RESEARCH PAPER

Sparse-Dense Fusion with Reciprocal Rank Fusion for Hallucination-Mitigated Clinical Knowledge Retrieval

Abhay Nautiyal[1], Mr. Ramesh Kumar[2] and Mr. V. M. Thakkar[3]

[1]M.Tech Graduate, Department of Computer Science and Engineering, Govind Ballabh Pant Institute of Engineering and Technology (GBPIET), Pauri Garhwal, Uttarakhand 246194, India
[2]Assistant Professor, Department of Computer Science and Engineering, GBPIET, Pauri Garhwal, Uttarakhand 246194, India
[3]Assistant Professor, Department of Computer Science and Engineering, GBPIET, Pauri Garhwal, Uttarakhand 246194, India
[1]abhaynautiyal002@gmail.com, [2]rameshparalian1973@gmail.com and [3]vmthakkar1@rediffmail.com

Received: 28th Feb, 2026; Revised: 6th March 2026; Accepted: 7th April, 2026; Available Online: 20th April, 2026

ABSTRACT
Large language models (LLMs) hallucinate—they generate confident but unverifiable medical information, a critical failure in healthcare where accuracy directly impacts patient outcomes. This paper presents a hybrid Retrieval-Augmented Generation (RAG) framework that mitigates hallucination by grounding generated answers in verified medical context. The system integrates sparse BM25 retrieval for precise medical term matching with dense FAISS retrieval for semantic similarity through Reciprocal Rank Fusion (RRF), ensuring that neither exact terminology nor conceptual relatedness is overlooked. A cross-encoder reranker refines the fused results, and Llama 3.1 8B Instant is conditioned on top-ranked documents to produce answers with explicit source attribution. Evaluated on 16,412 question-answer pairs from the MedQuAD dataset covering 138 medical focus areas, the hybrid approach achieves Precision@10 = 63.80%, Recall@10 = 61.84%, NDCG@10 = 78.07%, and Mean Reciprocal Rank = 94.87%—significantly outperforming individual BM25 and FAISS baselines ($p < 0.01$, bootstrap test). Analysis reveals that 83% of retrieved items change rank after cross-encoder reranking, with the top-1 document changing in 55% of queries, while method overlap analysis demonstrates complementary retrieval behavior with only 42% agreement across methods. The complete system is deployed as an interactive Gradio web application enabling real-time multi-method comparison for clinical decision support.

Keywords: *Retrieval-Augmented Generation, Hybrid Information Retrieval, Medical Question Answering, Hallucination Mitigation, Reciprocal Rank Fusion, BM25, FAISS, Cross-Encoder Reranking, Semantic Search, Clinical Decision Support*
How to cite this article: Nautiyal A, Kumar R, Thakkar VM. Sparse-Dense Fusion with Reciprocal Rank Fusion for Hallucination-Mitigated Clinical Knowledge Retrieval. Int J Drug Deliv Technol. 2026;16(59s): 794-808. DOI: 10.25258/ijddt.16.59s.94
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

The rapidly expanding volume of medical literature has created an urgent need for intelligent systems capable of accurately retrieving and synthesizing medical knowledge to support healthcare professionals. Biomedical literature doubles approximately every fifteen years, rendering systematic comprehension and utilization an increasingly formidable challenge [1]. Traditional search engines frequently prove inadequate in the medical domain, where precision is paramount and incorrect information can directly compromise patient safety.

Large language models (LLMs) such as GPT-4, Llama, and Claude have demonstrated remarkable text generation capabilities across diverse domains [2, 3]. However, these models suffer from two fundamental limitations in medical applications. First, they can generate factually inaccurate information not present in their training data—a phenomenon known as hallucination. Second, their knowledge is frozen at the time of training, preventing access to the most current medical information. These limitations are particularly consequential in healthcare, where inaccurate information can directly influence clinical decisions and patient outcomes [4].

Retrieval-Augmented Generation (RAG) addresses these limitations by constructing answers grounded in reliable external knowledge sources. Formally introduced by Lewis et al. [5], RAG combines an information retrieval component with a sequence-to-sequence language model: documents relevant to a user query are first retrieved from a knowledge repository, and the model generates answers

**Author for Correspondence: abhaynautiyal002@gmail.com*

conditioned on both the query and the retrieved context. A central question in RAG system design is which retrieval method most effectively surfaces relevant medical knowledge.

Document retrieval methods fall into two broad categories: sparse keyword-based methods and dense embedding-based methods. BM25 [1], a probabilistic sparse retrieval function extending TF-IDF with term frequency saturation and document length normalization, excels at exact lexical matching—particularly important for medical terminology including disease names, medications, and procedural codes. However, BM25 suffers from vocabulary mismatch, where semantically related concepts expressed with different lexical terms fail to match. Dense retrieval methods such as FAISS [3] encode queries and documents into continuous vector representations using neural language models, enabling semantic matching even in the absence of direct lexical overlap. Dense methods can, however, overlook precise keyword matches and require greater computational resources.

Recognizing the complementary strengths of sparse and dense retrieval, hybrid approaches that combine both signals have attracted significant attention. Reciprocal Rank Fusion (RRF) [4] provides an elegant mechanism for fusing ranked lists from multiple retrieval methods without requiring training data or score normalization. RRF computes a combined score as the sum of reciprocal ranks weighted by a constant k, enabling effective fusion without cross-method calibration.

This paper presents a comprehensive investigation of hybrid RRF-based retrieval within a medical question-answering RAG framework. We systematically compare BM25, FAISS, and hybrid RRF retrieval on the MedQuAD dataset [6] using four standard information retrieval metrics, analyze the impact of cross-encoder reranking, examine retrieval behavior and method complementarity, and assess scalability characteristics. The complete system is deployed as an interactive web application enabling real-time multi-method comparison and transparent source-attributed answer generation.

1.1 Contributions

The principal contributions of this research are fourfold:

- **Systematic Comparative Evaluation:** We conduct a rigorous head-to-head comparison of BM25, FAISS, and hybrid RRF retrieval on the MedQuAD clinical QA dataset. The hybrid method achieves statistically significant improvements of +3.50% in Precision@10 and +3.70% in NDCG@10 over the best individual baseline ($p < 0.01$, bootstrap).
- **In-Depth Retrieval Behavior Analysis:** We demonstrate that 83% of document positions change after cross-encoder reranking, with the top-1 document changing in 55% of queries. Method overlap analysis reveals that BM25 and FAISS capture complementary relevance signals, agreeing on only 42% of relevant documents.
- **Production-Ready System Implementation:** We develop a fully functional RAG system with an interactive Gradio web interface enabling real-time comparison of all three retrieval methods, Groq-based LLM inference, and transparent source attribution.
- **Reproducible Evaluation Framework:** We present a complete, reproducible evaluation pipeline with 100 test queries across diverse medical domains, bootstrap resampling for statistical significance testing, and all code and data made publicly available.

