

Research paper

# MedSumGraph: enhancing GraphRAG for medical QA with summarization and optimized prompts

DaeHo Kim<sup>a</sup>, SoYeop Yoo<sup>b</sup>, OkRan Jeong<sup>a,\*</sup><sup>a</sup> School of Computing, Gachon University, Seongnam-si, 13120, Republic of Korea<sup>b</sup> Dept. of Electrical Engineering and Computer Science, DGIST, Daegu, 42988, Republic of Korea

## ARTICLE INFO

## Keywords:

Medical decision support system  
 Knowledge graph  
 Large language model

## ABSTRACT

The rapid development of large language models (LLMs) has accelerated research into applying artificial intelligence (AI) to domains such as medical question answering and clinical decision support. However, LLMs face substantial limitations in medical contexts due to challenges in understanding specialized terminology, complex contextual information, hallucination issues (i.e., generating incorrect responses), and the black-box nature of their reasoning processes. To address these issues, methods like retrieval-augmented generation (RAG) and its graph-based variant, GraphRAG, have been proposed to incorporate external knowledge into LLMs. Nonetheless, these approaches often rely heavily on external resources and increase system complexity. In this study, we introduce MedSumGraph, a medical question-answering system that enhances GraphRAG by integrating structured medical knowledge summaries and optimized prompt designs. Our method enables LLMs to better interpret domain-specific knowledge without requiring additional training, and it enhances the reliability and interpretability of responses by directly embedding factual evidence and graph-based reasoning into the generation process. MedSumGraph achieves competitive performance on two out of eight multiple-choice medical QA benchmarks, including MedQA (USMLE), outperforming closed-source LLMs and domain-specific foundation models. Moreover, it generalizes effectively to open-domain QA tasks, yielding significant gains in reasoning over common knowledge and evaluating the truthfulness of answers. These findings demonstrate the potential of structured summarization and graph-based reasoning in enhancing the trustworthiness and versatility of LLM-driven medical AI systems.

## 1. Introduction

Large language models (LLMs) have rapidly advanced natural language processing capabilities across various domains, including healthcare. With models like GPT-4 [1] demonstrating remarkable proficiency in generating human-like text, there has been growing interest in applying these models to medical artificial intelligence (AI) tasks such as medical question answering (QA) and clinical decision support [2–4]. These applications aim to improve access to accurate medical knowledge, streamline clinical workflows, and support informed decision-making.

Despite these potential benefits, general-purpose LLMs face significant limitations in the medical domain. Hallucination—generating incorrect or fabricated content—remains a major risk when such models are used for clinical guidance [5]. Furthermore, the black-box nature of LLMs hinders transparency, making it challenging to verify or

comprehend their reasoning [6]. Most importantly, medicine requires an understanding of specialized terminology, evolving guidelines, and complex interrelationships among symptoms, diagnoses, and treatments—factors that generalist LLMs often fail to address adequately.

To mitigate these issues, several domain-specific models such as BioBERT [7], PubMedBERT [8], and Med-PaLM2 [9] have been introduced. These models demonstrate an improved understanding of biomedical language by being trained or fine-tuned on large-scale medical corpora. However, they are still limited by fixed knowledge (snapshot training data), lack of transparency in reasoning, and, in some cases, restricted accessibility due to proprietary constraints.

A promising direction to address the limitations of static and opaque models is the integration of external knowledge into the generation process [10,11]. Retrieval-augmented generation (RAG) [12] and its graph-based extension, GraphRAG [13], represent attempts to ground LLM outputs in relevant knowledge by retrieving textual or structured

\* Corresponding author.

E-mail addresses: [ikimdh91@gachon.ac.kr](mailto:ikimdh91@gachon.ac.kr) (D. Kim), [soyeop@dgist.ac.kr](mailto:soyeop@dgist.ac.kr) (S. Yoo), [orjeong@gachon.ac.kr](mailto:orjeong@gachon.ac.kr) (O. Jeong).<https://doi.org/10.1016/j.artmed.2025.103311>

Received 8 July 2025; Received in revised form 24 September 2025; Accepted 21 November 2025

Available online 24 November 2025

0933-3657/© 2025 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

data from external sources. While these approaches offer improvements in factual accuracy and reasoning depth, they also introduce challenges in managing the volume and complexity of retrieved content.

In this study, we propose MedSumGraph, an enhanced medical question-answering system designed to overcome the limitations of existing approaches. MedSumGraph integrates structured medical knowledge summarization and optimized prompt design within the GraphRAG framework. While GraphRAG effectively incorporates structural relations, it struggles with managing information overload from large-scale medical data and does not directly address prompt optimization, thereby increasing system complexity. Similarly, prompt-based approaches such as MedPrompt improve reasoning performance but remain limited in the depth of external knowledge integration. To address these challenges, MedSumGraph introduces a medical knowledge summarization module that condenses lengthy, multi-source clinical contexts into essential facts (e.g., definitions, causes, symptoms, treatments), thereby reducing information overload. In parallel, optimized prompt design enables the seamless integration of the knowledge graph's rich structural relations with the summarized key information into LLM inputs. This approach allows large language models to better interpret domain-specific medical knowledge, mitigates hallucination, and enhances both the reliability and interpretability of generated responses. Ultimately, MedSumGraph contributes to reducing the risk of incorrect clinical decision-making and improving patient safety.

To validate our approach, we evaluated MedSumGraph on the MedQA benchmark (USMLE) [4], comparing it with multiple domain-specific and retrieval-augmented baselines. As shown in Fig. 1, MedSumGraph achieves competitive performance on the MedQA benchmark, outperforming domain-specific LLMs and existing retrieval-augmented methods. This result highlights the effectiveness of combining knowledge summarization with graph-based reasoning in improving answer accuracy and reliability in the medical domain.

The main contributions of this paper are as follows:

**MedSumGraph Architecture:** We propose MedSumGraph, a novel medical QA system that synergistically combines retrieval-augmented LLM generation with knowledge graph reasoning and summarization. To our knowledge, this is the first framework to integrate GraphRAG with a summarization module specifically designed for medical knowledge, aiming to leverage both structured and unstructured information for enhanced answer generation.

**Knowledge Summarization Module:** We design a method for medical knowledge summarization that condenses relevant information from multiple sources (e.g., PubMed articles, clinical notes) into brief, coherent summaries. These summaries capture essential facts and provide high-quality context to the LLM, reducing information overload and focusing the model on important details.

**Optimized Prompt Design:** We develop an optimized prompt schema that effectively incorporates the knowledge graph triples and summary snippets into the LLM's input. Our prompt design addresses the complexity of terminology and ensures that the model interprets the provided facts correctly. It also improves transparency by structuring the prompt in a way that mirrors a logical reasoning path, making the model's subsequent answer more traceable.

**Empirical Performance and Analysis:** We conduct extensive experiments on medical question-answering benchmarks, including synthetic clinical scenarios and real-world medical exam questions. MedSumGraph is shown to outperform baseline approaches (including standard RAG and previous GraphRAG implementations) in answer accuracy and factuality. We also provide a thorough analysis of examples, illustrating how the integration of summarization and graph knowledge helps prevent LLM hallucinations and produces more trustworthy, explainable answers. The main objectives, secondary objectives, and methods of this paper are as follows.

**Primary Goal:** To enhance the accuracy and interpretability of medical question answering (QA) without requiring additional training or fine-tuning.

**Secondary Goal:** To demonstrate the generalizability of MedSumGraph beyond the medical domain to open-domain QA, while also validating the reliability of LLM-generated responses.

**Method:** We integrate medical knowledge summarization, graph-based retrieval-augmented generation (GraphRAG), and optimized prompt design—incorporating dynamic few-shot prompting, chain-of-thought (CoT) reasoning, and ensemble strategies—within a hybrid global-local retrieval pipeline.

In the following sections, we provide a detailed description of the design and implementation of MedSumGraph. Section 3 introduces the overall methodology, including the process of constructing a medical knowledge graph and generating responses using global-local retrieval and optimized prompting. Sections 4, 5, and 6 present the experiment setup and the results of quantitative and qualitative evaluations,

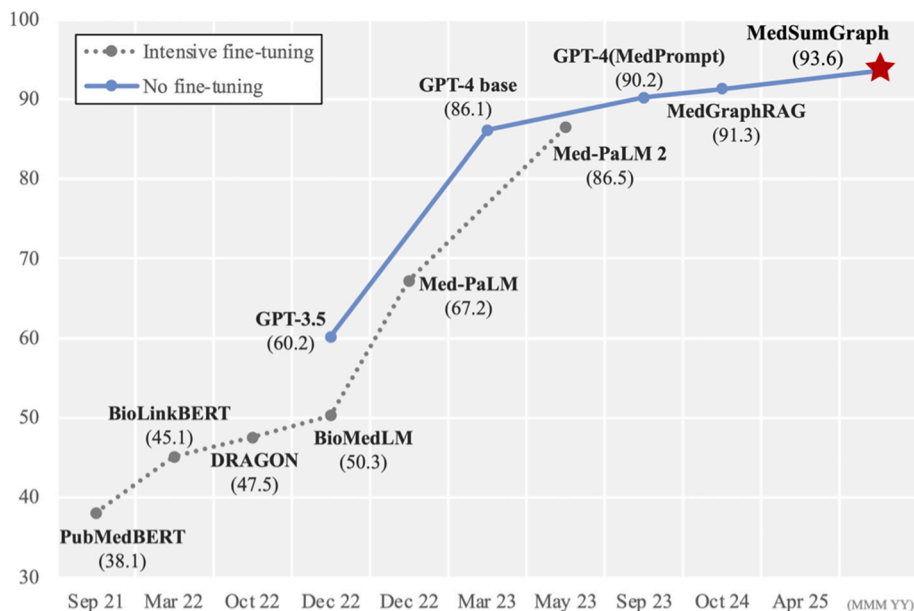


Fig. 1. Performance comparison on the MedQA (USMLE) benchmark.

comparing MedSumGraph with existing methods. Section 7 discusses the practical challenges, limitations, and implications of applying LLMs in the medical domain. Finally, Section 8 concludes the paper and outlines directions for future research.

