FLARE-Med: Enhancing Medical QA with Active Retrieval and Reranking

Isla Kong
Carnegie Mellon University
Pittsburgh, PA 15213
qingxuak@andrew.cmu.edu

Yimei Wu
Carnegie Mellon University
Pittsburgh, PA 15213
yimeiw@andrew.cmu.edu

Yunfan Yang
Carnegie Mellon University
Pittsburgh, PA 15213
yunfanya@andrew.cmu.edu

1 Introduction

Medical question answering is a vital application of artificial intelligence, where the accuracy and reliability of responses are crucial for clinical decision-making and patient outcomes. Existing models often struggle with complex medical queries that require multi-hop reasoning and extensive domain knowledge [7]. To address these challenges, we propose an enhanced retrieval-augmented generation (RAG) framework that integrates Forward-Looking Active Retrieval (FLARE) and follow-up questioning. FLARE enables the system to iteratively guide the retrieval based on intermediate output, progressively refining the evidence base. Follow-up questioning further enhances multi-hop reasoning by generating and addressing sub-questions, enabling the model to handle complex, nuanced medical questions with improved contextual understanding.

2 Dataset & Task

2.1 Dataset

To train and evaluate our medical question answering system, we utilize three established medical QA benchmarks selected from the MIRAGE suite, which is designed to benchmark RAG systems for complex medical reasoning. Due to computational constraints, we focus on the following sub-datasets:

- **MMLU-Med**: A subset of the Massive Multitask Language Understanding (MMLU) benchmark covering medical topics such as anatomy, clinical knowledge, and medical genetics. It includes challenging multiple-choice questions requiring factual recall and clinical reasoning.
- **MedQA-US**: A collection of multiple-choice questions drawn from the United States Medical Licensing Examination (USMLE), requiring multi-hop reasoning and deep understanding of medical concepts [2].
- **BioASQ-Y/N**: A yes/no biomedical question-answering dataset derived from the BioASQ challenge, targeting binary classification of factual questions based on evidence extracted from biomedical literature [6].

Dataset	Size	#Options	Avg. Tokens	Source
MMLU-Med	1,089	4	63	Examination
MedQA-US	1,273	4	177	Examination
BioASQ-Y/N	618	2	17	Literature

Table 1: Summary of medical QA datasets used in our experiments
(#Options = number of options; Avg. Tokens = average token length per question)

2.2 Task

The core task of our system is to answer medical questions using evidence retrieved from a trusted corpus and to generate reasoned outputs. Our system is designed to simulate multi-hop reasoning and improve factual grounding by performing the following subtasks:

- **Evidence Retrieval**: Retrieve the top-k most relevant documents from a medical corpus (Textbooks and StatPearls) using a dense retriever (MedCPT), conditioned on the input question.
- **Forward-Looking Retrieval (FLARE)**: Generate a partial answer using the LLM and use it to enhance the initial query, leading to improved second-stage retrieval.
- **Follow-up Questioning**: Automatically generate follow-up questions based on intermediate reasoning outputs. Each follow-up triggers additional retrieval to simulate multi-hop reasoning.
- **Context Integration and Reasoning to Generate Answer**: Aggregate the retrieved content and guide the LLM to generate a structured response based on step-by-step reasoning. Produce a final answer in structured JSON format with a rationale that selects from the given multiple-choice options.
- **Evaluate Accuracy:** Defined as the proportion of questions for which the final predicted answer matches the ground truth.

3 Related Work

RAG has emerged as a critical technique for enhancing large language models (LLMs) in knowledge-intensive domains such as healthcare and biomedicine [3]. Within the medical domain, Xiong et al. [7] proposed **MedRAG**, a framework that integrates various medical corpora and retrieval models to support factual and evidence-based clinical reasoning. Their findings highlighted that carefully selected retrievers and corpora significantly boost the performance of medical QA systems.

For document retrieval, dense retrievers have shown clear advantages over traditional lexical methods. MedCPT [8], a domain-specific dense retriever fine-tuned in clinical and biomedical texts, demonstrated state-of-the-art retrieval effectiveness for medical question-solving tasks. We adopt MedCPT in our system as the primary retriever to better align the retrieved content with the medical terminology and context.

Addressing the limitations of one-time retrieval, iterative retrieval methods have been proposed for multi-step reasoning. Jiang et al. [1] introduced Forward-Looking Active Retrieval Augmented Generation, which iteratively generates partial answers and refines retrieval queries based on intermediate outputs. This active retrieval strategy reduces hallucinations and improves the precision of the answer by dynamically expanding the evidence base during generation.