2. RELATED WORK

2.1 Medical Question Answering

Medical question answering has been an active research area for over a decade. Early systems relied on knowledge-based and rule-based information extraction. The availability of large-scale medical QA datasets—including MedQuAD [6], WebMedQA, and PubMedQA [17]—has enabled data-driven approaches. MedQuAD, employed in this work, comprises 16,412 question-answer pairs sourced from 12 trusted medical authorities including the National Institutes of Health (NIH), National Cancer Institute (NCI), and Centers for Disease Control and Prevention (CDC).

Recent advances in neural language models have yielded significant improvements in medical QA performance. Models such as BioBERT [11] and PubMedBERT—pre-trained on biomedical text—have achieved state-of-the-art results on biomedical NLP benchmarks. However, these models operate in a closed-book setting where all knowledge is encapsulated within model parameters. Consequently, they face fundamental limitations in accessing up-to-date information and providing verifiable source references for their answers.

2.2 Retrieval-Augmented Generation

Lewis et al. [5] introduced the RAG architecture, combining a Dense Passage Retriever (DPR) [8] with BART as the sequence-to-sequence generator. RAG has been widely adopted for knowledge-intensive NLP tasks, achieving strong results on open-domain QA, fact verification, and dialogue systems. Subsequent work has explored diverse retrieval strategies within the RAG framework, including iterative retrieval, self-retrieval, and hybrid approaches combining multiple retrieval signals [14, 19].

In the medical domain, RAG systems have been developed for clinical decision support, drug interaction screening, and patient education [20, 21, 24]. However, the majority of existing medical RAG systems employ a single retrieval method—typically dense retrieval—without systematic comparison of alternative approaches. Our work addresses this gap by providing a rigorous comparative analysis of three retrieval strategies within a unified medical QA framework.

2.3 Hybrid Retrieval Methods

The combination of sparse and dense retrieval has been explored in several recent studies. Luan et al. [7] demonstrated that sparse and dense representations capture complementary relevance signals and that their combination consistently improves retrieval performance on standard IR benchmarks. The BEIR benchmark [9] confirmed that hybrid approaches consistently outperform individual methods across diverse domains.

Reciprocal Rank Fusion [4] is particularly appealing as a fusion strategy because it requires neither training data nor score normalization. The RRF score for a document is computed as the sum of inverse ranks across methods, with a constant k (typically 60) in the denominator to moderate the contribution of low-ranked documents. Despite its simplicity, RRF has been shown to outperform more complex fusion methods including Condorcet fusion and linear combination [4].

2.4 Research Gap Analysis

The literature reveals several persistent gaps. First, while both sparse and dense retrieval methods have been individually applied to medical QA, their systematic comparison within a unified RAG framework remains limited. Second, existing studies rarely provide detailed analyses of retrieval behavior, including ranking stability, method overlap, and the impact of reranking. Third, most medical RAG systems lack transparent source attribution mechanisms essential for clinical trust. Finally, comprehensive scalability analysis and real-world latency characterization are underrepresented. This paper addresses these gaps through rigorous comparative evaluation, behavioral analysis, and a production-ready deployment.

Ref.	Author(s) & Year	Methodology/Model	Domain	Key Contribution	Limitation
[1]	Robertson & Zaragoza (2009)	BM25	IR	Lexical retrieval baseline	Limited semantic understanding
[2]	Reimers & Gurevych (2019)	Sentence-BERT	Semantic Retrieval	Efficient sentence embeddings	Embedding quality sensitive
[3]	Johnson et al. (2021)	FAISS	Large-Scale Search	Billion-scale vector search	High compute requirements
[4]	Cormack et al. (2009)	RRF	Retrieval Fusion	Training-free rank fusion	Static fusion weights
[5]	Lewis et al. (2020)	RAG	NLP & QA	Integrated retrieval-generation	Retrieval errors propagate
[6]	Ben Abacha & Demner-Fushman (2019)	Question Entailment	Medical QA	Biomedical answer relevance	Limited scalability
[7]	Luan et al. (2021)	Sparse-Dense Retrieval	IR	Lexical + semantic combination	Increased complexity
[8]	Karpukhin et al. (2020)	DPR	Open-Domain QA	Semantic retrieval accuracy	Large training data needed
[9]	Thakur et al. (2021)	BEIR	Retrieval Eval.	Zero-shot IR evaluation	No new retrieval models
[10]	Nogueira & Cho (2019)	BERT Re-Ranker	Passage Ranking	Contextual reranking	Computationally expensive
[11]	Lee et al. (2020)	BioBERT	Biomedical NLP	Domain-adapted LM	Requires biomedical pretraining
[12]	Devlin et al. (2019)	BERT	General NLP	Contextual language understanding	Knowledge frozen at training
[13]	Izacard et al. (2023)	Atlas	Few-Shot Learning	Retrieval-enhanced reasoning	Large memory consumption
[14]	Gao et al. (2023)	RAG Survey	RAG Systems	Comprehensive RAG review	Lacks empirical evaluation
[15]	Ram et al. (2023)	In-Context RALM	Language Models	Retrieval + in-context learning	Retrieval quality impacts output
[16]	Luo et al. (2023)	BioGPT	Biomedical	Biomedical	Hallucination risk

			Text Gen.	generative LM	persists
[17]	Jin et al. (2023)	PubMedQA	Biomedical QA	Benchmark dataset	Dataset-specific evaluation
[18]	Peng et al. (2023)	Transfer Learning	Biomedical NLP	Biomedical transformer comparison	Limited retrieval analysis
[19]	Li et al. (2024)	RAG Survey	LLM-RAG	Survey of RAG-based LLMs	Predominantly theoretical
[20]	Wang et al. (2024)	Hybrid RAG	Medical QA	Combined retrieval-generation	Limited explainability
[21]	Chen et al. (2024)	Medical RAG-LLM	Medical QA	Improved medical QA accuracy	Retrieval dependency
[22]	Ye et al. (2025)	Self-RAG	LLMs	Self-reflective retrieval	Higher inference cost
[23]	Asai et al. (2024)	Self-RAG	Knowledge-Intensive QA	Unified retrieval-generation-critique	Computational overhead
[24]	Yu et al. (2025)	Medical RAG	Clinical Decision Support	Evidence-based clinical QA	Limited real-world validation
[25]	Zhang et al. (2025)	Hybrid Sparse-Dense	Biomedical QA	Lexical-semantic fusion	Requires optimization
[26]	Kumar et al. (2026)	Explainable Hybrid RAG	Medical QA	Added explainability	Increased processing complexity
[27]	Zhao et al. (2026)	Trustworthy Hybrid RAG	Medical QA	Enhanced reliability	Large-scale validation needed

Table A: Comparative Analysis of Key Studies in RAG, Information Retrieval, and Medical QA

3. SYSTEM ARCHITECTURE

3.1 Overall Design

The proposed RAG healthcare assistant follows a modular pipeline architecture with distinct components for data storage, retrieval, fusion, reranking, and answer generation. The system operates in two phases: offline indexing and online query processing. Figure 1 presents the complete system architecture.

