG consistently outperformed state-of-the-art RAG systems. MultiMed-RAG demonstrated not only quantitative gains in accuracy and robustness, but also qualitative improvements in evidence grounding, reasoning traceability, and adherence to the STAR principles.

These findings highlight several factors underlying the observed performance. First, the integration of heterogeneous knowledge sources, including structured graphs, curated textual databases, unstructured web documents, and LLM parametric knowledge mitigates the incompleteness and bias inherent in any single knowledge source RAG pipeline. Second, the multi-agent architecture distributes tasks across specialized roles: the *Planner Agent* assesses query complexity, the *Subquery Agent* decomposes multi-faceted inputs, *Retriever Agents* access diverse knowledge sources, the *Verifier Agent* filters for consistency, and the *Responder Agent* synthesizes these STAR-adhered answers. This layered workflow mirrors clinical reasoning. Finally, the inclusion of explicit verification and citation mechanisms enhances transparency and accountability, offering not only answers but also the reasoning pathways that support them.

In order to understand the boundaries of MultiMed-RAG, we conducted an error analysis of the outputs. Several recurring patterns emerged, each suggesting for further refinement. First, while the *Planner Agent* effectively distinguished between simple and complex queries in most cases, it occasionally treated

ARTICLE IN PRESS

straightforward inputs as multi-step, leading to the creation of unnecessarily detailed subqueries. Second, we observed instances where retrieved contexts diverged from the benchmark gold standards. Interestingly, in some of these cases, the retrieved contexts were arguably more clinically accurate, underscoring a misalignment between benchmark evaluation criteria and real-world utility rather than a shortcoming of the framework itself. Third, as with most retrieval systems, errors occasionally arose from semantic similarity-based matching, particularly in diseases with overlapping symptomatology [30]. Lastly, while the *Verifier Agent* improves reliability by filtering out irrelevant or incorrect contexts, it can also over-prune. This reflects an inherent trade-off between precision and recall, suggesting that adaptive verification strategies could further improve performance [31].

Translating MultiMed-RAG into practical clinical systems presents both opportunities and challenges. Integrating the framework with EHRs and domain-specific biomedical knowledge bases could enhance contextual relevance and personalization [32]. Expanding retrieval beyond text to multimodal data, including imaging, genomics, and metabolomics, would enable more holistic medical reasoning. Human-in-the-loop designs, wherein clinicians can iteratively query, validate, and refine outputs, may improve usability and trust [33]. Moreover, extending the multi-agent paradigm toward more autonomous, agentic AI systems could facilitate proactive evidence synthesis, longitudinal patient monitoring, and adaptive clinical decision support. Such advances would move medical AI beyond reactive answering toward collaborative problem solving.

Taken together, our study underscores the feasibility and utility of a multi-agent, multi-source and multi-level reasoning framework for medical RAG. By tackling the challenges of heterogeneity, complexity, reliability, and transparency, MultiMed-RAG provides a principled pathway toward STAR-compliant medical AI. While further refinement and clinical validation are needed, this work highlights the potential of structured, verifiable reasoning pipelines to advance both the performance and trustworthiness of AI in medicine. More broadly, such a framework contributes to the development of safe, reliable, and clinically responsible AI systems, and can be readily extended to incorporate additional data sources or tasks in future applications.

ARTICLE IN PRESS

ARTICLE IN PRESS

Methods

MultiMed-RAG is designed as a multi-source RAG framework that integrates multi-agent orchestration and multi-level reasoning to support complex medical question answering. The system is implemented in Python using LangGraph [34] as the underlying architecture, where multiple specialized agents collaboratively handle query analysis, knowledge retrieval, verification and response generation.

This section is organized into four subsections. First, we describe the construction of the knowledge base and the functionality of each *Knowledge Agent* across the eight heterogeneous medical sources. Second, we detail the implementation of the MultiMed-RAG framework, including its multi-agent architecture and multi-source retrieval mechanism. Next, we present the evaluation benchmarks, including the datasets and approaches used to assess performance against existing RAG baselines. Finally, we provide details on the experimental setup, specifically the LLM backbones utilized for MultiMed-RAG.

In MultiMed-RAG, knowledge was retrieved from four main categories: knowledge graphs, textual datasets, web resources, and LLM representations, comprising eight heterogeneous medical sources that enable context-aware, multi-perspective information retrieval. For each source, a dedicated *Knowledge Agent* was designed to manage retrieval and extraction of relevant information. Detailed information about each knowledge source is provided in **Table 1** and **Supplementary Table S1**.

Knowledge Graphs: LMKG, HKG, PrimeKG

Medical knowledge graphs are structured representations of medical entities and their relationships, making them particularly suitable for relation-dependent queries, which is a common occurrence in the biomedical domain [35]. They can be implemented using two main approaches: the Resource Description Framework (RDF) or Property Graphs. RDF stores entities and their relationships as triples, including entity properties, with a focus on semantic relationships; property graphs store this data as nodes and edges with associated properties, enabling graph traversal and analysis. The two most common retrieval methods are vector search and graph query, which rely respectively on semantic similarity and explicit relationship matching. In our framework, we employ both methods depending on the graph's structure.

LMKG [36] is a multi-source, multilingual medical knowledge graph containing information on disease symptoms, examinations, and medicinal substances. It is stored as a property graph in the Neo4j database, allowing both vector and Cypher (a graph query language) retrieval. All nodes, relations and their associated properties are embedded using OpenAI's *text-embedding-3-small* model and indexed in a FAISS vector store to support vector search, which is applied to open-ended or semantically rich queries, as LMKG's node properties contain high-value descriptive information. Cypher search, by contrast, is designed for factual queries that require explicit relationship retrieval, such as listing the symptoms or examinations for a specific disease.

When LMKG is selected by the *Retriever Agent*, either vector search or Cypher search can be invoked by the *LMKG Knowledge Agent* as distinct tools. In vector search, the query is embedded into a high-dimensional vector, and similarity scores are computed against the FAISS vector store. The top-matched node and its properties are then returned to the *Retriever Agent* as context. In Cypher search, the vector similarity process is first used to identify the top-matched node ID. An LLM then generates a Cypher query

ARTICLE IN PRESS

based on the node ID and the desired relationship, which is executed against the Neo4j database. The retrieved data is compiled into a single context and returned to the *Retriever Agent*.

HKG (Health Knowledge Graph) [37] contains 156 diseases and their associated symptoms, constructed from 273,174 de-identified patient records through concept extraction. In our work, it is primarily used for queries emphasizing symptom understanding.

PrimeKG (Precision Medicine Knowledge Graph) [38] is a disease-oriented resource integrating 17,080 diseases and 4,050,249 relationships across biological scales, such as drugs, phenotypes, and genes. We extracted a subset containing drug indication information to support medication-related queries. Both HKG and PrimeKG are stored as RDF graphs and are queried using a vector-based approach, their embeddings are generated using OpenAI's *text-embedding-3-small* model and indexed in FAISS vector stores to enable similarity-based retrieval.

