Rationale-Guided Retrieval Augmented Generation for Medical Question Answering

Anonymous ACL submission

Abstract

Large language models (LLM) hold significant potential for applications in biomedicine, but they struggle with hallucinations and outdated knowledge. While retrieval-augmented generation (RAG) is generally employed to address these issues, it also has its own set of challenges: (1) LLMs are vulnerable to irrelevant or incorrect pieces of documents, (2) medical queries are often not well-targeted for helpful information, and (3) the retriever is prone to bias toward the specific source corpus it was trained on. In this study, we present RAG^2 (RAtionale-Guided RAG), a new framework for enhancing the reliability of RAG in biomedical contexts. RAG^2 incorporates three key innovations: a small surrogate model trained on perplexity-based labels of rationales, which selectively augments informative documents while filtering out distractors; LLM-generated rationales as queries to improve the utility of retrieved documents; a structure designed to retrieve documents evenly from a comprehensive set of four biomedical corpora, effectively mitigating retriever bias. Our experiments demonstrate that RAG^2 improves the state-of-the-art LLMs of varying sizes, with improvements of up to 12%, and it outperforms existing RAG models by up to 14.5% across three medical question-answering benchmarks.

1 Introduction

Large language models (LLM), including proprietary models like GPT-4 (OpenAI, 2023) and Med-Gemini (Saab et al., 2024), as well as open-source models such as Llama-3 (AI@Meta, 2024), have demonstrated remarkable performance across various tasks in biomedicine, including USMLE-style question-answering (QA) benchmarks (Jin et al., 2021).[1] Despite their state-of-the-art performance, LLMs still face major challenges that limit their widespread adoption in high-stakes areas. One major concern is hallucination, where the model produces information that sounds plausible but is incorrect (Maynez et al., 2020; Huang et al., 2023). This challenge is particularly critical in the medical field, where precise and accurate information is crucial. Another concern is outdated knowledge; updating the model's parametric knowledge is a resource-intensive process, making it challenging to ensure the model has up-to-date information. applications (Zhang et al., 2023; Kasai et al., 2024).

Retrieval-augmented generation (RAG) (Lewis et al., 2020) has emerged as a promising solution to address these limitations. RAG techniques aim to reduce hallucinations and ensure the model delivers up-to-date information by integrating relevant and trustworthy documents directly into its input. RAG in the biomedical domain poses unique challenges. First, LLMs are vulnerable to being misled by irrelevant or incorrect documents, which can undermine the primary objective of RAG—enhancing the model's reliability (Wu et al., 2024). Furthermore, the complexity and specialized terminology prevalent in biomedical texts introduce additional hurdles. These characteristics increase the risk of retrieving irrelevant/unhelpful information and complicate the generation of accurate outputs based on the retrieved documents. Second, medical queries often struggle to effectively target useful information. Some questions include extensive patient details, such as medical histories, symptoms, and test results, which are crucial for clinical reasoning. However, this abundance of information can overwhelm retrieval systems, making it difficult to identify the most critical diagnostic clues. Conversely, overly brief queries may lack the necessary context and rely on implicit medical knowledge that is not explicitly stated, complicating effective retrieval. Additionally, the complexity and specialized terminology prevalent in biomedical texts

[1] The USMLE or United States Medical Licensing Examination is a standardized test that is required for obtaining a medical license in the United States.

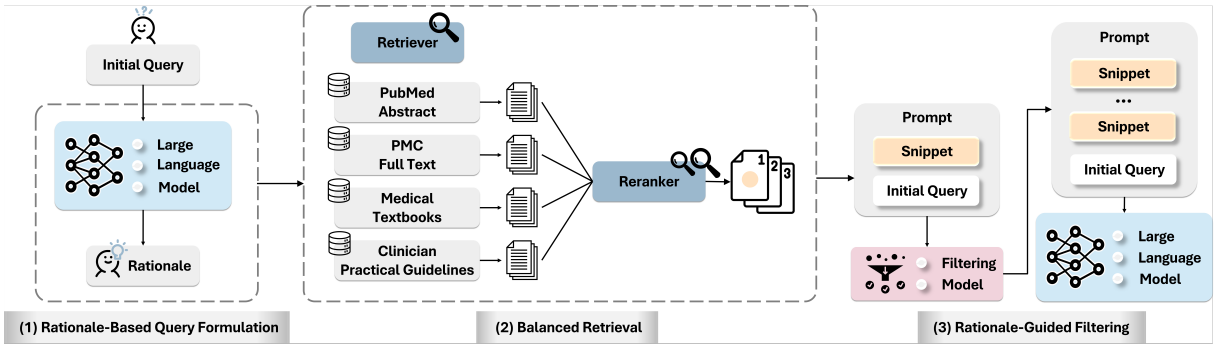

Figure 1: Our RAG^2 framework comprises the following three steps. (1) Rationale-guided filtering: A surrogate model determines whether incorporating the retrieved documents into the LLM prompt increases the confidence and accuracy of the LLM's rationale. (2) Rationale-based query formulation: The model-generated rationale is used as the query for document retrieval instead of the initial query. (3) Balanced retrieval: document documents are retrieved in equal amounts from four corpora: two large ones, PubMed and PMC, and two smaller but clinically focused corpora, textbooks, and clinical guidelines. This balanced approach ensures fair inclusion of document from all sources, preventing over-reliance on larger corpora, which could otherwise leave out valuable out-of-domain document. Subsequently, a reranker is used to rerank the retrieved documents by cross-encoding the initial query and each document.

further complicate retrieval by increasing the risk of retrieving irrelevant or unhelpful information and hindering the generation of accurate outputs based on the retrieved documents.

Retriever bias (Chen et al., 2021; Dai et al., 2024) also presents a significant challenge in biomedical information retrieval, particularly due to the vast size of certain medical corpora and the biases in training data. As a result, larger sources often overshadow smaller, specialized ones that may contain critical, up-to-date information, such as clinical guidelines. This issue is further exacerbated by dense retrievers which are trained to prioritize text from their training corpus, leading to the potential exclusion of relevant out-of-domain evidence. For instance, MedCPT (Jin et al., 2023) might favor PubMed documents, potentially neglecting specialized or out-of-domain sources that are crucial for accurately answering a query.