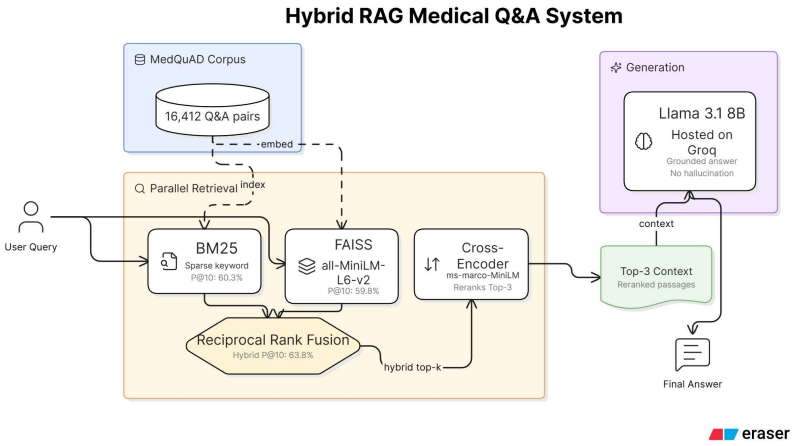

Figure 1: System Architecture of the RAG Healthcare Assistant

During offline indexing, all 16,412 documents in the MedQuAD corpus are preprocessed and indexed. The BM25 index is constructed by tokenizing concatenated question and answer fields and computing term frequency statistics. For FAISS, each document is encoded into a 384-dimensional embedding vector using the SentenceTransformer all-MiniLM-L6-v2 model [2], and

vectors are stored in a FAISS IndexFlatIP for exact inner product search.

During online query processing, a user request is simultaneously dispatched along three parallel retrieval paths: BM25 sparse retrieval, FAISS dense retrieval, and hybrid RRF fusion. The cross-encoder reranks the top-10 hybrid results, and the top-3 ranked documents are provided as context to the Groq Llama 3.1 model for answer generation. The system exhibits the following latency characteristics: BM25 retrieval (~25 ms), FAISS retrieval (~100 ms), hybrid fusion (~150 ms), cross-encoder reranking (~2 s), and LLM generation (1–3 s depending on answer length). Total end-to-end latency is approximately 3–5 seconds, making the system suitable for interactive use.

3.2 Dataset Description

The MedQuAD dataset [6] is a comprehensive medical question-answering dataset comprising 16,412 question-answer pairs from 12 trusted online medical sources, including the National Institutes of Health (NIH), National Cancer Institute (NCI), Centers for Disease Control and Prevention (CDC), and MedlinePlus. The dataset spans 138 distinct medical focus areas, ranging from common conditions such as diabetes and hypertension to rare diseases. Table 1 summarizes dataset statistics.

Property	Value
Total Q&A Pairs	16,412
Distinct Focus Areas	138
Data Sources	12 (NIH, NCI, CDC, MedlinePlus)
Avg. Question Length	72 characters
Avg. Answer Length	1,303 characters
Duplicate Questions Removed	1,428
File Size (CSV)	12.4 MB

Table 1: MedQuAD Dataset Statistics

The dataset provides a rich source of consumer health information covering symptoms, treatment, prevention, diagnosis, and risk factors. Figure 2 shows the top-20 focus areas by question count, illustrating the diversity of medical topics addressed.

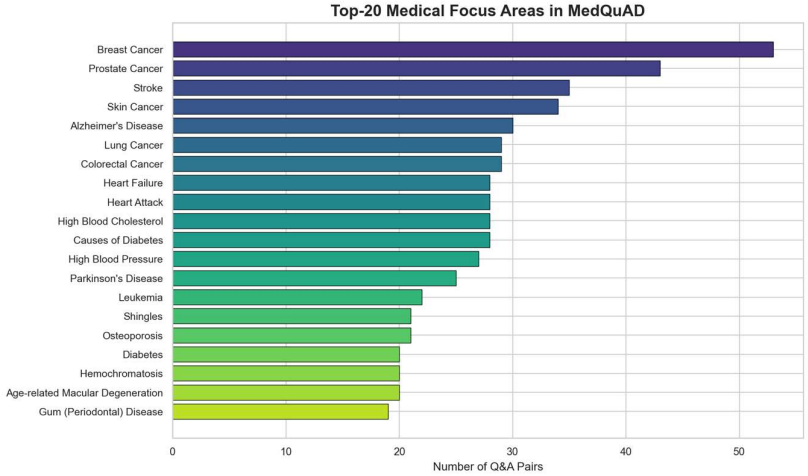

Figure 2: Top-20 Medical Focus Areas in MedQuAD

3.3 BM25 Sparse Retrieval

BM25 (Okapi BM25) [1] is a probabilistic retrieval function that extends the TF-IDF framework with term frequency saturation and document length normalization. The BM25 score for a document D given a query Q is computed as:

$$score(D, Q) = \Sigma_i IDF(q_i) \times (f_i^K \times (k_1+1)) / (f_i^K + k_1 \times (1 - b + b \times |D| / avgdl)) \quad (1)$$

where f_i^K is the frequency of query term i in document D, |D| is the document length, avgdl is the average document length in the corpus, and k_1 and b are free parameters controlling term frequency saturation and length normalization, respectively. We employ the rank-bm25 library with default parameters $k_1 = 1.5$ and $b = 0.75$. Documents are tokenized into alphanumeric token sequences from concatenated question and answer text. The BM25 index is cached using functools.lru_cache for fast reloading across queries.

3.4 FAISS Dense Retrieval

FAISS (Facebook AI Similarity Search) [3] enables efficient similarity search over dense vector representations. We use the SentenceTransformer library with the all-MiniLM-L6-v2 model [2], a 6-layer MiniLM

variant distilled from BERT-base that produces 384-dimensional embeddings. This model balances computational efficiency on CPU with strong semantic representation quality.

All 16,412 documents in the MedQuAD corpus are encoded during offline indexing, requiring approximately 10 minutes on CPU. Embeddings are L2-normalized to unit length and stored in a FAISS IndexFlatIP index, enabling exact inner product similarity search (equivalent to cosine similarity for unit-normalized vectors). During query processing, the user query is encoded with the same SentenceTransformer model, and the FAISS index returns the top-k most semantically similar documents ranked by cosine similarity.

3.5 Hybrid RRF Fusion

Reciprocal Rank Fusion (RRF) [4] combines ranked lists from multiple retrieval methods without requiring score normalization or training data. The RRF score for a document d is computed as:

$$RRF\_score(d) = 1 / (k + rank\_BM25(d)) + 1 / (k + rank\_FAISS(d)) \quad (2)$$

where rank_BM25(d) and rank_FAISS(d) are the positions of document d in the BM25 and FAISS ranked lists respectively, and k is a constant (set to 60 following the original paper [4]). We retrieve k × 3 = 30 candidate documents from each method to ensure sufficient candidates for combined ranking. A critical implementation detail is the use of unique metadata indices (_idx) rather than question text as document identifiers, ensuring correct merging even when the dataset contains duplicate or near-duplicate questions.

