

SearchRAG: Can Search Engines Be Helpful for LLM-based Medical Question Answering?

Yucheng Shi¹, Tianze Yang¹, Canyu Chen², Quanzheng Li³, Tianming Liu¹, Xiang Li³, Ninghao Liu⁴

¹University of Georgia

²Northwestern University ³Massachusetts General Hospital and Harvard Medical School

⁴The Hong Kong Polytechnic University

Abstract—Large Language Models (LLMs) have shown remarkable capabilities in general domains but often struggle with tasks requiring specialized knowledge. Conventional Retrieval-Augmented Generation (RAG) techniques typically retrieve external information from static knowledge bases, which can be outdated or incomplete, missing fine-grained clinical details essential for accurate medical question answering. In this work, we propose SearchRAG, a novel framework that overcomes these limitations by leveraging real-time search engines. Our method employs synthetic query generation to convert complex medical questions into search-engine-friendly queries and utilizes uncertainty-based knowledge selection to filter and incorporate the most relevant and informative medical knowledge into the LLM's input. Experimental results demonstrate that our method significantly improves response accuracy in medical question answering tasks, particularly for complex questions requiring detailed and up-to-date knowledge. We provide our code here¹.

Index Terms—RAG, LLMs, Medical QA, Search Engine.

I. INTRODUCTION

Large language models (LLMs) have shown proficiency in handling general-domain question answering tasks [1]. However, they often face challenges when dealing with specialized fields that require up-to-date or domain-specific knowledge, such as medical question answering [2], [3]. The limitation arises because the LLM parameters remain fixed after training, preventing the model from incorporating new medical knowledge essential for precise decision-making.

Retrieval-Augmented Generation (RAG) methods have been proposed to address this issue by incorporating external knowledge sources into the LLMs' input, thereby enhancing performance with additional context [4], [5]. Traditional RAG approaches typically retrieve information from static knowledge bases or knowledge graphs [6], [7]. However, these approaches face significant limitations. Recent studies [7] show that RAG performance is heavily dependent on the quality and coverage of the knowledge corpus used. In many instances, using RAG actually reduces accuracy because the used knowledge bases is outdated or incomplete, missing fine-grained details critical for accurate medical decision-making [7].

In this work, we propose to tackle these challenges by utilizing search engines to enhance RAG systems. Search engines play a vital role for medical professionals, who rely on them to access clinical and medical knowledge that supports informed medical decision-making [8]. By leveraging search

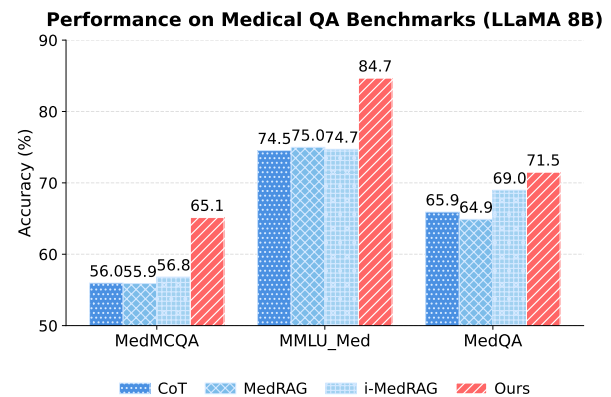


Fig. 1. Performance comparison of different methods on medical QA benchmarks using LLaMA 8B. Our proposed method consistently outperforms baseline approaches (CoT), conventional RAG (MedRAG), and iterative RAG (i-MedRAG) across all three datasets (MedMCQA, MMLU_Med, and MedQA), demonstrating significant improvements in accuracy.

engines, we can equip LLMs with real-time access to peer-reviewed research, clinical trial data, and treatment protocols with fine-grained details that are essential for effective medical decision support [9].