## 2. Related works

### 2.1. LLM for medicine

As artificial intelligence technology rapidly advances, LLMs are being actively researched in medical fields such as medical question-answering systems and clinical decision support systems [4,5]. However, large language models trained on general commonsense data find it difficult to understand knowledge in specialized fields such as law, medicine, and mathematics, and achieving adequate performance is not easy. Models such as BioBERT [7] and PubMedBERT [8] are representative language models used in the medical field, trained with medical data, and have achieved significant success in various biomedical tasks. However, fine-tuning large language models requires high-quality, well-curated data, consumes massive computing resources for model training, and still faces issues such as catastrophic forgetting, where the model forgets information from the original data after learning with new data.

To address these challenges, recent studies have introduced domain-specific large medical language models. Med-PaLM [14] and Med-PaLM 2 [9] are large-scale LLMs specialized for medical applications, designed to provide reliable medical reasoning and safe clinical responses. Similarly, MedLM [15], developed with large-scale clinical datasets, demonstrated strong performance in tasks such as medical question answering and summarization, showing that domain adaptation of LLMs is critical for trustworthy clinical usage. More recently, Med-Gemini [16] extended the capabilities of LLMs into the multimodal medical domain, integrating not only text but also medical imaging modalities such as chest X-rays and pathology slides. This direction highlights the potential of LLMs to become comprehensive clinical assistants that combine heterogeneous sources of medical knowledge.

In parallel, methods such as RAG (Retrieval Augmented Generation) [12] and GraphRAG [13] are being actively explored to enhance the ability of LLMs to understand and respond to complex and specialized knowledge without the need for costly additional fine-tuning. These hybrid approaches aim to leverage both parametric knowledge from LLMs and structured domain-specific knowledge sources, such as knowledge graphs, to improve reliability and explainability in medical AI systems.

### 2.2. Technologies advance to LLM

To overcome the limitations of LLMs in specialized domains such as medicine, recent research has focused on integrating LLMs with complementary technologies, including retrieval-augmented generation (RAG) and graph-based augmentation methods.

#### 2.2.1. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) [12] has emerged as a promising approach for enhancing medical question answering systems. By retrieving relevant external knowledge and conditioning language model generation on this information, RAG enables more accurate and contextually grounded responses, particularly in knowledge-intensive domains such as medicine. Lewis et al. [12] introduced the original RAG framework, which combines dense retrieval from a large text corpus with sequence-to-sequence generation, demonstrating substantial performance improvements on open-domain QA tasks. In the medical domain, RAG has been adapted to retrieve biomedical literature, clinical guidelines, and electronic health record (EHR) information to improve answer accuracy and relevance. Zhang et al. [17] applied RAG to the MedQA-USMLE dataset, integrating biomedical knowledge

sources to enhance the reasoning ability of language models. Their approach uses a dense retriever to select relevant passages from PubMed articles and clinical guidelines, which are then fed into a generative model for answer synthesis. This method outperforms standard LLM-based QA systems by providing evidence-grounded and interpretable responses, though challenges remain in handling long-context queries and ensuring retrieval precision. These studies illustrate the potential of RAG-based approaches for medical QA, highlighting their ability to combine retrieval of structured and unstructured biomedical knowledge with generative reasoning for improved performance and explainability. However, RAG struggles with multi-hop reasoning across documents and maintaining performance in domains with dense, structured terminology like medicine [18].

#### 2.2.2. Knowledge graphs (KGs)

Knowledge graphs (KGs) [19] have been increasingly utilized in medical question answering (Medical QA) to model structured relationships among diseases, drugs, genes, and other biomedical entities. By providing explicit semantic context, KGs enhance the reasoning capabilities of QA systems, enabling more accurate and interpretable responses in complex clinical scenarios. Recent studies have demonstrated the effectiveness of integrating KGs with large language models (LLMs) to improve performance on medical QA benchmarks. Yasunaga et al. [20] proposed QA-GNN, a model that combines question-context representations with knowledge graph reasoning to jointly infer answers. The model identifies relevant knowledge from the KG based on the question and candidate answers and propagates information through a graph neural network. QA-GNN achieved state-of-the-art performance on benchmarks including MedQA-USMLE, demonstrating that KG-guided reasoning improves accuracy over baseline language models and language model + KG methods. Despite its effectiveness, QA-GNN faces limitations related to KG completeness and scalability for large biomedical graphs. Sohn et al. [21] proposed Rationale-Guided RAG (RAG<sup>2</sup>), which incorporates rationale-guided filtering, LLM-generated rationale queries, and a balanced retrieval strategy from multiple biomedical corpora. RAG<sup>2</sup> selectively augments informative document snippets while filtering out distractors, leverages LLM-generated rationales as queries to improve retrieval utility, and ensures a balanced coverage of biomedical knowledge sources. These studies highlight the critical role of knowledge graph integration in medical QA, demonstrating that KG-based approaches can significantly enhance both reasoning and explainability when combined with large language models.

#### 2.2.3. GraphRAG

GraphRAG [13] enhances RAG by incorporating knowledge graph structure into the retrieval pipeline. It extracts entities and relations from retrieved texts to form graph contexts that the model can traverse for multi-hop reasoning, thereby improving interpretability and answer accuracy in domain-specific QA. Nonetheless, it incurs significant overhead in preprocessing steps like entity extraction, graph construction, and summarization, and may underperform when queries lack explicit entity mentions. MedGraphRAG [22] tailors GraphRAG for medical applications by constructing hierarchical triple graphs that link user documents, credible medical sources, and controlled vocabularies. It introduces a U-Retrieval mechanism combining top-down precise retrieval with bottom-up response refinement. Evaluated on multiple medical QA and fact-checking benchmarks, MedGraphRAG consistently outperforms baselines while providing explicit source citations to enhance safety and reliability in medical response generation. However, GraphRAG also requires various additional tasks such as knowledge graph extraction, graph community generation, and summarization, which necessitate high computing resources and time. Additionally, it does not function well for queries that do not directly mention related entities, which is a limitation of the technology.

#### 2.2.4. HybridRAG

HybridRAG [23] integrates vector similarity-based retrieval with graph-based structural reasoning to take advantage of both dense semantic retrieval and explicit relational context. This hybrid design has shown improved performance in knowledge-intensive tasks and enhanced domain adaptability by merging complementary retrieval modalities. However, it adds system complexity and may be sensitive to context ordering, potentially affecting stability and accuracy in generation.

Overall, recent advances in integrating complementary technologies with large language models (LLMs), including Retrieval-Augmented Generation (RAG), knowledge graphs (KGs), GraphRAG, and HybridRAG, have markedly enhanced both the performance and interpretability of medical question answering systems. These approaches collectively demonstrate that while retrieval- and graph-based augmentations substantially improve LLM capabilities in medical QA, significant challenges remain regarding computational efficiency, knowledge coverage, and the ability to handle complex or implicitly stated queries.

#### 2.3. Prompt strategies

Methods for improving the performance of language models include providing external contextual information for data in specific domains that LLMs cannot easily understand, as described in Section 2.2. However, efficiently designing the model's prompts also demonstrates visible performance improvements. The following introduces papers related to the latest prompting techniques, focusing on the prompt strategies included in MedPrompt [24]. The studies introduced demonstrate performance improvements achieved solely through optimal prompts, without requiring additional training, fine-tuning, or external knowledge.

Dong et al. [25] proposed a in-context learning (ICL) technique that improves a model's ability to solve new tasks by providing it with several examples. In the medical question-answering task, an example of ICL is when a question is given, other question examples similar to the question can be retrieved and provided as prompts. The number of examples used can significantly impact model performance, and ICL is already being actively utilized in other fields. Wei et al. [26] proposed CoT prompting, which breaks down complex problems into intermediate reasoning steps before generating an answer. This approach has been shown to improve LLM reasoning capabilities, and it is particularly effective in multi-step clinical and biomedical QA tasks. Le et al. [27] investigated instruction tuning combined with CoT prompts to enhance biomedical question-answering performance. Their experiments showed that instruction tuning significantly improves accuracy, while CoT prompts improve zero-shot reasoning ability. However, fine-tuning CoT prompts does not universally enhance performance and may sometimes degrade results for larger models. Nachane et al. [28] introduced AutoMedPrompt, a framework that leverages system prompt optimization to elicit medically relevant reasoning from LLMs. This method improves QA performance without additional model training or fine-tuning. Jenish et al. [29] presented OpenMedLM, a prompting platform that enables state-of-the-art performance for open-source LLMs on multiple medical QA benchmarks. By carefully designing prompts, OpenMedLM achieves superior reasoning and accuracy in specialized medical tasks. Ensemble [30] is a technique that combines the responses of multiple models to achieve more powerful and accurate results by combining individual results through functions such as averaging, consensus, or majority voting. During the process of generating responses, the 'temperature' parameter can be used to control the diversity of the model's generated answers to some extent. Additionally, by rearranging or shuffling the few-shot prompt components, it helps address the language model's sequence sensitivity issue, enabling more robust response generation. These studies illustrate the critical role of prompt engineering in improving both the reasoning and accuracy of

LLMs in domain-specific QA tasks.

We introduce MedSumGraph, which can improve the limitations mentioned earlier. The core of our method is an enhanced GraphRAG technique that combines Medical Question Summarization, graph-based context, and optimal prompts to enable LLMs to easily understand knowledge in specialized fields such as medicine without additional training, such as fine-tuning. Medical Question Summarization provides a concise summary of the entire query, enabling a global understanding, while graph-based context facilitates a more specific and deeper understanding through knowledge graphs that explicitly represent structural and relational information. Furthermore, by applying the most effective prompt design method, we have optimized performance not only in the medical domain but also in open domains.