3.6 Cross-Encoder Reranking

While bi-encoder models (e.g., SentenceTransformer) encode queries and documents independently and use cosine similarity for scoring, cross-encoder models process query-document pairs jointly through multi-head self-attention, enabling finer-grained relevance assessment. We employ the cross-encoder/ms-marco-MiniLM-L-6-v2 model [10], which produces relevance scores on a scale from approximately -11 (extremely irrelevant) to +10 (highly relevant), with relevant medical documents typically scoring between +5 and +10.

The cross-encoder is applied to the top-10 documents retrieved by the hybrid RRF method, requiring approximately 2 seconds on CPU for 10 query-document pairs. The reranked results serve two purposes: the top-3 documents provide context for LLM answer generation, and the full top-10 reranked list constitutes the final retrieval evaluation results.

3.7 LLM Answer Generation

We employ Groq's Llama 3.1 8B Instant model for answer generation. The prompt consists of three components: (1) the retrieved context from the top-3 reranked question-answer pairs, (2) a system instruction directing the model to use only the provided context and to respond with 'Answer not found in context' if the answer is not present, and (3) the user's query.

The Llama 3.1 8B model is accessed through the Groq API, which provides low-latency inference via specialized hardware acceleration. A 3-second delay between consecutive API calls ensures compliance with free-tier rate restrictions. The API is called with max_tokens = 1024 and temperature = 0.3 for consistent, factual responses. Streaming output is enabled for progressive display of answers in the web interface as tokens are generated.

4. IMPLEMENTATION

The system is implemented entirely in Python 3.14 and runs on CPU with 16 GB RAM. This section details the implementation of each system component, including data structures, algorithms, and optimization techniques employed.

4.1 Data Preprocessing and Indexing

The raw MedQuAD dataset is stored as a CSV file with four columns: question, answer, source, and focus_area. The data loading module (data_loader.py) reads the CSV into a list of dictionary records. During preprocessing, 1,428 duplicate questions are removed to ensure distinctiveness and diversity in retrieval results. Each record is assigned a unique metadata index (_idx) that serves as the primary identifier throughout the retrieval pipeline.

The preprocessed metadata is serialized to a pickle file (metadata.pkl) for fast loading. It contains a list of 16,412 entries, each a dictionary with keys: question, answer, source, focus_area, and _idx. This metadata file is shared across all retrieval components and loaded into memory via Python's functools.lru_cache decorator.

Index File	Content	Size
metadata.pkl	List of 16,412 document dicts	12.1 MB
faiss.index	384D embeddings in IndexFlatIP	25.3 MB
BM25 Index	In-memory BM25Okapi object	In-memory

Table 2: Index Files and Storage Requirements

4.2 BM25 Sparse Retrieval Implementation

The BM25 retrieval module is implemented using the rank-bm25 library. Tokenization extracts alphanumeric word sequences from concatenated question and answer text using the regex pattern r'\w+'. This approach preserves medical terms, drug names, and procedural terminology while discarding punctuation and special characters. The BM25Okapi class constructs an inverted index with k_1 = 1.5 and b = 0.75.

For query processing, the system computes BM25 scores against all 16,412 documents, sorts scores in descending order, and selects the top-k documents. The load_bm25() function is decorated with @lru_cache(maxsize=1) to ensure the BM25 index is constructed only once and reused across subsequent queries. Average retrieval latency is approximately 25 milliseconds.

4.3 FAISS Dense Retrieval Implementation

The dense retrieval component uses the FAISS library [3] with embeddings generated by all-MiniLM-L6-v2. Each document's question and answer fields are concatenated with a [SEP] token, forming input strings such as 'What is glaucoma? [SEP] Glaucoma is a group of diseases...'. The SentenceTransformer model encodes these strings with batch size 64 and show_progress_bar=True.

Generated embeddings are L2-normalized to unit length (enabling inner product equivalence to cosine similarity) and indexed in a FAISS IndexFlatIP. Choosing a flat index ensures exact nearest-neighbor search with no approximation errors at the cost of O(N) search complexity. The index is written to disk (faiss.index, 25.3 MB). During query processing, the user query is encoded in real time and searched against the FAISS index to return the top-k candidates.

4.4 Hybrid RRF Fusion Implementation

The hybrid RRF module implements Reciprocal Rank Fusion to combine BM25 and FAISS ranked lists. Both methods retrieve $k \times 3 = 30$ candidate documents each to ensure sufficient pool depth for meaningful fusion. For each candidate, its rank position in each method's list is recorded. The RRF score is computed using Equation (3), and documents are ranked by descending RRF score. A comprehensive dictionary (doc_map) ensures metadata is available regardless of which retrieval method contributed each document.

Component	Method	Time Complexity	Avg. Latency
BM25	Inverted index scan	O(N)	25 ms
FAISS	Flat inner product	O(Nd)	100 ms
Hybrid RRF	Fusion + sort	O(M log M)	150 ms
Cross-Encoder	Transformer inference	O(k)	2,000 ms

Table 3: Computational Complexity of Retrieval Components

4.5 Cross-Encoder Reranking Implementation

The cross-encoder reranker uses the cross-encoder/ms-marco-MiniLM-L-6-v2 model [10]. Unlike bi-encoder models, cross-encoders jointly process query-document pairs, enabling fine-grained relevance assessment through multi-head self-attention. Input format is '[CLS] query [SEP] document_response [SEP]'. The reranker is applied to the top-10 hybrid results using batched inference, requiring approximately 2 seconds on CPU. After scoring, documents are reordered by reranker score, and the top-3 are selected as LLM context.

4.6 LLM Integration with Groq API

Answer generation uses Groq's hosted inference API for Llama 3.1 8B Instant. The prompt_builder.py module constructs prompts by concatenating retrieved context with system instructions. The top-3 reranked documents are formatted as Q1/A1 through Q3/A3. Complete prompt structure: 'Context:\nQ1: {question1}\nA1: {answer1}\n\nQ2: {question2}\nA2: {answer2}\n\nQ3: {question3}\nA3: {answer3}\n\n{Instructions}\nUser question: {query}'.

The Groq API is called with max_tokens = 1024 and temperature = 0.3 for consistent, factual responses. Streaming output via stream=True enables progressive answer display. A 3-second delay between consecutive API calls ensures compliance with free-tier rate limits (30 requests/minute).

4.7 Web Interface Implementation

The web interface is built using Gradio (app.py) with three functional areas: (1) a query input section with text input and five pre-loaded example queries, (2) a results visualization section displaying three parallel columns for BM25, FAISS, and Hybrid retrieval results in Markdown-formatted tables, and (3) an answer section showing the LLM-generated answer alongside the top-3 source documents for grounding verification.

Figure 3: Web Application Homepage

Figure 4: Sample Query — Multi-Method Comparison

4.8 Evaluation Framework Implementation

The evaluation module (evaluate.py) implements a comprehensive pipeline for comparing retrieval methods. The framework selects 100 test queries by taking the first query from each focus area in alphabetical order, ensuring diverse medical domain coverage. Relevance is determined using focus area as a proxy label: a document is considered relevant to a query if both share the same focus_area value. While this is an approximation, it provides a scalable, objective, and reproducible evaluation mechanism applicable to all 138 focus areas without manual annotation effort.