Building upon these insights, we propose RAG^2 (RAtionale-Guided Retrieval Augmented Generation), a novel framework improving the reliability of RAG in biomedical applications (see Figure 1). First, we introduce a rationale-guided filtering method, training a Flan-T5 model (Chung et al., 2024) using labels derived from the differences in perplexity between rationales with and without retrieved documents. This model estimates the documents' informativeness for the base LLM, selectively augmenting useful information while filtering out potential distractors. Next, our method substitutes direct medical questions with LLM-generated rationales as queries using chain-of-thought prompting (Wei et al., 2022; Jiang et al., 2023b; Kang et al., 2024). By using rationale-based queries, which involve systematically breaking down and solving problems step by step, key components of complex queries can be identified, while also enabling query expansion for brief questions (Jagerman et al., 2023; Wang et al., 2023a). Finally, we implement a balanced retrieval strategy that retrieves documents equally from diverse biomedical corpora. This approach mitigates the potential bias present in dense retrievers by promoting a more comprehensive retrieval process across multiple information sources, regardless of their size or level of training exposure.

We evaluate our RAG^2 framework using three closed-book medical QA benchmarks—MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MMLU-Med (Hendrycks et al., 2021)—where no oracle documents are present in the retrieval corpus. Our method significantly enhances the average accuracy of state-of-the-art LLMs of varying sizes, including Llama-3-8B (AI@Meta, 2024), Meerkat-7B (Kim et al., 2024), and GPT-4o (OpenAI, 2024). Additionally, it consistently surpasses the performance of a recently developed medical RAG framework, MedRAG (Xiong et al., 2024a), notably achieving a 14.5% improvement in the performance of the Llama-3-8B model.

Our contributions are summarized as follows:

1. We introduce RAG^2, A novel RAG framework for medical QA, enhancing filtering, querying, and corpus indexing processes.

2. Our novel rationale-guided filtering method employs a small model trained with data from rationale perplexity to assess document relevance, enhancing efficiency and novelty and address medical domain annotation scarcity.

3. In closed-book medical QA benchmarks, we empirically demonstrate that the proposed RAG^2 framework significantly outperforms the previous medical qa methods, including large commercial, small open-source, and medical-specialized models, with an average accuracy improvement of 10.1%, 3.8%, and 0.3%, respectively.

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023) simply append retrieved documents to inputs, often failing to deliver consistently high-quality information. Several advanced approaches involve training LLMs to adaptively perform RAG (Zhang et al., 2024; Gan et al., 2024). Self-RAG (Asai et al., 2023), for instance, trains an LLM with synthetic data from GPT-4 to determine whether to use retrieved document or not. CRAG (Yan et al., 2024) uses an iterative approach with an evaluator model and ChatGPT for query rewriting. However, training LLMs is computationally expensive, particularly when using two LLMs, as seen in FILCO (Wang et al., 2023b). In contrast, our RAG^2 framework utilizes a small model with 770M parameters to assess document utility, allowing for efficient filtering without the need to fine-tune the entire LLM.

Other approaches focus on training smaller models separately from the base LLM. For instance, ARES (Saad-Falcon et al., 2024) employs three small language models trained on a mix of synthetic and human-labeled data, though it still involves substantial costs. Adaptive-RAG (Jeong et al., 2024b) assesses query complexity based on changes in the correctness of an LLM's response, but this method only applies to queries that are answered correctly, overlooking the nuanced effects of retrieved documents beyond basic accuracy. In contrast, our method measures perplexity differentials in the rationales generated by the base LLM, providing the filtering model with more detailed training signals without requiring additional annotation efforts.

Speculative RAG (Wang et al., 2024b) and other concurrent methods leverage rationales to enhance RAG. It generates multiple pseudo answers and uses a large verifier model to select the best rationale. RAT (Wang et al., 2024a) iteratively retrieves and refines rationales in multiple steps. However, these approaches come with significant downsides, such as high computational costs due to the use of large models for training and multiple LLM calls, as well as increased latency from the iterative retrieval and decision-making processes.

Uncertainty-based approaches have gained attraction to dynamically guide the usage of RAG. FLARE (Jiang et al., 2023b) actively decides when and what to retrieve based on iterative prediction for the next sentence, using low confidence as a guideline for retrieval decisions. SEAKR (Yao et al., 2024) similarly utilizes uncertainty to determine when and how to retrieve information by re-ranking documents based on uncertainty levels. However, they both rely on numerous iterative pseudo-generations, leading to significant computational costs as well. Our approach also incorporates uncertainty but limits its use to only the training stage, thereby avoiding costly iterations at the inference stage.

2.2 Medical RAG

Researchers have proposed several methods to enhance biomedical information retrieval and RAG frameworks. MedCPT (Jin et al., 2023) introduced off-the-shelf retriever and reranker models tailored for biomedical information retrieval. It overcame the limited availability of query-article annotations in medicine domain by using 255 million user click logs from PubMed. MedRAG (Xiong et al., 2024a) is the first RAG toolkit in this field to comprehensively integrate indexing, retrieval, and reranking processes. It extracts documents from the MedCorps corpus using a hybrid approach that integrates both sparse and dense retrievers. Self-BioRAG (Jeong et al., 2024a) adopted the concept of Self-RAG to the domain to handle complex medical queries more effectively.

Concurrent with our work, Bailicai (Long et al., 2024) introduced notable features, including self-knowledge boundary identification and directed acyclic graph-based task decomposition. Additionally, i-MedRAG (Xiong et al., 2024b) integrates iterative follow-up query generation. We would like to note that, with the exception of MedRAG, all other approaches necessitate fine-tuning LLMs

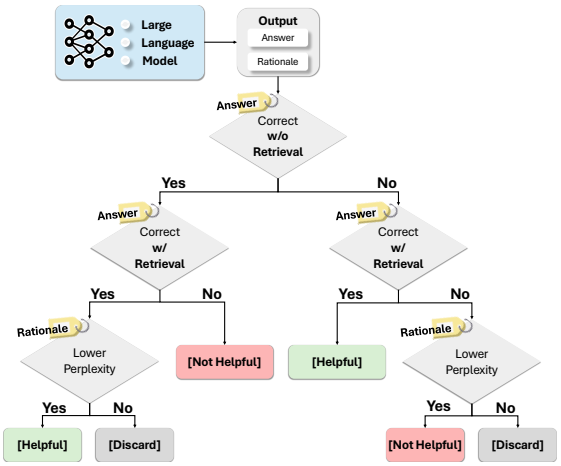

Figure 2: Illustration of the data annotation process used to train our rationale-guided filtering model.

or performing multiple iterative RAG executions. In contrast, our method highlights improved training efficiency and run-time performance by training only a small filtering model and relying on single-step retrieval.