### 3. Methodology

We propose MedSumGraph, a two-phase framework comprising (1) medical knowledge-graph construction (§3.1) and (2) question answering (§3.2). Fig. 2 presents the end-to-end architecture from the input query to the final answer. In Phase 1 (left of Fig. 2), the system extracts medical entity candidates (CUI, name, type, definition) from an external source, UMLS, integrates multi-source evidence using Wikipedia, summarizes salient clinical facts, and induces relations to form triples that populate the MedSumGraph. In Phase 2 (right of Fig. 2), given a user query, MedSumGraph performs hybrid retrieval—local (entity-driven) and global (summary-driven)—followed by embedding-based re-ranking to select a compact evidence subgraph. The selected triples and summaries are then assembled into optimized prompts (dynamic few-shot, chain-of-thought, and ensemble techniques) for the LLM, which generates an answer with cited evidence and a concise reasoning trace.

#### 3.1. Phase 1: knowledge graph construction

Fig. 3 illustrates the overall process of constructing a medical knowledge graph. This process consists of four substeps: (1) medical entity candidate extraction, (2) multi-source knowledge integration, (3) knowledge summarization, and (4) relationship extraction.

##### 3.1.1. Medical entity candidates extraction

To define candidate medical entities to be used as nodes in the knowledge graph, we utilized the Unified Medical Language System (UMLS) Metathesaurus [31]. UMLS is a standardized medical terminology system developed by the U.S. National Library of Medicine that integrates diverse medical vocabularies and classification systems. In this study, we extracted medical entities  $E = \{e_1, e_2, \dots, e_n\}$  and each entity  $e$  has Concept Unique Identifier (CUI), Name, and Semantic Type:  $e = \{cui, name, type\}$ .

##### 3.1.2. Multisource knowledge integration

In the knowledge integration step, we extracted and integrated definition information from the UMLS [31] and summarized pages with object name information from Wikipedia [32] based on the candidate medical objects defined earlier. To collect the definition information from the UMLS, we retrieved the definition information based on the CUI information of the medical term. We used the Wikipedia data, employing the WikipediaRetriever [33] library provided by Langchain, to extract the summary paragraphs for relevant terms. Information from both sources was then integrated into a single context as:  $C_e = Definition_{UMLS} + Definition_{wiki}$ . Thus, all entities are constructed as  $e = \{cui, name, type, context\}$ .

##### 3.1.3. Knowledge summarization

In the knowledge summarization step, we extracted the summarized context information by instructing the  $\mathcal{L}_{sum}^G$  which is  $\mathcal{L}^G$  with sum-

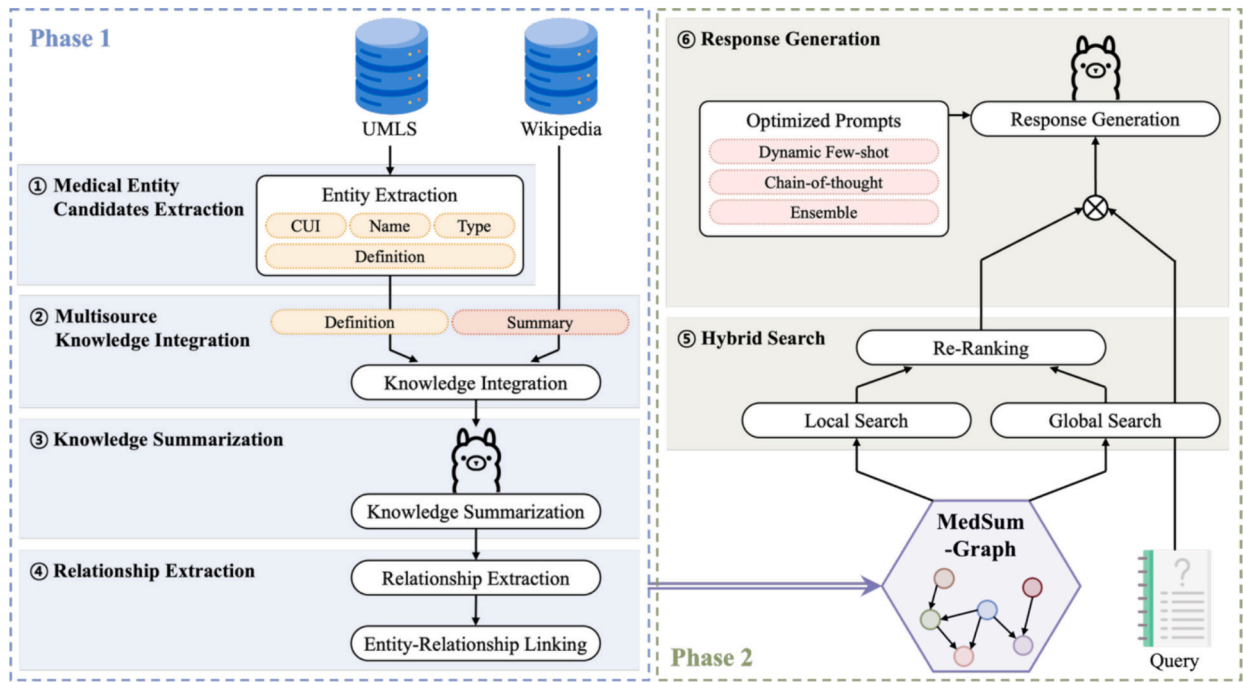
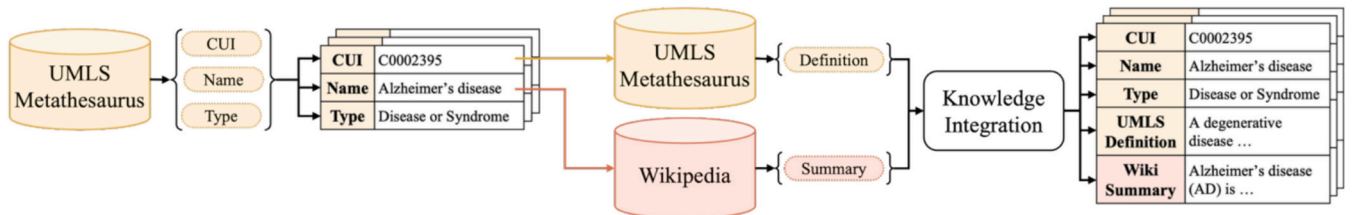


Fig. 2. System Overview of MedSumGraph.

① Medical Entity Candidates Extraction

② Multisource Knowledge Integration



③ Knowledge Summarization

④ Relationship Extraction

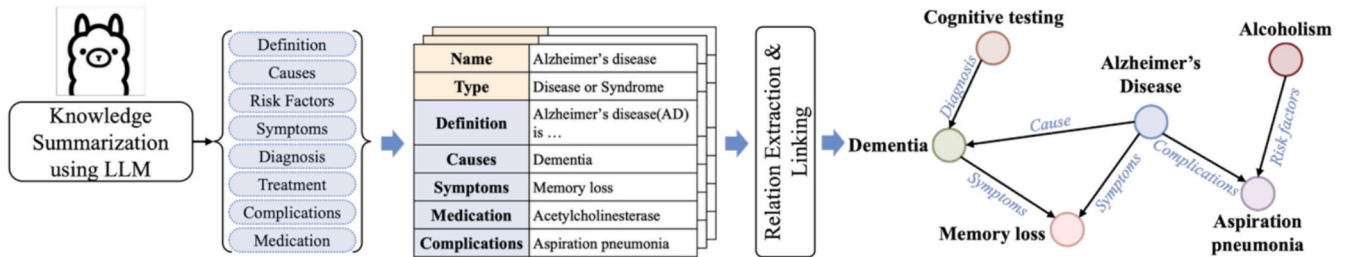


Fig. 3. The medical knowledge graph construction phase of MedSumGraph.

marization prompt to extract only the key information from the long-sentence context information collected for each medical term. This step is necessary because the collected context is considerably large; therefore, it is necessary to select only the important information required to build a knowledge graph. Other studies have utilized retrieved documents for graph construction without refining them. However, the larger the document size, the more unnecessary entity and relationship information may be extracted, and providing unnecessarily large graph information as prompt input to the LLM in the response generation stage may degrade the performance of the LLM and waste resources for graph construction [34,35]. Therefore, in this study, we summarized the collected context documents to include only core information using LLM to efficiently build medical knowledge graphs. We

defined the core information for medical terms as definitions, causes, risk factors, symptoms, diagnosis, treatment, complications, medication, and design prompts for the LLM  $\mathcal{L}_{sum}^G$  to extract the above information as:  $C_{sum} = \mathcal{L}_{sum}^G(C_e)$ .

3.1.4. Relationship extraction

The relationship extraction step instructs the  $\mathcal{L}_{rel}^G$  to extract medical term objects and inter-object relationship information from the summarized medical term context information  $C_{sum}$ . As the summarized context information generated in the previous step has already excluded the core information associated with each medical term, the relationships extracted in this step include additional information such as the

definition, cause, and symptoms of the medical term:  $R_{e_j}^{e_i} = \{e_i^1, r_{ij}, e_j^1 \mid r_{ij} = \mathcal{L}_{rel}^G(C_{sum}^i)\}$ . For example, the knowledge graph related to Alzheimer’s is as follows: {Alzheimer’s disease, *cause*, dementia}, {Alzheimer’s disease, *complications*, aspiration pneumonia}, {Alzheimer’s disease, *symptoms*, memory loss}. This graph visually represents the various aspects of Alzheimer’s disease, illustrating the relationships between its definition, causes, symptoms, and treatment.