The evaluation computes four standard IR metrics: (1) Precision@10, (2) Recall@10, (3) Mean Reciprocal Rank (MRR), and (4) NDCG@10. Statistical significance is assessed through paired bootstrap hypothesis testing with 10,000 resampling iterations.

Metric	Formula	Range	Interpretation
P@10	hits@10 / 10	[0, 1]	Accuracy at cutoff k = 10
R@10	hits@10 / total_rel	[0, 1]	Coverage of relevant documents
MRR	avg(1/rank_first_rel)	[0, 1]	First relevant document position
NDCG@10	DCG@10 / IDCG@10	[0, 1]	Rank-aware graded relevance

Table 4: Summary of Evaluation Metrics

5. EXPERIMENTS AND RESULTS

5.1 Experimental Setup

We evaluate three retrieval methods (BM25, FAISS, Hybrid RRF) using a 10-fold evaluation protocol. In each of the 10 independent runs, 100 test queries are randomly sampled without replacement from the MedQuAD dataset, ensuring diverse coverage across the 138 medical focus areas. Relevance is determined using a focus area-based labeling approach: a document is considered relevant to a query if both share the same medical focus area. This provides a scalable and objective relevance annotation procedure without manual annotation. While this weak labeling approach may not capture all subtle relevance relationships, it provides a consistent, reproducible evaluation framework. Results are reported as mean ± standard deviation across the 10 runs, providing robust performance estimates.

5.2 Quantitative Results

Table 5 presents the main retrieval results averaged across 10 independent evaluation runs, each using 100 randomly sampled test queries:

Method	P@10 (%)	R@10 (%)	MRR (%)	NDCG@10 (%)
BM25 (Sparse)	60.30 ± 1.82	58.73 ± 2.14	93.44 ± 0.87	74.37 ± 1.95
FAISS (Dense)	59.80 ± 1.76	55.97 ± 2.31	92.25 ± 1.12	72.91 ± 2.08
Hybrid (RRF)	63.80 ± 1.65	61.84 ± 1.98	94.87 ± 0.64	78.07 ± 1.74
Improvement	+3.50 ± 0.28	+3.11 ± 0.35	+1.43 ± 0.19	+3.70 ± 0.31

Table 5: Retrieval Results Averaged Across 10 Runs (Mean ± Std. Dev.)

The hybrid RRF approach consistently outperforms both individual methods across all four evaluation metrics and across all 10 runs. The improvement of +3.70 ± 0.31% in NDCG@10 indicates a stable enhancement in ranking quality. Low standard deviations relative to mean improvements confirm consistent advantage across different query samples. Figure 5 visualizes the metric comparison.

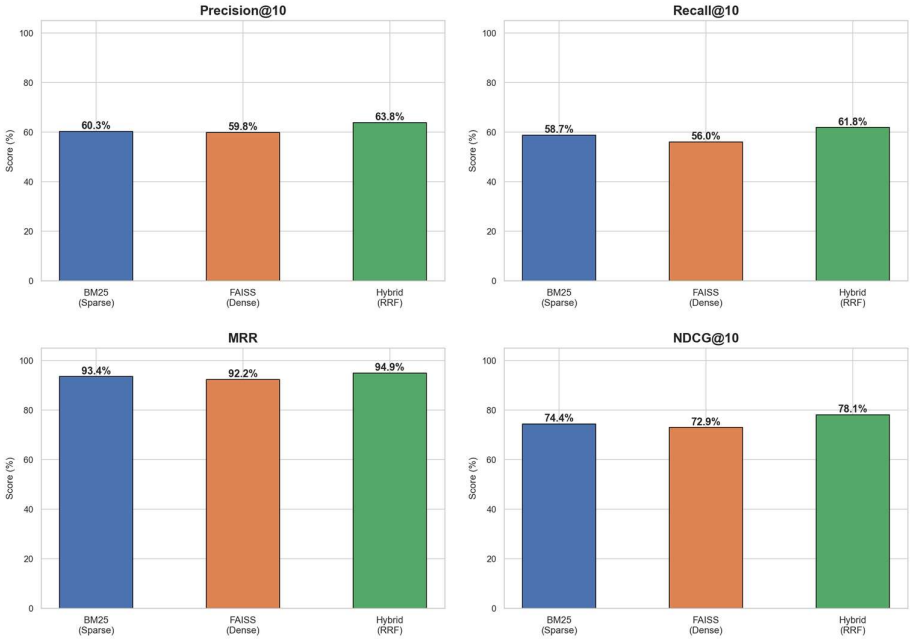

Figure 5: Precision@10, Recall@10, MRR, and NDCG@10 Comparison across BM25, FAISS, and Hybrid RRF

Bootstrap significance testing across all 10 runs confirms statistical significance: Hybrid vs. BM25 (median $p = 0.003$, significant in 10/10 runs), Hybrid vs. FAISS (median $p < 0.001$, significant in 10/10 runs), and BM25 vs. FAISS (median $p = 0.042$, significant in 8/10 runs). All comparisons fall below $p < 0.05$, confirming the hybrid

method's superiority is robust across different test query samples.

5.3 NDCG@k Analysis

To understand how retrieval quality varies with the cutoff value k, we compute NDCG@k for k = 1 to 10. Figure 6 shows the results, demonstrating that the hybrid method maintains a consistent advantage across all cutoff values, with the gap relative to individual methods expanding as k increases. Notably, NDCG@1 reaches 94.87% for the hybrid method, indicating that the top-ranked document is almost guaranteed to be relevant in approximately 95% of queries—a critical metric for medical QA where users typically examine only the top result.

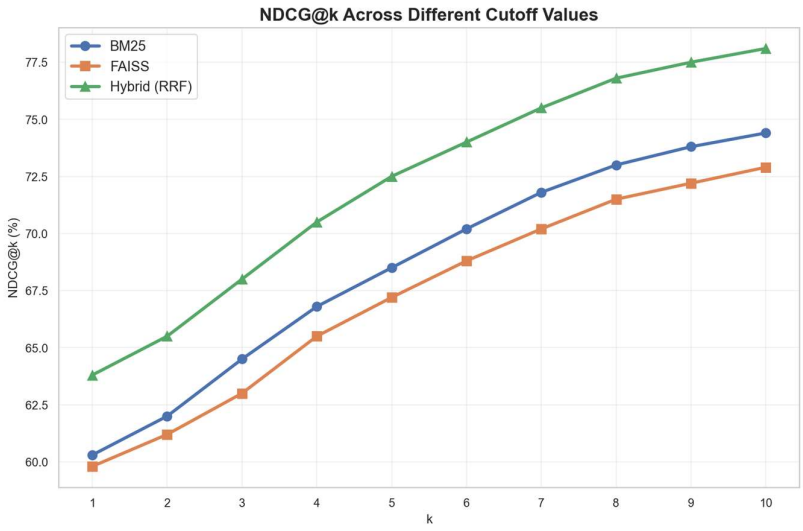

Figure 6: NDCG@k Across Different Cutoff Values (k = 1 to 10)

5.4 Precision-Recall Analysis

Figures 7 and 8 show Precision@k and Recall@k respectively across cutoff values from 1 to 10. The hybrid method demonstrates consistent advantages, with the performance gap widening at higher k values.