3 Method

We start by explaining the training objective of RAG models. Next, we provide a detailed breakdown of the pipeline, focusing on three key steps: (1) rationale-guided filtering, (2) rationale-based query formulation, and (3) balanced retrieval. Figure 1 presents an overview of the entire model.

3.1 Objective

Let $\mathbf{x} = [x_1, \ldots, x_N]$ represent a natural language query consisting of N tokens, and let $\mathbf{y} = [y_1, \ldots, y_M]$ represent the output sequence consisting of M tokens. An LLM is trained to minimize the probability of generating the output sequence $\mathbf{y}$ by maximizing the likelihood of the following conditional probabilities for each token y_t as follows:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{D}^*) = \prod_{t=1}^{M} p(y_t \mid \mathbf{y}_{<t}, \mathbf{x}, \mathbf{D}^*), \quad (1)$$

where $\mathbf{y}_{<t}$ represents the sequence of tokens generated up to time step $t-1$, and $\mathbf{D}^* = [\mathbf{d}_1, \ldots, \mathbf{d}_k]$ represents the top-k retrieved documents that are most relevant to the query. The retrieval process involves embedding both the query and documents in a shared latent space using a query encoder f_q and an document encoder f_d. The most relevant documents are then selected based on their inner product as follows:

$$\mathbf{D}^* = \text{Top-k}\left(\langle f_q(\mathbf{x}), f_d(\mathbf{d})\rangle | \mathbf{d} \in \mathcal{C}\right), \quad (2)$$

where $\mathcal{C}$ represents the set of candidate documents.

3.2 Rationale-Guided Filtering

Retrieval mechanisms do not always pick the most relevant documents as similarity alone does not ensure a positive contribution to the model's output. To filter out distractors and obtain only useful document, we train a filtering model to evaluate the retrieved documents before they are presented to the LLM. Our method begins by analyzing the LLM's outputs when answering questions under two conditions: (1) using its intrinsic parametric knowledge, and (2) using external information retrieved from a corpus. If the model successfully answers a question with the aid of retrieved documents but fails to do so independently, the documents are labeled as "helpful"; otherwise, they are labeled as "not helpful.". This labeling strategy was similarly adopted in previous works (Jeong et al., 2024b), but it fails to fully capture the document utility, particularly in cases where the model's accuracy remains unchanged regardless of whether the document is included.

To address this, we calculate the difference in perplexity, ΔPPL, to determine the document's impact on the model's confidence as follows:

$$\Delta\text{PPL} = \text{PPL}(\mathbf{x}) - \text{PPL}(\mathbf{x}, \mathbf{d}) \geq \tau, \quad (3)$$

where $\mathbf{d}$ is the target document and τ represents a threshold set to select the top percentage of perplexity differentials. The perplexity scores for the input query $\mathbf{x}$ and the document $\mathbf{d}$ are calculated as follows:

$$\text{PPL}(\mathbf{x}) = \exp\left(-\frac{1}{L}\sum_{i=0}^{L-1} \log P(x_i \mid \mathbf{x}_{<i})\right),$$
$$\text{PPL}(\mathbf{x}, \mathbf{d}) = \exp\left(-\frac{1}{L}\sum_{i=0}^{L-1} \log P(x_i \mid \mathbf{x}_{<i}, \mathbf{d})\right). \quad (4)$$

It is important to note that while lower perplexity (i.e., higher model confidence) generally correlates with higher accuracy, this relationship is not always consistent. Perplexity alone cannot reliably guarantee accuracy, making it necessary to introduce a threshold for more precise labeling. Through our validation process, we determined that setting the

threshold value τ to the top 25% of perplexity differentials consistently yielded the best performance and was therefore fixed across all our experiments.

Figure 2 illustrates the data annotation process. By applying this systematic labeling approach, we generate a perplexity-stratified dataset derived from the base model's outputs, which is subsequently used to train a Flan-T5-large model as our filtering model. This method ensures that the filtering model effectively assesses the utility of retrieved documents for the base model, allowing only relevant and beneficial information to be integrated into the LLM.

3.3 Rationale-Based Query Formulation

As we discussed in the introduction, the initial query, $\mathbf{x}$, can be overly lengthy or too brief, increasing the complexity of medical QA. To address this, we use model-generated rationales as queries to search for relevant information (Wang et al., 2023a; Kang et al., 2024). Following Kim et al. (2024), we extract the LLM's rationales using the following chain-of-thought prompt:

> *The following are multiple choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output your explanation and single option from the given options as the final answer.*
> *Here is the question*: `[initial_query]`

Here, we replace `[initial_query]` with the initial query $\mathbf{x}$ and obtain the model's response. We search for document documents solely using the rationale, excluding the initial query. Including both the initial query and the rationale exceeds the maximum length of the retriever and, as confirmed by our initial experiments, leads to suboptimal performance. Also, we note that the same LLM is used both for rationale generation and QA. In our experiments, we show that rationale queries can still be effective even with smaller open-source models (see Table 2 in Experiments).

3.4 Balanced Retrieval

Searching document from a broad corpus can provide high coverage, but it can lead the retriever to favor dominant corpora and popular information (Chen et al., 2021; Kim et al., 2023). As a result, smaller yet essential corpora may be underrepresented. To address this issue, we use balanced retrieval, a simple yet effective method. This approach extracts an equal number of documents from each corpus, ensuring that all corpora are represented more evenly compared to existing methods (Xiong et al., 2024a). After the balanced retrieval stage, we further refine the selection of relevant documents. We use MedCPT (Jin et al., 2023), an off-the-shelf reranker, which encodes the original query along with each document to determine a relevance score. This process prioritizes documents that are more closely aligned with the query by reranking the results initially retrieved from multiple sources in a balanced manner, enhancing overall performance.

Dataset	Training	Validation	Test
MedQA	10,178	1,272	1,273
MedMCQA	182,822	4,183	6,150
MMLU-Med	-	-	1,089

Table 1: The number of question-answer pairs in the three medical QA datasets, MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and MMLU-Med (Hendrycks et al., 2021).