Unlike GraphRAG [13], we do not perform an explicit graph community construction step. While community-based graph summarization is effective in general domains, it incurs a high computational cost. We found that medical domain data is typically well-structured and concise, making explicit community construction less necessary. Instead, we leverage global search using question summarization in the response generation phase (Section 3.2), thereby achieving similar benefits while simplifying the pipeline and reducing system complexity.

### 3.2. Phase 2: response generation

As illustrated in Fig. 4, the response generation phase in MedSumGraph comprises two core components: (1) hybrid knowledge retrieval, and (2) response generation using an optimized, medical-specific prompt. This hybrid process allows the system to effectively leverage both textual summaries and structured graph data when generating answers to medical queries.

#### 3.2.1. Hybrid search via global search and local search with Re-ranking

To support precise and context-rich responses, we extend the standard GraphRAG response generation approach with a hybrid search mechanism:

- ⑩ **Global Search** is an approach for understanding questions from a high-level perspective, where a given question is summarized by LLM, and a response is generated based on the implied knowledge representation. The advantage of this approach is that the LLM provides a broader and more comprehensive understanding of the query.
- ⑩ **Local Search** is a method to break down a complex query and understand it. Given a query, it extracts medical entity names and generates responses by obtaining graphs related to the entity names in the query, based on the extracted entity names.

A global search provides a comprehensive understanding of a query from a broader perspective, whereas a local search offers a detailed understanding of complex queries. By combining these two methods, MedSumGraph enhances both the breadth and depth of understanding, enabling the model to interpret complex clinical questions from multiple perspectives and produce more accurate and comprehensive answers. Moreover, we applied re-ranking technology to select only triples similar to the given query. The reasons for needing re-ranking technology are: (1) preventing LLM performance degradation due to irrelevant triples, (2) saving context size to effectively convey more qualitative information, and (3) reducing irrelevant facts or potential errors to generate consistent answers. We utilized the SentenceTransformer (sentence-transformers/all-MiniLM-L6-v2) model to embed both the query and the extracted triples, then measured cosine similarity to select only the top 20 triples most relevant to the query.

#### 3.2.2. Optimized prompts for the medical domain

To further enhance performance, we incorporate medical-specific prompting strategies inspired by MedPrompt [24], which demonstrates that carefully designed prompts can significantly boost LLM performance on domain-specific tasks without fine-tuning. Our prompt design integrates three key techniques:

- ⑩ **Dynamic Few-Shot Prompting:** For each query, we retrieve the five most semantically similar examples from a labeled medical QA dataset using cosine similarity over embeddings. These examples are dynamically inserted into the prompt as few-shot demonstrations, tailoring the context to each new question. This approach differs from the original MedPrompt, which uses k-NN clustering.
- ⑩ **Chain-of-Thought Reasoning:** To improve interpretability and accuracy on multi-step medical questions, we apply a structured reasoning format in the prompt. This encourages the model to generate step-by-step logical explanations before producing an answer.
- ⑩ **Ensembling:** To mitigate variability in LLM outputs, we use an ensemble strategy in which the model generates multiple candidate responses, and the final answer is chosen based on majority voting across outputs.

These prompt engineering techniques collectively enable high-quality medical question answering while avoiding the computational overhead of fine-tuning. Our prompting method performs effectively in

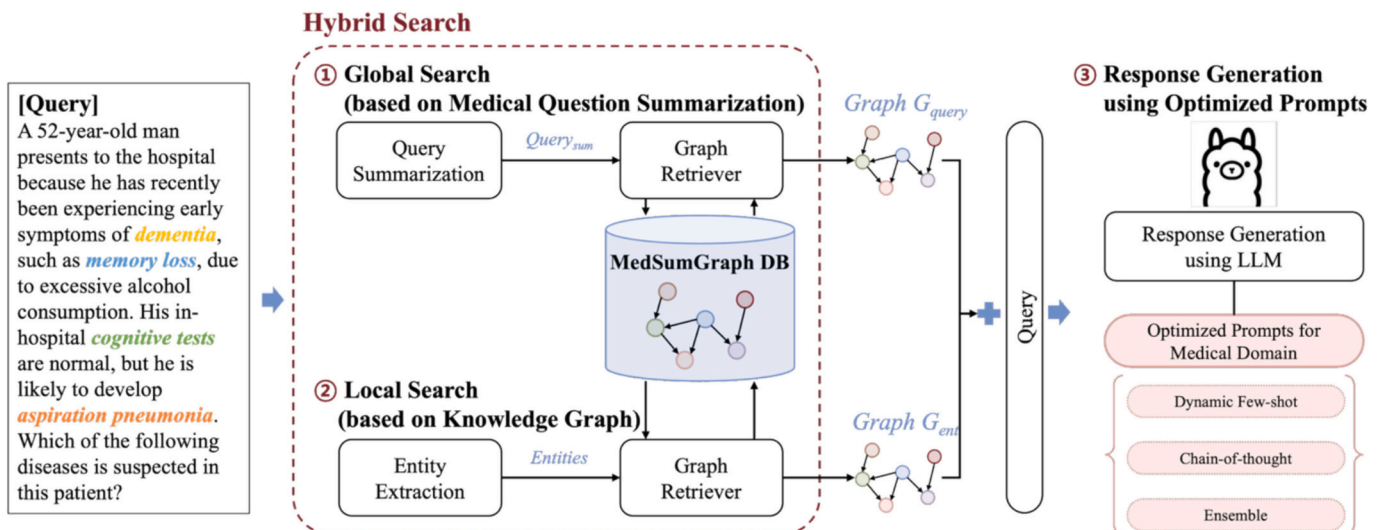


Fig. 4. The response generation phase of MedSumGraph.

specialized domains, reducing the time and expensive computing resources required for fine-tuning.

## 4. Experiments

### 4.1. Dataset

#### 4.1.1. Knowledge graph construction datasets

To build our own knowledge graph specialized in medical data, we collected data from the Unified Medical Language System (UMLS) [31], a unified medical terminology system, and Wikipedia [32], which contains an enormous amount of knowledge data contributed by global users. UMLS information was obtained from the Concept Unique Identifier (CUI), Semantic Network, and Specialist Lexicon, which together accounted for approximately 3.42 million medical terms. Among these, we extracted and used only the Unique Identifier, Name, Semantic Type, and Definition information. The Wikipedia data collection was also based on 3.42 million medical terms extracted from UMLS.

We used Langchain’s WikipediaRetriever API [33] to extract summarized information from Wikipedia pages. Because Wikipedia data encompasses information from various domains, including medical data, it can enhance the performance of medical and open-domain QA tasks. The performance improvement is demonstrated in the open-domain experimental results. However, the WikipediaRetriever in LangChain operates by querying the Wikipedia API in real time to retrieve documents. In other words, the retriever directly accesses and returns the most up-to-date Wikipedia content available through the API, rather than relying on pre-downloaded or fixed-date snapshots (dumps). While this characteristic ensures that the system can immediately reflect the most recent information, it also implies that the results for the same query may vary over time, which introduces limitations in reproducibility.

#### 4.1.2. Test data

The test data used in our experiments consisted of eight multiple-choice medical QA datasets, one LLM reliability validation dataset, and one open-domain dataset, as presented in Table 1. The medical QA dataset comprises MedQA [4], PubMedQA [36], and six MMLUs [37]. The LLM reliability validation dataset includes TrustfulQA [38]. Additionally, the open-domain dataset encompasses OpenbookQA [39], published by AllenAI. The evaluation metric was accuracy, and all datasets were collected using the Huggingface Dataset library.

**Table 1**  
Test data.

Category	Name	Count	Description
Medical QA	MedQA	1273	US Medical licensing exam
	PubMedQA	500	Closed-domain question answering
	MMLU College-Medicine	173	College medicine
	MMLU College-Biology	144	College biology
	MMLU Professional-medicine	272	Professional medicine
	MMLU Anatomy	135	Anatomy
	MMLU Genetics	100	Medical genetics
	MMLU Clinical Knowledge	265	Clinical knowledge
Open-domain QA	OpenbookQA	500	Open-domain question answering
LLM reliability validation	TrustfulQA	790	Measuring the reliability of LLMs

### 4.2. Experiment setting

#### 4.2.1. Medical QA setup

We compared our proposed method with two open-source baselines: standardRAG [12] and GraphRAG [13] developed by Microsoft. The objective of this experiment is to evaluate the performance of LLMs in answering complex medical queries using different types of external knowledge integration and prompt designs. Effective prompt construction and knowledge conditioning are crucial for enhancing LLM comprehension in the medical domain.

To this end, we have tested various combinations of contextual inputs: (1) structured medical knowledge graphs, and (2) medical question summarizations with patient-specific information. These contexts are delivered through different prompt templates and reasoning strategies. Table 2 outlines the specific types of LLM contexts (knowledge graph, summarization) and MedPrompt techniques (few-shot, chain-of-thought, ensembling) [24] applied to each experimental condition.

Fig. 5 presents the prompt templates used in our experiments. Fig. 5 (a) shows the baseline prompt template, which simply appends the question and multiple-choice options. In contrast, Fig. 5(b) shows our enhanced template used in MedSumGraph. It integrates dynamic few-shot examples, contextual summaries, and chain-of-thought prompting, followed by a final answer generation step. The shaded components highlight the additional reasoning and adaptation capabilities not present in the baseline.

We have utilized Llama3.1-70B [40] with 4-bit quantization using Ollama library [41], an open-source LLM, for response generation across all settings on two Nvidia Titan GPUs. This quantization model is lightweight and can also operate on a single GPU with 20-25GB of VRAM, providing greater flexibility in hardware requirements. For MedSumGraph, we use the model for both response generation and knowledge graph construction, while the SentenceTransformer library (sentence-transformers/all-MiniLM-L6-v2) [42] is used to encode all questions and select the top 5 examples for few-shot prompting. We have set the ensemble size to 5 across all CoT generations to increase answer robustness. All hyperparameters can be found in Appendix A.

