From Retrieval to Generation: Unifying External and Parametric Knowledge for Medical Question Answering

Lei Li

Gaoling School of Artificial Intelligence, Renmin University of China Beijing, China leil@ruc.edu.cn

> Yingying Zhang Tencent Jarvis Lab Beijing, China ninzhang@tencent.com

Xiao Zhou*
Gaoling School of Artificial Intelligence, Renmin
University of China
Beijing, China
xiaozhou@ruc.edu.cn

Xian Wu*
Tencent Jarvis Lab
Beijing, China
kevinxwu@tencent.com

Abstract

Medical question answering (QA) requires extensive access to domainspecific knowledge. A promising direction is to enhance large language models (LLMs) with external knowledge retrieved from medical corpora or *parametric knowledge* stored in model parameters. Existing approaches typically fall into two categories: Retrieval-Augmented Generation (RAG), which grounds model reasoning on externally retrieved evidence, and Generation-Augmented Generation (GAG), which depends solely on the model's internal knowledge to generate contextual documents. However, RAG often suffers from noisy or incomplete retrieval, while GAG is vulnerable to hallucinated or inaccurate information due to unconstrained generation. Both issues can mislead reasoning and undermine answer reliability. To address these challenges, we propose MEDRGAG, a unified retrieval-generation augmented framework that seamlessly integrates external and parametric knowledge for medical QA. MEDRGAG comprises two key modules: Knowledge-Guided Context Completion (KGCC), which directs the generator to produce background documents that complement the missing knowledge revealed by retrieval; and Knowledge-Aware Document Selection (KADS), which adaptively selects an optimal combination of retrieved and generated documents to form concise yet comprehensive evidence for answer generation. Extensive experiments on five medical QA benchmarks demonstrate that MedRGAG achieves a 12.5% improvement over MedRAG and a 4.5% gain over MedGENIE, highlighting the effectiveness of unifying retrieval and generation for knowledge-intensive reasoning. Our code and data are publicly available at https://anonymous.4open.science/r/MedRGAG.

CCS Concepts

• Information systems \rightarrow Retrieval models and ranking.

Keywords

Medical Question Answering, Retrieval-Augmented Generation, Generation-Augmented Generation

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a broad spectrum of natural language understanding

and question answering tasks [4, 39, 44, 53]. However, their performance remains constrained by the limitations of internal parametric knowledge, which often results in *hallucinations*—plausible yet factually incorrect generations [17, 20]. These factual inconsistencies pose substantial challenges for medical question answering (QA), a knowledge-intensive task that demands precise reasoning and a high degree of factual reliability. In clinical settings, even minor hallucinations can compromise medical validity, underscoring the necessity for QA systems to generate trustworthy, evidence-grounded responses [15, 51, 55].

To enhance the reliability and factual accuracy of LLM-based question answering, recent research has increasingly emphasized knowledge-augmented generation methods, which ground model reasoning on external knowledge sources [9, 57]. Among these approaches, Retrieval-Augmented Generation (RAG) has emerged as a representative paradigm that follows a retrieval-then-read framework (see Figure 1 (a)). Specifically, RAG first retrieves relevant information from structured or unstructured medical corpora (e.g., PubMed, Wikipedia, and UMLS) and then conditions the LLM's answer generation on the retrieved evidence [5, 49, 56]. By incorporating up-to-date and verifiable documents, RAG effectively grounds the model's reasoning process, thereby reducing hallucinations and enhancing answer transparency [28, 31]. Despite its advantages, the effectiveness of RAG remains constrained by two major factors: (1) retrieved documents are typically chunked into fixed-length passages, often containing noisy or irrelevant information that distracts reasoning [5, 56]; and (2) retrieval alone may fail to provide sufficient knowledge coverage, leaving critical information gaps that hinder accurate medical QA [38, 50].

In parallel with retrieval-based approaches, another line of research explores *knowledge-augmented generation* by exploiting the *parametric knowledge* embedded within LLMs [6, 34]. This paradigm, known as Generation-Augmented Generation (GAG), follows a *generate-then-read* framework (see Figure 1 (b)), in which a generator first produces several contextual documents conditioned on the input question, and a reader subsequently leverages these generated contexts to infer the final answer [14, 41, 52]. This approach eliminates the reliance on external corpora and enables the model to construct query-specific contexts that more precisely align with the semantic intent of the question. Nevertheless, because GAG relies entirely on generated documents as its knowledge source, it

^{*}Xiao Zhou and Xian Wu are corresponding authors.

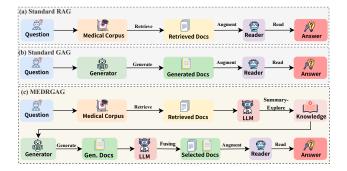


Figure 1: Comparison of MedRGAG with standard RAG and GAG. Figure (a) illustrates RAG, where the retriever extracts external knowledge to enhance the reader's answer generation. Figure (b) depicts GAG, where the generator activates its internal parametric knowledge to assist the reader in producing answers. Figure (c) presents MedRGAG, which enables the retriever and generator to jointly unify external and parametric knowledge, providing comprehensive evidence to enhance the reader's answer reliability.

remains susceptible to hallucinated or inaccurate content within those contexts, which can mislead reasoning and ultimately lead to incorrect answers in question answering [42, 54].

Recent studies have further explored combining retrieved and generated knowledge to harness the complementary strengths of external and parametric sources [12, 54]. These approaches primarily aim to mitigate knowledge conflicts that arise during the fusion of retrieved and generated documents, marking a valuable step toward bridging the two paradigms. Nevertheless, fully addressing the intrinsic limitations of each paradigm, including the restricted knowledge coverage of RAG and the potential overreliance on generation-based contexts in GAG, remains a challenging and unresolved problem in knowledge-intensive medical QA.

To address these challenges, we propose MEDRGAG, a unified retrieval-generation augmented framework that seamlessly integrates external and parametric knowledge for medical question answering. As illustrated in Figure 1 (c), MEDRGAG follows a retrieval-generation-then-read paradigm: it first employs a retriever to obtain relevant documents from large-scale medical corpora, then leverages a generator to produce complementary background documents, and finally adaptively fuses both sources of evidence, enabling the reader model to generate reliable answers. This design allows MedRGAG to unify retrieval and generation in a coherent pipeline, bridging the gap between external and parametric knowledge sources. Specifically, MedRGAG comprises two core modules: (1) Knowledge-Guided Context Completion (KGCC)—this module summarizes and analyzes the retrieved documents to identify missing knowledge required for answering the question, and subsequently guides the generator to produce targeted background documents that complement the retrieved evidence; and (2) Knowledge-Aware Document Selection (KADS)-this module groups retrieved and generated documents according to the knowledge requirements of the question and selects a diverse, comprehensive, and non-redundant subset of evidence for downstream reasoning.

Through the coordinated operation of these two modules, Medragas substantially improves knowledge completeness while alleviating overreliance on hallucination-prone generated contexts, thereby enhancing both factual reliability and reasoning robustness in medical QA.

