MedRAG: A Hybrid Retrieval-Augmented AI Framework for Healthcare Research

Sagar Y
M.Tech Artificial Intelligence and Data Science
School of Computing and Artificial Intelligence (SCAI)
VIT Bhopal University, Bhopal, India
Email: sagar.25mas10035@vitbhopal.ac.in

Dr. Pon Harshavardhanan
Dean, School of Computing and Artificial Intelligence (SCAI)
VIT Bhopal University, Bhopal, India
Email: pon.harshavardhanan@vitbhopal.ac.in

***Abstract*—Large Language Models (LLMs) have demonstrated strong generative and reasoning capabilities; however, their applicability to healthcare research remains constrained by static pretraining data and post-training knowledge cutoff limitations. This paper presents MedRAG, a hybrid Retrieval-Augmented Generation (RAG) framework designed to support evidence-grounded, up-to-date, and research-oriented analysis in AI-driven healthcare domains.**

MedRAG integrates a large-scale vectorized corpus of recent AI and healthcare research publications (2023–2025) with a multi-stage pipeline comprising dense semantic retrieval, fusion-based multi-document synthesis, lightweight generative refinement, and medical safety validation. Experimental evaluation demonstrates a 22.9% improvement in response quality over retrieval-based baselines, particularly in terms of completeness and coherence. The results indicate that hybrid RAG architectures provide a practical and resource-efficient alternative to continual model retraining for rapidly evolving healthcare research environments.

***Index Terms*—Retrieval-Augmented Generation, Large Language Models, AI in Healthcare, FAISS, Fusion-in-Decoder, Knowledge Grounding.**

I. Introduction

Large Language Models (LLMs) have become foundational components of modern artificial intelligence systems, enabling advances in language understanding, summarization, reasoning, and generation across a wide range of domains [1], [2]. In the healthcare domain, LLMs have been explored for applications such as clinical documentation, diagnostic reasoning support, patient education, and biomedical literature analysis [3], [4], [5]. These developments indicate strong potential for augmenting healthcare research workflows.

Despite these advances, the applicability of LLMs to healthcare research is fundamentally constrained by their reliance on static pretraining corpora. Most state-of-the-art models are trained on large but temporally fixed datasets, resulting in post-training knowledge cutoffs that limit their ability to reflect recent clinical studies, evolving guidelines, and emerging AI methodologies [3], [6]. In research-oriented settings, such gaps can reduce the reliability and relevance of generated outputs.

A straightforward solution to the knowledge cutoff problem is continual retraining or fine-tuning of language models. However, this approach is computationally expensive and often impractical in academic or institutional environments [2]. Moreover, retraining alone does not inherently guarantee improved factual grounding or mitigation of hallucinated content, particularly when models are required to reason across multiple heterogeneous research sources [7].

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm to address these challenges by enabling dynamic access to external knowledge during inference [8]. By retrieving relevant documents and conditioning generation on retrieved evidence, RAG systems improve factual accuracy and reduce hallucination compared to standalone LLMs [9], [10]. Nevertheless, naïve RAG implementations often struggle with coherent multi-document synthesis, especially for complex healthcare research queries.

In addition, AI-driven healthcare research increasingly demands cross-domain reasoning that spans natural language processing, medical imaging, clinical analytics, and ethical governance [6], [4]. Existing tools are often designed for narrow tasks or limited datasets and do not adequately support holistic research exploration.

Motivated by these challenges, this paper introduces **MedRAG**, a hybrid Retrieval-Augmented Generation framework designed for post-cutoff healthcare research. MedRAG integrates large-scale dense retrieval, fusion-based multi-document synthesis, lightweight generative refinement, and explicit safety validation to provide grounded, coherent, and research-oriented responses while remaining computationally efficient.

II. Related Work

This section reviews prior research relevant to the proposed MedRAG framework. The discussion is organized into thematic subsections covering large language models in healthcare, retrieval-augmented generation paradigms, medical RAG systems, and safety and trust considerations. This structure highlights existing capabilities while identifying gaps addressed by the proposed work.

A. Large Language Models in Healthcare

Large language models have been increasingly investigated for healthcare-related tasks, including diagnostic reasoning, clinical text summarization, patient interaction, and biomedical literature analysis [3], [4], [5]. Empirical studies demonstrate that LLMs can capture substantial medical knowledge and perform competitively on standardized medical benchmarks.