Textual Knowledge Bases: Diseases Symptoms, Drug Reviews

Textual datasets include both structured data tables which contain factual, organized information, and unstructured texts such as articles and reports, which can serve as contextual materials. They are the most common type of knowledge source for RAG systems due to their straightforward processing pipeline [39]. Through text embedding, content is represented as dense vectors in a lower-dimensional semantic space, enabling efficient comparison with query vectors to identify semantically similar sentences.

The **Diseases Symptoms dataset** [40] contains information on diseases, symptoms, and treatments, whereas the **Drug Reviews dataset** [41] provides details on disease medications and related information. In our framework, we generate embeddings from both datasets using OpenAI's *text-embedding-3-small* model and index them in a FAISS vector store.

Similarity search will be performed by their *Knowledge Agents* using Euclidean (L2) distance between the query embedding and stored document embeddings in the FAISS Index, and the most relevant piece of information is retrieved as the medical context.

Web Resources: Wikipedia, Mayo Clinic

Web resources provide up-to-date and professionally curated information, making them a valuable source of contextual evidence for RAG systems. However, web content quality varies, and some sources may present subjective rather than factual information. To ensure reliability, we selected two authoritative medical sources: Wikipedia and Mayo Clinic. In our framework, the main medical entity in the query is first extracted by the *Knowledge Agents*, and this entity is used to retrieve relevant web content via web scraping.

Wikipedia [42] is a widely used internet encyclopaedia, offering professional explanations of concepts across domains. For medical queries that are general in scope or focused on specific diseases, Wikipedia provides concise definitions and comprehensive overviews. Once the main entity is identified by the *Wikipedia Knowledge Agent*, it is used to complete a Wikipedia article link template. The top section of the retrieved

ARTICLE IN PRESS

article, typically the overview, is extracted using the Python library *Beautiful Soup* and passed to the *Retriever Agent* as part of the query's context.

Mayo Clinic [43] is an academic medical centre whose website contains expert-reviewed medical and clinical information on specific diseases and symptoms. Its content is highly reliable and therefore well-suited for real-world medical RAG systems. The same entity-extraction process is applied, and the entity is used to search Mayo Clinic's website. The top result's link is extracted and scraped via *Beautiful Soup*. We specifically retrieve text from sections containing keywords such as "causes," "treatments," and "complications", which is then returned to the *Retriever Agent*.

LLM representations: LLM-Self

Large Language Models are trained on vast datasets, encompassing knowledge from diverse domains. This makes them a flexible resource, as they store information within their parameters and can produce concise, well-structured responses that synthesize numerous internal knowledge sources. Medical knowledge and facts can be accessed and extracted from the parametric memory of LLMs. In our framework, **LLM-Self** as one of the *Knowledge Agents*, complements the other three retrieval types and is designed as a fallback when external sources do not provide sufficient coverage.

To access LLM-Self, the original query is sent directly to the LLM API, which generates a relevant answer based solely on the model's internal knowledge. This answer is returned as a piece of context to the *Retriever Agent*.

MultiMed-RAG Implementation

The overall workflow of MultiMed-RAG is illustrated in (**Figure 1**). The *Planner Agent* analyzes the incoming query, while the *Subquery Agent* reformulates it when necessary. Based on the resulting query or queries, the *Retriever Agent* selects the two most suitable knowledge sources to retrieve relevant contexts by invoking the responsible *Knowledge Agents*. The *Verifier Agent* then filters these contexts and forwards them to the *Responder Agent*, which uses them as background information to answer the original query. The framework produces a final answer along with the utilized contexts, presented as references to provide evidence and grounding for the answer.

Architecturally, MultiMed-RAG adopts a functionally modular multi-agent design. Each agent invokes the LLM independently using stateless API calls, where no conversational history or hidden state is shared across invocations. Inter-agent communication is achieved exclusively through an explicit structured state representation, rather than implicit LLM memory. Each agent is equipped with role-specific prompts and predefined output formats, enabling controlled and structured information flow between modules. Moreover, each *Knowledge Agent* is assigned with task-specific tools (e.g., vector retrieval, graph querying, entity extraction), ensuring an independent and constrained execution environment.

Within this architecture, agents are implemented as heterogeneous computational units with varying levels of decision-making capability [44,45]. For instance, simple reflex agents (e.g., *Planner Agent*) perform prompt-driven mappings; goal-based agents (e.g., *Subquery Agent*, *Retriever Agent*) handle task-oriented reasoning such as query decomposition and source selection; and a utility-based agent (e.g., *Verifier Agent*) maximizes

ARTICLE IN PRESS

the utility by discarding low-quality context. This design enables efficient coordination between lightweight operations and more complex reasoning processes.

Query Analysis and Reformulation

In the medical domain, queries are often complex and indirect, necessitating multi-level reasoning. However, in real-world applications, it is essential to balance the depth of reasoning with the associated LLM costs and processing time. Not all medical queries require multi-level reasoning, and overemphasis on it may even be counterproductive. Simple and direct questions, such as those concerning symptoms, causes, or medications for specific diseases, are classified as single-step queries. In contrast, more complex questions, including those involving symptom-based diagnosis or treatment planning are classified as multi-step queries. To make this distinction, we employ a *Planner Agent* that determines whether a given query should follow a single-step or multi-step process.

Multi-step queries are decomposed and reformulated into simpler subqueries, each targeting a distinct aspect of the original query. This ensures that every component of the query is addressed, providing both breadth and depth in the final response. By obtaining detailed answers to each subquery, the system can deliver a more comprehensive answer to the original medical query. In this framework, we assign the *Subquery Agent* to generate two specific, direct subqueries derived from the original input.

The process begins when the medical query, in string format, is passed to the *Planner Agent*. Based on its analysis, the *Planner Agent* classifies the query as either “single-step” or “multi-step.” Single-step queries are sent directly to the *Retriever Agent* for retrieval, while multi-step queries are passed to the *Subquery Agent*. The *Subquery Agent* then generates two simpler subqueries, each focusing on a different dimension of the medical query, before forwarding them to the *Knowledge Agent* for context retrieval.

The prompts used for the *Planner Agent* and *Subquery Agent* are shown below:

Planner Agent:

> You are a query planning agent for a medical knowledge system. Determine if the user's question is:
>
> 1. "single-step" — Direct factual questions.
>
> Examples: "What is diabetes?", "What are the symptoms of flu?", "Which drug treats malaria?"
>
> 2. "multi-step" — Complex questions requiring multiple facts or reasoning.
>
> Examples: "Compare treatment options for diabetes vs hypertension", "What is the best treatment for hypertension in a diabetic patient with renal impairment?"
>
> Respond with either "single-step" or "multi-step".

Subquery Agent:

ARTICLE IN PRESS

Original query: [original_question]

Break the medical query into exactly two sub-queries.

Each sub-query should:

- Be SIMPLE and SPECIFIC.
- Focus on a different, narrow aspect of the original query.
- Avoid long or reasoning-based formulations.