However, directly using the original medical questions or LLM-written queries as search engine input often yields suboptimal results. This is because original medical questions typically contain redundant contextual information that can mislead the search focus. Moreover, current LLMs are not trained to align with search engines; therefore, they lack the inherent capability to generate queries optimized for search engines, which require specific structures and operators like Boolean logic and MeSH terms to function effectively [10], [11]. This misalignment between the LLM-written queries and the requirements of search engines leads to the retrieval of irrelevant or low-quality information, ultimately degrading the RAG performance [12]. This raises a critical research question: **How to help current LLMs with search engines for effective medical knowledge retrieval?**

To address this challenge, we propose **SearchRAG**, a novel retrieval-augmented generation framework that enhances the alignment between LLMs and search engines in inference-time through *synthetic query generation* and *uncertainty-based query selection*. Our approach first utilizes an LLM

¹<https://github.com/sycny/SearchRAG>

to generate a large and diverse set of search queries by repeatedly sampling with high temperature, based on the original medical question. These synthetic queries are designed to capture different aspects of the medical question in a search-engine friendly form. While not all generated queries are optimal, the diversity and scale of queries ensure that some candidates will effectively capture the clinical intent. Then, to identify the most effective queries, our framework employs an uncertainty-based selection mechanism that evaluates the knowledge snippets retrieved by each query. By measuring the LLM’s uncertainty reduction when incorporating different knowledge snippets, we retain only those that contribute the most to improving model confidence.

Through extensive experiments on medical QA benchmarks (Figure 1), we demonstrate three key contributions:

- **Search Engine-Aligned Query Generation:** We propose SearchRAG, a novel method that bridges the gap between LLM capabilities and search engine requirements through synthetic query generation, specifically optimized for medical information retrieval.
- **Uncertainty-Based Query Selection:** We develop a selection mechanism that uses the LLM’s internal entropy to identify and prioritize the most clinically relevant information from multiple retrieval candidates.
- **Effective Performance Improvement:** We empirically show that our framework improves the LLaMA 8B model by an average 12.69% compared to the non-RAG baseline in answer accuracy in medical QA tasks.

The paper is structured as follows: Section II details our dual-component architecture for query generation and knowledge selection, Section III presents comparative evaluations across multiple medical benchmarks, Section IV discusses related work, and Section V concludes our paper.

II. METHODOLOGY

In this section, we introduce SearchRAG, an innovative RAG framework aimed at enhancing LLMs’ ability to effectively use search engines for knowledge retrieval, especially when addressing complex and specialized medical questions.

A. Problem Formulation and Overview

Retrieval-Augmented Generation (RAG) integrates external knowledge into the input of LLMs to enhance their responses [4], [5]. Formally, given an input prompt x , RAG retrieves a set of knowledge snippets $\mathcal{K} = \{\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_n\}$ using a retrieval function $f_{\text{ret}}: \mathcal{K} = f_{\text{ret}}(x)$, and generates a response y using model f_{LLM} conditioned on both the input and the retrieved knowledge:

$$y = f_{\text{LLM}}(x, \mathcal{K}). \quad (1)$$

However, when dealing with complex and verbose medical questions, such as the one shown in Figure 2, directly using x for retrieval can lead to suboptimal results. This is because such questions often include extraneous details that can mislead search engines and result in the retrieval of irrelevant information [6], [7], [13]. To overcome these limitations,

our proposed SearchRAG modifies the RAG framework by introducing two key components:

- A *Synthetic Query Generation Module* that generates candidate queries by transforming original questions into search-engine-friendly form.
- An *Uncertainty-Based Query Selection* mechanism that leverages the LLM’s internal uncertainty to select the most informative knowledge snippets.

By integrating these components, we enhance the retrieval of pertinent information to answer questions, ultimately improving response accuracy and relevance.

B. Synthetic Query Generation Module

Medical questions are often too complex to be directly used as search engine queries. To address this challenge, SearchRAG proposes to use synthetic queries generated by LLMs to improve the retrieval of knowledge pertinent to answering the original question. The need for synthetic queries stems from two fundamental challenges in medical question answering:

- **Complexity of Medical Questions:** Medical questions often contain specialized terminology and multi-hop clinical logic [13]. A single condition may be referred to by different medical terms (e.g., “myocardial infarction” and “heart attack”), making it challenging to match relevant content comprehensively. Additionally, critical retrieval keywords may not appear explicitly in the original question, which hinders effective retrieval.
- **Constraints of Search Engine:** Existing search engines employ algorithms and ranking systems that favor well-structured, concise queries with clear intent. Medical questions usually fail to meet these requirements, thus preventing the retrieval of relevant medical content.