However, multiple studies have highlighted significant limitations that restrict real-world applicability. A recurring concern is the tendency of LLMs to generate hallucinated or unsupported statements, particularly when queries require precise medical knowledge [11]. Additionally, static pretraining leads to outdated responses that may not reflect current clinical guidelines or emerging research findings [6].

These findings underscore the need for mechanisms that dynamically incorporate authoritative and recent medical knowledge into the generation process.

B. Retrieval-Augmented Generation Paradigms

Retrieval-Augmented Generation was proposed to mitigate the limitations of parametric language models by incorporating external knowledge during inference [8]. By retrieving relevant documents and conditioning generation on retrieved evidence, RAG systems have demonstrated improved factual grounding on knowledge-intensive tasks.

Subsequent research introduced fusion-based architectures that improve multi-document reasoning by jointly attending to multiple retrieved passages during decoding [9], [10]. These approaches have shown significant gains over single-context generation, particularly for questions requiring synthesis across multiple sources.

Despite these advances, most RAG research has focused on general-domain NLP benchmarks. Healthcare-specific challenges, such as domain terminology, safety constraints, and cross-disciplinary integration, are often not explicitly addressed [6], [7].

C. Medical RAG Systems and Domain-Specific Applications

Several studies have applied RAG-based pipelines to biomedical and clinical tasks, including clinical question answering, trial data analysis, and medical information retrieval [6], [12]. These systems demonstrate reduced hallucination compared to standalone LLMs and highlight the value of evidence-grounded generation in medical contexts.

However, existing medical RAG systems are often constrained by limited datasets, narrow task definitions, or short temporal coverage [6]. Many systems are optimized for specific use cases such as patient education or decision support, rather than broader healthcare research exploration.

Recent reviews emphasize the need for large-scale, time-aware biomedical corpora and research-oriented evaluation protocols. These observations motivate the design of systems that support cross-domain synthesis and continuous incorporation of newly published research.

D. Safety, Trust, and Responsible AI in Healthcare

Safety and trustworthiness are critical requirements for AI systems operating in healthcare domains. Prior work highlights risks associated with unvalidated medical claims, lack of transparency, and overreliance on generative outputs [11], [13].

Explainability and validation mechanisms have been proposed to improve trust in medical AI systems, including rule-based filtering and post-generation validation [11]. These concerns motivate the integration of explicit safety validation layers within RAG pipelines for healthcare research.

These findings motivate the inclusion of explicit safety validation layers within RAG pipelines. By embedding responsible AI considerations directly into the generation workflow, systems can improve reliability while maintaining utility for research and educational purposes.

E. Summary and Positioning of MedRAG

In contrast to existing approaches, MedRAG is designed as a research-oriented hybrid RAG framework that integrates large-scale post-2022 biomedical literature with fusion-based synthesis, generative refinement, and explicit safety validation. By addressing limitations related to knowledge cutoff, multi-document reasoning, and responsible output generation, MedRAG fills a critical gap in AI-assisted healthcare research systems.

III. Proposed System Design

The MedRAG framework is designed as a hybrid Retrieval-Augmented Generation (RAG) system that integrates dense document retrieval, multi-document evidence synthesis, generative refinement, and safety validation into a unified pipeline. The design emphasizes modularity, scalability, and responsible deployment for healthcare research applications. Figure 1 presents an overview of the system architecture.

A. Design Objectives and Rationale

The primary objective of MedRAG is to enable healthcare researchers to query recent scientific literature in a manner that is both evidence-grounded and analytically coherent. Unlike standalone language models, which rely solely on parametric knowledge, MedRAG explicitly separates knowledge access from language generation. This separation allows the system to incorporate newly published research without requiring continual retraining of large language models.

A secondary objective is to support cross-domain reasoning across multiple AI and healthcare subfields, including natural language processing, medical imaging, clinical analytics, and ethical considerations. To achieve this, MedRAG is designed to retrieve and synthesize information from multiple heterogeneous sources rather than relying on a single contextual passage.

Finally, the system is designed with responsible AI principles in mind. Given the sensitivity of healthcare information, MedRAG includes explicit safeguards to prevent unsafe medical claims and to ensure that generated responses remain suitable for research and educational use rather than clinical decision-making.

B. FAISS-Based Dense Retrieval Strategy

Efficient and accurate retrieval of relevant literature is a critical component of the MedRAG pipeline. To this end, the system employs FAISS (Facebook AI Similarity Search) for dense semantic retrieval over a large corpus of biomedical and AI-focused research documents.