4 Experiments

4.1 Datasets

We utilize three widely recognized medical QA datasets. (1) MedQA (Jin et al., 2021) consists of USMLE-style questions curated by medical examination experts from various medical question banks. (2) MedMCQA (Pal et al., 2022) comprises exam questions from the two Indian affiliations, AIIMS (All India Institute of Medical Sciences) and NEET PG (National Eligibility cum Entrance Test for Post Graduate courses). (3) MMLU-Med, a subset of the MMLU dataset (Hendrycks et al., 2021), consists of questions for six biomedical subjects—clinical knowledge, medical genetics, anatomy, professional medicine, college biology, and college medicine—spanning from high school to professional-level knowledge. All of these datasets consist of multiple-choice questions with four answer options. Detailed statistics about the datasets can be found in Table 1.

4.2 Models

Large Language Models We used state-of-the-art baseline models from three categories: commercial LLMs (GPT-3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2023), GPT-4o), open-source LLMs (Llama-2-7B (Touvron et al., 2023), Mistral-

Model	MedQA	MedMCQA	MMLU-Med	Average
Commercial LLMs (zero- or few-shot)				
GPT-3.5 (OpenAI, 2022) (5-shot)	53.6	51.0	67.3	57.3
GPT-4 (OpenAI, 2023) (5-shot)	81.4	72.4	87.1	80.3
GPT-4o (0-shot)	88.5	76.7	**92.8**	86.0
+ MedCPT (Jin et al., 2023) ($k = 1$)	86.6	72.5	90.1	83.1
+ MedCPT w/ Rationale Query ($k = 1$)	87.3	74.7	90.2	84.1
+ MedRAG (Xiong et al., 2024a)	88.3	75.9	92.4	85.5
+ InstructRAG-ICL (2-shot) (Wei et al., 2024)	87.7	73.5	90.0	85.6
+ RAG^2 (**Ours**)	**89.8**	**76.8**	92.2	**86.3**
Open-source LLMs (zero- or few-shot)				
Llama-2-7B (Touvron et al., 2023) (3-shot)	35.2	36.3	46.3	39.3
Mistral-7B (Jiang et al., 2023a) (0-shot)	41.1	40.2	55.8	45.7
Llama-3-8B (AI@Meta, 2024) (0-shot)	57.7	53.5	69.5	60.2
+ MedCPT (Jin et al., 2023) ($k = 1$)	55.3	51.3	65.8	57.5
+ MedCPT w/ Rationale Query ($k = 1$)	58.0	52.1	70.3	60.1
+ MedRAG (Xiong et al., 2024a)	56.4	56.6	69.2	60.7
+ InstructRAG-ICL (2-shot) (Wei et al., 2024)	55.5	58.1	71.9	61.8
+ RAG^2 (**Ours**)	**64.6**	**59.4**	**74.8**	**66.3**
Medical LLMs (fine-tuned)				
MediTron-7B (Chen et al., 2023)	50.2	57.9	56.7	54.9
BioMistral-7B (Labrak et al., 2024)	54.3	61.1	64.6	60.0
Meerkat-7B (Kim et al., 2024)	71.2	60.8	73.8	68.6
+ MedCPT (Jin et al., 2023) ($k = 1$)	71.8	57.9	74.0	67.9
+ MedCPT w/ Rationale Query ($k = 1$)	73.3	58.4	75.7	69.1
+ MedRAG (Xiong et al., 2024a)	67.9	60.6	76.1	68.2
+ InstructRAG-ICL (2-shot) (Wei et al., 2024)	65.8	53.2	63.7	60.9
+ RAG^2 (**Ours**)	**75.6**	**63.0**	**78.7**	**72.4**

Table 2: Performance (accuracy) of LLMs and RAG models on the three medical QA benchmarks. MedRAG and RAG^2 were applied to the underlying models, specifically GPT-4o, Llama-3-8B, and Meerkat-7B. Note that Mistral-7B and Llama-3-8B are instruction-tuned, whereas Llama-2-7B is not. MedCPT used only the top-1 document, while MedRAG and RAG^2 utilized the optimal top-k values, which were determined through validation. The best scores for each benchmark are highlighted in bold.

7B-instruct (Jiang et al., 2023a), Llama-3-8B-instruct (AI@Meta, 2024)), and medical LLMs (MediTron-7B (Chen et al., 2023), BioMistral-7B (Labrak et al., 2024), Meerkat-7B (Kim et al., 2024)). From these categories, we selected the top-performing models based on benchmark performance as our backbone models. (1) GPT-4o is the latest version of OpenAI's closed-source LLM, demonstrating state-of-the-art performance across various tasks. (2) Llama-3-8B-instruct[2], part of the Llama series (Touvron et al., 2023), is the state-of-the-art open-source model. We use this model as-is, without further fine-tuning on the target datasets. For simplicity, we refer to it as Llama-3-8B. (3) Meerkat-7B is initialized using the Mistral-7B weights (Jiang et al., 2023a) and further instruction-tuned with rationales generated by GPT-4. The model is then fine-tuned specifically on the MedQA and MedMCQA datasets. We applied our RAG method to these models to assess whether our RAG method can further enhance their high performance without the need for fine-tuning.

RAG Methods We used the following baselines. (1) MedCPT (Jin et al., 2023) is a retriever pre-trained using user search logs from PubMed. (2) "MedCPT w/ Rationale Query" refers to the MedCPT retriever enhanced with model-generated rationale queries. (3) MedRAG (Xiong et al., 2024a) employs four different sparse and dense retrievers to gather information from a collection of corpora called MedCorp, then reranks the retrieved results using Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). (4) InstructRAG (Wei et al., 2024) uses in-context learning to explicitly denoise retrieved documents and generate rationales. It first prompts a language model to generate explanatory rationales for training samples, then uses these rationales as demonstrations for in-context learning during inference. This approach enables the model to learn explicit denoising and reasoning strategies