Our solution leverages the same LLM we aim to enhance through RAG to generate synthetic queries, eliminating the need for additional models. We design a specialized prompt template p that guides the LLM to decompose and transform the original medical question into search queries. More specifically, the LLM first analyzes the medical question to extract key clinical entities, reformulates them using standardized medical terminology, and finally structures the queries with search-friendly syntax. Given an original prompt x and our template p , we generate a set of synthetic queries through repeated sampling:

$$\{q_1, q_2, \dots, q_m\} = \bigcup_{i=1}^m f_{\text{LLM}}(x, p). \quad (2)$$

Here, we employ a high-temperature sampling strategy to generate a *large and diverse* set of candidate queries $\{q_1, q_2, \dots, q_m\}$. Example queries for a question x are shown in Figure 2 with $\{q_1: \text{“ototoxic effects of platinum-based drugs”}; q_2: \text{“mechanisms cisplatin hearing loss”}; q_3: \text{“Proteasome inhibitor drug effects on inner ear”}\}$. These queries contain concise keywords suitable for search engines and cover

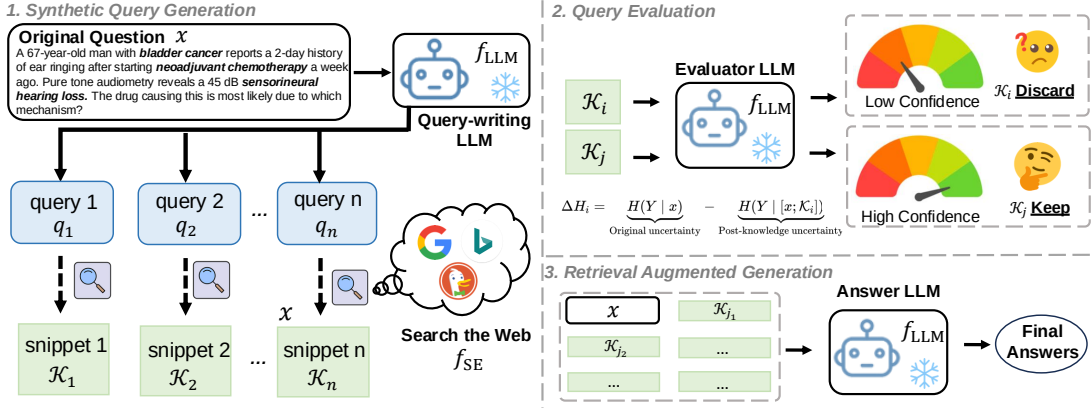


Fig. 2. Overview of SearchRAG: Our framework first transforms complex medical questions into search-optimized synthetic queries through repeated sampling. The retrieved knowledge snippets are then filtered using uncertainty-based selection to identify the most relevant information for enhancing LLM responses.

multiple aspects of question x . Among them, q_2 is the most informative one².

Our strategy is inspired by inference-time scaling law: while not every query can retrieve optimal information, some candidates will be more effective than others at retrieving useful knowledge. The expanded candidate pool increases the probability of obtaining high-quality knowledge snippets. To fully leverage this variability, our subsequent uncertainty-based selection mechanism (detailed in the next section) will identify and utilize the most effective ones.

C. Align Search Engines to LLMs via Uncertainty-Based Query Selection

While we generate a large number of synthetic queries, identifying the most effective ones is a non-trivial task. To address this, we propose to leverage the LLM’s internal uncertainty as a guiding signal to guide query selection. Specifically, we assess the uncertainty reduction associated with each retrieved snippet, treating it as a measure of information gain [14]. Queries that retrieve knowledge leading to greater uncertainty reduction are prioritized, ensuring that only the most informative snippets contribute to RAG.

Our approach aims to align search engine to LLMs by helping LLMs effectively use web search to retrieve knowledge to answer given questions. By using uncertainty as a signal, our system selects queries that maximize knowledge utility, refining the alignment process on a per-query basis. Unlike tuning-based alignment [15], our method optimizes in inference-time without modifying model parameters, making it robust to diverse retrieval sources and evolving topics.