We conduct extensive experiments on five widely used medical QA benchmarks, including MedQA [22], MedMCQA [32], MMLU-Med [16], PubMedQA* [23], and BioASQ [45]. We quantitatively compare performance against representative RAG-based methods such as MedRAG [49] and GAG-based methods such as MedGE-NIE [14]. To ensure robustness, we evaluate our framework under three different reader architectures: Qwen2.5-7B-Instruct [43], LLaMA-3.1-8B-Instruct [13], and Ministral-8B-Instruct [21]. Experimental results demonstrate that MEDRGAG achieves an average accuracy gain of 12.5% over MedRAG and 4.5% over MedGENIE across all datasets and reader settings. Further analyses show that our framework not only generates more effective complementary background documents but also successfully recovers low-similarity yet highly informative retrieved evidence.

In summary, our contributions are threefold:

- We propose MedRGAG, a unified retrieval—generation augmented framework that bridges external and parametric knowledge through a coherent retrieval—generation—then—read paradigm for medical question answering.
- We design two core modules: Knowledge-Guided Context Completion (KGCC), which generates complementary documents to fill missing knowledge, and Knowledge-Aware Document Selection (KADS), which adaptively selects a reliable combination of retrieved and generated evidence.
- We conduct extensive experiments on five medical QA benchmarks, demonstrating that MEDRGAG consistently outperforms a wide range of baselines. Further analyses verify that its two core modules effectively generate complementary background contexts and recover useful evidence.

2 Related Work

2.1 Medical Question Answering

Medical question answering (QA) has long been a fundamental task in biomedical natural language processing [25, 58]. Early studies primarily rely on BERT-based pretrained models [11], which achieve strong performance across various medical benchmarks [1, 29]. With the rapid advancement of large language models (LLMs) such as GPT-4 [3] and LLaMA [44], their exceptional reasoning and comprehension capabilities are also extended to medical QA. To further strengthen domain-specific expertise, a common approach is to continue pretraining these models on large-scale biomedical corpora (e.g., HuatuoGPT [53] and PMC-LLaMA [47]). However, this strategy demands substantial computational resources and high-quality annotated data, which greatly limits its practical applicability. Consequently, knowledge-augmented QA attracts increasing attention as a more flexible and interpretable alternative [9, 57]. This paradigm leverages external knowledge as auxiliary reference material, with two representative approaches: Retrieval-Augmented Generation (RAG) and Generation-Augmented Generation (GAG).

2.2 Medical Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) enhances large language models by grounding their reasoning on external knowledge sources retrieved from large-scale corpora [30, 35]. MedRAG [49] establishes a comprehensive RAG benchmark and toolkit that integrates hybrid retrieval for medical corpora. Building on this foundation, i-MedRAG [50] introduces iterative follow-up query generation to refine retrieval quality, while Self-BioRAG [19] incorporates selfreflective retrieval to better handle complex multi-hop medical reasoning. More recently, Omni-RAG [8] proposes a source-planning optimization strategy to retrieve information from diverse medical resources. Despite these advances, RAG-based methods still rely heavily on the relevance and completeness of retrieved documents, which are often noisy and lack critical knowledge. To address these limitations, our framework leverages the intrinsic knowledge of large models to generate complementary background documents and adaptively recover useful evidence while filtering out irrelevant content, thereby constructing a more comprehensive and reliable evidence set for medical QA.

2.3 Medical Generation-Augmented Generation

Generation-augmented methods prompt large language models to generate intermediate contexts for question answering, thereby exploiting their internal parametric knowledge [33, 36]. Representative studies such as GenRead [52] and CGAP [41] demonstrate that LLMs can serve as strong context generators in open-domain QA. MedGENIE [14] applies the generate-then-read paradigm to produce multi-view artificial contexts for medical QA. GRG [2] and COMBO [54] further combine generated and retrieved contexts through simple fusion strategies. Despite these advances, these approaches suffer from a critical limitation: the reader often relies heavily on generated documents that may contain hallucinated or inaccurate information. To address this issue, our framework performs multi-step reasoning to generate background documents and integrates trustworthy retrieved evidence to improve the factuality and completeness of the final context, thereby ensuring more reliable answer generation in medical QA.

3 Methodology

3.1 Problem Formulation

Definition 3.1.1 (Medical QA). Given a medical multiple-choice question q, which consists of a question stem and an answer set $A = \{a_1, \ldots, a_{|A|}\}$, the goal of medical question answering is to identify the most appropriate option \hat{a} as the correct answer. In an LLM-based framework, a reader model \mathcal{M}_r takes the question q together with a task-specific prompt \mathcal{P}_r as input and generates the predicted answer:

$$\hat{a} = \mathcal{M}_r(q, \mathcal{P}_r \mid \theta_r), \tag{1}$$

where θ_r denotes the parameters of the reader model. This basic formulation assumes that the model generates answers solely based on its internal parametric knowledge without external evidence. **Definition 3.1.2 (RAG).** Retrieval-Augmented Generation (RAG) integrates external knowledge into the input of large language models to enhance their reasoning capability and factual accuracy. Formally, given a medical corpus $C = \{d_1, \ldots, d_{|C|}\}$ and a retriever

 \mathcal{R} parameterized by $\theta_{\mathcal{R}}$, the retriever identifies the top-k most relevant documents D_r with respect to a question q:

$$D_r = \{d_1^r, \dots, d_k^r\} = \mathcal{R}(q, C \mid \theta_{\mathcal{R}}). \tag{2}$$

The reader model \mathcal{M}_r then takes the question q together with the retrieved document set D_r and a task-specific prompt \mathcal{P}_r as input to generate the predicted answer:

$$\hat{a} = \mathcal{M}_r(q, D_r, \mathcal{P}_r \mid \theta_r). \tag{3}$$

Definition 3.1.3 (GAG). In contrast, Generation-Augmented Generation (GAG) employs a large language model as the generator \mathcal{M}_g to produce k tailored background documents D_g for a given question q:

$$D_q = \{d_1^g, \dots, d_k^g\} = \mathcal{M}_q(q, \mathcal{P}_q \mid \theta_q), \tag{4}$$

where \mathcal{P}_g denotes the generation prompt designed to encourage diversity and factual reliability, and θ_g represents the parameters of the generator. The reader model \mathcal{M}_r then utilizes these generated documents as auxiliary context to infer the final answer:

$$\hat{a} = \mathcal{M}_r(q, D_q, \mathcal{P}_r \mid \theta_r). \tag{5}$$

3.2 MEDRGAG

We present the overall architecture of our proposed Medraga in Figure 2, which unifies retrieval and generation to integrate external and parametric knowledge for medical question answering. The framework operates through three sequential stages:

- Source-Balanced Evidence Retrieval. This stage retrieves relevant medical documents from multiple heterogeneous sources. By adopting a source-balanced retrieval strategy, it ensures fair and diverse evidence coverage while reducing the bias toward dominant information sources.
- Knowledge-Guided Context Completion. Building upon the retrieved evidence, this stage identifies missing or incomplete knowledge and generates complementary background documents. Through multi-step reasoning, it enriches the context with accurate and diverse information necessary for answering complex medical questions.
- Knowledge-Aware Document Selection. In the final stage, the framework integrates both retrieved and generated documents and adaptively selects the most relevant and reliable subset as the final evidence for the reader. This selection strikes a balance between factuality and relevance, ensuring the completeness of the supporting knowledge.