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

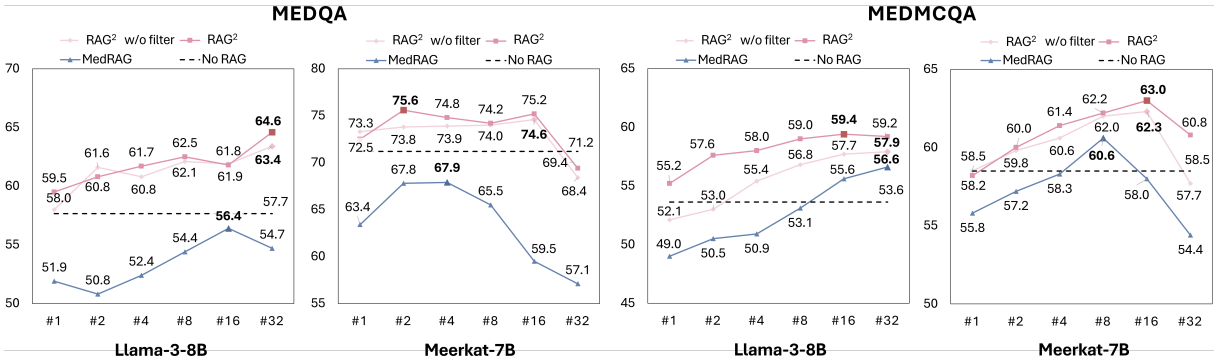

Figure 3: Accuracy (y-axis) of Llama-3-8B and Meerkat-7B with/without different RAG models when varying the number of top-k documents (x-axis). The performance on the MMLU-Med dataset can be found in the Appendix.

without fine-tuning. We chose not to include Self-BioRAG (Jeong et al., 2024a) in our experiments because our goal is to demonstrate an efficient RAG approach that focuses on training only a small filtering model, without the need to train the LLM itself. While Self-BioRAG utilizes GPT-4 to create instruction-following datasets and then trains the Llama-2 7B and 13B models, we instead train a filtering model based on Flan-T5-large, which has only 770M parameters (Chung et al., 2024). This smaller model can be trained on a single RTX 3090 24G GPU.

Implementation Details We trained our filtering model using QA pairs from MedQA and MedMCQA. We used only 40K of the MedMCQA dataset, a subset of the full training data. Since MMLU-Med does not provide training data, we used the MedMCQA training data to train the filtering model. Further details on the validation process and hyperparameter settings are provided in the appendix. Details of the retrieval corpus and index are described in the appendix.

4.3 Main Results

Table 2 presents the accuracy of various LLMs, both with and without MedRAG and RAG^2, across three medical QA benchmarks. The results indicate that our RAG^2 method consistently enhances the baseline LLMs, yielding an average score improvement of 0.3%, 10.1%, and 3.8% for GPT-4o, Llama-3-8B-instruct, and Meerkat-7B, respectively. In contrast, MedCPT, MedCPT w/ Rationale Query, and MedRAG show inconsistent results, with average scores declining. This suggests that without advanced retrieval and filtering modules, RAG frameworks do not always guarantee improved performance, especially in the medical domain. Additionally, the performance improvement on MMLU-Med was 5.3% for Llama-3-8B and 4.9% for Meerkat-7B, demonstrating the transferability of our filtering model to out-of-domain datasets.

Method	Med-QA	Med-MCQA
Llama-3-8B (MedCPT w/o Filtering)	55.3	51.3
+ Adaptive-RAG (Jeong et al., 2024b)	57.3	53.1
+ GPT-4o	58.5	**55.8**
+ Rationale-guided Filtering (**Ours**)	**58.6**	**55.8**

Table 3: Performance (accuracy) of Llama-3-8B depending on the filtering method applied. The top-1 documents are used.

5 Analysis

5.1 Ablation Study

The Number of Retrieved documents Figure 3 illustrates the performance of various RAG methods across different top-k values, using Llama-3-8B and Meerkat-7B as backbone LLMs. MedRAG demonstrates inconsistent performance, often underperforming compared to LLMs without retrieval capabilities. In contrast, the use of rationale queries generally enhances the performance of the baseline LLMs, leading to improvements of up to 5.7% on MedQA and 3.4% on MedMCQA for Llama-3-8B, and up to 4.3% on MedQA and 3.8% on MedMCQA for Meerkat-7B. Further application of rationale-guided filtering results in additional performance gains, with Llama-3-8B achieving up to 6.2% on MedQA and 5.8% on MedMCQA, and Meerkat-7B reaching up to 4.4% on MedQA and 4.5% on MedMCQA.

For Llama-3-8B, performance tends to improve as the top-k value increases, while Meerkat-7B shows a peak in performance followed by a decline. This difference could be attributed to the models' trained context lengths: Llama-3-8B is trained with

Method	MedQA	MedMCQA	MMLU-Med
LLama-3-8B			
+ MedRAG	51.9	49.0	65.1
+ Ours	**55.3**	**51.3**	**65.8**
Meerkat-7B			
+ MedRAG	63.2	56.0	73.1
+ Ours	**71.8**	**57.9**	**74.0**

Table 4: Comparison of MedRAG (Xiong et al., 2024a) and our balanced retrieval. The top-1 documents are used.

a maximum context length of 8K, whereas Meerkat-7B is instruction-tuned with a 2K maximum length. This likely explains the performance drop when longer document is provided. Nonetheless, across all top-k values, our RAG^2 method consistently outperforms other methods.

Filtering Methods We investigated how the performance of the baseline model, Llama-3-8B, is affected by different filtering methods. To isolate the impact of filtering in this experiment, we fixed the top-k value at 1 and used the original query instead of a rationale-based query. Our baseline approach was inspired by the state-of-the-art Adaptive-RAG method (Jeong et al., 2024b). We labeled training set queries as "simple" if the model could answer them correctly without retrieval and "complex" if it could only answer them correctly with retrieval. We then trained a small binary classifier using these labeled queries as input. The "GPT-4o" filtering determines whether a retrieved document contributes to answering the question. On the other hand, our filtering method uses the perplexity of the LLM's rationale to annotate question-document pairs. As a result, this approach significantly improves Llama-3-8B's performance, surpassing Adaptive-RAG by 1.3% on MedMQ and 2.7% on MedMCQA. Additionally, it achieves results similar to GPT-4o while avoiding its API costs.

Effect of Balanced Retrieval Table 4 demonstrates that our balanced retrieval consistently outperforms MedRAG, highlighting the effectiveness of extracting information in a balanced manner from each corpus. We used the same initial query and top-1 document, without applying other methods, to solely compare the retrieval component. Additionally, further studies with corpus ablation, as well as ablation studies of balanced retrieval, are provided in the appendix to demonstrate how this approach mitigates retriever bias.