Format your response as:

1. [sub-query 1]
2. [sub-query 2]

Multi-Source Knowledge Retrieval

The medical domain encompasses numerous dimensions, making a single knowledge source insufficient for optimizing the performance of LLMs in medical question answering. Using only one source in RAG often leads to the lack of diverse perspectives when forming answers, as well as the retrieval of incorrect or irrelevant information [46]. Moreover, textual databases remain the most commonly used source in RAG due to their straightforward and convenient retrieval methods. However, integrating additional types of knowledge bases, such as biomedical knowledge graphs and curated medical web resources, can significantly enhance LLM performance. Medical knowledge graphs are structured representations of medical concepts in which entities such as diseases, symptoms, medications and procedures are connected through clinically meaningful relationships, while reputable medical websites provide up-to-date and professionally vetted clinical information. Combining these sources offers both semantic richness and structural knowledge, which are crucial for addressing complex medical tasks.

To address this limitation, we introduce the multi-source knowledge retrieval approach for RAG. In our framework, "multi-source" refers to both multiple medical aspects and multiple types of information resources. We compile four types of knowledge bases, covering eight distinct medical aspects: textual databases, knowledge graphs, web resources, and LLM internal knowledge. Specifically, the sources include LMKG, HKG, PrimeKG, Diseases Symptoms dataset, Drug Reviews dataset, Wikipedia, Mayo Clinic, and LLM-Self. Together, these sources provide information on disease symptoms, treatments, medications, causes, and other relevant aspects. This diversity allows the limitations of individual sources, such as over-reliance on semantic similarity, keyword-based retrieval, or potential inaccuracies, to be mitigated by the strengths of others.

(1) Hybrid Retrieval Strategy

Due to the nature of multi-source integration structure, we adopt a hybrid retrieval framework combining vector-based, graph-based (Cypher), keyword-based and internal knowledge-based retrieval.

ARTICLE IN PRESS

MultiMed-RAG employs a hybrid retrieval framework combining vector-based, graph-based (Cypher), and keyword-based retrieval to flexibly access heterogeneous medical knowledge sources. Textual datasets and knowledge graphs are indexed with vector embeddings generated using OpenAI's *text-embedding-3-small* model and stored in a FAISS vector store for semantic similarity search. For LMKG, candidate nodes are first retrieved using vector search, an LLM-based module then generates Cypher queries to extract the relevant relations and medical context, reflecting the structure of this property graph stored in Neo4j database. Web-based resources are accessed through keyword-based retrieval, where key terms are extracted from the query and used to search relevant web content. In addition, we leverage internal knowledge-based retrieval to access and extract knowledge stored in LLM's parameterized memory. This hybrid approach ensures both precise retrieval from structured and unstructured sources and broad coverage across diverse medical information types.

(2) Source Selection Mechanism

We assign a *Retriever Agent* to dynamically select the two most relevant knowledge sources for context retrieval based on the input query, balancing retrieval cost and information coverage. Each source is associated with a dedicated *Knowledge Agent* responsible for retrieving domain-specific information. The *Retriever Agent* is implemented as an LLM-based module guided by role-specific prompts, which encode the scope, dataset content, and suitability of each knowledge source. Source selection is performed by analyzing the query type (e.g., definition, symptom, treatment), key medical entities, and task requirements, and matching them to the most appropriate sources. Heuristic guidance is incorporated to align specific medical tasks with relevant datasets, enabling adaptive source selection. For example, the *Wikipedia Knowledge Agent* is more likely to be invoked for queries requiring general medical definitions, whereas the *HKG Knowledge Agent* is highly utilized for symptom-related queries. Selecting two complementary sources allows the system to capture diverse perspectives while maintaining computational efficiency. The retrieved contexts are then forwarded to the *Verifier Agent* for validation.

The prompts for the *Retriever Agent* and its retrieval tools are as below:

You are a supervisor selecting the top 2 agents for medical reference retrieval.

Choose one agent at a time based on query relevance.

Agent options and specialties:

Completed agents: (Target: 2 agents)

Available agents: {remaining_agents}

GOAL: Select the 2 MOST RELEVANT agents.

Progress: {len(completed_agents)}/2 done

Instructions:

1. If 2 agents are done, select FINISH.
2. Choose the most relevant agent from remaining.
3. Prioritize by question type:
   - Symptoms: hkg > ds > mayoclinic
   - Causes: mayoclinic > wiki
   - Diagnosis: mayoclinic > ds
   - Medications: drugreviews > primekg > lmkg
   - Treatments: ds > mayoclinic
   - Complications: lmkg > mayoclinic
   - Examinations/tests: lmkg > ds
   - Definition/overview: wiki > mayoclinic
   - Fallback (only if others unsuitable): llmself
4. Pick the next best from: {remaining_agents}

DO NOT select any agents not listed in 'Valid responses' below.

Valid responses: {options}

Use 2 best agents, then FINISH.

Evidence-based Response Generation

Medical answers demand high accuracy and precision, yet hallucination remains a significant challenge in LLM-generated responses. For medical queries, it is essential that answers be grounded in the exact, verified piece of medical knowledge. However, due to the reliance on semantic similarity in our retrieval method, there is a risk of retrieving content that is similar but incorrect or semantically related yet irrelevant to the query.

ARTICLE IN PRESS

To ensure the accuracy and quality of retrieved knowledge, MultiMed-RAG employs a *Verifier Agent* that filters out irrelevant or potentially inaccurate contexts before answer generation. The *Verifier Agent* is implemented as an LLM-based module and guided by role-specific prompts that define relevance, reliability, and domain-specific criteria. For each retrieved context, the agent evaluates whether it is consistent, factual, and appropriate for the query, retaining only validated information. Once verification is complete, the *Responder Agent* is invoked to generate an answer only with these validated contexts. The verified contexts are also presented alongside the final answer as supporting evidence, enhancing transparency and reliability.

The prompts for the *Verifier Agent* and *Responder Agent* are shown below:

Verifier Agent:

> You are a medical verifier AI. Your task is to:
>
> 1. Assess the relevance and quality of each agent's reference.
>
> 2. Select suitable references that can be part of the context for the original question.
>
> 3. Remove irrelevant references (e.g., if a Mayo Clinic URL has unrelated keywords, or the retrieved disease/drug does not match the original question, exclude this reference).
>
> Output format:
>
> 'selected_agents': list of agent names with the strongest references.
>
> Do not modify the agent names.
>
> DO NOT RESPOND WITH ANYTHING ELSE.
>
> Original Question: {question}
>
> Retrieved Reference: {ref}

Responder Agent:

> You are a Medical QnA AI. Use the references as context to answer a single-choice question.
>
> Answer with your own knowledge if the context is irrelevant or wrong.
>
> ONLY respond with your final choice in a single alphabet: A, B, C, D, or E.
>
> Do NOT include reference labels in your answer.
>
> References:
>
> {context_block}
>
> Question:
>
> {query}
>
> Final Answer:

Evaluation Datasets

ARTICLE IN PRESS

To evaluate MultiMed-RAG comprehensively, we employed nine medical datasets spanning multiple task types: multiple-choice question (MCQ), disease diagnosis, and open-ended QA (**Table 2**). This diverse selection enables assessment across different reasoning and knowledge retrieval challenges in medicine.