3.3 Source-Balanced Evidence Retrieval

The first stage focuses on retrieving medical documents relevant to the query from multiple heterogeneous knowledge sources. A naïve solution is to merge all sources into a single unified corpus and apply a dense retriever to evaluate query–document similarity, subsequently selecting the top-k documents as evidence. Although such a strategy provides broad coverage, previous studies reveal that retrieving from an overly aggregated corpus can bias the retriever toward dominant data sources or frequently occurring information, thereby compromising the fairness and diversity of the retrieved evidence [7, 8].

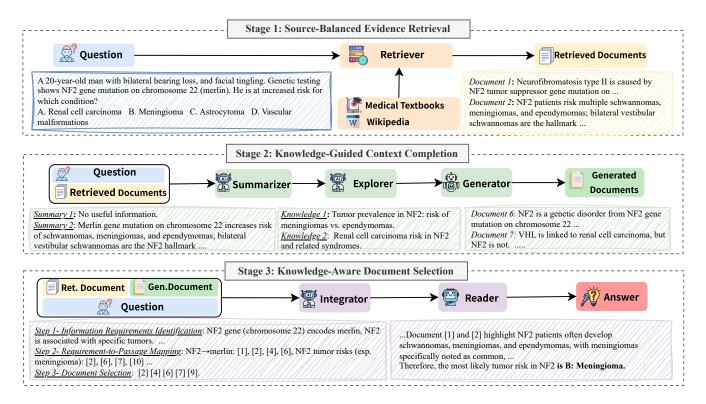


Figure 2: Framework of MedRGAG, which comprises three stages: source-balanced evidence retrieval, knowledge-guided context completion, and knowledge-aware document selection.

To address this issue, we employ a *source-balanced retrieval* strategy inspired by Sohn et al. [40]. Formally, given a question q and a collection of source corpora $\{C_1, C_2, \ldots, C_s\}$, the retriever independently selects an equal number of top-ranked documents from each source to construct an intermediate candidate set:

$$\tilde{D}_r = \bigcup_{i=1}^s \mathcal{R}(q, C_i \mid \theta_{\mathcal{R}}), \tag{6}$$

thereby ensuring a more balanced and representative distribution of information across different corpora compared with conventional unified-retrieval approaches [30].

After this balanced retrieval step, the candidate set \tilde{D}_r , containing $k \times s$ documents, is subsequently refined through a reranking process. We employ MedCPT-Reranker [24], an off-the-shelf crossencoder model that jointly encodes the question and each candidate document to estimate fine-grained relevance scores. The top-k documents with the highest scores in \tilde{D}_r are then selected to form the final retrieved evidence set D_r for the downstream stages.

3.4 Knowledge-Guided Context Completion

While retrieved documents provide valuable external evidence, they still exhibit two major limitations. First, the coverage of medical knowledge is often incomplete: relevant documents may capture only partial aspects of a question and thus fail to deliver comprehensive reasoning support. Second, the retrieved set frequently contains

noisy or irrelevant content, which can mislead the reader and ultimately degrade answer accuracy. To address these challenges, we propose a Knowledge-Guided Context Completion (KGCC) module that generates complementary background documents through a three-step process, effectively addressing the problem of insufficient and incomplete knowledge in retrieval-based evidence.

The KGCC process comprises three key steps: summarization of retrieved knowledge, exploration of missing knowledge, and generation of complementary background documents.

Summarization Prompt \mathcal{P}_s

You are a professional medical expert. Given the question and a retrieved document, summarize only the useful knowledge points for answering the question. If the document is irrelevant, return "No useful information".

Step 1: Summarization of Retrieved Knowledge. We employ a large language model as a summarizer \mathcal{M}_s , which extracts and condenses the essential information from each retrieved document with respect to the given medical question. The summarizer generates concise knowledge summaries that retain only the content useful for answering the question while filtering out noisy or irrelevant details. If a document contains no relevant information, the model explicitly outputs "No useful information." To further improve the quality of summarization, we adopt an in-context learning approach,

providing exemplar demonstrations retrieved from the training corpus to guide the model's summarization behavior.

Step 2: Exploration of Missing Knowledge. Building upon the summarized evidence, we evaluate its adequacy in addressing the medical question. An explorer model \mathcal{M}_e is introduced to identify the most critical missing knowledge points that are necessary yet absent from the current summaries, ensuring that the forthcoming generation stage focuses on complementary information. Formally, the explorer produces a set of missing knowledge items:

$$\mathcal{K} = \{k_1, \dots, k_m\} = \mathcal{M}_e(q, D_s, \mathcal{P}_e \mid \theta_e), \tag{7}$$

where \mathcal{K} denotes the set of missing knowledge points and \mathcal{P}_e represents the exploration prompt. Documents labeled as "No useful information" in Step 1 are excluded to improve efficiency.

Exploration Prompt \mathcal{P}_e

You are a professional medical expert. Given the question and the several pieces of useful information, identify the most important missing knowledge required to answer the question thoroughly.

Step 3: Generation of Background Documents. Leveraging the missing knowledge points \mathcal{K} , the generator \mathcal{M}_g produces background documents that complement the retrieved evidence. Each knowledge point $k_i \in \mathcal{K}$ serves as a guiding signal for generating a corresponding background document:

$$d_i^g = \mathcal{M}_q(q, k_i, \mathcal{P}_q \mid \theta_q), \tag{8}$$

This process yields m background documents. When m < k, additional documents are generated directly from the question q to ensure a complete set of k background documents $D_q = \{d_1^g, \ldots, d_k^g\}$.

This design provides two major advantages. First, conditioning the generation process on distinct knowledge points encourages diversity among the generated documents. Second, generating supplementary documents directly from the question introduces a broader perspective that enriches the overall background knowledge while avoiding overfitting to localized knowledge.

3.5 Knowledge-Aware Document Selection

Given the retrieved and generated documents, the central challenge lies in identifying the most relevant and reliable evidence while eliminating redundant or noisy content. Previous approaches, such as GenRead [52], simply concatenate the retrieved and generated documents into a single input. While this approach increases knowledge coverage, it also introduces two major drawbacks: (1) irrelevant or low-quality documents are retained, diluting the useful information and distracting the reasoning process; and (2) the enlarged input length substantially increases the contextual load of the reader model, leading to inefficient in reasoning. To address these limitations, we introduce the Knowledge-Aware Document Selection (KADS) module, which adaptively selects a compact yet informative subset of documents, ensuring that the reader model operates on evidence that is both useful and comprehensive.