#### 4.2.2. LLM reliability validation and open-domain QA setup

An LLM reliability validation experiment was conducted to demonstrate the reliability of the answers generated by the proposed model. An open-domain QA experiment was conducted to prove the effectiveness of our model in both the medical and open domains. The experiments were conducted using the Llama3.1-70B model and compared with the Llama3.1-70B baseline and MedSumGraph. Notably, the TrustfulQA [38] dataset is not suitable for few-shot prompts; therefore, we did not apply prompts in our experiments. As in the medical QA experiment, we utilized Llama3.1-70B for knowledge graph construction and the SentenceTransformer library for the embedding model. For knowledge graph construction, we utilized only Wikipedia data to build the knowledge graph, as UMLS is specialized in the medical domain among existing knowledge graph sources.

## 4.3. Results

### 4.3.1. Medical QA benchmark results

In this study, we focused on medical QA experiments and compared

**Table 2**  
Context and prompting methods used in each model.

Category	LLM Context type		MedPrompt		
	RAG	KG	Few-shot	CoT	Ensembling
Baseline	-	-	-	-	-
RAG	✓	-	-	-	-
GraphRAG	-	✓	-	-	-
MedSumGraph	✓	✓	✓	✓	✓

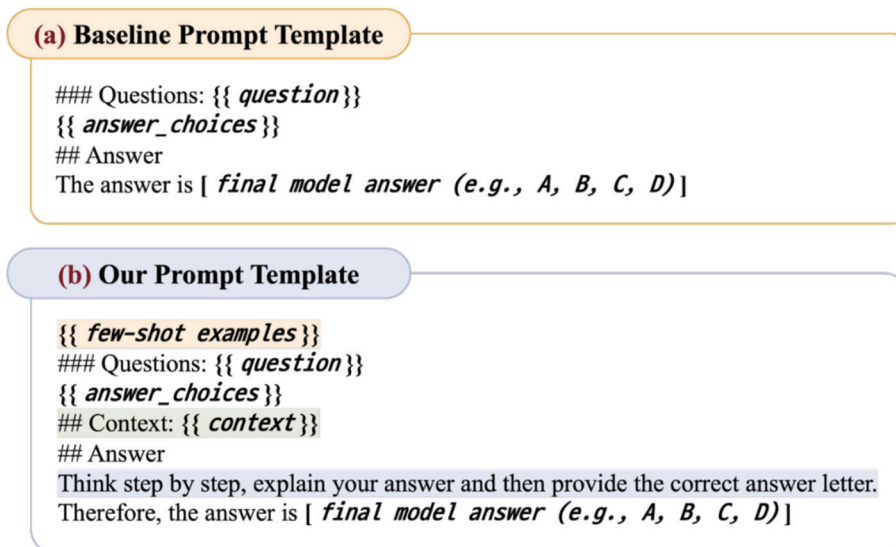


Fig. 5. Prompt templates for medical QA: (a) Baseline template, (b) MedSumGraph prompt.

them for the medical domain without any training. We compared our proposed model with standard RAG and GraphRAG methods on eight representative medical QA benchmark datasets and validated their performance on open-source language models, such as Llama3 (8B, 70B).

The results are presented in Table 3, demonstrating that MedSumGraph applied to the Llama3.1-70B model had the highest performance, with an average performance improvement of 25 percentage points (pps) compared to the baseline model. The highest-performing Llama3.1-70B model achieved performance gains of 17 pps, 14 pps, and 11pps compared to the Baseline, RAG, and GraphRAG models, respectively. This result suggests that the untuned model requires additional external knowledge to effectively understand problems in specialized domains such as the medical domain, further demonstrating that performance varies depending on how external knowledge is integrated into the LLM.

We also found that the performance of the model with a small number of RAGs was comparable to that of the baseline model (as indicated by the red text in Table 3). This suggests that representing information as a graph that captures rich structural relational information is more effective when the context is drawn from text that is not relevant to the query. We additionally present the statistical analysis results for the baseline model, RAG, GraphRAG, and MedSumGraph on the MedQA dataset in Table 7 of Section Appendix B, where each model is evaluated on five different random seeds.

#### 4.3.2. Comparison results with the SOTA model

We presented comparative experiments with medical SOTA models, including GPT-4 [1] and MedPaLM2 [6], and the results are shown in

**Table 3**  
Medical QA benchmark results by method. The best accuracy is highlighted in bold.

Method (Model)	MedQA	Pub MedQA	MMLU Col-Med	MMLU Col-Bio	MMLU ProMed	MMLU Ana	MMLU Gen	MMLU Clinic	Average
<b>Baselines</b>									
(Llama3-8B)	53.97	59.20	63.01	73.61	75.74	67.41	65.00	66.42	65.55
(Llama3-70B)	70.78	61.40	73.41	90.97	87.13	78.52	85.00	79.62	78.35
<b>RAG</b>									
(Llama3-8B)	56.09	64.20	67.63	70.14	76.84	70.37	72.00	70.19	68.43
(Llama3-70B)	72.43	65.20	77.46	93.06	86.76	78.52	91.00	81.13	80.70
<b>GraphRAG</b>									
(Llama3-8B)	64.81	66.40	68.79	81.94	78.31	74.81	74.00	73.58	72.83
(Llama3-70B)	77.06	66.80	80.92	94.44	89.34	82.96	88.00	82.64	82.77
<b>MedSumGraph</b>									
(Llama3-8B)	84.76	76.20	88.44	93.75	91.18	85.93	90.00	90.19	87.56
(Llama3-70B)	<b>93.57</b>	<b>84.20</b>	<b>90.17</b>	<b>95.83</b>	<b>93.75</b>	<b>90.37</b>	<b>91.00</b>	<b>94.72</b>	<b>91.70</b>

Table 4. Four models (MedGraphRAG-GPT4, MedGraphRAG-Llama3.1-70B, MedPrompt-GPT4, and MedPaLM2) published in three papers (MedGraphRAG [22], MedPrompt [24], and Med-PaLM2 [6]) were selected for comparison. The results indicated that our proposed model achieved competitive performance on the MedQA and PubMedQA datasets and that MedPrompt-GPT4 had the highest performance on the other datasets. Furthermore, when averaging the performance of all eight datasets, our proposed model ranked third, behind MedPrompt-GPT4 and MedGraph-GPT4.

However, as noted in the Med-PaLM2 [6], the comparison with GPT-4 is not straightforward, given that GPT-4 is a proprietary system and the extent of overlap between the evaluation data and its training data remains unknown. Nevertheless, our proposed model offers a novel and efficient method for incorporating external knowledge into LLMs without requiring expensive language model fine-tuning. In experiments using the same language model (Llama3.1-70B) except GPT4, which is a commercial language model, our proposed model demonstrated superior performance, even outperforming MedPaLM2, which is fine-tuned for medical tasks.

#### 4.3.3. LLM reliability and open-domain experiment results

This experiment was conducted to demonstrate that the proposed model is effective in both the medical and open domains, as well as LLM reliability tasks. The results are presented in Table 5. The results indicated that the accuracy increased by approximately 7.8 pps in the open-domain QA task and by 9.62 pps in the LLM reliability task compared to the baseline.

This performance improvement is likely attributed to the impact of

**Table 4**

Results compared to SOTA model. The best accuracy is highlighted in bold.

Model	MedQA	Pub MedQA	MMLU Col-Med	MMLU Col-Bio	MMLU Pro-Med	MMLU Ana	MMLU Gen	MMLU Clinic
Med-GraphRAG (GPT4)	90.20	82.00	89.00	97.90	95.20	89.60	98.00	95.80
Med-GraphRAG (Llama3-70B)	88.40	83.80	91.40	96.50	93.20	89.80	91.00	94.10
MedPrompt (GPT4)	91.30	83.30	<b>91.50</b>	<b>98.10</b>	<b>95.80</b>	<b>93.20</b>	<b>98.50</b>	<b>96.40</b>
MedPaLM2	86.50	81.80	83.20	95.80	95.20	84.40	92.00	88.70
MedSumGraph (Llama3-70B)	<b>93.57</b>	<b>84.20</b>	90.17	95.83	93.75	90.37	92.00	94.72

**Table 5**

LLM reliability and open-domain experiment results. The best accuracy is highlighted in bold.

Category	Dataset	Model	Accuracy
Open-domain QA	OpenbookQA	Baselines	85.2 %
		MedSumGraph	<b>93.0 %</b>
LLM Reliability	TrustfulQA	Baselines	71.0 %
		MedSumGraph	<b>80.6 %</b>

the contextual information obtained from Wikipedia and the well-designed MedPrompt, which was used to build the knowledge graph in this study. Wikipedia is a suitable external knowledge source, as it contains data from various domains and can be easily extended to other domains. According to the MedPrompt [24], we believe that MedPrompt’s prompt design method is not only effective in the medical field but can also be applied to other domains, which significantly influences the performance of open-domain QA and the reliability of LLMs.

## 5. Case study

In this section, we demonstrated the MedSumGraph model for specialized medical queries using a medical QA case, as shown in Fig. 6. We present the case of a 62-year-old stroke patient who developed a purulent cough and cold one week after being admitted to the hospital. The clinical task is to identify the causative microorganisms responsible for the patient’s symptoms. The correct answer to this question is *D. Staphylococcus aureus*. Identifying the correct answer involves guessing the disease based on the patient’s symptoms and determining the causative organism for the given disease. In the case of the patient shown in Fig. 6, the following reasoning process is required: ‘symptoms of purulent cough and cold → pneumonia’, ‘pneumonia developed 1 week after hospitalization → hospital-acquired pneumonia (HAP)’, ‘causative microorganism of HAP → *Staphylococcus aureus*’.