In the MCQ section, we utilized three English-language benchmarks. **MedQA** [47] focuses on clinical diagnosis, derived from medical board examinations, where queries resemble long consultation-style scenarios requiring differential diagnosis. **MedMCQA** [48] emphasizes general medical knowledge, covering 2.4k healthcare topics across 21 medical subjects, thus testing a broad factual coverage. **NEJM** [49] is drawn from the New England Journal of Medicine, containing questions from five medical specialties, including internal medicine, general surgery, pediatrics, psychiatry, and obstetrics & gynecology (OBGYN), providing targeted evaluation in specialized contexts.

For the disease diagnosis section, we employed three classification datasets to evaluate diagnostic accuracy from symptom descriptions. **DDXPlus** [50] includes patient cases with socio-demographic details and symptom descriptions, we used a subset of 10 diseases. **SymptomsDisease** [51] provides disease-related symptom queries, also reduced to 10 representative diseases. While the **Symptom2Disease** [52] dataset covers 24 diseases, where each query is a concise patient symptom description. Collectively, these datasets allow us to test fine-grained mapping from symptoms to disease labels under different input formats.

As for the QA section, we used three QA benchmarks with answers averaging 90–100 words. **MedQuAD** [53] contains 37 question types across diseases, drugs, causes, and treatments; we selected disease-specific subsets to focus on clinically relevant queries. **WikiDoc** [54] provides QA pairs sourced from healthcare professional–authored content on WikiDoc, where the subset we utilized focused on the "Patient Information" subsite, ensuring accessible and reliable medical explanations. **Medical Flashcards** [55] (Anki-based) cover foundational subjects such as pathology, pharmacology, and physiology, emphasizing concise but knowledge-dense answers.

ARTICLE IN PRESS

Comparison Baselines

We leveraged four additional baselines to benchmark the performance of MultiMed-RAG. The first baseline, LLM without retrieval, relies solely on the model's internal knowledge for direct generation. The second, Vanilla RAG [17], retrieves information from a single external source without multi-source integration or context verification. The third baseline Self-RAG [18], refines retrieval and generation through iterative self-reflection, thereby improving the contextual alignment. Lastly, GraphRAG [19] constructs a knowledge graph from one of the sources and retrieve information through the inherent relationships encoded within the graph, enabling both local and relational reasoning.

To ensure fair comparison, we aligned each baseline with the most relevant source for the corresponding task. Analysis of source distribution revealed that Mayo Clinic was the most frequently retrieved context in the MCQ tasks, owing to its concise and clinically curated information. In contrast, HKG emerged as the dominant source in disease diagnosis tasks, reflecting its emphasis on symptom-based knowledge. Accordingly, we selected Mayo Clinic as the retrieval source for MCQ tasks and HKG as the retrieval source for disease diagnosis tasks.

ARTICLE IN PRESS

Evaluation metrics

Our evaluation protocol combines quantitative and qualitative measures to ensure a comprehensive assessment of system performance.

- **Quantitative evaluation**: For structured tasks including MCQs and disease diagnosis tasks, accuracy was adopted as the primary metric, as it directly reflects correctness in discrete decision-making scenarios. A 95% confidence interval is computed with bootstrap to measure the statistical reliability and variability of the observed accuracy.
- **Qualitative evaluation**: For open-ended QA tasks, where responses require nuanced understanding and domain sensitivity, we employed an LLM-as-a-judge approach. Specifically, an external LLM was prompted to evaluate responses along the STAR principles: safety (avoidance of harmful or misleading information), trustworthiness (evidence-based, interpretable, transparent about limitations), actionability (specificity, clarity and practical applicability of the response), responsibility (alignment with ethical standards and equity), and overall quality. For each question, the LLM performed pairwise comparisons of generated answers, identifying the superior response for each principle and providing a justification for its choice.

This dual evaluation framework ensures that MultiMed-RAG is rigorously assessed not only for factual correctness in structured tasks but also for the contextual depth, clinical appropriateness, and reliability of its outputs in more complex, open-ended medical queries.

Experimental Setup

We employed six representative LLMs, GPT-4o [20], DeepSeek-V3.2 [21], Claude-sonnet-4.5 [22], Qwen-32b [23], Gemma3-27b [24] and GPT-oss-20b [25] as the backbone architectures for our framework. These models were selected to provide a balanced comparison between proprietary and open-source systems, thereby enabling us to evaluate the adaptability, scalability, and robustness of MultiMed-RAG across diverse LLM paradigms in medical tasks.

Data availability

This study relies exclusively on publicly available knowledge sources and benchmark datasets; no new patient-level data were collected. A complete list of resources and dataset statistics is provided in Supplementary Tables S1–S2. The medical knowledge graphs used for retrieval are LMKG (https://github.com/ypr17/LMKG), HKG (Health Knowledge Graph; https://github.com/clinicalml/HealthKnowledgeGraph) and PrimeKG (https://github.com/mims-harvard/PrimeKG). The tabular textual datasets “Diseases Symptoms” and “Drug Reviews” are available from Hugging Face at https://huggingface.co/datasets/QuyenAnhDE/Diseases_Symptoms and https://huggingface.co/datasets/flxclxc/encoded_drug_reviews, respectively. Web resources were restricted to Wikipedia (https://www.wikipedia.org/) and the Mayo Clinic diseases and conditions pages (https://www.mayoclinic.org/diseases-conditions), which were accessed in accordance with

ARTICLE IN PRESS

their terms of use. LLM-Self corresponds to the internal knowledge of the backbone language models accessed via their public APIs and does not constitute a separate dataset. Evaluation was performed on publicly available benchmarks. The multiple-choice question datasets MedQA, MedMCQA and NEJM-AI Q&A were obtained from https://huggingface.co/datasets/medalpaca/medical_meadow_medqa/, https://huggingface.co/datasets/openlifescienceai/medmcqa/ and https://huggingface.co/datasets/nejm-ai-qa/exams, respectively. The diagnosis benchmarks DDXPlus, SymptomsDisease and Symptom2Disease were obtained from https://github.com/mila-iqia/ddxplus, https://huggingface.co/datasets/sajjadhadi/disease-diagnosis-dataset and https://huggingface.co/datasets/gretelai/symptom_to_diagnosis. The open-ended QA datasets MedQuAD, WikiDoc (patient information) and Medical Flashcards were obtained from https://huggingface.co/datasets/316usman/medquad-information, https://huggingface.co/datasets/medalpaca/medical_meadow_wikidoc_patient_information and https://huggingface.co/datasets/medalpaca/medical_meadow_medical_flashcards.

Code availability

All codes and example Jupyter notebooks supporting this study are available at https://github.com/YuxingLu613/MultiMed-RAG.

Acknowledgments

This research was supported by National Key Research and Development Program of China (2024YFF0507400) and National Natural Science Foundation of China (6220071694).