The proposed knowledge-aware document selection module adaptively integrates retrieved and generated evidence by aligning them with the specific knowledge requirements of each question. A large language model, referred to as the integrator \mathcal{M}_i , is prompted to reason over all 2k candidate documents and select a compact yet informative subset that best supports answer generation. Concretely, \mathcal{M}_i executes the following three reasoning operations:

- Knowledge Requirement Identification. The model first analyzes the question to determine the essential knowledge components necessary for a complete and accurate answer.
- Knowledge-to-Document Mapping. Each candidate document is then associated with one or more identified knowledge components based on its content relevance.
- Balanced Evidence Selection. Finally, the model evaluates each knowledge group and selects the top-k documents that collectively maximize knowledge coverage while minimizing redundancy.

Considering the strong interdependence among the three reasoning steps, we design a carefully constructed prompt that allows a single LLM to perform the entire selection process in an integrated manner. Formally, the final evidence set D_f is derived as:

$$D_f = \mathcal{M}_i(q, D_r \cup D_q, \mathcal{P}_i \mid \theta_i), \tag{9}$$

where \mathcal{P}_i denotes the selection prompt, and θ_i represents the parameters of the integrator model \mathcal{M}_i .

Selection Prompt \mathcal{P}_i

You are a medical expert. Given a question and retrieved documents, select the most useful ones for answering. Step 1. Identify key knowledge points required to answer the question. Step 2. Map each passage to these knowledge points; assign irrelevant ones to a "No Useful Information" group. Step 3. From all groups, select up to 5 passages that ensure coverage and avoid redundancy.

By dynamically balancing retrieved and generated evidence through knowledge-aware grouping, this module effectively reduces noise and redundancy, thereby enhancing reasoning efficiency and answer accuracy.

The reader model \mathcal{M}_r subsequently takes the refined evidence set as input to generate the final answer:

$$\hat{a} = \mathcal{M}_r(q, D_f, \mathcal{P}_r \mid \theta_r). \tag{10}$$

Algorithm 1 MEDRGAG Framework

- 1: **Input:** Question set Q, Corpus C, Retriever \mathcal{R} , Generator \mathcal{M}_g , Reader \mathcal{M}_r , Summarizer \mathcal{M}_s , Explorer \mathcal{M}_e , Integrator \mathcal{M}_i , Prompts set \mathcal{P} and model parameters θ
- 2: **Output:** Predict Answer $\hat{\mathcal{A}}$
- 3: **for** each question $q \in Q$ **do**
- 1: $D_r = \mathcal{R}(q, C \mid \theta_{\mathcal{R}})$ # Retrieve relevant documents
- 5: $D_s = \mathcal{M}_s(q, D_r, \mathcal{P}_s \mid \theta_s)$ # Summary useful information
- 6: $\mathcal{K} = \mathcal{M}_e(q, D_s, \mathcal{P}_e \mid \theta_e)$ # Explore missing knowledge
- 7: $D_q = \mathcal{M}_q(q, \mathcal{K}, \mathcal{P}_q \mid \theta_q)$ # Generate background documents
- 8: $D_f = \mathcal{M}_i(q, D_r \cup D_g, \mathcal{P}_i \mid \theta_i)$ # Select documents
- 9: $\hat{a} = \mathcal{M}_r(q, D_f, \mathcal{P}_r \mid \theta_r)$ # Produce the final answer
- 10: end for

Reader	Method	MedQA-US	MedMCQA	MMLU-Med	PubMedQA*	BioASQ-Y/N	Average
	Direct Response	62.37	55.53	75.76	37.40	73.30	60.87
	Vanilla RAG	61.12	57.33	75.94	34.20	72.01	60.12
	MedRAG	62.61	58.64	76.12	38.00	76.38	62.35
	i-MedRAG	66.76	59.12	76.48	40.00	77.07	63.89
Qwen2.5-7B	GENREAD	68.89	60.24	78.97	46.00	79.29	66.68
	MedGENIE	69.36	59.91	78.60	49.20	81.88	67.79
	GRG	69.84	60.89	78.24	46.60	81.39	67.39

60.22

62.13

58.45

58.38

59.32

60.63

60.89

60.24

60.89

60.60

61.77

51.23

54.96

56.80

57.82

59.96

59.45

60.08

60.00

62.13

79.34

81.63

75.48

75.85

76.95

77.31

78.15

78.24

76.77

78.60

80.90

71.63

71.44

76.22

76.58

77.04

78.51

77.41

76.95

80.72

47.80

51.60

55.40

50.20

52.00

53.80

54.60

56.60

57.00

58.20

57.80

31.80

26.20

29.80

31.20

41.60

42.00

42.60

42.60

44.60

82.36

84.79

76.38

73.30

75.24

76.74

78.32

80.26

82.36

79.94

82.04

72.65

72.49

74.11

75.59

78.80

82.36

81.55

81.07

84.14

68.41

71.14

66.64

64.99

66.39

67.82

68.41

68.83

69.20

69.80

71.43

56.92

57.15

59.84

61.53

65.35

66.07

65.78

65.81

69.31

72.35

75.57

67.48

67.24

68.42

70.62

70.07

68.81

68.97

71.64

74.63

57.27

60.64

62.29

66.42

69.36

68.03

67.24

68.42

74.94

Table 1: Main experiment results. Best results are in bold and second-best ones are underlined.

To further clarify the overall workflow, we present the complete MEDRGAG framework in pseudo-code, as shown in Algorithm 1.

CGAP

MedRGAG

Vanilla RAG

MedRAG

Llama3.1-8B

Ministral-8B

i-MedRAG

GENREAD

MedGENIE

MedRGAG

Vanilla RAG

MedRAG

i-MedRAG

GENREAD

MedGENIE

MedRGAG

GRG

CGAP

Direct Response

GRG

CGAP

Direct Response

4 Experiments

4.1 Datasets

We evaluate MEDRGAG on five widely used medical question answering benchmarks: MedQA [22], MedMCQA [32], MMLU-Med [16], PubMedQA* [23], and BioASQ-Y/N [45]. All datasets are formulated as multiple-choice QA tasks, where each question includes two to four candidate answers. The overall dataset statistics are summarized in Table 2, with additional details presented in Appendix B. Following standard evaluation practice [14, 49], we report accuracy as the primary performance metric.

4.2 Baselines

To comprehensively evaluate the effectiveness of Medraga, we compare it with three representative baseline categories. (1) **Direct response methods**, where the LLM directly answers questions. (2) **RAG-based methods**, including Vanilla RAG [30], MedRAG [49], and *i*-MedRAG [50], which retrieve relevant evidence from external medical corpora. (3) **GAG-based methods**, such as GenRead [52], MedGENIE [14], GRG [2], and CGAP [41], which synthesize auxiliary contexts from the model's internal parametric knowledge. A detailed description of each baseline is provided in Appendix C.

Table 2: The statistics of datasets. #A.: numbers of options; Avg. L: average token counts in each question.