For each query q_i , our search engine function f_{SE} retrieves structured knowledge snippets:

$$\mathcal{K}_i = f_{SE}(q_i), \quad i = 1, 2, \dots, m. \quad (3)$$

²A Google search for q_2 returns: “...cisplatin is retained for months to years. It can cause DNA damage, inhibit protein synthesis...”, which corresponds to the correct answer: “Cross-linking of DNA”. Though not mentioned in the original question, *cisplatin* is a well-known chemotherapy drug.

The retrieved snippets are concise extracts from search engine results, including titles and short descriptions for each webpage. Each retrieved snippet augments the original prompt through concatenation: $x'_i = [x; \mathcal{K}_i]$.

We estimate the LLM’s uncertainty in generating responses based on these augmented inputs using Shannon entropy [16]. Let Y be the random variable representing possible responses y generated by LLMs. Since computing the full response distribution is infeasible [14], [16], we approximate uncertainty using the entropy of the first token, defined as:

$$H(Y | X) = - \sum_{w \in \mathcal{V}} p(w | X) \log_2 p(w | X), \quad (4)$$

where $p(w | X)$ is the probability of generating token w as the first token across the vocabulary $\mathcal{V}$. This approximation is suitable because medical questions can be naturally formulated as multiple-choice questions. The first token entropy directly reflects the model’s confidence in different answer choices when the model directly outputs option labels (ABCD) as the first token, which makes this uncertainty approximation both computationally efficient and semantically meaningful for our task. For each knowledge snippet, we compute the reduction in conditional entropy using the probability of output tokens before and after retrieval:

$$\Delta H_i = \underbrace{H(Y | x)}_{\text{Original uncertainty}} - \underbrace{H(Y | x'_i)}_{\text{Post-retrieval uncertainty}}. \quad (5)$$

A positive value of ΔH_i indicates that the knowledge snippet $\mathcal{K}_i$ has reduced the model uncertainty. The most informative snippets are aggregated into a final knowledge set:

$$\mathcal{K}^* = \bigcup \{ \mathcal{K}_i \mid \Delta H_i > 0 \}. \quad (6)$$

This inference-time alignment leverages the model internal states to ensure that only the knowledge snippets capable of increasing prediction certainty are used for the final generation. In conclusion, our framework samples diverse search queries from the LLM, retrieves relevant knowledge, evaluates snippets by their uncertainty reduction, and aggregates confident information for final answer generation.

III. EXPERIMENTS

A. Experimental Settings

We evaluate our method, SearchRAG, against a comprehensive set of baselines on three medical question-answering benchmarks. This section details the evaluation benchmarks, the methods we compare against, our implementation specifics, and the retrieval configurations used. All experiments were conducted on NVIDIA A6000 GPUs (48GB memory).

1) *Evaluation Benchmarks*: All methods are evaluated on three established multiple-choice medical QA benchmarks. We follow the standard evaluation protocols as used in MedRAG and i-MedRAG [7], [17], including: **MedQA** [2], **MMLU_Med** [18], and **MedMCQA** [3].

2) *Compared Methods*: We compare SearchRAG against a wide array of models to ensure a thorough evaluation.

a) *Non-RAG Baseline*.: We include Chain-of-Thought (CoT) prompting [19] as a representative baseline without any retrieved knowledge enhancement. 1. RAG-based Comparisons. We benchmark against several state-of-the-art RAG models: (1) **MedRAG** [7], which uses a static corpus of textbooks and a large PubMed subset with the MedCPT retriever; (2) **i-MedRAG** [17], an iterative version of MedRAG; and (3) recent general-domain RAG techniques including **Adaptive-RAG** [20], **Rationale-Guided RAG** [21], and **InstructRAG** [22]. 2. Domain-specific Medical LLMs. We evaluate against strong LLMs pre-trained or fine-tuned on a large amount of medical corpora: **MEDITRON** (70B) [23], **PMC-LLaMA** (13B) [24], **MedLLaMA3** (8B) [25], **AdaptLLM-v2** (8B) [26], and **BioMistral** (7B) [27]. 3. Instruction-finetuned RAG Models. We also compare with strong RAG models specifically fine-tuned for retrieval augmented applications: **Self-RAG** (13B) [28], **ChatQA1.5** (8B/70B) [29], and **SimRAG** (27B) [30].