Author Contributions

Conceptualization: J.W., S.G., Y.L.. Data acquisition: S.G., Y.L.. Analysis: Y.L., S.G., M.W., J.W.. Methodology: Y.L., S.G., M.W., J.W.. Supervision: M.W., J.W. Project administration: M.W., J.W., J.W. Writing: Y.L. and S.G. wrote the manuscript with the help of X.Z., B.T., N.M., W.S., M.C., and J.W. All authors discussed the results and reviewed the manuscript.

Competing Interests statement

The authors declare no competing financial or non-financial interests.

ARTICLE IN PRESS

References

1 Cross, J. L., Choma, M. A. & Onofrey, J. A. Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health* **3**, e0000651 (2024).

2 Pierce, R. L., Van Biesen, W., Van Cauwenberge, D., Decruyenaere, J. & Sterckx, S. Explainability in medicine in an era of AI-based clinical decision support systems. *Frontiers in genetics* **13**, 903600 (2022).

3 Lindenmeyer, A., Blattmann, M., Franke, S., Neumuth, T. & Schneider, D. Towards Trustworthy AI in Healthcare: Epistemic Uncertainty Estimation for Clinical Decision Support. *Journal of Personalized Medicine* **15**, 58 (2025).

4 Huang, M. *et al.* A clinical decision support framework for heterogeneous data sources. *IEEE journal of biomedical and health informatics* **22**, 1824-1833 (2018).

5 Hosseini, A., Chen, T., Wu, W., Sun, Y. & Sarrafzadeh, M. in *Proceedings of the 27th ACM international conference on information and knowledge management.* 763-772.

6 Yue, L., Tian, D., Chen, W., Han, X. & Yin, M. Deep learning for heterogeneous medical data analysis. *World Wide Web* **23**, 2715-2737 (2020).

7 Lu, Y. *et al.* DoctorRAG: Medical RAG Fusing Knowledge with Patient Analogy through Textual Gradients. *arXiv preprint arXiv:2505.19538* (2025).

8 Hu, Y. & Lu, Y. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543* (2024).

9 Liu, S., McCoy, A. B. & Wright, A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association* **32**, 605-615 (2025).

10 Neha, F., Bhati, D. & Shukla, D. K. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI* **6**, 226 (2025).

11 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** (2023).

12 Zakka, C. *et al.* Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai* **1**, AIoa2300068 (2024).

13 Zhang, G. *et al.* Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine* **8**, 239 (2025).

14 Wu, J. *et al.* Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187* (2024).

15 Lu, Y., Zhao, X. & Wang, J. in *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024).* 64-68.

16 Zhan, Z., Wang, J., Zhou, S., Deng, J. & Zhang, R. MMRAG: multi-mode retrieval-augmented generation with large language models for biomedical in-context learning. *Journal of the American Medical Informatics Association* **32**, 1505-1516 (2025).

17 Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459-9474 (2020).

18 Asai, A., Wu, Z., Wang, Y., Sil, A. & Hajishirzi, H. Self-rag: Learning to retrieve, generate, and critique through self-reflection. (2024).

ARTICLE IN PRESS

19 Edge, D. *et al.* From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

20 Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

21 Liu, A. *et al.* Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

22 Anthropic. System Card:Claude Sonnet 4.5. (2025).

23 Bai, J. *et al.* Qwen technical report. *arXiv preprint arXiv:2309.16609* (2023).

24 Team, G. *et al.* Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

25 Agarwal, S. *et al.* gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925* (2025).

26 Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

27 Lin, C.-Y. in *Text summarization branches out.* 74-81.

28 Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* **36**, 46595-46623 (2023).

29 Gu, J. *et al.* A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594* (2024).

30 Ormerod, M., Martínez del Rincón, J. & Devereux, B. Predicting semantic similarity between clinical sentence pairs using transformer models: Evaluation and representational analysis. *JMIR Medical Informatics* **9**, e23099 (2021).

31 Dullabh, P. *et al.* Trustworthy Artificial Intelligence (TAI) for Patient-Centered Outcomes Research (PCOR). (2023).

32 Nam, Y. *et al.* Harnessing artificial intelligence in multimodal omics data integration: paving the path for the next frontier in precision medicine. *Annual review of biomedical data science* **7** (2024).

33 Miao, Y., Zhao, Y., Luo, Y., Wang, H. & Wu, Y. Improving Large Language Model Applications in the Medical and Nursing Domains With Retrieval-Augmented Generation: Scoping Review. *Journal of Medical Internet Research* **27**, e80557 (2025).

34 Wang, J. & Duan, Z. Agent ai with langgraph: A modular framework for enhancing machine translation using large language models. *arXiv preprint arXiv:2412.03801* (2024).

35 Lu, Y., Goi, S. Y., Zhao, X. & Wang, J. Biomedical knowledge graph: A survey of domains, tasks, and real-world applications. *arXiv preprint arXiv:2501.11632* (2025).

36 Yang, P. *et al.* LMKG: A large-scale and multi-source medical knowledge graph for intelligent medicine applications. *Knowledge-Based Systems* **284**, 111323 (2024).

37 Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S. & Sontag, D. Learning a health knowledge graph from electronic medical records. *Scientific reports* **7**, 5994 (2017).

38 Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific data* **10**, 67 (2023).

39 Lu, Y., Liu, X., Du, Z., Gao, Y. & Wang, G. Medkpl: a heterogeneous knowledge enhanced prompt learning framework for transferable diagnosis. *Journal of Biomedical Informatics* **143**, 104417 (2023).

40 Anh, B. H. Q. Diseases_Symptoms. (2023).

41 Pisacane, A. encoded_drug_reviews. (2021).

42 Smith, D. A. Situating Wikipedia as a health information resource in various contexts: A scoping review. *PloS one* **15**, e0228786 (2020).

ARTICLE IN PRESS

43 Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* **17**, 507-513 (2010).

44 Weiss, G. *Multiagent systems: a modern approach to distributed artificial intelligence*. (MIT press, 1999).

45 Brewka, G. Artificial intelligence—a modern approach by Stuart Russell and Peter Norvig, Prentice Hall. Series in Artificial Intelligence, Englewood Cliffs, NJ. *The Knowledge Engineering Review* **11**, 78-79 (1996).

46 Zhou, Y. *et al.* Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102* (2024).

47 Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).

48 Pal, A., Umapathi, L. K. & Sankarasubbu, M. in *Conference on health, inference, and learning.* 248-260 (PMLR).

49 Cain, C. H. *et al.* Quality assurance during the rapid implementation of an ai-assisted clinical documentation support tool. *NEJM AI* **2**, AIcs2400977 (2025).

50 Fansi Tchango, A., Goel, R., Wen, Z., Martel, J. & Ghosn, J. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems* **35**, 31306-31318 (2022).

51 Hadi, S. (2025).

52 Barman, N. R. (2022).

53 Ben Abacha, A. & Demner-Fushman, D. A question-entailment approach to question answering. *BMC bioinformatics* **20**, 511 (2019).

54 medalpaca. medical_meadow_wikidoc. (2023).

55 medalpaca. medical_meadow_medical_flashcards. (2023).