Let the document corpus be represented as $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, where each document chunk d_i is mapped to a dense embedding vector $\mathbf{v}_i \in \mathbb{R}^d$ using a sentence embedding model. Given a user query q, the query is similarly encoded as $\mathbf{v}_q \in \mathbb{R}^d$.

Semantic similarity between the query and document embeddings is computed using inner-product similarity:

$$\text{sim}(\mathbf{v}_q, \mathbf{v}_i) = \mathbf{v}_q^\top \mathbf{v}_i \tag{1}$$

The retrieval module selects the top-K document chunks with the highest similarity scores:

$$\mathcal{R}_K(q) = \arg\max_{d_i \in \mathcal{D}}^{K} \text{sim}(\mathbf{v}_q, \mathbf{v}_i) \tag{2}$$

In the current implementation, the value of K is empirically set to $K = 8$, balancing contextual coverage and computational efficiency. Dense retrieval is preferred over sparse keyword-based methods such as BM25 because it captures semantic similarity beyond exact term overlap, which is particularly important in healthcare literature where synonymous terminology and implicit relationships are common.

FAISS enables scalable approximate nearest-neighbor search, allowing MedRAG to efficiently operate over a corpus containing approximately 100,000 document chunks without significant degradation in retrieval latency.

C. Fusion-in-Decoder Multi-Document Synthesis

Retrieved document chunks often contain complementary or partially overlapping information. To effectively integrate evidence across multiple sources, MedRAG employs a Fusion-in-Decoder (FiD) architecture for multi-document synthesis.

Given the retrieved document set $\mathcal{R}_K(q) = \{d_1, d_2, \ldots, d_K\}$, each document is independently encoded using a shared encoder:

$$\mathbf{h}_i = \text{Encoder}(q, d_i), \quad i \in \{1, \ldots, K\} \tag{3}$$

Unlike traditional encoder-decoder models that concatenate documents at the input level, FiD defers fusion until the decoding stage. The decoder jointly attends over all encoded representations to generate an output sequence $y = (y_1, \ldots, y_T)$:

$$P(y \mid q, \mathcal{R}_K) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, \mathbf{h}_1, \ldots, \mathbf{h}_K) \tag{4}$$

This design allows the model to reason across multiple retrieved passages while preserving passage-level independence during encoding. FiD is particularly well-suited for healthcare research queries that require synthesis of findings across multiple recent studies rather than extraction from a single document.

D. Generative Refinement and Narrative Structuring

While the FiD model produces a grounded synthesis of retrieved evidence, its outputs are often dense and minimally structured. To improve readability and analytical clarity, MedRAG incorporates a lightweight generative refinement stage that restructures the synthesized content into a coherent research-style narrative.

This refinement stage focuses on organizing information into logical sections, improving transitions between concepts, and ensuring that technical terminology is presented consistently. Importantly, the refinement model does not introduce new factual content; instead, it operates as a post-processing layer that enhances clarity while preserving the grounding established by the retrieval and synthesis stages.

By decoupling synthesis from narrative refinement, MedRAG achieves improved output quality without relying on large, computationally expensive generative models.

E. Safety Validation and Responsible Output Control

Healthcare research systems must adhere to strict safety and ethical standards. To mitigate the risk of unsafe or misleading outputs, MedRAG integrates a safety validation module as a final processing stage.

This module evaluates generated responses for the presence of disallowed medical claims, such as direct diagnoses or treatment recommendations. Additionally, it enforces the inclusion of appropriate disclaimers and filters language that may be misinterpreted as clinical advice.

By incorporating safety validation directly into the generation pipeline, MedRAG ensures that improvements in analytical quality do not come at the expense of responsible AI practices. This design choice is essential for deploying research-oriented AI systems in sensitive domains such as healthcare.