3) *Retrieval Configurations*: The retrieval sources and methods differ between our approach and key baselines.

a) *Baseline Retrieval (Static Corpora)*.: The MedRAG and i-MedRAG baselines retrieve from static, offline corpora. This includes a collection of **18 medical textbooks** [13] and a curated **PubMed subset** of 23.9 million articles [7]. For these methods, we follow their reported setup, using a dense retriever with the MedCPT bi-encoder [7] to fetch the top- k most similar document chunks, with $k = 32$.

b) *SearchRAG Retrieval (Dynamic Web Search)*.: Our method, SearchRAG, utilizes dynamic retrieval via a **live web search**. We use the Google Search engine through a third-party API (<https://serper.dev/>). For each generated query, we extract knowledge from the top 3 organic search results (title and snippet) as well as the knowledge graph or answer box, if available. The retrieved text snippets are typically 50–200 words long. We deliberately do not scrape full web pages and explicitly exclude any AI-generated overview content from the search results, relying only on the concise, source-attributed snippets for context augmentation.

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON MEDICAL QA DATASETS.

Method	MedMCQA	MMLU_Med	MedQA	Avg. Accuracy
Proprietary LLMs				
GPT-3.5 [1]	58.04	75.48	66.61	66.71
GPT-4 [31]	72.02	87.24	82.80	80.69
Medical LLMs				
PMC-LLaMA (13B) [24]	48.29	52.53	42.58	47.80
MEDITRON (70B) [23]	52.67	65.38	49.57	55.87
AdaptLLM-v2 (8B) [26]	42.74	51.24	43.13	45.70
BioMistral(7B) [27]	32.20	47.47	32.52	37.40
MedLLaMA3 (8B) [25]	61.18	77.13	61.43	66.58
Finetuned LLMs for RAG				
Self-RAG (13B) [28]	44.00	53.90	48.60	48.83
ChatQA1.5 (8B) [29]	46.97	61.40	42.36	50.24
ChatQA1.5 (70B) [29]	62.54	80.51	68.89	70.65
SimRAG (27B) [30]	64.16	81.63	63.63	69.80
RAG with LLaMA 3.1-8B				
CoT w/o RAG [19]	55.96	74.52	65.91	65.46
Adaptive-RAG [20]	55.75	76.22	66.06	66.01
Rationale-Guided RAG [21]	59.50	78.05	69.60	69.05
InstructRAG [22]	58.81	77.50	65.59	67.30
MedRAG (Textbooks) [7]	55.89	74.98	64.89	65.25
MedRAG (PubMed) [7]	50.87	71.28	60.41	60.85
MedRAG (SearchEngine) [7]	58.16	76.40	63.63	66.06
i-MedRAG [17]	56.80	74.70	68.97	66.82
Our RAG	65.14	84.67	71.49	73.77
RAG with LLaMA 3.1-70B				
CoT w/o RAG [19]	69.33	87.53	78.08	78.31
Adaptive-RAG [20]	69.57	88.34	78.00	78.64
Rationale-Guided RAG [21]	70.05	89.62	80.44	80.04
InstructRAG [22]	70.38	87.79	77.77	78.65
MedRAG (Textbooks) [7]	69.23	86.33	79.34	78.30
MedRAG (PubMed) [7]	69.95	86.80	79.26	78.67
MedRAG (SearchEngine) [7]	69.69	87.70	77.61	78.33
i-MedRAG [17]	69.23	87.17	79.58	78.66
Our RAG	74.40	90.95	83.34	82.90