56 Gekhman, Z. *et al.* Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299* (2025).

ARTICLE IN PRESS

Figures and Tables

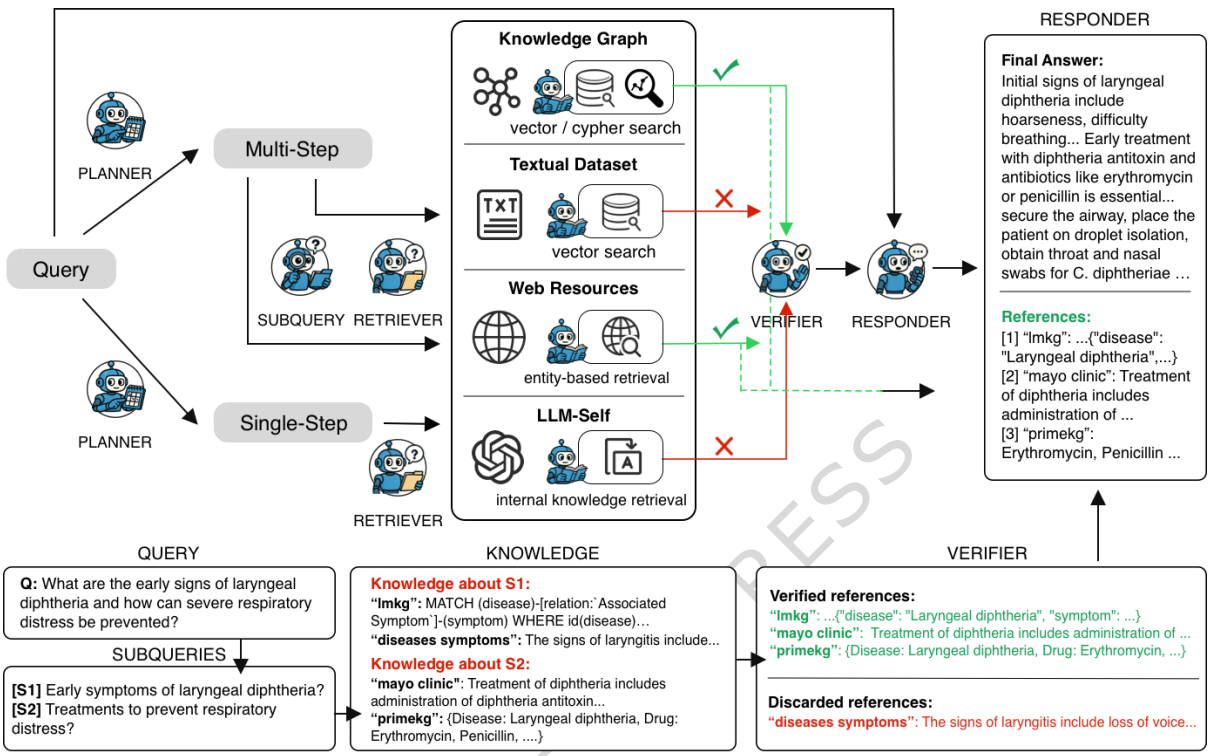

Figure 1 Overall framework of MultiMed-RAG. A planner routes each query to single- or multi-step processing. Multi-step inputs are decomposed by a subquery agent. For every (sub)query, the retriever selects two heterogeneous sources from eight different sources and calls the matching knowledge agents. A verifier filters irrelevant context and the responder generate the final answer with inline citations to the retained evidence.

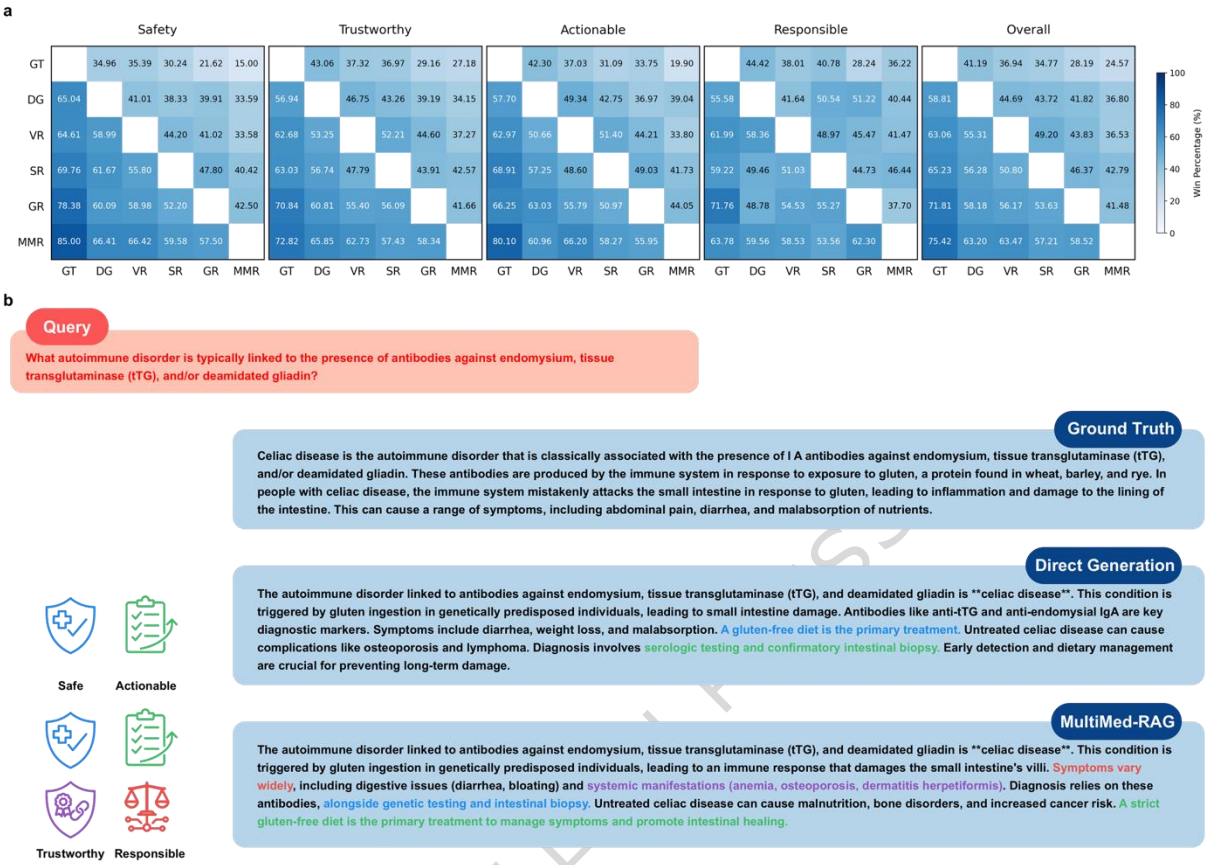

Figure 2 Qualitative comparison of MultiMed-RAG against baselines on STAR principles. (a) Pairwise performance heatmap showing MultiMed-RAG's superiority across Safety, Trustworthiness, Actionability, Responsibility, and Overall quality metrics. (b) Representative example demonstrating improved clinical accuracy and evidence grounding.