In parallel, leveraging LLMs' latent knowledge for retrieval has gained traction. Ma et al. [4] demonstrated that fine-tuned LLaMA models can serve effectively as dense retrievers and rerankers, significantly improving retrieval quality without the need for large annotated datasets. Furthermore, Shen et al. [5] explored using LLM directly as zero-shot retrievers, achieving competitive performance by augmenting queries internally without external retrievers.

These advances collectively suggest that the integration of dense retrievers, forward-looking retrieval mechanisms, and iterative reasoning strategies is crucial to building robust medical question answering systems. Motivated by these insights, our work extends the MedRAG architecture with FLARE-style partial answer generation and follow-up question expansion, aiming to enhance multi-hop factual reasoning in complex medical QA settings.

4 Approach

4.1 Baseline Methodology

Our baseline system follows a MedRAG-inspired RAG framework, designed to support the answer of factual and explainable medical questions through evidence-based reasoning. The approach consists of two key stages: document retrieval and answer generation.

4.1.1 Corpora Collection from Multiple Sources

To construct the retrieval corpus, we collect content from two medical sources: clinical textbooks and StatPearls. StatPearls consists of peer-reviewed reference articles authored by medical professionals, while the textbooks provide curated material across a wide range of clinical specialties.

Each document is segmented into overlapping textual snippets using a fixed-length sliding-window approach. This preprocessing allows for fine-grained semantic retrieval by enabling the system to match short, focused evidence segments to user queries. The final corpus of snippets serves as the knowledge base for all retrieval operations.

Corpus	#Documents	#Snippets	Avg. Snippet Length	Domain
StatPearls	9.3k	301.2k	119 tokens	Clinical reference
Textbooks	18	125.8k	182 tokens	Medical education

Table 2: Statistics of corpora used in our medical retrieval system.

4.1.2 Dense Retrieval with Domain-Specific Embeddings

To retrieve relevant evidence, we use a dense retriever that encodes both questions and snippets into a shared embedding space. Specifically, we adopt a medical-domain embedding model trained to preserve clinical semantics. Given a question, the retriever computes its vector representation and returns the top-k snippets with the highest similarity scores. This dense retrieval mechanism improves upon traditional keyword-based methods by enabling semantic matching in biomedical contexts.

4.1.3 Zero-Shot Answer Generation via In-Context Learning

The retrieved evidence is concatenated into a prompt together with the original question and passed to a LLM. We adopt a zero-shot prompting approach, where the model is guided by structured instructions but not further fine-tuned. The LLM reasons over the retrieved context and produces a step-by-step explanation followed by a final answer selection. This design allows flexible adaptation to new medical questions while grounding responses in retrieved evidence.

4.2 Main Methods

To improve the depth of reasoning and evidence alignment in RAG, we propose two methodological enhancements: Forward-Looking Active Retrieval Augmented Generation and its extension with follow-up questioning. These methods aim to overcome the limitations of one-shot retrieval by introducing a proactive and iterative interaction between generation and retrieval.

4.2.1 Forward-Looking Active Retrieval Augmented Generation (FLARE)

FLARE introduces a generation step prior to retrieval, where the language model produces a partial response conditioned on the initial query and minimal context. This partial output is treated as a forward-looking signal that captures the model's expectations or the latent reasoning trajectory.

Figure 1 illustrates this process. The LLM first produces a look-ahead response based on the initial question. This is then used to guide the retrieval through MedCPT. The retrieved snippets, conditioned on the enriched query, are passed back into the model to produce the final answer. Figure 1 illustrates this process. The LLM first produces a look-ahead response based on the initial question. This is then used to guide the retrieval through MedCPT. The retrieved snippets, conditioned on the enriched query, are passed back into the model to produce the final answer.