B. RQ1: Main Results

Table I reports the performance of SearchRAG across three medical QA benchmarks with both 8B and 70B LLaMA backbones. We summarize three key findings. **First**, traditional RAG methods show limited effectiveness, particularly with smaller models. For instance, MedRAG with PubMed retrieval underperforms the non-RAG CoT baseline on MedMCQA (50.87% vs. 55.96%), highlighting the limitations of static corpora in capturing fine-grained, up-to-date knowledge. Notably, even when MedRAG uses web search as the retrieval source, its performance remains suboptimal (e.g., 58.16% on MedMCQA), indicating that simply replacing the knowledge source with a search engine is insufficient without tailored query formulation and filtering. **Second**, SearchRAG achieves consistent gains over all RAG baselines, reaching 73.77% average accuracy with the 8B model and 82.90% with the 70B model, surpassing all other RAG baselines. Our improvements stem from better alignment between queries and search engines and more precise filtering via uncertainty reduction. **Third**, SearchRAG outperforms all fine-tuned medical LLMs in the domain, including MEDITRON (70B), PMC-LLaMA (13B), and ChatQA1.5 (70B), despite using a general-purpose LLM without additional training. This demonstrates the strength of our proposed retrieval without reliance on domain-specific

pertaining. In summary, the results validate SearchRAG’s two-stage strategy: query optimization and uncertainty-based selection jointly improve retrieval quality. The consistent gains across datasets and model sizes highlight its effectiveness and generalizability.

C. RQ2: Effectiveness of Knowledge Selection

To validate our uncertainty-based knowledge selection mechanism, we conduct experiments on 350 randomly sampled instances from each dataset using both model variants. Contrary to the intuition that unfiltered knowledge provides richer context, Table II reveals that selective filtering consistently improves performance, particularly for smaller models. There are two key insights as follows: (1) The LLaMA 8B model shows 4-7% absolute gains across all datasets, indicating smaller models’ vulnerability to irrelevant information: unfiltered knowledge introduces distracting noise that misleads reasoning. (2) The 70B model is more robust to noise but still improves with our proposed selection techniques, indicating even large models struggle with irrelevant context. These results confirm that our uncertainty-based selection effectively reduces distraction and enhances reasoning by retaining only relevant knowledge. To further assess the quality of retrieved content, we consulted licensed medical professionals, who confirmed that most of the information retrieved via Google Search was factually accurate.

D. RQ3: Impact of the Number of Total Generated Queries

We conducted an evaluation to understand how the number of generated queries affects model performance, using 350 randomly sampled instances. The configurations tested included using 0 (the original question), 4, 16, and 32 generated queries. Figure 3 illustrates the impact of varying the number of queries on performance. Our analysis reveals two key findings: First, using only the original questions (0 query) resulted in limited RAG effectiveness, with performance even degrading compared to the non-RAG baseline, achieving only 62.9% on the MedQA. **This suggests unmodified medical questions fail to effectively leverage search engines for retrieving relevant information**, and the low performance confirms that these medical questions do not have direct matches in existing online knowledge sources. Second, introducing synthetic query generation led to improvements across all datasets. Performance scaled with the number of generated queries, peaking with 32 queries. This clear correlation between query count and accuracy indicates that a larger query set helps capture more diverse and relevant knowledge to enhance the RAG process.

IV. RELATED WORK

While large language models excel at question answering, they struggle with specialized medical knowledge [6]. Traditional RAG systems have been proposed to mitigate these limitations by integrating external knowledge sources into the LLMs’ input, thereby enhancing their ability to generate

accurate and contextually relevant responses [1], [4]. Nonetheless, these early RAG approaches predominantly relied on retrieving information from structured databases or knowledge graphs, which can be insufficient for handling the complexity and nuanced requirements of specialized medical queries [7], [13]. Recent advancements in RAG systems have focused on improving the accuracy and relevance of retrieved medical information. Xiong et al. [17] propose i-MedRAG, which uses iterative follow-up questions to help LLMs better understand complex medical queries through dynamic information seeking. Recent work has also explored uncertainty estimation in RAG systems for various purposes [32], [33]. Despite these advancements, existing RAG frameworks still struggle with retrieving fine-grained and up-to-date knowledge, particularly for complex medical questions. This is because traditional knowledge bases and graphs often lack the granularity and timeliness needed for specialized medical queries. Our approach addresses these limitations by utilizing search engines for real-time knowledge retrieval.