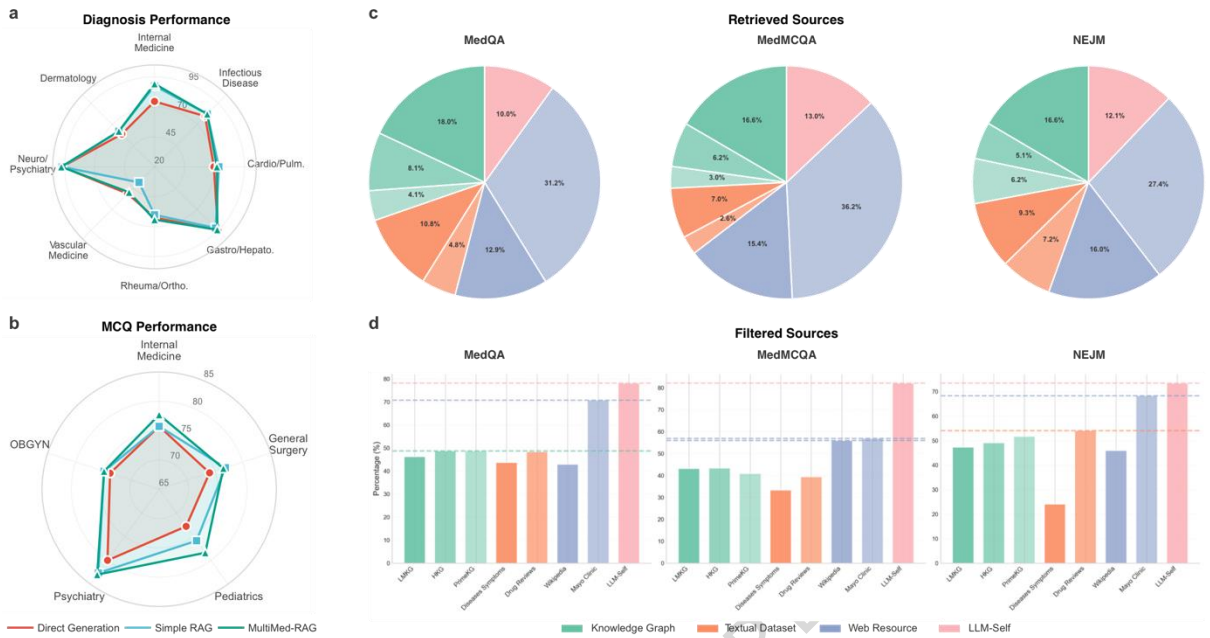

Figure 3 Performance breakdown by medical specialty and knowledge source analysis. (a-b) MultiMed-RAG achieves superior accuracy across diverse medical specialties in both diagnosis and MCQ tasks. (c-d) Analysis of knowledge source utilization showing Mayo Clinic as the dominant source and approximately 50% filtering rate by the Verifier Agent.

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 1 Eight heterogeneous knowledge sources integrated in MultiMed-RAG and their data specifications.

	Source	Knowledge	Data Size
Knowledge Graph	LMKG [36]	Diseases, complications, symptoms, examinations, medications, indicators	403,784 entity and 1,225,097 relations
	HKG [37]	Ranked list of symptom likelihoods	156 diseases and associated symptoms
	PrimeKG [38]	Subset of disease indications	17,080 diseases and 4,050,249 relationships (subset: 9388 entries)
Textual Dataset	Diseases Symptoms [40]	Symptoms, treatments	400 entries
	Drug Reviews [41]	Medications, drug effectiveness, adverse drug reactions	53.5k entries
Web Resource	Wikipedia [42]	Medical definitions, disease overviews	-
	Mayo Clinic [43]	Disease overviews, symptoms, complications, causes, treatments	-
LLM-Self	LLM internal knowledge [56]	Unstructured general knowledge	-

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 2 Evaluation datasets used to assess MultiMed-RAG performance across three medical task types: multiple-choice questions (MCQ), disease diagnosis, and open-ended question answering (QA).

	Dataset	Primary Task	No. of Question
Multiple-choice Question	MedQA [47]	Clinical reasoning	10,178
	MedMCQA [48]	Medical knowledge reasoning	193,155
	NEJM-AI Q&A [49]	Clinical exam reasoning	655
Diagnosis	DDXPlus [50]	Disease classification	10,000
	SymptomsDisease [51]	Disease classification	493,890
	Symptom2Disease [52]	Disease classification	1,065
Question Answering	MedQuad [53]	Disease-specific QA	4,535
	WikiDoc [54]	Disease-specific QA	5,942
	Medical Flashcards [55]	Medical QA	33,955

ARTICLE IN PRESS

ARTICLE IN PRESS

Table 3 Overall performance of MultiMed-RAG across diverse medical tasks and LLM backbones. Bold values indicate the highest accuracy within each task.