We described how the proposed model addresses this problem. First, we used medical question summarization to summarize the patient’s information, symptoms, test results, and diagnosis. The summary states that the probable diagnosis is HAP and the causative organism is *Streptococcus pneumoniae*. Consequently, we observed that the inference regarding the diagnosis (HAP) is correct; however, the inference regarding the causative organism (*Streptococcus pneumoniae*) is incorrect. The graph shows that *Staphylococcus aureus* is the cause of HAP and *Streptococcus pneumoniae* is the cause of community-acquired pneumonia (CAP). Therefore, by synthesizing all the contextual information, the correct diagnosis is inferred from the query summary.

However, the incorrect cause of the symptoms is predicted, and the additional graph information, which indicates that *Staphylococcus aureus* causes HAP, can be used to correct this error and finally derive the correct answer. Thus, in a field that requires complex terminology and reasoning skills, such as medical QA, the proposed hybrid method of medical query summarization and graph-based retrieval is effective in providing a multifaceted approach to professional medical queries from both global and local perspectives while incorporating relevant external knowledge into LLMs. Specifically, fact-based structural relationship information in the knowledge graph can support the reasoning basis of the LLM. This suggests that the knowledge graph can mitigate the black-

box and hallucination issues associated with LLM, thereby facilitating the development of more reliable and explainable medical AI.

## 6. Ablation study

In this section, we analyze the contribution of the main components of MedSumGraph as shown in Fig. 7. Experiments are conducted on MedQA with Llama-3.1-70B. The left panel of Fig. 7 compares Baseline, RAG, and GraphRAG in terms of Accuracy (%). The right panel reports incremental gains on the Baseline as we progressively add the main methods used in MedSumGraph. Bars show absolute accuracy; stacked segments indicate percentage-point (%p) improvements, and the red line tracks the cumulative score. Starting from a 70.78 % baseline, summarization yields a 6.83 % increase, GraphRAG adds 8.72 %, and MedPrompt contributes 7.24 %, reaching an overall 93.57 %. These results indicate that external, query-relevant knowledge is crucial for medical QA, and graph-based retrieval that encodes relational structure yields substantial gains, while prompt optimization further consolidates these improvements.

## 7. Discussions

The use of large language models (LLMs) in healthcare is expanding rapidly, with applications in clinical decision support [2] and medical question answering [3]. However, significant challenges remain before these models can be safely and reliably applied in real-world clinical environments [4]. Studies have found that while general-purpose LLMs can handle some medical queries, they often fail when tasked with understanding specialized terminology or performing complex reasoning. These findings highlight the need for domain-specific strategies that enhance the transparency and factual reliability of LLM-generated responses.

One key challenge is the reliance on external knowledge sources. Retrieval-Augmented Generation (RAG) [12] has emerged as a promising solution by combining LLMs with non-parametric retrieval systems. However, its effectiveness depends heavily on the quality and relevance of the retrieved information. In parallel, prompt-based approaches, such as MedPrompt [24], have demonstrated that LLMs can be adapted to specialized domains without requiring fine-tuning. Nonetheless, these strategies alone may not be sufficient to ensure reliable or interpretable outputs in high-stakes fields such as medicine.

Our proposed method addresses these limitations by integrating knowledge graph construction, summarization, and optimized prompting. However, this study has several limitations. First, clinical knowledge evolves rapidly, making it challenging to maintain and update the medical knowledge graph, which may affect the generalizability of the proposed approach. Second, the effectiveness of cross-domain summarization may vary depending on the specific medical context, potentially leading to inconsistent performance across different tasks. Third, the framework is sensitive to the order of input context, which may influence the quality of the generated responses. Fourth, the system requires substantial computational resources, including the Llama3.1-70b model and knowledge graph construction, which may limit real-time applicability and cost-effectiveness in clinical settings. Fifth, proposed model not only demonstrates strong performance on medical QA benchmarks such as MedQA, but also shows improved results in open-domain QA

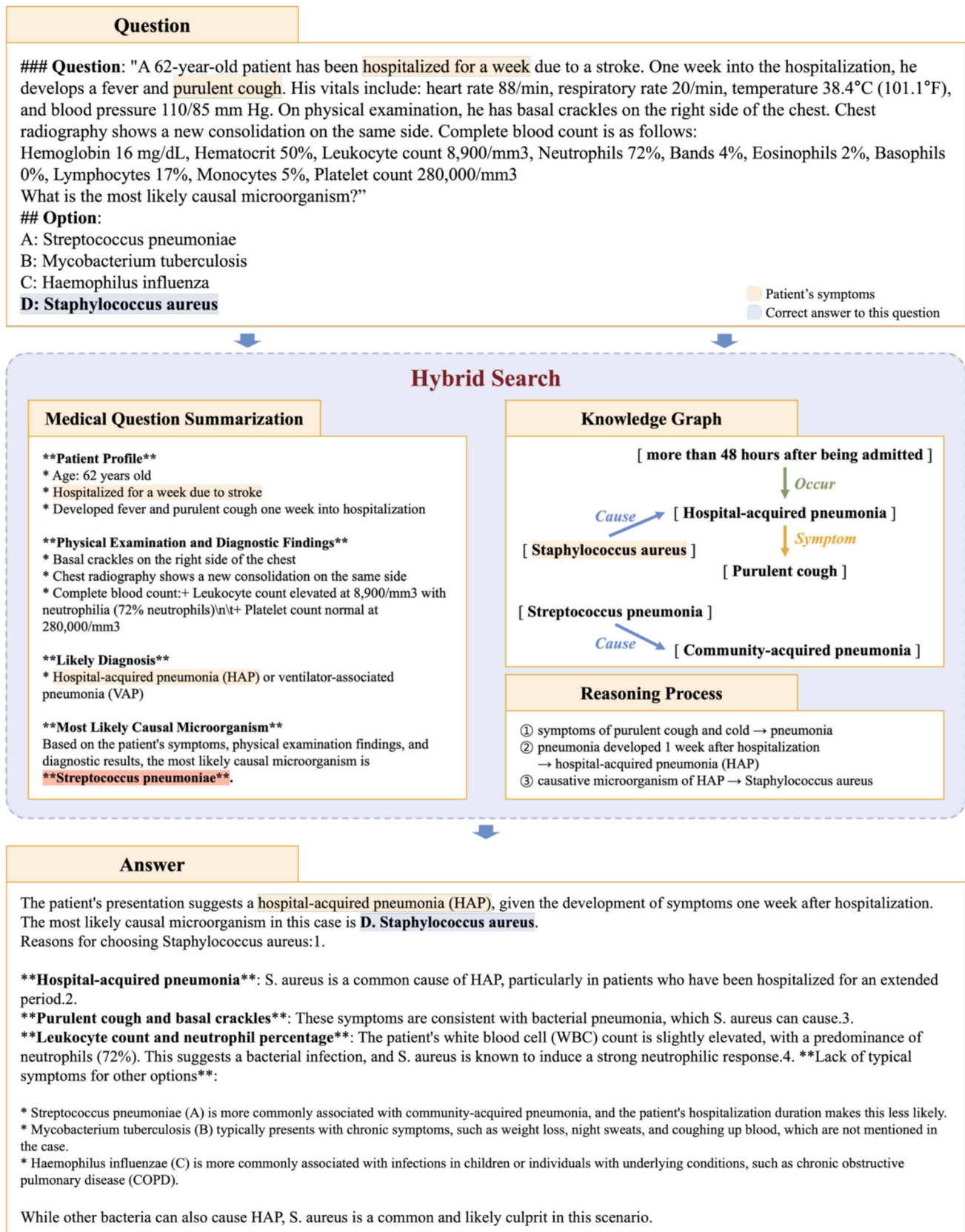


Fig. 6. Example of generating answers in MedSumGraph.

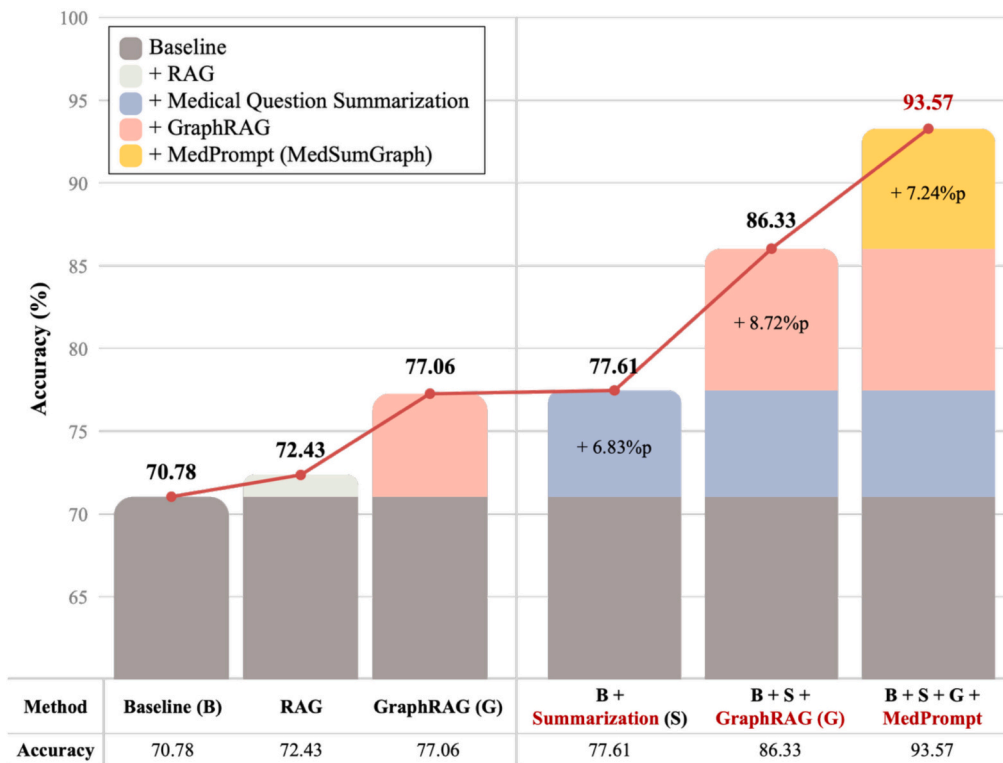


Fig. 7. Performance improvement by component in MedSumGraph.

tasks. However, these improvements may primarily stem from the integration of Wikipedia-derived knowledge in a graph-structured form, suggesting the need for additional ablation studies and more detailed analyses. Nevertheless, we argue that our approach—retrieving knowledge from Wikipedia most relevant to the given queries in open-domain QA, summarizing the inherently noisy content, and then structuring it into a knowledge graph for integration into the LLM’s context—likely contributed substantially to the observed performance gains. Future work should explore dynamic knowledge updating mechanisms, evaluate robustness across diverse medical domains, mitigate input order sensitivity, and develop strategies to improve computational efficiency.