V. CONCLUSION

We introduced SearchRAG, a retrieval-augmented generation framework that enhances large language models for complex medical QA. By combining synthetic query generation with uncertainty-based knowledge selection, SearchRAG optimizes retrieval from search engines. Experiments show notable gains in accuracy and coherence over traditional RAG methods, demonstrating its effectiveness in overcoming their limitations.

REFERENCES

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [2] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, “What disease does this patient have? a large-scale open domain question answering dataset from medical exams,” *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.
- [3] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering,” in *Proceedings of the Conference on Health, Inference, and Learning*, ser. Proceedings of Machine Learning Research, G. Flores, G. H. Chen, T. Pollard, J. C. Ho, and T. Naumann, Eds., vol. 174. PMLR, 07–08 Apr 2022, pp. 248–260. [Online]. Available: <https://proceedings.mlr.press/v174/pal22a.html>
- [4] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval augmented language model pre-training,” in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.
- [5] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, “Kepler: A unified model for knowledge embedding and pre-trained language representation,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 176–194, 2021.
- [6] Y. Shi, S. Xu, Z. Liu, T. Liu, X. Li, and N. Liu, “Mededit: Model editing for medical question answering with external knowledge bases,” *arXiv preprint arXiv:2309.16035*, 2023.
- [7] G. Xiong, Q. Jin, Z. Lu, and A. Zhang, “Benchmarking retrieval-augmented generation for medicine,” *arXiv preprint arXiv:2402.13178*, 2024.
- [8] M. Lykke, S. Price, and L. Delcambre, “How doctors search: A study of query behaviour and the impact on search results,” *Information processing & management*, vol. 48, no. 6, pp. 1151–1170, 2012.
- [9] O. Yoran, T. Wolfson, O. Ram, and J. Berant, “Making retrieval-augmented language models robust to irrelevant context,” *arXiv preprint arXiv:2310.01558*, 2023.

TABLE II
PERFORMANCE COMPARISON WITH/WITHOUT KNOWLEDGE FILTERING ACROSS MODEL SCALES AND DATASETS.

Model	MedMCQA		MMLU_Med		MedQA	
	Unfiltered	Filtered	Unfiltered	Filtered	Unfiltered	Filtered
8B	60.57	64.86 (+7.08%)	81.14	84.86 (+4.58%)	67.71	72.29 (+6.76%)
70B	70.86	74.57 (+5.24%)	88.29	89.71 (+1.61%)	82.57	83.43 (+1.04%)

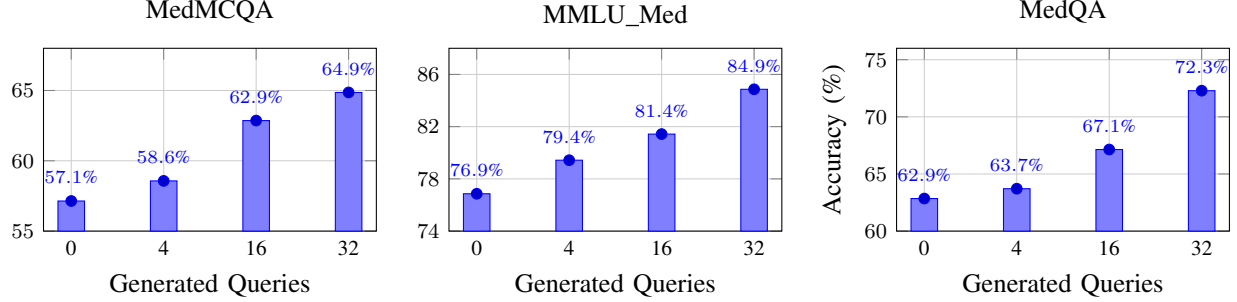


Fig. 3. Comparison of model accuracies for MedMCQA, MMLU_Med, and MedQA across different numbers of generated queries. Note that the 0-query baseline directly uses the original question for retrieval. As the number of generated queries increases, the performance of RAG improves significantly.