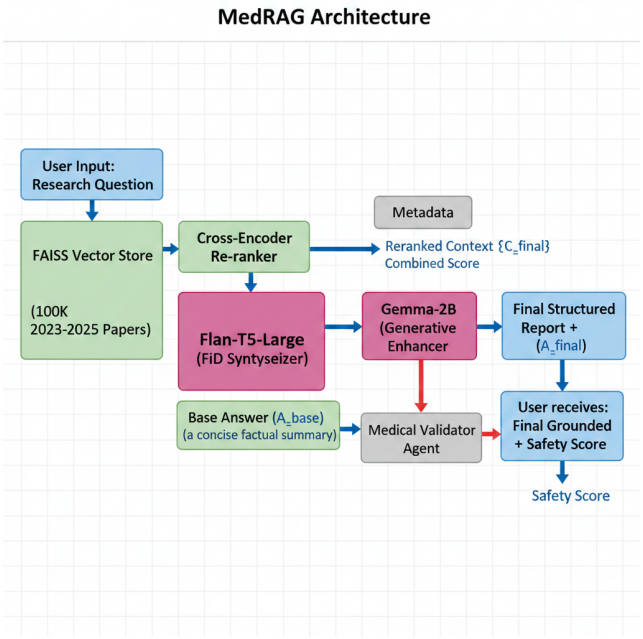

Fig. 1. MedRAG hybrid architecture integrating retrieval, synthesis, refinement, and validation.

IV. Implementation and Dataset

The MedRAG framework is implemented using a modular Python-based architecture that separates retrieval, synthesis, refinement, and validation into independent components. This modular design improves reproducibility, facilitates ablation

studies, and enables future extensions without re-engineering the entire pipeline.

A. Dataset Collection and Preprocessing

The knowledge base used in MedRAG consists of approximately 100,000 research papers published between 2023 and 2025, focusing on the intersection of artificial intelligence and healthcare. The corpus includes studies related to clinical decision support, medical natural language processing, medical imaging, predictive analytics, and ethical considerations in healthcare AI.

Each document is preprocessed to remove non-informative sections such as references and appendices. Documents are then segmented into overlapping text chunks to preserve contextual continuity while ensuring compatibility with embedding models. Chunk sizes are selected to balance semantic completeness and retrieval granularity, enabling precise matching for complex research queries.

B. Embedding and Index Construction

Each document chunk is embedded using a sentence-transformer model that maps text into a fixed-dimensional dense vector space. These embeddings are indexed using FAISS to support efficient similarity search. The indexing strategy allows MedRAG to scale to large corpora while maintaining acceptable query-time latency.

The embedding and indexing pipeline is executed offline, ensuring that online inference costs are dominated by retrieval and generation rather than preprocessing. This design choice supports efficient deployment in academic environments with limited computational resources.

C. System Integration and Execution Flow

At inference time, a user query is processed sequentially through retrieval, synthesis, refinement, and safety validation stages. Intermediate outputs are logged to enable analysis of system behavior and to support future debugging and optimization. Prompt constraints are applied during generation to enforce an academic tone and discourage speculative or unsupported claims.

V. Results and Analysis

This section evaluates the performance of MedRAG and analyzes the impact of its hybrid design on response quality and reliability.

A. Evaluation Methodology

The system is evaluated using a set of complex, research-oriented queries representative of real-world information needs in AI-driven healthcare research. These queries span technical implementation questions, clinical research analysis, methodological comparisons, and ethical considerations.

MedRAG is compared against a retrieval-only baseline that retrieves relevant documents but does not apply fusion-based synthesis, generative refinement, or safety validation. This comparison isolates the contribution of each additional module in the proposed architecture.

Responses are evaluated using a human-in-the-loop assessment protocol. Each response is rated on a five-point Likert scale across multiple dimensions, including relevance, completeness, coherence, factual grounding, and overall usefulness for research purposes.

B. Quantitative and Qualitative Results

TABLE I
Performance Comparison Between Baseline Retrieval and Proposed MedRAG Framework

Evaluation Metric	Baseline Retrieval	Proposed MedRAG
Relevance Score (0–5)	3.6	**4.3**
Completeness Score (0–5)	2.9	**4.4**
Coherence Score (0–5)	3.1	**4.1**
Factual Grounding	Moderate	**High**
Hallucination Tendency	Medium	**Low**
Average Response Time (s)	1.7	31.7
Overall Quality Score (0–5)	3.27	**4.02**

C. Quantitative and Qualitative Results

The baseline system achieves an average overall quality score of 3.27 out of 5.0. In contrast, MedRAG achieves an average score of 4.02, corresponding to a relative improvement of 22.9%. The largest gains are observed in completeness and coherence, indicating that the fusion and refinement stages significantly enhance the integration and presentation of retrieved information.

Qualitative analysis further reveals that baseline responses often contain fragmented explanations and limited contextual integration. MedRAG responses, by contrast, present structured narratives that synthesize findings across multiple sources. This improvement is particularly pronounced for queries that require comparison of multiple research approaches or discussion of emerging trends.