5.2 Case Study

Figure A4 (appendix) presents an example from the MedQA dataset involving a patient with a severe COPD exacerbation. Initially, the Meerkat-7B model recognized the severity of the situation but failed to prioritize less invasive options like BiPAP (Option A), which would be appropriate before considering intubation (Option C). This outcome highlights the limitations of relying solely on the model's parametric knowledge. When applying RAG with the rationale query, the model was presented with a document containing information about tension pneumothorax and needle decompression. Although this information was medically accurate, it was irrelevant to the patient's condition of COPD exacerbation. The presence of this extraneous data caused the model to incorrectly infer the possibility of a tension pneumothorax, leading to an inappropriate recommendation for needle decompression (Option D). After applying our filtering method to remove the irrelevant information about tension pneumothorax, the model correctly identified the patient's condition as severe COPD exacerbation. As a result, it selected BiPAP (Option A) as the appropriate next step. This example underscores the critical role of document filtering in reliable RAG: the initial error stemmed from the model's distraction by irrelevant information, but proper filtering led to the correct diagnosis and management plan. This case highlights how our method could improve the base LLM's ability to deliver more accurate clinical guidance by prioritizing relevant document and minimizing the risk of misdiagnosis in real-world scenarios.

6 Conclusion

We present RAG^2, a novel framework designed to improve the reliability and performance of LLMs in medical question-answering tasks. Our approach integrates rationale queries, balanced retrieval, and a rationale-guided filtering method. This triad of enhancements serves to thoroughly refine and optimize the entire RAG pipeline. Experimental results consistently demonstrate that RAG^2 significantly improves the accuracy of various state-of-the-art LLMs across multiple benchmarks, marking a significant step toward more reliable AI-assisted medical decision-making.

7 Limitation

Our RAG^2 framework has several limitations. First, it has only been tested in the biomedical domain, and its applicability to general domains remains unexplored. Second, we used only one size of the Flan-T5 model as the surrogate filtering model; experimenting with different sizes or architectures could lead to further improvements. Lastly, due to the Flan-T5 model's limited context length, it can only filter evidence documents one by one, potentially missing the collective impact of multiple relevant documents, though it still proves effective in practice.

References

AI@Meta. 2024. Llama 3 model card.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

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)*, pages 4472–4485.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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*, pages 758–759.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024. Neural retrievers are biased towards llm-generated content. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Chunjing Gan, Dan Yang, Binbin Hu, Hanxiao Zhang, Siyuan Li, Ziqi Liu, Yue Shen, Lin Ju, Zhiqiang Zhang, Jinjie Gu, et al. 2024. Similarity is not all you need: Endowing retrieval augmented generation with multi layered thoughts. *arXiv preprint arXiv:2405.19893*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024a. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024b. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

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):6421.

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):btad651.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: what's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. 2024. Small language models learn enhanced reasoning skills from medical textbooks. *arXiv preprint arXiv:2404.00376*.

Hyunjae Kim, Jaehyo Yoo, Seunghyun Yoon, and Jaewoo Kang. 2023. Automatic creation of named entity recognition datasets by querying phrase representations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7148–7163.

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*, pages 611–626.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

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*, 33:9459–9474.

Cui Long, Yongbin Liu, Chunping Ouyang, and Ying Yu. 2024. Bailicai: A domain-optimized retrieval-augmented generation framework for medical applications. *arXiv preprint arXiv:2407.21055*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 1.

OpenAI. 2024. Hello gpt-4o.

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*, pages 248–260. PMLR.

Lin CY Rouge. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, volume 5.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.

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):172–180.

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*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024a. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, et al. 2024b. Speculative rag: Enhancing retrieval augmented generation through drafting. *arXiv preprint arXiv:2407.08223*.

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*, 35:24824–24837.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation with explicit denoising. *arXiv preprint arXiv:2406.13629*.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *arXiv preprint arXiv:2404.10198*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *arXiv preprint arXiv:2408.00727*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *arXiv preprint arXiv:2406.19215*.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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*, pages 8289–8311.

A Appendix

A.1 Ablation Study of Top-k documents on MMLU-Med

Figure A1 shows the performance of baseline models with and without RAG methods as the number of retrieved documents is varied. Note that our filtering model is trained using MedMCQA since the MMLU-Med training set does not exist. Similar to the results observed in MedQA and MedMCQA (Figure 3), our RAG^2 model achieves the best performance, demonstrating the effectiveness of our approach even in out-of-domain distributions.

A.2 Datasets

MedQA (Jin et al., 2021) MedQA is a multilingual multiple-choice QA dataset derived from professional medical board exams. It encompasses three languages: English, simplified Chinese, and traditional Chinese. The English subset, MedQA-US, specifically comprises USMLE questions.

MedMCQA (Pal et al., 2022) MedMCQA is a comprehensive multiple-choice question-answering dataset designed for medical domain evaluation. It comprises over 194K curated multiple-choice questions derived from AIIMS and NEET PG entrance exams. These questions span across 2.4k healthcare topics and 21 medical subjects, offering a broad coverage of the medical field.

MMLU-Med (Hendrycks et al., 2021) MMLU-Med is a specialized subset of the Massive Multitask Language Understanding (MMLU) benchmark, which is first introduced by Singhal et al. (2023). While the full MMLU benchmark encompasses 57 diverse tasks, MMLU-Med is subset of six biomedical-related domains: anatomy, clinical knowledge, professional medicine, human genetics, college medicine, and college biology.

A.3 Implementation Details

Retrieval For our information retrieval tasks, we used the MedCPT retriever and the reranker. The corpus employed in this study is the same as that used in Self-BioRAG (Jeong et al., 2024a), an earlier work that preceded MedRAG. This corpus includes PubMed abstracts, PMC full texts, and clinician practical guidelines Chen et al. (2023). (used for training Meditron, with 8 out of 16 publicly available), and 18 medical textbooks. Unlike MedRAG, which utilizes a different corpus called MedCorp—comprising PubMed, textbooks, StatPearls,[3] and Wikipedia—we chose to retain the Self-BioRAG corpus due to its already proven effectiveness. Although PubMed provides only abstracts, which lack detailed discussions or results, we supplemented this with PMC full texts to provide a more comprehensive dataset. To ensure comprehensive coverage and prevent the inadvertent truncation of context, we applied a sliding window mechanism with overlap for chunking the documents.