- [10] Google, “How we’re improving search results when you use quotes,” 2022, accessed: 2025-02-13. [Online]. Available: <https://blog.google/products/search/how-were-improving-search-results-when-you-use-quotes/>
- [11] Ahrefs, “Google advanced search operators: The complete list (42 advanced operators),” <https://ahrefs.com/blog/google-advanced-search-operators/>, 2024, accessed: 2024-06-17.
- [12] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri, “The power of noise: Redefining retrieval for rag systems,” *arXiv preprint arXiv:2401.14887*, 2024.
- [13] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, “What disease does this patient have,” *A Large-scale Open Domain Question Answering Dataset from Medical Exams. arXiv [cs. CL]*, 2020.
- [14] Y. Shi, Q. Tan, X. Wu, S. Zhong, K. Zhou, and N. Liu, “Retrieval-enhanced knowledge editing in language models for multi-hop question answering,” in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024, pp. 2056–2066.
- [15] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, “Query rewriting for retrieval-augmented large language models,” *arXiv preprint arXiv:2305.14283*, 2023.
- [16] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.
- [17] G. Xiong, Q. Jin, X. Wang, M. Zhang, Z. Lu, and A. Zhang, “Improving retrieval-augmented generation in medicine with iterative follow-up questions,” in *Biocomputing 2025: Proceedings of the Pacific Symposium*. World Scientific, 2024, pp. 199–214.
- [18] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, “Measuring massive multitask language understanding,” *arXiv preprint arXiv:2009.03300*, 2020.
- [19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [20] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park, “Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity,” *arXiv preprint arXiv:2403.14403*, 2024.
- [21] J. Sohn, Y. Park, C. Yoon, S. Park, H. Hwang, M. Sung, H. Kim, and J. Kang, “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)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 12 739–12 753. [Online]. Available: <https://aclanthology.org/2025.naacl-long.635/>
- [22] Z. Wei, W.-L. Chen, and Y. Meng, “InstructRAG: Instructing retrieval-augmented generation via self-synthesized rationales,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=P1qhkp8gQT>
- [23] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami *et al.*, “Meditron-70b: Scaling medical pretraining for large language models,” *arXiv preprint arXiv:2311.16079*, 2023.
- [24] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, “Pmc-llama: toward building open-source language models for medicine,” *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1833–1843, 2024.
- [25] J. S. Labs, “Jsl-medllama-3-8b-v2.0,” <https://huggingface.co/johnsnowlabs/JSI-MedLlama-3-8B-v2.0>, 2024, model available under CC BY-NC-ND 4.0 license.
- [26] D. Cheng, Y. Gu, S. Huang, J. Bi, M. Huang, and F. Wei, “Instruction pre-training: Language models are supervised multitask learners,” *arXiv preprint arXiv:2406.14491*, 2024.
- [27] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, “Biomistral: A collection of open-source pretrained large language models for medical domains,” *arXiv preprint arXiv:2402.10373*, 2024.
- [28] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “Self-rag: Learning to retrieve, generate, and critique through self-reflection,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [29] Z. Liu, W. Ping, R. Roy, P. Xu, C. Lee, M. Shoenybi, and B. Catanzaro, “Chatqa: Surpassing gpt-4 on conversational qa and rag,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 15 416–15 459, 2024.
- [30] R. Xu, H. Liu, S. Nag, Z. Dai, Y. Xie, X. Tang, C. Luo, Y. Li, J. C. Ho, C. Yang, and Q. He, “SimRAG: Self-improving retrieval-augmented generation for adapting large language models to specialized domains,” 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)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 11 534–11 550. [Online]. Available: <https://aclanthology.org/2025.naacl-long.575/>
- [31] OpenAI, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: <https://arxiv.org/abs/2303.08774>
- [32] S. Ozaki, Y. Kato, S. Feng, M. Tomita, K. Hayashi, R. Obara, M. Oyama, K. Hayashi, H. Kamigaito, and T. Watanabe, “Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain,” *arXiv preprint arXiv:2412.20309*, 2024.
- [33] Z. Li, J. Xiong, F. Ye, C. Zheng, X. Wu, J. Lu, Z. Wan, X. Liang, C. Li, Z. Sun *et al.*, “Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation,” *arXiv preprint arXiv:2410.02719*, 2024.