Our approach achieves competitive performance on both medical and open-domain QA benchmarks; however, its effectiveness may be limited when noise is introduced into the knowledge graph or when query-relevant triples are not adequately included. To address these limitations, we first developed a domain-specific medical knowledge graph and conducted a framework evaluation on medical QA tasks. Subsequently, we explored the broader applicability of multi-source knowledge integration in open-domain settings by incorporating general-purpose resources such as Wikipedia.

### 8. Conclusion

In this study, we proposed MedSumGraph, a novel hybrid framework that enhances GraphRAG for medical question answering by combining structured knowledge graphs, multi-source medical knowledge summarization, and optimized prompting strategies. Our approach enables general-purpose LLMs to understand domain-specific terminology and respond to complex medical queries without fine-tuning.

We demonstrated that MedSumGraph outperforms existing methods on multiple medical QA benchmarks, as well as in open-domain and factuality-focused evaluations. These results validate the effectiveness of integrating structured and unstructured knowledge for generating

accurate and interpretable LLM outputs.

While our work offers a promising pathway for applying LLMs in healthcare, further research is necessary to improve reliability, transparency, and domain generalizability. We emphasize the importance of developing explainable AI tools that support clinical reasoning and provide users—both clinicians and patients—with fact-based, trustworthy answers. Through this study, we hope to contribute to the advancement of medical AI systems that enhance the quality and accessibility of clinical knowledge.

### CRedit authorship contribution statement

**DaeHo Kim:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **SoYeop Yoo:** Writing – original draft, Visualization, Software, Resources, Methodology, Data curation. **OkRan Jeong:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgements

This research was supported by the Electronics and Telecommunications Research Institute research operation support project (basic project), and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (Nos. 25ZT1100, RS-2025-25443681).

## Appendix A. Hyameters

In this section, we discuss the hyperparameters utilized in the medical benchmark. Hyperparameters are broadly categorized into model and prompt, with details provided in Table 6 below.

**Table 6**  
Hyperparameters.

Category	Hyperparameter	Value	Description
Model	temperature	0.5	Controls randomness; set to deterministic for QA
	top-k	50	Nucleus sampling threshold
	top-p	0.95	Candidate pool size for top-k sampling
	repetition_penalty	1.1	Penalty factor to reduce token repetition
	max_tokens	512	Maximum tokens generated
Prompt	Dynamic few-shot top-k	5	Dynamic Few-Shot Prompting selects the most relevant example for each test query, enabling the model to leverage contextually similar cases and achieve performance gains.
	Ensembling	5	Ensembling is a technique that combines the outputs of multiple model runs to derive more robust or accurate results
	Reranker top-k	20	A variable that selects only the top-k triples most relevant to the query by re-ranking them among countless triples.

## Appendix B. Statistical validation

In this section, we analyze the results of baseline, RAG, GraphRAG, and MedSumGraph (Ours) models on the MedQA dataset, each evaluated over five different random seeds. Beyond simple accuracy, we computed the mean and standard deviation for each model to assess the central tendency and variability of performance. We further calculated the 95 % confidence interval (CI) to quantify the uncertainty around the mean estimates and conducted paired *t*-tests to determine whether the observed differences between baseline and other models were statistically significant. Finally, we computed the effect size (Cohen's *d*) to quantify the magnitude of performance improvements.

The statistical analysis results are summarized in Table 7, demonstrating that the MedSumGraph (Ours) model outperforms the baseline. Examining the mean, standard deviation, and 95 % confidence intervals, MedSumGraph (Ours) exhibits high accuracy with low variability across seeds. Furthermore, the paired *t*-test shows a *p*-value smaller than 0.001, indicating a highly significant improvement, and the Cohen's *d* value confirms a substantial effect size compared to the baseline, supporting consistent performance gains across repeated experiments. These results indicate that the improvements provided by MedSumGraph (Ours) are not due to chance but reflect robust and reliable performance across multiple random seed settings.

**Table 7**  
Statistical validation results in the Medical QA benchmark.

Method	Accuracy (mean $\pm$ std)	95 % CI	p-value	Cohen's <i>d</i>
Baselines	70.024 $\pm$ 0.71	[69.147, 70.901]	–	–
RAG	72.263 $\pm$ 0.14	[72.088, 72.437]	<i>p</i> < 0.01	3.467
GraphRAG	76.996 $\pm$ 0.06	[76.917, 77.075]	<i>p</i> < 0.001	9.926
MedSumGraph	92.996 $\pm$ 0.65	[92.184, 93.809]	<i>p</i> < 0.001	25.052

## Appendix C. Example of framework performing a medical QA task

Table 8 shows examples of which entities and triples the proposed framework generates during the response generation step using local search and global search, respectively. A key point here is that when using only local search (a common research approach), the important entity (“Pulmonary embolism”) needed to match the correct answer was not discovered. However, it can be seen that this entity was included among those extracted by global search. In other words, we propose a method that appropriately fuses both local search and global search techniques, and its effectiveness can be confirmed through the examples in Table 8.

**Table 8**  
Example of our framework performing a Medical Question Answering Task.

Question	A 65-year-old woman comes to the physician for a routine bone mineral density screening. She does not have any children. Menopause was at age 55. Her mother died of breast cancer at the age of 48 years. She has hypertension for which she takes ramipril. Dual-energy x-ray absorptiometry at the femoral neck shows a T-score of $-2.7$ SD. The physician considers treatment with raloxifene. This medication would put her at increased risk of developing which of the following conditions?	
Options	<b>A: Pulmonary embolism (Answer)</b> B: Cellulitis C: Breast cancer D: Hypercholesterolemia	
Method	Local Search	Global Search
Entities	["65-year-old woman", "No children", "Menopause at age 55", "Family history of breast cancer", "Breast cancer", "Hypertension", "Bone mineral density screening", "Dual-energy X-ray absorptiometry", "Femoral neck", "T-score $-2.7$ ", "Ramipril", "Raloxifene"]	[Hypertension, Breast cancer, Dual-energy X-ray absorptiometry (DEXA), T-score, Raloxifene (SERM), Ramipril (ACE inhibitor), Pulmonary embolism (PE)]

(continued on next page)





Table 8 (continued)

Question	A 65-year-old woman comes to the physician for a routine bone mineral density screening. She does not have any children. Menopause was at age 55. Her mother died of breast cancer at the age of 48 years. She has hypertension for which she takes ramipril. Dual-energy x-ray absorptiometry at the femoral neck shows a T-score of -2.7 SD. The physician considers treatment with raloxifene. This medication would put her at increased risk of developing which of the following conditions?	
Options	<b>A: Pulmonary embolism (Answer)</b> B: Cellulitis C: Breast cancer D: Hypercholesterolemia	
Method	Local Search	Global Search
Triples	("Hypertension", "treated_with", "Ramipril"), ("Bone mineral density screening", "includes", "Dual-energy X-ray absorptiometry"), ("Dual-energy X-ray absorptiometry", "site", "Femoral neck"), ("T-score - 2.7 SD", "indicates", "Osteoporosis"), ("Raloxifene", "benefit", "Osteoporosis treatment & breast cancer risk reduction")	("Osteoporosis", "defined_as", "Low bone mineral density") ("Hypertension", "treated_with", "Ramipril"), ("Breast cancer", "risk_factor", "Family history") ("Osteoporosis", "diagnosed_by", "DEXA"), ("DEXA", "result", "T-score") ("Osteoporosis", "treated_with", "Raloxifene"), ("Raloxifene", "class", "SERM") ("Raloxifene", "reduces_risk", "Fragility fractures") ("Osteoporosis", "complication", "Fragility fractures") ("Raloxifene", "complication", "Venous thromboembolism") ("Raloxifene", "risk_increase", "Pulmonary embolism")
Visualization		
Question	A 4-year-old child is rushed to the emergency department after developing sudden abdominal pain followed by nausea, vomiting, and dark, almost black-colored stool. Prior to the onset of symptoms, he was seen playing with his mother's purse containing a bottle of vitamin supplements, which she takes for chronic microcytic anemia. Which of the following medications is the treatment for this patient's intoxication?	
Options	<b>B: Deferoxamine (Answer)</b> A: Dimercaprol C: Protamine D: Succimer	
Method	Local Search	Global Search
Entities	["4-year-old child", "Sudden abdominal pain", "Nausea", "Vomiting", "Dark, almost black-colored stool", "Emergency department", "Vitamin supplements", "Chronic microcytic anemia", "Iron", "Intoxication / Iron poisoning", "Deferoxamine"]	["Sudden abdominal pain", "Nausea", "Vomiting", "Dark, almost black-colored stool / Melena", "Vitamin supplements", "Iron", "Acute iron intoxication", "Deferoxamine", "Gastrointestinal bleeding", "Metabolic acidosis", "Organ damage", "Shock"]
Triples	("Child", "exposed_to", "Vitamin supplements") ("Vitamin supplements", "contains", "Iron") ("Child", "condition", "Iron intoxication") ("Iron intoxication", "treated_with", "Deferoxamine")	("Child", "exposed_to", "Vitamin supplements") ("Vitamin supplements", "contains", "Iron") ("Iron overdose", "causes", "Acute iron intoxication") ("Acute iron intoxication", "diagnosed_by", "Clinical history and symptoms") ("Acute iron intoxication", "treated_with", "Deferoxamine") ("Acute iron intoxication", "may_cause", "Gastrointestinal bleeding")
Visualization		

### Appendix D. Qualitative error analysis

In this section, we present the results of qualitative error analysis that can occur in medical QA, as shown in Table 9. To perform qualitative error analysis, we analyzed Missing information, Factual Error, and Logical/Reasoning Error using one representative example. Missing information refers to cases where no triples were retrieved from the knowledge graph. Factual Error occurs when triples containing incorrect information are retrieved. Logical/Reasoning Error occurs when appropriate triples are retrieved but the LLM fails to understand them correctly. The results observed through qualitative error analysis indicate that our proposed method contains sufficient triples for medical QA and refines triples similar to the query. This enabled us to achieve competitive performance across various medical benchmarks.