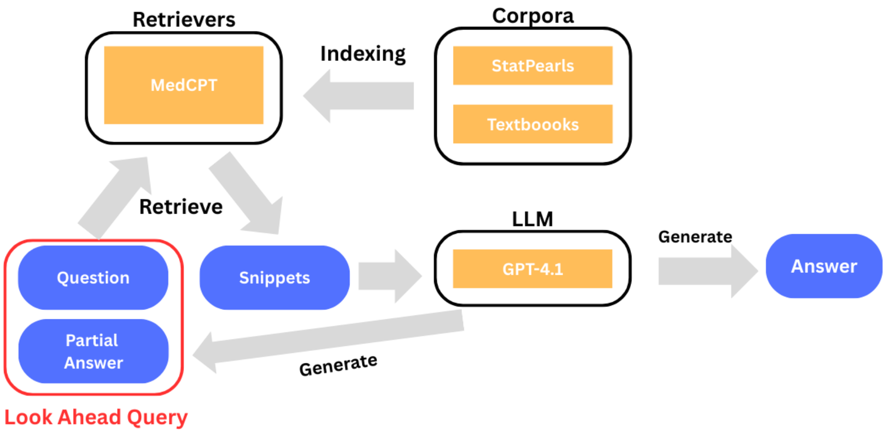

Figure 1: FLARE pipeline

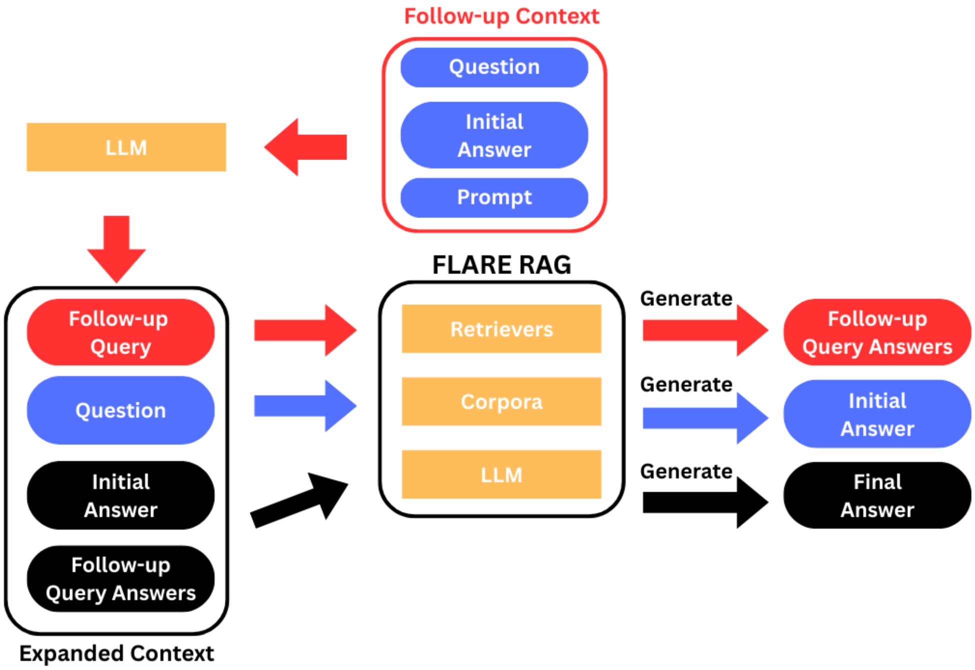

Figure 2: FLARE with Follow-up Questioning

This method transforms the RAG pipeline from a passive one-shot retrieval to an active, generation-informed search process. The resulting retrieved documents tend to be more focused and contextually relevant, enhancing the grounding of final responses and reducing hallucination.

4.2.2 FLARE with Follow-up Questioning

We further extend the FLARE framework by incorporating an iterative follow-up question mechanism. After generating the initial answer, the model is prompted to produce clarification or exploratory sub-questions that capture missing details or support multi-step reasoning. Each follow-up query is processed via the same FLARE pipeline: a look-ahead generation, retrieval, and answer synthesis.

As shown in Figure 2, the model begins with an initial answer and uses it to generate follow-up queries. Each query is resolved using retrieval and generation, and its answer is integrated into an evolving context. This context is used to inform subsequent queries and ultimately the final answer.

This method transforms RAG into a multi-turn process, allowing the model to explicitly resolve uncertainties, expand its evidence base, and build a more complete response over time. The iterative loop concludes either after a predefined number of follow-up turns or when a confident final answer is detected.

5 Experiments

5.1 Environment

We build FLARE-Med with follow-up questioning on top of the MedRAG framework, which includes three main components: the document corpus, the retriever, and the language model. For the corpus, we use two medical sources: a subset of medical textbooks and articles from StatPearls. These documents are indexed for retrieval. We use MedCPT as the retriever, which supports dense retrieval and is specialized for medical question answering. The language model used is GPT-4.1 nano, accessed through the OpenAI API. All experiments are run on an AWS EC2 `g5.2xlarge` instance with one NVIDIA A10G GPU (24 GB), 8 vCPUs, and 32 GB of RAM.