Dataset	Size	#A.	Avg. L	Source
MedQA-US	1,273	4	177	Examination
MedMCQA	4,183	4	26	Examination
MMLU-Med	1,089	4	63	Examination
PubMedQA*	500	3	24	Literature
BioASQ-Y/N	618	2	17	Literature

In our experiments, we employ BM25 [37] as the retriever, which retrieves candidate documents from multi-corpus: medical text-books [22] and Wikipedia articles [49]. We adopt LLaMA-3.1-8B-Instruct [13] as the generator to produce background contexts. The reader is instantiated with Qwen2.5-7B-Instruct [43], LLaMA-3.1-8B-Instruct [13], and Ministral-8B-Instruct [21]. Additionally, we use GPT-40-mini [18] for other LLM-based modules in MedRGAG, including the summarizer, explorer, and integrator. For fair comparison, all retrieval, generation, and fusion stages are standardized to produce a final top-5 set of documents delivered to the reader. Further implementation details are provided in Appendix A.

Table 3: Ablation result on	Owen2.5-7B-Instruct reader.
-----------------------------	-----------------------------

Method	MedQA-US	MedMCQA	MMLU-Med
Direct Response	62.37	55.53	75.76
MedRGAG	75.57	62.13	81.63
w/o Generation	63.47	58.69	77.96
w/o Retrieval	72.90	61.15	80.35
w/o KGCC	72.27	61.30	79.43
w/o KADS	74.00	61.10	80.90

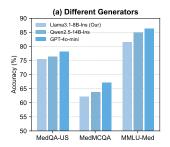
4.3 Main Results

The experimental results are presented in Table 1. Overall, Medragaga attains the highest average accuracy across all five medical QA benchmarks, consistently outperforming all baseline models. We highlight three key findings below.

- (1) Effectiveness over direct response methods. Medraga achieves remarkable improvements compared with the direct-response setting, where the reader model answers without any external knowledge. On average, it yields more than a 15% absolute gain, demonstrating that providing relevant background knowledge significantly enhances answer accuracy. Moreover, this advantage persists across three distinct reader architectures, underscoring the generalizability and robustness of Medraga.
- (2) Superiority to retrieval-augmented methods. Compared to retrieval-based frameworks such as Vanilla RAG, Medraga achieves over 9% improvement on average across all reader models. Although advanced RAG methods such as Medraga and *i*-Medraga enhance retrieval quality via refined selection or iterative querying, their performance gains remain limited. This modest improvement reflects the insufficient knowledge coverage of existing medical corpora, where retrieved documents often fail to encompass all essential medical knowledge. In contrast, Medraga leverages its generator to supplement missing knowledge beyond the retrieval corpus, leading to markedly improved answer accuracy.
- (3) Beyond generation-augmented methods. Medragas surpasses generation-augmented QA frameworks such as MedGE-NIE, GenRead, GRG, and CGAP, achieving an average improvement of 4.5% over MedGENIE. These results show that integrating retrieved evidence into generation-based contexts effectively alleviates the influence of hallucinated documents and yields more accurate reasoning. Furthermore, Medragas outperforms GRG, which directly merges retrieved and generated documents, demonstrating that adaptive document selection enables more precise, question-specific evidence aggregation.

4.4 Ablation Study

We conduct ablation studies to assess the contribution of each key component within MedRGAG. The variants are defined as follows: (1) w/o Generation removes the generation module and relies solely on retrieved documents for answering. (2) w/o Retrieval removes external retrieval and depends only on generated documents. (3) w/o KGCC disables the knowledge-guided context completion module, generating background documents directly without identifying missing knowledge. (4) w/o KADS removes the knowledge-aware



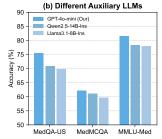


Figure 3: Results of different model scales on the Qwen2.5-7B-Instruct reader. Figures (a) and (b) show different LLM scales used as generators and summarizer-explorer-integrator.

document selection module, instead directly ranking the ten candidate documents to select the top-5 without multi-step reasoning.

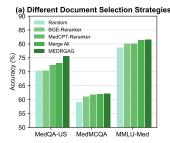
As shown in Table 3, removing either the generation or retrieval component results in a substantial performance drop, indicating that both knowledge sources are indispensable for achieving comprehensive and reliable reasoning. The decline is more pronounced when removing generation, suggesting that generated documents capture question-specific knowledge more effectively than retrieved evidence. Disabling the KGCC module further decreases accuracy, confirming that knowledge-guided completion yields more informative and targeted contexts. Likewise, eliminating the KADS module also degrades performance, demonstrating that adaptive document selection is crucial for identifying useful evidence for each question.

4.5 Performance Analysis

Effect of Generator Scale. We analyze how the generator's scale influences overall performance by progressively increasing its parameter size from LLaMA3.1-8B-Instruct to Qwen2.5-14B-Instruct and GPT-4o-mini. As illustrated in Figure 3 (a), accuracy consistently rises with larger generator models, indicating that more capable generators can produce higher-quality and broader-coverage knowledge. Compared with the reasoning-intensive clinical dataset MedQA-US, the improvement is particularly pronounced on MedM-CQA and MMLU-Med, which focus on fundamental biomedical knowledge, suggesting that larger models possess richer foundational medical understanding.

Effect of Auxiliary LLM Scale. We further assess how the scale of auxiliary LLMs (serving as the summarizer, explorer, and integrator) impacts overall QA performance. In the main configuration, these components are implemented with GPT-40-mini, and we additionally evaluate smaller models, including Qwen2.5-14B-Instruct and LLaMA3.1-8B-Instruct. As shown in Figure 3 (b), accuracy declines as the auxiliary model size becomes smaller. The 8B and 14B models perform comparably yet remain substantially below GPT-40-mini, revealing that smaller LLMs struggle with the complex subtasks of summarization, knowledge exploration, and document selection. Their weaker instruction-following and reasoning capabilities tend to accumulate subtle but compounding errors across stages, ultimately diminishing overall answer accuracy.

Effect of Document Selection Strategies. We further analyze how different document selection strategies affect the final answer



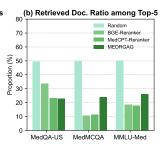


Figure 4: Results of different document selection strategies on the Qwen2.5-7B-Instruct reader. Figure (a) shows different document selection strategies on accuracy. Figure (b) shows the proportion of retrieved documents within the final top-5 across different selection strategies.

accuracy. For each question, we use five retrieved and five generated documents produced by MedRGAG and compare several selection methods: Random, which randomly samples five documents from ten candidates. BGE-Reranker-Large [48] and MedCPT-Reranker [24], which re-rank all candidates and select the top five documents. Merge All, which provides all ten documents to the reader without any filtering. As shown in Figure 4 (a), our adaptive selection strategy achieves the best overall performance across three datasets. While MedCPT slightly outperforms BGE-Reranker due to its medical-domain optimization, both remain inferior to our approach. Notably, MedRGAG attains comparable or even higher accuracy than the Merge All setting while feeding the reader with only half as many documents (5 instead of 10), demonstrating that adaptive document selection effectively retains useful evidence, filters out irrelevant content, and enhances reasoning efficiency. **Analysis of Document Source Preference.** We further examine how different document selection strategies balance the contributions of retrieved and generated evidence. For each method, we compute the proportion of retrieved documents among the final top-5 evidence, as illustrated in Figure 4 (b). Overall, generated documents are more favored, aligning with prior findings that ranking models tend to assign higher similarity scores to generative content [14, 42]. In contrast, our adaptive selection strategy markedly increases the proportion of retrieved documents, suggesting that it successfully recovers valuable retrieval-based evidence that would otherwise be overshadowed by high-similarity generated passages. An exception arises in the MedQA dataset, where our method does not increase the proportion. This is likely because its lengthy and complex question stems cause the retrieved passages to be only partially relevant, rather than directly informative. Overall, these results demonstrate that MedRGAG effectively identifies genuinely

4.6 Case Study

To further demonstrate the complementary strengths of retrieved and generated knowledge in MEDRGAG, we present a representative simplified example in Table 4. For clarity, only the first two retrieved and generated documents are displayed here, while more detailed examples are provided in Appendix E.

useful documents from both sources to enhance QA performance.