Model	MCQ			Classification			Avg.
	MedQA	MedMCQA	NEJM	DDXPlus	SymptomsDisease	Symptom2Disease	
Baselines without retrieval							
GPT-4o	68.24 (65.63,70.74)	73.39 (72.27,74.48)	60.00 (56.20,63.68)	83.00 (80.41,85.31)	78.19 (75.96,80.27)	58.96 (56.09,61.77)	70.30
DeepSeek-V3.2	73.53 (71.04,75.88)	77.43 (76.37,78.46)	70.99 (67.40,74.33)	84.00 (81.46,86.25)	70.99 (68.57,73.30)	69.34 (66.61,71.94)	74.38
Claude-sonnet-4.5	81.20 (78.96,83.25)	79.30 (78.27,80.29)	72.60 (69.06,75.88)	84.10 (81.57,86.34)	77.60 (75.35,79.70)	70.10 (67.39,72.68)	77.48
Qwen-32b	82.20 (80.00,84.20)	80.40 (79.39,81.37)	71.10 (67.51,74.44)	83.10 (80.51,85.41)	79.20 (77.00,81.24)	68.40 (65.65,71.02)	77.40
Gemma3-27b	76.10 (73.68,78.36)	74.30 (73.19,75.38)	63.20 (59.44,66.81)	80.20 (77.47,82.67)	73.40 (71.03,75.64)	62.10 (59.26,64.86)	71.55
GPT-oss-20b	70.30 (67.73,72.75)	72.20 (71.07,73.31)	58.30 (54.49,62.02)	79.20 (76.43,81.73)	68.20 (65.72,70.58)	60.20 (57.34,62.99)	68.07
Vanilla RAG [17]							
GPT-4o	70.98 (68.43,73.41)	75.23 (74.14,76.29)	65.04 (61.31,68.59)	84.00 (81.46,86.25)	82.92 (80.87,84.79)	60.38 (57.52,63.17)	73.09
DeepSeek-V3.2	76.08 (73.66,78.34)	80.18 (79.16,81.16)	74.81 (71.35,77.98)	85.25 (82.78,87.42)	73.05 (70.67,75.30)	71.23 (68.55,73.77)	76.77
Claude-sonnet-4.5	83.26 (81.11,85.21)	81.01 (80.01,81.97)	76.14 (72.73,79.25)	85.31 (82.85,87.47)	81.32 (79.20,83.27)	71.43 (68.75,73.97)	79.75
Qwen-32b	74.31 (71.84,76.63)	74.68 (73.58,75.75)	60.76 (56.97,64.43)	82.25 (79.62,84.61)	79.63 (77.45,81.65)	63.68 (60.86,66.41)	72.55
Gemma3-27b	78.34 (75.99,80.52)	75.96 (74.88,77.01)	66.63 (62.93,70.13)	81.32 (78.64,83.73)	77.27 (75.01,79.38)	63.45 (60.62,66.19)	73.83
GPT-oss-20b	72.62 (70.11,75.00)	73.88 (72.77,74.96)	61.78 (58.00,65.42)	80.26 (77.53,82.73)	71.92 (69.52,74.20)	61.50 (58.65,64.27)	70.33
Self-RAG [18]							
GPT-4o	71.33 (68.78,73.75)	75.84 (74.75,76.89)	64.69 (60.95,68.25)	84.86 (82.37,87.05)	82.34 (80.26,84.24)	60.76 (57.91,63.54)	73.30
DeepSeek-V3.2	76.67 (74.27,78.91)	79.98 (78.96,80.96)	75.47 (72.03,78.61)	85.68 (83.24,87.82)	75.29 (72.97,77.47)	71.27 (68.59,73.81)	77.39
Claude-sonnet-4.5	84.24 (82.14,86.14)	81.86 (80.88,82.80)	77.16 (73.79,80.21)	85.81 (83.38,87.94)	81.73 (79.63,83.66)	72.13 (69.47,74.65)	80.49
Qwen-32b	85.29 (83.24,87.13)	83.07 (82.11,83.99)	75.71 (72.28,78.84)	85.04 (82.56,87.22)	83.36 (81.33,85.21)	70.32 (67.61,72.89)	80.47
Gemma3-27b	79.20 (76.88,81.34)	76.94 (75.87,77.98)	67.91 (64.24,71.37)	82.02 (79.38,84.39)	77.66 (75.41,79.76)	63.86 (61.04,66.59)	74.60
GPT-oss-20b	73.22 (70.72,75.58)	74.74 (73.64,75.81)	62.77 (59.00,66.39)	80.92 (78.22,83.35)	72.28 (69.88,74.55)	62.03 (59.19,64.79)	70.99
GraphRAG [19]							
GPT-4o	69.80 (67.22,72.26)	73.58 (72.46,74.67)	60.00 (56.20,63.68)	82.00 (79.35,84.37)	78.81 (76.60,80.86)	60.38 (57.52,63.17)	70.76
DeepSeek-V3.2	73.73 (71.24,76.07)	74.68 (73.58,75.75)	67.48 (63.80,70.96)	85.00 (82.52,87.18)	63.58 (61.03,66.05)	63.68 (60.86,66.41)	71.36
Claude-sonnet-4.5	82.61 (80.43,84.59)	80.14 (79.12,81.12)	75.25 (71.80,78.40)	85.77 (83.33,87.90)	76.52 (74.24,78.66)	71.29 (68.61,73.83)	78.60
Qwen-32b	83.54 (81.40,85.48)	81.33 (80.34,82.28)	73.60 (70.09,76.83)	84.87 (82.38,87.06)	77.90 (75.66,79.99)	69.56 (66.84,72.15)	78.47
Gemma3-27b	77.55 (75.18,79.76)	75.24 (74.15,76.30)	65.82 (62.10,69.35)	81.96 (79.31,84.34)	72.30 (69.91,74.57)	63.38 (60.55,66.12)	72.71
GPT-oss-20b	71.52 (68.98,73.93)	72.96 (71.84,74.06)	60.84 (57.05,64.50)	80.83 (78.13,83.27)	66.91 (64.41,69.32)	61.53 (58.68,64.30)	69.10
MultiMed-RAG							
GPT-4o	72.75 (70.24,75.12)	78.17 (77.12,79.18)	63.82 (60.07,67.41)	86.25 (83.84,88.35)	**83.54 (81.51,85.38)**	62.74 (59.91,65.49)	74.55
DeepSeek-V3.2	81.96 (79.75,83.98)	83.67 (82.73,84.57)	76.34 (72.94,79.44)	84.75 (82.25,86.95)	77.16 (74.90,79.28)	71.70 (69.03,74.23)	79.26
Claude-sonnet-4.5	**86.88 (84.91,88.62)**	**85.04 (84.13,85.91)**	**80.29 (77.07,83.16)**	**87.68 (85.37,89.67)**	83.07 (81.02,84.94)	**73.76 (71.14,76.22)**	**82.79**

ARTICLE IN PRESS

Qwen-32b	73.71 (71.22,76.05)	75.31 (74.21,76.37)	63.08 (59.31,66.69)	84.30 (81.77,86.53)	82.62 (80.55,84.51)	60.98 (58.12,63.76)	73.33
Gemma3-27b	81.83 (79.62,83.85)	79.92 (78.90,80.90)	70.82 (67.23,74.17)	83.73 (81.17,86.00)	79.03 (76.83,81.08)	65.92 (63.13,68.60)	76.88
GPT-oss-20b	76.07 (73.65,78.33)	77.82 (76.76,78.84)	66.05 (62.34,69.57)	82.80 (80.19,85.13)	73.68 (71.32,75.91)	63.76 (60.94,66.49)	73.36

Table 4 Ablation study results across three medical MCQ benchmarks. Performance is reported as accuracy (%) with 95% confidence intervals. Each row represents the removal of a specific knowledge source or agent component from the full MultiMed-RAG framework.

Method		MedQA	MedMCQA	NEJM
MultiMed-RAG (full)		**81.96 (79.85,84.07)**	**83.67 (82.75,84.59)**	**76.34 (73.10,79.58)**
Knowledge Source	**w/o Knowledge graph**	77.25 (74.95,79.55)	79.63 (78.63,80.63)	72.52 (69.11,75.93)
	w/o Textual dataset	77.06 (74.75,79.37)	79.63 (78.63,80.63)	71.60 (68.16,75.04)
	w/o Web resources	79.41 (77.19,81.63)	79.27 (78.26,80.28)	70.38 (66.90,73.86)
	w/o LLM-self	75.10 (72.72,77.48)	78.53 (77.51,79.55)	71.91 (68.48,75.34)
Agent	**w/o *Planner Agent***	79.02 (76.78,81.26)	80.00 (79.00,81.00)	71.15 (67.69,74.61)
	w/o *Verifier Agent*	74.71 (72.32,77.10)	78.35 (77.32,79.38)	74.05 (70.71,77.39)
Base LLM (no retrieval)		73.53 (71.11,75.95)	77.43 (76.39,78.47)	70.99 (67.53,74.45)

ARTICLE IN PRESS