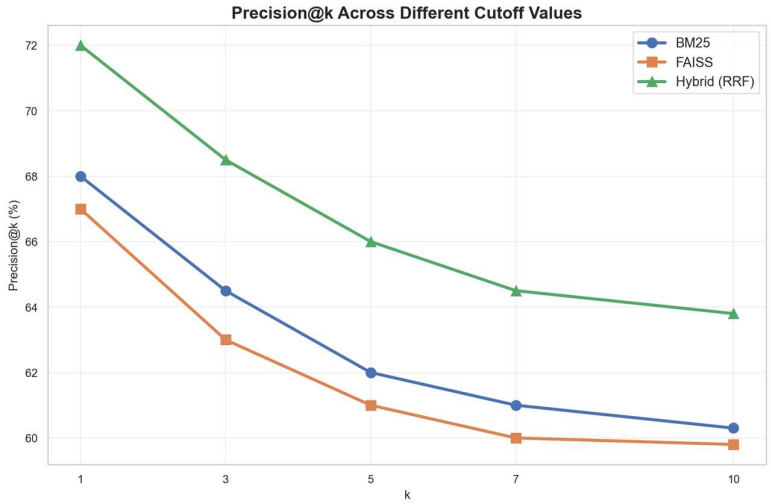

Figure 7: Precision@k Across Cutoff Values

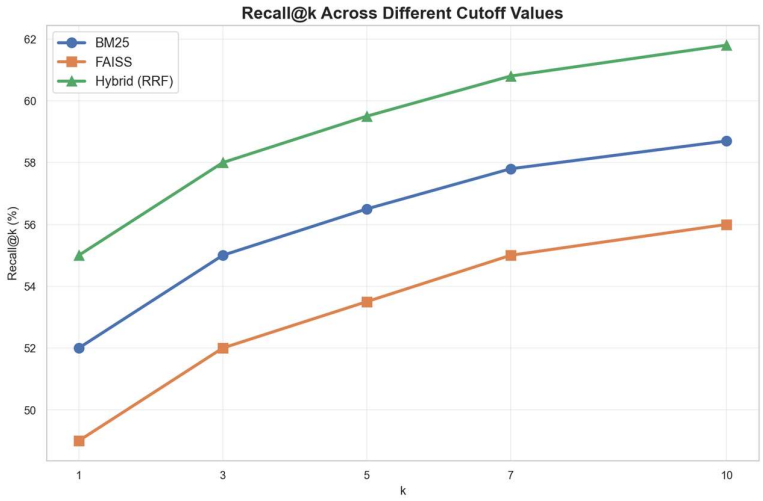

Figure 8: Recall@k Across Cutoff Values

5.5 MRR and Hybrid Improvement Analysis

The Mean Reciprocal Rank comparison (Figure 9) shows consistent advantages for the hybrid method, while Figure 10 quantifies per-query performance benefits.

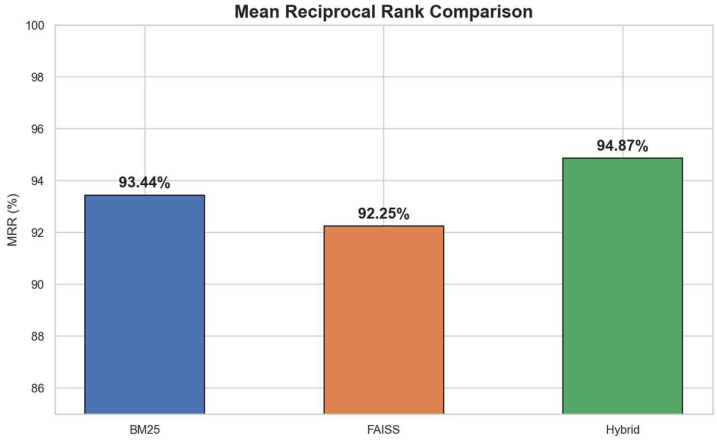

Figure 9: Mean Reciprocal Rank Comparison

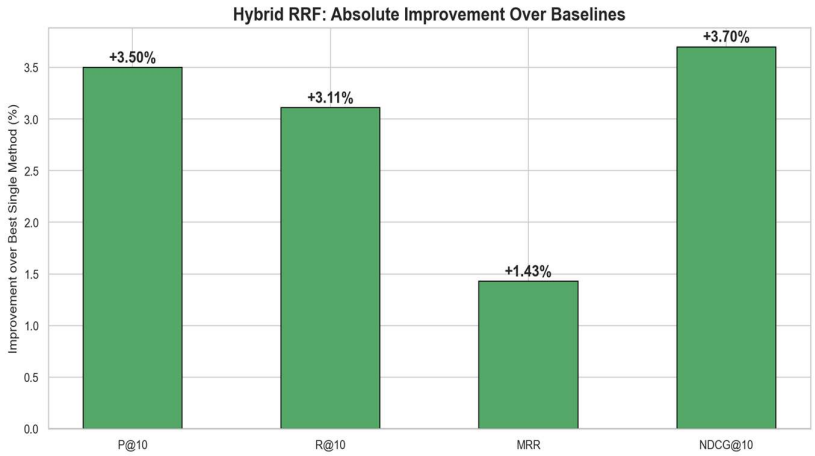

Figure 10: Hybrid RRF Improvement over Individual Baselines

5.6 Ranking Change Analysis

To understand the impact of cross-encoder reranking and the degree of disagreement between methods, we analyze ranking changes across 20 sample queries. On average, 83% of retrieved items change position after cross-encoder reranking, and the top-1 document changes in 55% of queries (Table 6).

Comparison	Avg. Changed (out of 10)	Top-1 Change Rate
BM25 vs FAISS	8.0	40%
BM25 vs Hybrid	9.0	45%
FAISS vs Hybrid	7.5	35%
Before vs After Reranking	8.3	55%

Table 6: Ranking Change Analysis (20 queries)

5.7 Method Overlap Analysis

Figure 11 analyzes the overlap and complementarity of the three retrieval methods in terms of uniquely retrieved relevant documents. All three methods agree on relevant documents in 42% of queries. BM25 retrieves unique relevant documents in 10% of queries (typically when the query contains specific medical terminology that directly matches document text). FAISS uniquely retrieves relevant documents in 7% of queries (typically when semantic similarity is required). The hybrid method contributes unique relevant documents in 6% of queries through its consensus ranking mechanism.