Table 4: Case study showing how generated knowledge complements retrieved evidence in MedRGAG.

Question: A 20-year-old man with bilateral hearing loss and facial tingling. Genetic testing shows NF2 gene mutation on chromosome 22 (merlin). He is at increased risk for which condition?

Options: A. Renal cell carcinoma B. Meningioma C. Astrocytoma D. Vascular malformations.

Retrieved Doc. 1: Neurofibromatosis type II is caused by NF2 tumor suppressor gene mutation on chromosome 22. The gene encodes merlin, which normally regulates growth factors...

Retrieved Doc. 2: NF2 patients risk multiple meningiomas, and ependymomas; bilateral vestibular schwannomas are the hallmark. NF2 involves loss of merlin gene on chromosome 22 ...

Generated Doc. 1: NF2 is a genetic disorder from NF2 gene mutation on chromosome 22. Patients develop bilateral vestibular schwannomas causing hearing loss, and are at increased risk for meningiomas ...

Generated Doc. 2: VHL is linked to renal cell carcinoma, but NF2 is not. NF2 patients develop vestibular schwannomas and may also develop meningiomas and ependymomas...

Selected Docs: [Ret_2], [Gen_1], [Gen_2] **Final Answer:** B. Meningioma ✓

The question is: "A 20-year-old male with an NF2 gene mutation is at increased risk for which disease?" Among the retrieved documents, Doc 1 describes the genetic mechanism of the NF2 mutation, which is relevant but does not specify the associated pathologies, whereas Doc 2 mentions that NF2 patients are predisposed to meningiomas but lacks further explanation. In contrast, the generated documents enrich the reasoning context: Gen 1 elaborates on the etiology of NF2 and its characteristic tumor spectrum, while Gen 2 distinguishes NF2 from VHL-related syndromes, thereby ruling out the distractor option "renal cell carcinoma." Finally, MedRGAG successfully selects the three most informative documents, [Ret_2], [Gen_1], and [Gen_2], to support the reader model in producing the correct answer: "B. Meningioma". This case illustrates how MedRGAG identifies missing knowledge in retrieved evidence, generates complementary documents, and selects the most relevant contexts for accurate reasoning.

5 Conclusion

In this paper, we present Medraga, a unified retrieval—generation augmented framework that bridges external (retrieved) and parametric (generated) knowledge for medical question answering. Unlike traditional RAG and GAG paradigms that that rely on a single knowledge source, Medraga combines both through two key modules: Knowledge-Guided Context Completion (KGCC), which identifies and fills knowledge gaps revealed by retrieval, and Knowledge-Aware Document Selection (KADS), which adaptively selects the most useful evidence for reasoning. Extensive experiments on five benchmark datasets and multiple reader architectures show consistent performance gains over strong retrieval- and generation-based baselines, demonstrating robustness and generalizability. Our analyses and case studies further indicate that Medraga mitigates the

limitations of incomplete retrieval by recovering valuable retrievalbased evidence while curbing reliance on hallucination-prone generated contexts. In future work, we will extend this framework to broader knowledge-intensive QA tasks and more complex medical scenarios. A promising direction is to unlock large models' internal knowledge and reasoning ability, integrating it with verifiable retrieved evidence to support trustworthy clinical reasoning.

References

- Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In Proceedings of the 18th bioNLP workshop and shared task. 370–379.
- [2] Abdelrahman Abdallah and Adam Jatowt. 2023. Generator-retriever-generator approach for open-domain question answering. arXiv preprint arXiv:2307.11278 (2023).
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023).
- [4] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023).
- [5] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. (2024).
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems 33 (2020), 1877–1901.
- [7] Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating Entity Disambiguation and the Role of Popularity in Retrieval-Based NLP. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 4472-4485.
- [8] Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. 2025. Towards Omni-RAG: Comprehensive Retrieval-Augmented Generation for Large Language Models in Medical Applications. arXiv preprint arXiv:2501.02460 (2025).
- [9] Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. 2025. A survey on knowledge-oriented retrieval-augmented generation. arXiv preprint arXiv:2503.10677 (2025).
- [10] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. 758–759.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 4171–4186.
- [12] Xinkai Du, Quanjie Han, Chao Lv, Yan Liu, Yalin Sun, Hao Shu, Hongbo Shan, and Maosong Sun. 2025. Improving Generated and Retrieved Knowledge Combination Through Zero-shot Generation. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 1–5.
- [13] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv e-prints (2024), arXiv-2407.
- [14] Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To Generate or to Retrieve? On the Effectiveness of Artificial Contexts for Medical Open-Domain Question Answering. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 9878–9919.
- [15] Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In Proceedings of the 2022 conference on empirical methods in natural language processing. 5770–5793.
- [16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020).
- [17] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Transactions on Information Systems

- (2024).
- [18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-40 system card. arXiv preprint arXiv:2410.21276 (2024).
- [19] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics* 40, Supplement_1 (2024), i119–i129.
- [20] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM computing surveys 55, 12 (2023), 1–38.
- [21] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825
- [22] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences 11, 14 (2021), 6421.
- [23] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2567–2577.
- [24] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. Bioinformatics 39, 11 (2023), btad651.
- [25] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: a survey of approaches and challenges. ACM Computing Surveys (CSUR) 55, 2 (2022), 1–36.
- [26] Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. Scientific Data 10, 1 (2023), 170.
- [27] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In Proceedings of the 29th symposium on operating systems principles. 611–626.
- [28] Philippe Laban, Alexander R Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. arXiv preprint arXiv:2407.01370 (2024).
- [29] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
- [30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems 33 (2020), 9459–9474.
- [31] Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. arXiv preprint arXiv:2403.10446 (2024).
- [32] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on health, inference, and learning. PMLR, 248–260.
- [33] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases?. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2463–2473.
- [34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.
- [35] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics 11 (2023), 1316–1331.
- [36] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 5418–5426.
- [37] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval 3, 4 (2009), 333–389.