**Table 9**  
Qualitative error analysis.

Question	A 62-year-old patient has been hospitalized for a week due to a stroke. One week into the hospitalization, he develops a fever and purulent cough. His vitals include: heart rate 88/min, respiratory rate 20/min, temperature 38.4 °C (101.1 °F), and blood pressure 110/85 mmHg. On physical examination, he has basal crackles on the right side of the chest. Chest radiography shows a new consolidation on the same side. Complete blood count is as follows: Hemoglobin 16 mg/dL Hematocrit 50 % Leukocyte count 8900/mm3 Neutrophils 72 % Bands 4 % Eosinophils 2 % Basophils 0 % Lymphocytes 17 % Monocytes 5 % Platelet count 280,000/mm3 What is the most likely causal microorganism?			
Options	A:Streptococcus pneumoniae B:Mycobacterium tuberculosis C:Haemophilus influenzae D:Staphylococcus aureus			
Gold Answer	D:Staphylococcus aureus			
Context (Triple)	-	{Streptococcus pneumoniae, cause, Pneumonia}	{Staphylococcus aureus, cause, Hospital-acquired pneumonia}	{Staphylococcus aureus, cause, Hospital-acquired pneumonia}
Error Type	Missing Information	Factual Error	Logical/Reasoning Error	Correct (Our)
Predict	A:Streptococcus pneumoniae	A:Streptococcus pneumoniae	A:Streptococcus pneumoniae	D:Staphylococcus aureus
				
Analysis	Difficulty distinguishing HAP from CAP due to insufficient information such as antibiotic use history, immune status, and mechanical ventilation status	An LLM generates an answer based on a triple of incorrect information and it may select <i>Streptococcus pneumoniae</i> instead of <i>Staphylococcus aureus</i> or <i>Pseudomonas</i> , which are common pathogens in hospital-acquired pneumonia (HAP).	Despite providing the LLM with a triple containing relevant information as context, it failed to capture the patient’s hospitalization duration expressed in the query and instead selected “the most common CAP pathogen” → ignoring that it was hospital-acquired pneumonia (HAP).	The proposed model receives context containing relevant external knowledge related to the query from our knowledge graph, and the LLM accurately determines the correct answer by thoroughly understanding and evaluating the given context.
Qualitative evaluation	Accuracy ↓ Factual consistency ↓	Accuracy ↓ Factual consistency ↓	Accuracy ↓ Logic/Reasoning Quality ↓	Accuracy ↑ Factual consistency ↑ Logic/reasoning quality ↑

### References

[1] GPT-4 Technical Report. arXiv.Org. 2023. <https://doi.org/10.48550/arXiv.2303.08774>.

[2] Brown T, et al. Language models are few-shot learners. *Adv Neural Inf Proces Syst* 2020;33:1877–901.

[3] Clusmann J, et al. The future landscape of large language models in medicine. *Commun Med* 2023;3(1):141.

[4] Jin D, et al. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl Sci* 2020;11(14):6421. <https://arxiv.org/abs/2009.13081>.

[5] Sutton RT, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3(1):17. <https://www.nature.com/articles/s41746-020-0221-y>.

[6] Thirunavukarasu AJ, Ting DSJ, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(8):1930–40. <https://doi.org/10.1038/s41591-023-02448-8>.

[7] Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234–40. <https://doi.org/10.1093/bioinformatics/btz682>.

[8] Gu Y, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc* 2021;3(1):1–23. <https://doi.org/10.1145/3458754>.

[9] Singhal K, et al. Toward expert-level medical question answering with large language models. *Nat Med* 2025;1-8. <https://www.nature.com/articles/s41591-024-03423-7#article-info>.

[10] Kirkpatrick J, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci* 2017;114(13):3521–6. <https://doi.org/10.1073/pnas.1611835114>.

[11] Huang L, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst* 2025;43(2):1–55. <https://arxiv.org/abs/2311.05232>.

[12] Lewis P, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inf Proces Syst* 2021;33:9459–74. <https://arxiv.org/abs/2005.11401>.

[13] Edge D, et al. From local to global: a graph rag approach to query-focused summarization. <https://arxiv.org/abs/2404.16130>; 2025.

[14] Singhal Karan, et al. Large language models encode clinical knowledge. *Nature* 2023;620(7972):172–80.

[15] Yagnik Niraj, et al. Medlm: exploring language models for medical question answering systems. In: *arXiv preprint arXiv:2401.11389*; 2024.

[16] Saab Khaled, et al. Capabilities of gemini models in medicine. In: *arXiv preprint arXiv:2404.18416*; 2024.

- [17] Zhang X, Li Y, Chen L, Liu X. Biomedical retrieval-augmented generation for medical question answering. In: arXiv preprint arXiv:2401.01234; 2024. <https://arxiv.org/abs/2401.01234>.
- [18] Chen Zhuo, et al. Improving retrieval augmented open-domain question-answering with vectorized contexts. In: arXiv preprint arXiv:2404.02022; 2024.
- [19] Fensel Dieter, et al. Introduction: what is a knowledge graph?. In: Knowledge graphs: methodology, tools and selected use cases; 2020. p. 1–10.
- [20] Yasunaga M, Ren H, Bosselut A, Liang P, Leskovec J. QA-GNN: reasoning with language models and knowledge graphs for question answering. In: Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies (NAACL-HLT); 2021. p. 4316–28. <https://aclanthology.org/2021.naacl-main.45/>.
- [21] Sohn Jiwoong, et al. Rationale-guided retrieval augmented generation for medical question answering. In: arXiv preprint arXiv:2411.00300; 2024.
- [22] Wu Junde, et al. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. In: arXiv preprint arXiv:2408.04187; 2024.
- [23] Sarmah B, et al. Hybridrag: integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction. In: Proceedings of the 5th ACM international conference on AI in finance; 2024. p. 608–16. <https://arxiv.org/abs/2408.04948>.
- [24] Nori H, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. <https://arxiv.org/abs/2311.16452>; 2023.
- [25] Brown Tom, et al. Language models are few-shot learners. *Adv Neural Inf Proces Syst* 2020;33:1877–901.
- [26] Wei Jason, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Proces Syst* 2022;35:24824–37.
- [27] Le Chenqian, et al. Instruction tuning and cot prompting for contextual medical qa with llms. In: arXiv preprint arXiv:2506.12182; 2025.
- [28] Sandeep Nachane Saeel, et al. Few shot chain-of-thought driven reasoning to prompt LLMs for open ended medical question answering. In: arXiv e-prints. arXiv:2403; 2024.
- [29] Maharjan Jenish, et al. OpenMedLM: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Sci Rep* 2024;14(1):14156.
- [30] Wang Xuezi, et al. Self-consistency improves chain of thought reasoning in language models. In: arXiv preprint arXiv:2203.11171; 2022.
- [31] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(suppl\_1):D267–D270. [https://academic.oup.com/nar/article/32/suppl\\_1/D267/2505235](https://academic.oup.com/nar/article/32/suppl_1/D267/2505235).
- [32] Wikipedia. [https://en.wikipedia.org/wiki/Main\\_Page](https://en.wikipedia.org/wiki/Main_Page); 2024.
- [33] WikipediaRetriever. 2023. <https://python.langchain.com/docs/integrations/retrievers/wikipedia/>.
- [34] Guo Kai, et al. Empowering graphrag with knowledge filtering and integration. In: arXiv preprint arXiv:2503.13804; 2025.
- [35] Zeng Qiming, et al. How significant are the real performance gains? an unbiased evaluation framework for GraphRAG. In: arXiv preprint arXiv:2506.06331; 2025.
- [36] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. <https://arxiv.org/abs/1909.06146>; 2019.
- [37] Hendrycks D, et al. Measuring massive multitask language understanding. <https://arxiv.org/abs/2009.03300>; 2021.
- [38] Lin S, Hilton J, Evans O. TruthfulQA: measuring how models mimic human falsehoods. <https://arxiv.org/abs/2109.07958>; 2022 (2021).
- [39] Mihaylov T, Clark P, Khot T, Sabharwal A. Can a suit of armor conduct electricity? a new dataset for open book question answering. <https://arxiv.org/abs/1809.02789>; 2018.
- [40] Touvron Hugo, et al. Llama: open and efficient foundation language models. In: arXiv preprint arXiv:2302.13971; 2023.
- [41] Marcondes Francisco S, et al. “Using ollama.” *natural language analytics with generative large-language models: A practical approach with Ollama and open-source LLMs*. Cham: Springer Nature Switzerland; 2025. p. 23–35.
- [42] Reimers Nils, Gurevych Iryna. Sentence-bert: Sentence embeddings using siamese bert-networks. In: arXiv preprint arXiv:1908.10084; 2019.