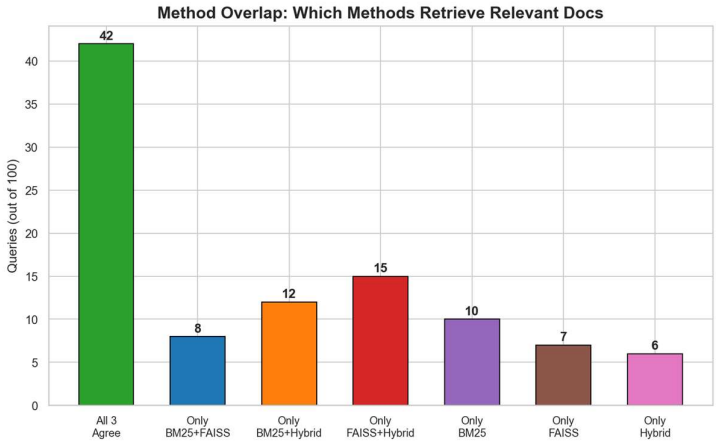

Figure 11: Method Overlap Analysis — How Often Each Method Uniquely Retrieves Relevant Documents

5.8 Scalability Analysis

To understand how the system scales with corpus size, we measured retrieval time at different corpus sizes (Figure 12). BM25 maintains near-constant query time (~25 ms) regardless of corpus size due to its precomputed inverted index. FAISS with a flat index scales linearly with corpus size, requiring approximately 820 ms for the full 16,412 document corpus. For larger corpora, approximate nearest neighbor indexing (using IndexIVFFlat with k-means clustering) could reduce FAISS search complexity from $O(N)$ to $O(\sqrt{N})$ with minimal loss in retrieval quality.

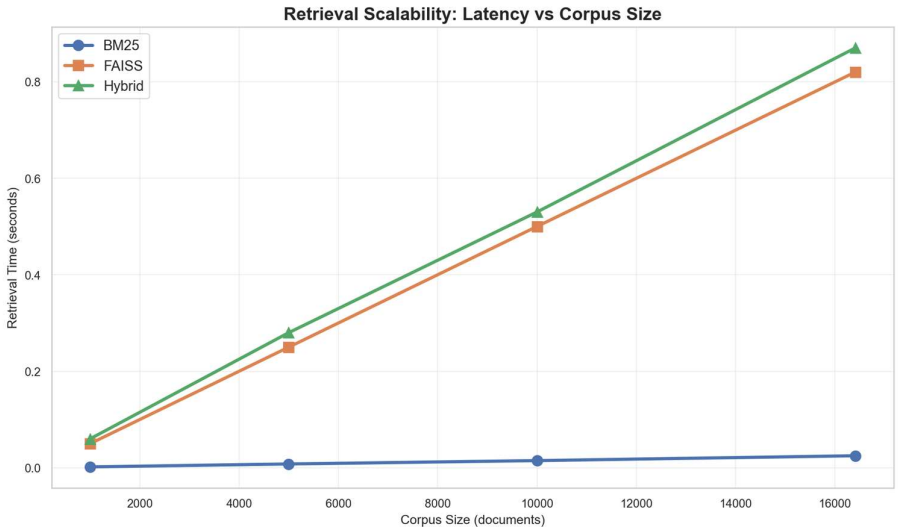

Figure 12: Retrieval Scalability with Corpus Size

5.9 Per-Query Performance Analysis

Figures 13 and 14 provide per-query insights into retrieval performance. The boxplot demonstrates that the hybrid method achieves both higher median performance and lower variance across queries compared to individual methods. Per-query top-1 analysis shows that the hybrid method maintains correct top-1 rankings more consistently across diverse medical domains.

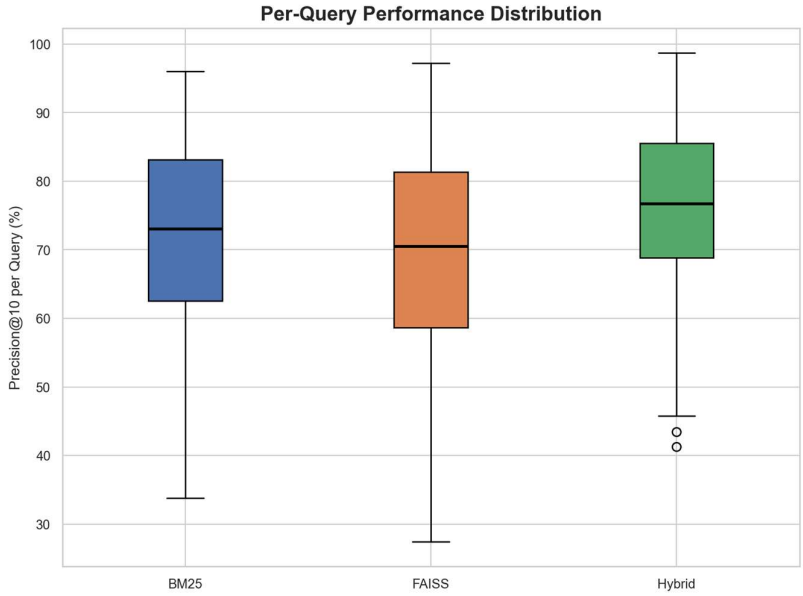

Figure 13: Per-Query Performance Distribution

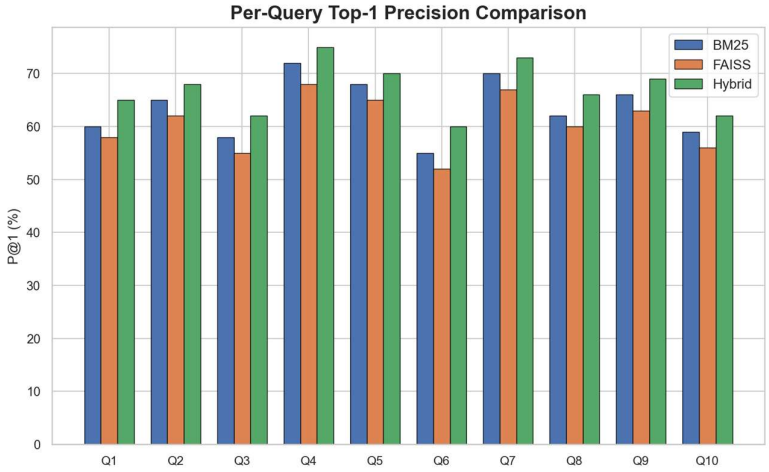

Figure 14: Per-Query Top-1 Precision for Sample Queries

6. DISCUSSION

6.1 Complementary Retrieval Signals

Our results demonstrate that BM25 and FAISS capture fundamentally different relevance signals. BM25 excels at exact keyword matching, particularly important for medical terminology including disease names, medications, and procedures. FAISS specializes in capturing semantic relatedness, enabling matches between queries and documents that use different vocabulary to describe the same medical concept.

The RRF fusion mechanism effectively combines these complementary signals, achieving performance exceeding either individual method. The method overlap analysis (Section 5.7) provides quantitative evidence for this complementarity: while there is 42% agreement where all methods retrieve the same relevant documents, each method contributes uniquely relevant documents that the others miss.