- [38] Yucheng Shi, Tianze Yang, Canyu Chen, Quanzheng Li, Tianming Liu, Xiang Li, and Ninghao Liu. 2025. SearchRAG: Can Search Engines Be Helpful for LLM-based Medical Question Answering? arXiv preprint arXiv:2502.13233 (2025).
- [39] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.
- [40] Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2025. Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 12739–12753.
- [41] Dan Su, Mostofa Patwary, Shrimai Prabhumoye, Peng Xu, Ryan Prenger, Mohammad Shoeybi, Pascale Fung, Animashree Anandkumar, and Bryan Catanzaro. 2023. Context Generation Improves Open Domain Question Answering. In Findings of the Association for Computational Linguistics: EACL 2023. 793–808.
- [42] Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6207–6227.
- [43] Qwen Team et al. 2024. Qwen2 technical report. arXiv preprint arXiv:2407.10671 2 (2024), 3.
- [44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
- [45] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC bioinformatics 16. 1 (2015), 138.
- [46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35 (2022), 24824–24837.
- [47] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. PMC-LLaMA: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association* 31, 9 (2024), 1833–1843.
- [48] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval. 641–649.
- [49] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In Findings of the Association for Computational Linguistics ACL 2024. 6233–6251.
- [50] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In Biocomputing 2025: Proceedings of the Pacific Symposium. World Scientific, 199–214.
- [51] Yifan Yang, Qiao Jin, Robert Leaman, Xiaoyu Liu, Guangzhi Xiong, Maame Sarfo-Gyamfi, Changlin Gong, Santiago Ferrière-Steinert, W John Wilbur, Xiaojun Li, et al. 2024. Ensuring safety and trust: Analyzing the risks of large language models in medicine. arXiv preprint arXiv:2411.14487 (2024).
- [52] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. arXiv preprint arXiv:2209.10063 (2022).
- [53] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, et al. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In Findings of the Association for Computational Linguistics: EMNLP 2023. 10859–10885.
- [54] Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging Generated and Retrieved Knowledge for Open-Domain QA. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 4710–4728.
- [55] Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How Do Large Language Models Capture the Ever-changing World Knowledge? A Review of Recent Advances. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 8289–8311.
- [56] Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In Proceedings of the ACM on Web Conference 2025. 4442–4457.
- [57] Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu. 2023. Knowledge-augmented methods for natural language processing. In Proceedings of the sixteenth ACM international conference on web search and

- data mining. 1228-1231.
- [58] Pierre Zweigenbaum. 2003. Question answering in biomedicine. In Proceedings Workshop on Natural Language Processing for Question Answering, EACL, Vol. 2005. Citeseer. 1–4.

A Implementation Details

We construct the retrieval corpus from two sources: a Textbook [22] collection with 125.8K snippets and a Wikipedia [49] corpus with 29.9M snippets. We employ BM25 [37] as the retriever to retrieve top-32 documents from each source, followed by the MedCPT-Reranker [24] to rerank the top-5 most relevant documents.

For each question, three missing knowledge points are identified. The generator, LLaMA3.1-8B-Instruct, operates in a zero-shot setting to produce up to 256 tokens per document. Two additional documents are generated directly conditioned on the question, resulting in five generated contexts. The reader employs a Chain-of-Thought (CoT) [46] reasoning prompt to derive the final answer. We use GPT-40-mini for other auxiliary LLMs, including the summarizer, explorer, and integrator. Temperature is set to 1.2 for the generator and explorer to encourage diversity, and 0.2 for the reader to maintain stability.

All baselines share identical retriever, corpus, and generator configurations for fairness. MedRAG [49] follows its original setup with four retrievers and Reciprocal Rank Fusion [10] for aggregation. GenRead [52] retrieves the top-1 document per question in training data and forms five in-context learning (ICL) clusters to generate diverse pseudo-contexts for test data. CGAP [41] generates five contexts and applies majority voting for the final prediction. GRG [2] generates ten candidate documents, from which MedCPT-Reranker selects the top-3 for document fusion.

We adopt the vLLM [27] engine for efficient batched inference and memory optimization. GPT-40-mini is accessed through the OpenAI API for closed-source generation components.

B Dataset Details

We evaluate MEDRGAG on five representative medical QA datasets covering clinical, biomedical, and professional knowledge domains:

- MedQA [22] MedQA is derived from the United States Medical Licensing Examination (USMLE), focusing on diagnosis, treatment, and clinical reasoning. We use the English test subset (1,273 four-choice questions).
- MedMCQA [32] MedMCQA contains Indian medical entrancestyle questions across 21 subjects and 2,400 topics. Since its test labels are unavailable, the official dev set (4,183 questions) is chosen.
- MMLU-Med [16] MMLU-Med comprises biomedical questions from six subfields, including Anatomy, College Biology, College Medicine, Clinical Knowledge, Human Genetics, and Professional Medicine, totaling 1,089 test samples.
- PubMedQA* [23] PubMedQA is built from PubMed abstracts with 1,000 expert-annotated questions. To match the RAG setting, we remove original contexts and adopt the 500 official test samples for evaluation.
- **BioASQ-Y/N** [26, 45] We extract all Yes/No questions from machine reading comprehension (Task B) tracks' gold-standard test sets over the five most recent years (2019–2023), resulting in 618 questions.

C Baseline Details

We compare MEDRGAG with a series of representative baselines from both retrieval-augmented (RAG) and generation-augmented (GAG) paradigms:

- Direct Response The LLM directly answers each question without external knowledge, evaluating its intrinsic reasoning capability.
- Vanilla RAG [30] A standard retrieval-then-read framework where a retriever retrieves top-k relevant documents from a corpus using vector similarity, and a reader model subsequently generates an answer.
- MedRAG [49] A medical RAG benchmark combining sparse and dense retrieval to obtain domain-specific evidence from MedCorp, followed by Reciprocal Rank Fusion (RRF) [10] to enhance retrieval robustness and coverage.
- i-MedRAG [50] An iterative extension of MedRAG that allows the LLM to issue follow-up queries, progressively refining retrieval quality and improving reasoning for multihop medical questions.
- **GenRead** [52] A standard *generate-then-read* framework that first prompts an LLM to synthesize contextual documents based on the input question, and then employs a reader model to generate an answer.
- MedGENIE [14] The first generation-augmented framework for medical QA, which produces multi-view artificial contexts via option-focused and option-free context to enhance reasoning diversity and factual grounding.
- GRG [2] A generator-retriever-generator pipeline where the LLM first generates hypothetical documents, retrieves supporting evidence, and then generates the final answer using both sources.
- CGAP [41] A two-stage framework performing context generation and answer prediction entirely within a single LLM, emphasizing efficiency and adaptability without relying on external corpora.

We also employ several representative large language models, retrievers, and rerankers in our experiments:

- Qwen2 [43] A Mixture-of-Experts model family (0.5B-72B) with strong multilingual reasoning capabilities, pretrained on large-scale corpora and instruction-tuned for general understanding.
- LLaMA3 [13] The latest Meta release trained on expanded data with extended context windows, achieving improved coherence and domain adaptability.
- Mistral [21] An efficient open-weight model employing sliding-window attention for longer context handling with superior quality-efficiency trade-offs.
- GPT-4 [18] OpenAI's flagship model trained with reinforcement learning from human feedback (RLHF), offering exceptional reasoning and instruction-following ability.
- BM25 [37] A classical lexical retriever using TF-IDF weighting, implemented with Pyserini to index and retrieve relevant medical snippets.
- MedCPT-Reranker [24] A biomedical reranker pre-trained on 255M PubMed click logs, enabling precise semantic matching for domain-specific retrieval.