5.2 Accuracy

Table 3 reports the performance of different methods across three medical question answering benchmarks: MMLU-Med, MedQA-US, and BioASQ-Y/N. We compare the standard Chain-of-Thought prompting (CoT), the MedRAG baseline, and our proposed FLARE-Med system enhanced with follow-up questioning. The CoT approach already performs strongly, achieving a precision of 77—- 86% across the datasets, reflecting the internalized domain knowledge pre-trained LLM. The MedRAG baseline, while using external documents, does not consistently outperform CoT, indicating that retrieval alone may not suffice when the context corpus is limited or misaligned. In contrast, our FLARE-Med system, which incorporates both look-ahead generation and iterative follow-up retrieval, consistently outperforms both baselines. This demonstrates the effectiveness of our multi-hop retrieval augmentation strategy in expanding relevant context and supporting complex reasoning.

Method	**MMLU-Med**	**MedQA-US**	**BioASQ-Y/N**
CoT	86.87 ± 1.02	77.14 ± 1.18	83.01 ± 1.51
MedRAG	85.12 ± 1.08	77.30 ± 1.17	72.49 ± 1.80
FLARE with Follow-up	**87.88** ± 0.99	**83.03** ± 1.05	**75.24** ± 1.74

Table 3: Accuracy (%) of different methods across datasets. All methods use the same retriever (MedCPT, $k = 8$).

6 Code Overview

Our implementation extends the MedRAG architecture to incorporate forward-looking retrieval and iterative refinement for medical question answering. The key components are organized as follows:

- **System Initialization**: The `MedRAG` class initializes the retrieval system, language model, prompting templates, and assigns the answering strategy based on settings for FLARE and follow-up question generation (lines 0–249). Different retrievers (e.g., MedCPT) and LLMs (e.g., GPT-4, Gemini) can be configured.
- **Baseline RAG**: The `medrag_answer` method (lines 341–485) performs standard one-shot retrieval and generation. It retrieves a fixed number of documents, concatenates them as context, and uses zero-shot prompting to generate an answer without iterative refinement.
- **Forward-Looking Retrieval (FLARE)**: The `_flare_without_follow_up` method (lines 744–919) enhances standard RAG by generating a partial answer (look-ahead response) based on the initial retrieval. This partial answer is then used to reformulate a more informative query for a second-round retrieval, improving the relevance of retrieved documents.
- **FLARE with Follow-up Questioning**: The `_flare_with_follow_up` method (lines 921–1149) enables iterative refinement. After answering the initial question, the model generates

multiple follow-up questions, retrieves new evidence for each, and accumulates context over several rounds. Final answer generation is based on this expanded, multi-hop evidence set.

- **Flexible Answering Pipeline**: Depending on whether FLARE and/or follow-up generation are enabled, the system dynamically selects between standard RAG, forward-looking retrieval, or multi-turn iterative answering. This modular design supports flexible experimentation across different reasoning strategies.

Overall, the codebase is structured to progressively enhance retrieval quality and reasoning depth, enabling more accurate and explainable medical question answering through iterative context refinement.

7 Timeline

Task	**Hours Spent**
Related work research	8
Understanding MedRAG codebase	6
Designing improvement framework	6
Adjust baseline model and run experiments	8
Implementing Flare and follow-up question model	15
Writing scripts and running experiments	15
Compiling results and visualizations	4
Writing the report and slides	10
Total	**72 hours**

Table 4: Project timeline and estimated effort

8 Research Log

We began by replicating MedRAG and running it on MMLU-Med and MedQA-US. Although MedRAG improved factual grounding over zero-shot prompting, its accuracy remained close to or below Chain-of-Thought (CoT) prompting, suggesting retrieval was not fully leveraged due to shallow reasoning or limited corpus size.

This prompted our further exploration of forward-looking retrieval strategies. A key challenge was designing a pipeline that automatically produces meaningful follow-up queries. We addressed this by generating partial answers using the LLM, then prompting it to propose relevant sub-questions, which were used to guide additional retrieval rounds.

Integrating follow-up retrieval with context expansion required careful token management and truncation control. Over several iterations, we refined the prompting format, stop conditions, and fallback logic to ensure a coherent final output.

9 Conclusion

We present FLARE-Med, an iterative RAG framework that simulates multi-hop reasoning by combining partial answer generation with dynamic follow-up questioning. Empirical results across MMLU-Med, MedQA-US, and BioASQ-Y/N show that FLARE-Med outperforms baseline MedRAG and CoT prompting, especially in scenarios requiring deeper factual reasoning. Future directions include using confidence estimation to dynamically trigger follow-up, and integrating agentic planning to learn adaptive retrieval paths for different query types.

References

[1] Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

[2] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*, 2023.

[5] Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[6] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Georgios Paliouras, and Ion Androutsopoulos. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

[7] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.

[8] Yifan Zhang, Zonghai Xu, and et al. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *arXiv preprint arXiv:2307.00589*, 2023.