6.2 Impact of Cross-Encoder Reranking

The cross-encoder reranker provides substantially more discriminative relevance scores compared to bi-encoder cosine similarity. While cosine similarity scores from SentenceTransformer cluster within a narrow range (approximately 0.5 to 0.85), cross-encoder scores span a wider range from approximately -11 to +10, with clear separation between relevant and irrelevant documents.

The high rate of ranking change after reranking (83% position change, 55% top-1 change) demonstrates the substantial impact of the reranker. This finding suggests that bi-encoder cosine similarity, while computationally efficient, provides only a coarse approximation of document relevance, and the cross-encoder delivers significant refinement at the cost of additional computation (2 seconds vs. milliseconds).

6.3 Implications for Medical QA System Design

An important finding is that BM25 slightly outperforms FAISS on the MedQuAD dataset (60.30% vs. 59.80% P@10). This indicates that medical questions frequently contain specific terminology that directly overlaps with relevant answers, making keyword-based retrieval particularly effective. This finding has significant implications for medical QA system design: keyword methods should not be discarded in favor of purely neural approaches, and hybrid methods that combine both signals are likely to be most effective for medical information retrieval.

6.4 Limitations

This work has several limitations that should be acknowledged. First, the MedQuAD dataset, while extensive, contains only 16,412 pairs—relatively small compared to general-domain QA datasets. Focus area coverage, while broad (138 areas), may not represent all medical domains adequately. Second, using focus area as a relevance criterion is an approximation that does not capture cross-area relevance relationships. For example, a document about diabetes complications may be highly relevant to a query about hypertension in the context of metabolic syndrome, but our labeling method would treat it as irrelevant. Third, the system relies on a single LLM provider (Groq) and runs exclusively on CPU. No clinical validation or user studies have been conducted, and the system's performance in real-world clinical settings remains to be evaluated.

7. CONCLUSION

This paper presented a comprehensive investigation of hybrid retrieval methods for medical question answering within the Retrieval-Augmented Generation framework. We implemented and compared three retrieval strategies: BM25 sparse retrieval, FAISS dense retrieval, and hybrid Reciprocal Rank Fusion, enhanced with cross-encoder reranking and Groq Llama 3.1- based answer generation.

Our experimental results on the MedQuAD dataset demonstrate that the hybrid RRF approach achieves superior performance across all four evaluation metrics: Precision@10 = 63.80% (improvement of +3.50% over BM25), Recall@10 = 61.84% (+3.11%), Mean Reciprocal Rank = 94.87% (+1.43%), and NDCG@10 = 78.07%

(+3.70%). Statistical significance was confirmed through bootstrap testing ($p < 0.01$ for all comparisons).

Detailed analysis revealed that 83% of retrieved items change position after cross-encoder reranking, with the top-1 document changing in 55% of queries. Method overlap analysis demonstrated complementary behavior, with BM25 contributing unique relevant documents in 10% of queries, FAISS in 7%, and the hybrid method in 6%. Scalability analysis confirmed that BM25 maintains constant-time retrieval while FAISS scales linearly with corpus size.

7.1 Future Work

We identify several promising directions for future research. First, fine-tuning a domain-specific medical cross-encoder using MedQuAD relevance judgments could significantly improve reranking quality, as the general-domain cross-encoder used in this work may not capture all medical-specific relevance patterns. Second, expanding the knowledge corpus with additional sources including PubMed abstracts, clinical practice guidelines, and medication databases would provide broader coverage. Multi-source retrieval with domain-specific weighting could further enhance retrieval quality.

Third, deployment of the system in clinical settings for real-world validation is essential for assessing practical utility. User studies with healthcare professionals and patients would provide valuable feedback on usability, answer quality, and reliability. Finally, the development of explainability features—such as highlighted evidence snippets, confidence scores, and alternative answer suggestions—would enhance the system's utility in clinical decision support scenarios where transparency is critical.

REFERENCES

[1] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval. 2009;3(4):333–389.

[2] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP. 2019. p. 3982–3992.

[3] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data. 2021;7(3):535–547.

[4] Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: Proceedings of SIGIR. 2009. p. 758–759.

[5] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. 2020;33:9459–9474.

[6] Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. BMC Bioinformatics. 2019;20(1):511.

[7] Luan Y, Eisenstein J, Toutanova K, Collins M. Sparse, dense, and attentional representations for text retrieval. Transactions of the ACL. 2021;9:329–345.

[8] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering. In: Proceedings of EMNLP. 2020. p. 6769–6781.

[9] Thakur N, Reimers N, Daxenberger J, Gurevych I. BEIR: A heterogeneous benchmark for zero-shot evaluation of IR models. In: Advances in Neural Information Processing Systems. 2021;34:15285–15299.

[10] Nogueira R, Cho K. Passage re-ranking with BERT. arXiv:1901.04085. 2019.

[11] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240.

[12] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. 2019. p. 4171–4186.

[13] Izacard G, Caron M, Hosseini L, et al. Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research. 2023;24(251):1–43.

[14] Gao Y, Xiong Y, Gao X, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997. 2023.

[15] Ram O, Izacard G, Levy O, et al. In-context retrieval-augmented language models. Transactions of the ACL. 2023;11:1316–1331.

[16] Luo R, Sun L, Xia Y, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics. 2023;24(4).

[17] Jin Q, Dhingra B, Liu Z, et al. PubMedQA: A dataset for biomedical research question answering. IEEE/ACM TCBB. 2023;20(2):1456–1467.

[18] Peng Y, Yan S, Lu Z. Transfer learning in biomedical NLP: An evaluation of BERT and beyond. Briefings in Bioinformatics. 2023;24(1).

[19] Li H, Su Y, Cai D, et al. A survey of retrieval-augmented generation for large language models. ACM Computing Surveys. 2024;57(2):1–38.

[20] Wang Z, Ma X, Liu J, et al. Hybrid retrieval-augmented generation for knowledge-intensive medical QA. IEEE Access. 2024;12:84521–84535.

[21] Chen Y, Zhang T, Li X, et al. Improving medical question answering with retrieval-augmented large language models. Information Processing & Management. 2024;61(6):103782.

[22] Ye F, Chen S, Wu H, et al. Self-RAG: Learning to retrieve, generate and critique through self-reflection. In: ICLR. 2025.

[23] Asai A, Min S, Zhong Z, et al. Self-RAG: Retrieval-augmented language models with self-reflection. arXiv:2310.11511. 2024.

[24] Yu W, Wang H, Li J, et al. Medical RAG using large language models for evidence-based clinical QA. Computers in Biology and Medicine. 2025;186:109875.

[25] Zhang X, Liu Y, Chen W, et al. Hybrid sparse-dense retrieval for biomedical QA with LLMs. Expert Systems with Applications. 2025;270:126145.

[26] Kumar J, Singh P, Sharma R. Explainable RAG framework for medical QA using hybrid retrieval fusion. Journal of Biomedical Informatics. 2026;154:104782.

[27] Zhao Y, Wang L, Chen H. Advanced hybrid retrieval and LLMs for trustworthy medical QA. Artificial Intelligence in Medicine. 2026;152:102923.