D. Latency and Resource Considerations

The enhanced response quality achieved by MedRAG is accompanied by increased inference latency. While the baseline system produces responses in under two seconds, the full MedRAG pipeline requires approximately thirty seconds per query. This trade-off is acceptable for research-oriented use cases, where analytical depth and evidence grounding are prioritized over real-time interaction.

VI. Limitations

Despite its advantages, the MedRAG framework has several limitations. First, system performance is inherently dependent on the quality and coverage of the underlying document corpus. Gaps or biases in the dataset may affect retrieval accuracy and downstream synthesis.

Second, the evaluation relies primarily on qualitative human judgment, which, while effective for capturing nuanced aspects

of response quality, limits statistical generalizability. Future work will incorporate standardized automated metrics and larger-scale evaluations.

Finally, although safety validation mechanisms are integrated, MedRAG is intended strictly for research and educational use. The system is not designed to provide clinical diagnoses or treatment recommendations, and its outputs should not be interpreted as medical advice.

VII. Conclusion

This paper presented MedRAG, a hybrid Retrieval-Augmented Generation framework designed to address the limitations of static pretraining in large language models for healthcare research. By integrating FAISS-based dense retrieval, Fusion-in-Decoder multi-document synthesis, generative refinement, and explicit safety validation, MedRAG enables grounded, coherent, and up-to-date research assistance across AI and healthcare domains.

Experimental results demonstrate a substantial improvement in response quality over retrieval-only baselines, particularly in terms of completeness and coherence. Importantly, these gains are achieved without reliance on continual model retraining or large proprietary language models, highlighting the practicality of hybrid RAG architectures in academic environments.

In future work, we will extend MedRAG toward multimodal retrieval, incorporating medical images, tables, and structured clinical data. Additionally, adaptive retrieval and generation strategies will be explored to dynamically balance latency and response quality. These directions form the foundation for subsequent stages of this research and future thesis-level contributions.

References

[1] T. Brown, B. Mann, and N. Ryder, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, 2020.

[2] H. Touvron, L. Martin, and K. Stone, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] K. Singhal, T. Tu, and J. Gottweis, "Large language models in healthcare: Applications, challenges, and future directions," *Nature Medicine*, 2023.

[4] S. Wang, H. Li, and Y. Zhang, "Large language models for medical text mining: A survey," *Artificial Intelligence in Medicine*, 2024.

[5] L. Rasmy and Y. Xiang, "Clinical decision support using large language models," *Journal of the American Medical Informatics Association*, 2023.

[6] Z. Guo, D. Yang, and X. Chen, "Retrieval-augmented generation in biomedical applications: A systematic review," *Journal of Biomedical Informatics*, 2024.

[7] D. Alvarez-Melis and T. Jaakkola, "On the robustness of retrieval-augmented generation," *Empirical Methods in Natural Language Processing*, 2023.

[8] P. Lewis, E. Perez, and A. Piktus, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, 2020.

[9] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *International Conference on Learning Representations*, 2021.

[10] Y. Zhang, Z. Sun, and Q. Liu, "Fusion-based retrieval-augmented generation for multi-document reasoning," *ACM Transactions on Information Systems*, 2023.

[11] R. Singh and M. Kaur, "Safety and trustworthiness of large language models in healthcare," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[12] X. Li and H. Chen, "Dense retrieval for biomedical question answering," *Bioinformatics*, 2022.

[13] M. Ribeiro and S. Singh, "Why should i trust you? explaining the predictions of any classifier," *ACM SIGKDD*, 2016.

[14] G. Izacard and E. Grave, "Distilling knowledge from reader to retriever for question answering," *International Conference on Learning Representations*, 2022.

[15] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with faiss," *IEEE Transactions on Big Data*, 2019.

[16] U. Khandelwal and O. Levy, "Generalization through memorization: Nearest neighbor language models," *International Conference on Learning Representations*, 2020.

[17] B. Shickel and P. Tighe, "Deep learning in clinical text analytics: A review," *Journal of Biomedical Informatics*, 2021.

[18] I. Beltagy, M. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[19] J. Wei, X. Wang, and D. Schuurmans, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, 2022.

[20] Y. Sagar and P. Harshavardhanan, "Medrag: A hybrid retrieval-augmented generation framework for healthcare research," *Unpublished Manuscript*, 2025, under review.