 BGE-Reranker [48] A general-purpose cross-encoder reranker from BAAI that performs full query-document attention for robust semantic ranking across domains.

D Prompts Design

In this section, we present the detailed prompt templates used in MEDRGAG, covering all key components of the framework—including the summarizer, explorer, generator, integrator, and reader.

Summarization Prompt \mathcal{P}_s

You are a professional medical expert. Given the following question and a retrieved document, distill the useful information that can assist in answering the question. Focus only on details directly supported by evidence from the document, and avoid including irrelevant or speculative content. If the document does not contain relevant information, return "No useful information." Do not attempt to answer the question—only summarize the essential knowledge needed for answering it accurately.

Input: Retrieved Document: {documents} Question: {question}

Output: Useful Information: {useful_information}

Exploration Prompt \mathcal{P}_e

You are a professional medical expert. Given the question and several pieces of useful information extracted from retrieved documents, identify the most important missing knowledge required to answer the question thoroughly. Analyze the question to determine key knowledge components, compare them with the provided information, and identify the gaps. Select the three most critical and non-redundant missing knowledge points, each expressed as a concise conceptual title rather than a full sentence.

Input: Useful Information: {information} Question: {question}

Output Format: - Reasoning: [Detailed explanation] - Knowledge 1: [Conceptual title 1] - Knowledge 2: [Conceptual title 2] - Knowledge 3: [Conceptual title 3]

Generation Prompt \mathcal{P}_q

You are a professional medical expert. Given the following medical question and a single knowledge point, generate a concise background document that provides relevant explanations or context strictly based on the given knowledge point. Do not infer or guess the correct answer, and avoid mentioning any answer options. Write in English and keep the content within 256 words.

Input: Question: {question} Knowledge Point: {knowledge_point}

Output: Background Document: {generated_document}

E More Case

Selection Prompt \mathcal{P}_i

You are a professional medical expert. Given a medical question and ten candidate passages (each labeled with an identifier [id]), select the top-5 most useful passages for answering the question accurately.

Follow the reasoning steps below: 1. Information Requirements Identification: Identify the key knowledge points necessary to answer the question thoroughly. 2. Requirement-to-Passage Mapping: Match each passage to the corresponding knowledge point(s) and classify irrelevant ones into a "No Useful Information" group. 3. Document Selection for Completeness and Conciseness: Choose up to five passages that together provide comprehensive coverage of the key knowledge points while minimizing redundancy.

Input: Documents: {documents} Question: {question}
Output Format: - Reasoning: [Detailed explanation] Final Selection: [id1] [id2] [id3] [id4] [id5]

Answering Prompt \mathcal{P}_r

You are a professional medical expert. Given the question and several retrieved or generated documents, reason step-by-step and provide the final answer. First, extract and utilize useful information from the documents; if insufficient, rely on your medical knowledge to complete the reasoning. Return your output in JSON format containing both reasoning and the final answer choice.

Input: Retrieved Documents: {documents} Question: {question}

Output Format: {"reasoning": "explanation", "answer_choice": "A/B/C/..."}

Table 5: Case study showing how generated knowledge complements retrieved evidence in MedRGAG.

Example: #0012 in MedQA

Question: A 20-year-old man comes to the physician because of worsening gait unsteadiness and bilateral hearing loss for 1 month. He has had intermittent tingling sensations on both cheeks over this time period. He has no history of serious medical illness and takes no medications. Audiometry shows bilateral sensorineural hearing loss. Genetic evaluation shows a mutation of a tumor suppressor gene on chromosome 22 that encodes merlin. This patient is at increased risk for which of the following conditions?"

Options: A. Renal cell carcinoma B. Meningioma C. Astrocytoma D. Vascular malformations.

Retrieved Doc. 1: Neurofibromatosis. Neurofibromatosis type II is caused by a mutation on chromosome 22. The mutation falls on the NF2 tumor suppressor gene. The gene normally encodes a cytoplasmic protein known as merlin... The loss of function leads to increased activity of growth factors normally regulated by merlin, leading to the formation of the tumors associated with NF2.

Retrieved Doc. 2: Pathology_Robbins. Most schwannomas are sporadic, but about 10% are associated with familial neurofibromatosis type 2

(NF2). NF2 patients are at risk of developing multiple schwannomas, meningiomas, and ependymomas (the latter are described in Chapter 23). The presence of bilateral vestibular schwannomas is a hallmark of NF2; despite the name, neurofibromas (described later) are not found in NF2 patients...

Summary Doc. 1: No useful information.

Summary Doc. 2: Patients with a mutation of the merlin gene on chromosome 22 are at risk of developing multiple schwannomas, meningiomas, and ependymomas. The presence of bilateral vestibular schwannomas is a hallmark of neurofibromatosis type 2 (NF2).

Explore Knowledge 1: Specific tumor prevalence in NF2 patients (e.g., risk of meningiomas vs. ependymomas).

Explore Knowledge 2: Risk of renal cell carcinoma in NF2 and related syndromes.

Generated Doc. 6: **Neurofibromatosis Type 2 (NF2) ** Neurofibromatosis Type 2 (NF2) is a rare genetic disorder characterized by the development

of non-cancerous (benign) tumors in the nervous system. The condition is caused by mutations in the NF2 gene, also known as the merlin tumor suppressor gene, which is located on chromosome 22. **Clinical Features:** NF2 patients often present with bilateral vestibular schwannomas (tumors of the nerves responsible for hearing and balance) leading to sensorineural hearing loss...

Generated Doc. 7: **Von Hippel-Lindau Syndrome and Neurofibromatosis Type 2 (NF2) Association with Renal Cell Carcinoma ** Von Hippel-Lindau syndrome (VHL) is a genetic disorder characterized by the development of benign and malignant tumors, particularly hemangioblastomas of the cerebellum, retina, and spinal cord. The NF2 gene encodes for merlin (also known as schwannomin), a tumor suppressor protein that plays a critical role in the regulation of cellular growth and division...

Selected Docs: Step 1: Information Requirements Identification: NF2 gene (chromosome 22) encodes merlin, NF2 is associated with specific tumors. Step 2: Requirement-to-Passage Mapping: NF2—merlin: [1], [2], [4], [6], [9]; NF2 tumor risks (esp. meningioma): [2], [6], [7], [9], [10]; Other tumors (e.g., ependymoma/astrocytoma): [2], [5], [4], [6], [7] Step 3: Document Selection for Completeness and Conciseness: Prioritize NF2—meningioma evidence; ensure RCC exclusion. Final Selection: [2] [4] [6] [7] [9].

Final Answer: B. Meningioma ✓