Corpus	# Docs	# documents	Index Size
MedCorp (Xiong et al., 2024a)			
PubMed	23.9M	23.9M	211GB
Wikipedia	6.5M	29.9M	262GB
StatPearls	9.3k	301.2k	3.0GB
Textbooks	18	125.8k	1.1GB
Total	30.4M	54.2M	477.2GB
Ours			
PubMed	36.5M	69.7M	400GB
PMC	1.1M	46.3M	160GB
CPG	35.7k	607.0k	3.5GB
Textbooks	18	134.0k	0.7GB
Total	37.6M	116.7M	564.2GB

Table A1: The statistics of MedCorp (Xiong et al., 2024a) and our retrieval corpus. CPG represents the clinical guidelines collected by Chen et al. (2023). Note that the number of documents can vary even for the same corpus, as it depends on how documents are divided into documents. Additionally, the indexing for MedCorp was done using four different retrievers, while we only used a single retriever, MedCPT, which resulted in a difference in index size.

Training We trained the filtering model on a single NVIDIA H100 GPU with 80GB memory over 40 epochs, using a learning rate of 3e-5 and a per-device batch size of 16. We selected a few candidate models from the validation set, as performance converged after certain epochs, and evaluated them on the test set.

Inference For inference, the same GPU was used with vLLM (Kwon et al., 2023) for faster processing with LLMs. Greedy decoding was employed with a temperature set to 0 to minimize randomness, but some variability persisted due to hardware/software factors. Notably, GPT-4o still exhibited some randomness under these conditions.

[3] https://www.statpearls.com/

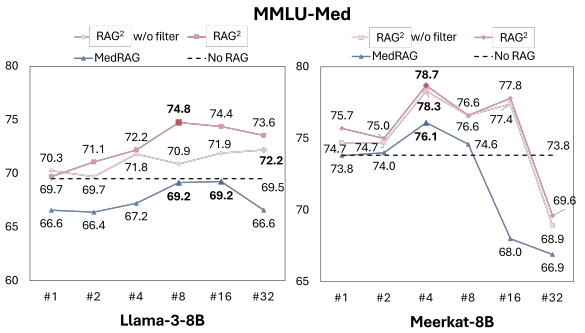

Figure A1: Accuracy (y-axis) of Llama-3-8B-Instruct and Meerkat-7B with/without different RAG models when varying the number of top-k documents (x-axis) on the MMLU-Med dataset. Since MMLU-Med only consists of the test set, we used the filtering model trained on MedMCQA during evaluation on MMLU-Med. The accuracy results indicate that the effective filter trained from a different dataset (MedMCQA) performs well across other dataset (MMLU-Med).

	Rationale Query of		
Model	**Llama-3-8B-Instruct**	**Meerkat-7B**	**GPT-4o**
Llama-3-8B	63.4	71.5	**73.6**
Meerkat-7B	71.3	74.6	**78.8**
GPT-4o	87.4	88.3	**89.8**

Table A2: Performance comparison of backbone models using rationale queries generated by Llama-3-8B, Meerkat-7B, and GPT-4o. The results indicate that GPT-4o-generated rationales consistently yield the highest performance improvements across all models, followed by Meerkat-7B and Llama-3-8B.

A.4 Quality of Rationale Queries

Table A2 shows the performance changes based on which model generated the rationale query. Consistently across all backbone models, performance improves most significantly with rationales generated by GPT-4o, followed by Meerkat and Llama-3-Instruct. Although commercial models like GPT-4o provide higher-quality rationales and can further boost performance, relying on them introduces additional dependencies and increases API usage costs.

A.5 Open-ended Clinical Questions

We have tested our method on multiple-choice QA datasets, including MedQA, MedMCQA, and MMLU-Med. While these benchmarks are valuable for assessing the medical knowledge of LLMs, they do not fully capture the complexity of real-world clinical scenarios. To address this, we further evaluated our method in a setting that more closely resembles actual clinical environments. We used the ClinicalQA25 dataset (Zakka et al., 2024), which consists of 25 queries from clinicians.

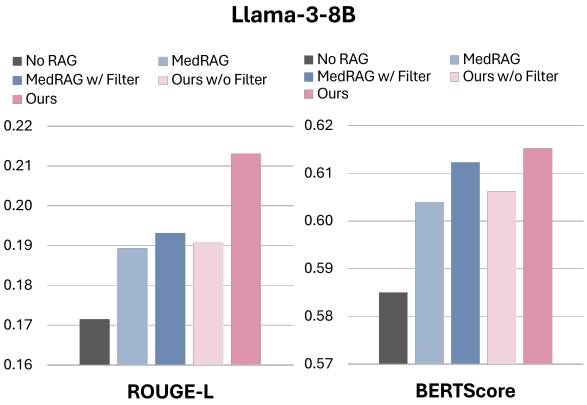

Figure A2: Performance of Llama-3-8B using different RAG pipelines on ClinicalQA25, which represents a real-world clinical scenario.

A.5.1 Metrics

We used the following two metrics to measure the models' long-form responses: ROUGE-L and BERTScore.

ROUGE-L (Rouge, 2004) ROUGE-L is based on the longest common subsequence (LCS) between the candidate text C and the reference text R, and is computed as follows:

$$\text{ROUGE-L Precision}(C, R) = \frac{LCS(C, R)}{|C|},$$
$$\text{ROUGE-L Recall}(C, R) = \frac{LCS(C, R)}{|R|},$$
$$\text{ROUGE-L F1}(C, R) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

BERTScore (Zhang et al., 2020) BERTScore computes precision, recall, and F1 score based on cosine similarity between token embeddings, defined as follows:

$$\text{Precision}(C, R) = \frac{1}{|C|} \sum_{c \in C} \max_{r \in R} \text{cosine_similarity}(c, r),$$
$$\text{Recall}(C, R) = \frac{1}{|R|} \sum_{r \in R} \max_{c \in C} \text{cosine_similarity}(r, c),$$
$$\text{BERTScore}(C, R) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

BERTScore is based on the similarity of word embeddings, allowing it to capture semantic meaning more effectively than ROUGE-L.

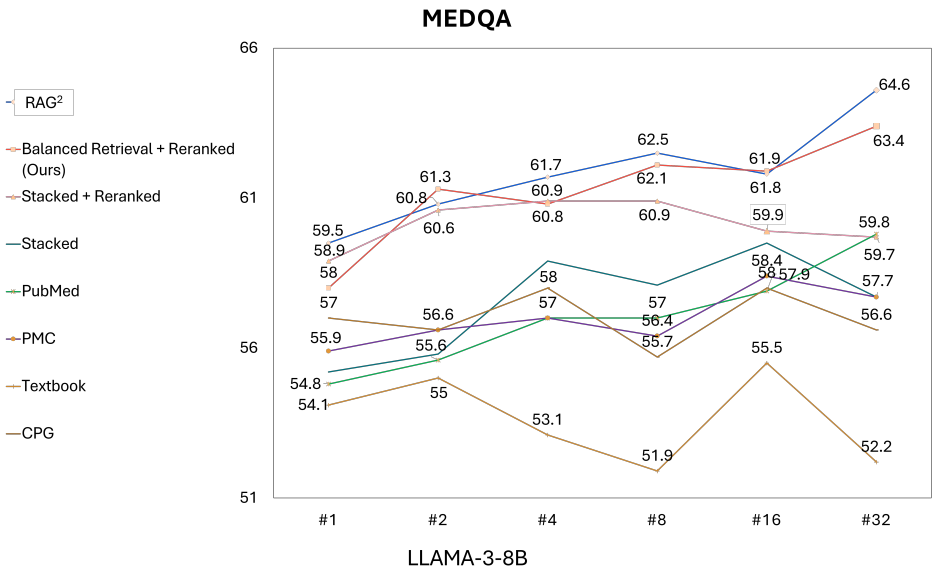

Figure A3: Comparison of retrieval strategies. We evaluate four approaches: independent corpus retrieval, stacked retrieval (following MedRAG), reranked stacked retrieval, and our proposed balanced retrieval. The balanced approach consistently outperforms others, highlighting the importance of considering corpus-specific properties in medical information retrieval.

A.5.2 Results

Figure A2 demonstrates that RAG, including both our proposed approach and MedRAG, outperforms the model-alone baseline. This suggests that RAG is effective in retrieving accurate and relevant information for open-ended questions. Additionally, the integration of our filtering mechanism seems to further improve performance, potentially enhancing retrieval accuracy. This experiment demonstrates that our filtering model, although trained on multiple-choice QA, still shows effectiveness in long-form and real-world clinical queries. In future work, we plan to further validate the broader applicability of our approach.

A.5.3 Balanced Retrieval

We evaluated our retrieval approach through multiple strategies. Beginning with independent corpus analysis, we assessed the utility of individual sources. Following MedRAG, we implemented a stacking method, combining all corpora, selecting top documents based on MIPS scores. We then applied reranking to these stacked retrieval results to potentially improve document relevance. Finally, we implemented our proposed balanced retrieval method both with and without filtering, addressing the potential overshadowing of crucial out-of-domain information inherent in the stacked approach. This comprehensive analysis compares our method against existing approaches, highlighting its effectiveness in leveraging diverse knowledge sources.

Our experiments revealed several key findings:

- Utilizing each corpus independently yielded suboptimal results, indicating that no single source contained all necessary information.
- The stacked approach, while comprehensive, underperformed compared to our balanced retrieval method indicates that indiscriminate corpus stacking, without considering source-specific characteristics, yields suboptimal results.
- Even after reranking, the stacked approach did not outperform our balanced retrieval method.
- Our balanced retrieval approach, particularly with filtering, consistently outperformed other methods, demonstrating its effectiveness in leveraging diverse information sources while maintaining utility.

These results underscore the importance of a thoughtful, balanced approach to multi-corpus retrieval in medical question answering tasks, particularly one that considers the unique properties of each corpus.

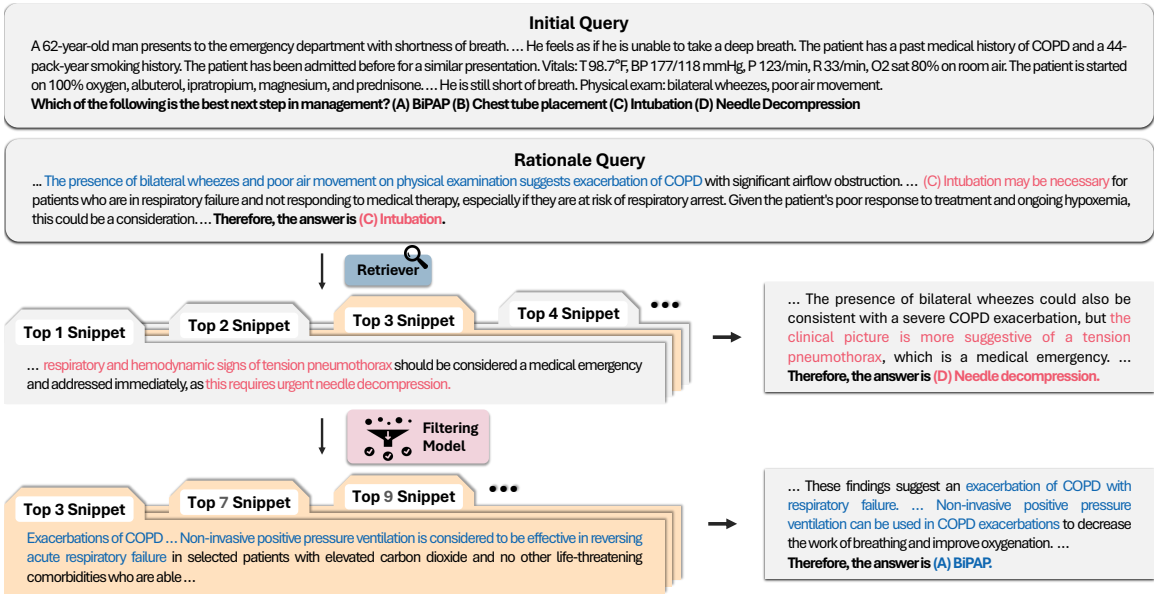

Figure A4: This case study demonstrates the effectiveness of our rationale-guided filtering approach (see the main text for detailed descriptions). Some content has been omitted due to space limitations.

A.6 Case Study

Figure A4 illustrate the effectiveness of our framework. Please see Section 5.2 for detailed descriptions of this example.