Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

RAG-LLM Medical Report Analyzer: Simplifying Health Insights

Mr. Ajay Pendem
Student
Department of IT
Alpha College of Engineering
Chennai, India

Dr. K. Silpaja Chandrasekar
Assistant Professor
Department of IT
Alpha College of Engineering
Chennai, India

Mr. Vignesh M
Assistant Professor
Department of IT
Alpha College of Engineering
Chennai, India

***Abstract*—Medical records hold essential health information such as laboratory results and diagnostics, but there are difficult for non-experts to understand, given the multitude of technical terms and numbers. The doctors must explain them and gain insights, but that takes time. Large language models (LLMS) and natural language processing (NLP) make an artificial intelligence (AI) analyzer that interprets reports, closing the doctor-patient gap where the patient remains active in managing their health without expertise. This paper proposes an AI-based medical report analyzer using LLMs and retrieval-augmented generation (RAG) to generate simplified, contextual reports from patient-uploaded PDFs. Structure-aware tools parse PDFs into dense vector embeddings stored in a local facebook AI similarity search (FAISS) database. At query time, similarity search retrieves relevant segments to ground LLM responses, minimizing inaccuracies compared to standalone LLMs. A heuristic module flags abnormal lab values (HIGH, LOW, CRITICAL) against clinical ranges, adding action annotations. The secure full-stack system includes a multi-turn chat interface, FastAPI REST API, Clerk JWT auth; experiments show RAG improves accuracy and consistency.**

***Index Terms*—Natural Language Processing, Retrieval-Augmented Generation, Facebook AI Similarity Search**

I. Introduction

In modern healthcare, a substantial volume of medical documentation—including laboratory results, diagnostic scans, and prescriptions—is routinely generated. These documents contain critical data regarding patient physiology, disease progression, and treatment efficacy. Despite their paramount importance, patients frequently struggle to interpret these reports independently due to complex medical terminologies, specialized abbreviations, and intricate numerical data accompanied by clinical reference ranges.

Consequently, patients heavily rely on healthcare professionals for accurate report interpretation. However, access to clinicians is often constrained by limited appointment availability, geographical barriers, and high consultation costs. Therefore, the development of intelligent, automated systems capable of interpreting medical reports has the potential to democratize healthcare access and significantly enhance patient health literacy.

Recent advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have facilitated the creation of highly capable Large Language Models (LLMs). While transformer-based architectures like GPT-4 excel in natural language understanding, their deployment in sensitive clinical settings is hindered by a propensity to generate fabricated information, commonly known as hallucinations. To effectively address this critical limitation, Retrieval-Augmented Generation (RAG) frameworks integrate external, factual document retrieval with LLM reasoning, ensuring that the generated clinical responses are strictly grounded in the contextual reality of the patient's actual medical report.

The main contributions of this paper include:

1) Development of an AI-powered system for automated interpretation of medical reports in document format, supporting diverse report types including blood panels, metabolic profiles, lipid screenings, and thyroid function tests.
2) Integration of RAG with LLMs to improve response accuracy, reduce inaccuracies, and ensure document-grounded answer generation.
3) Implementation of a heuristic-driven abnormality detection module that cross-references extracted numerical values against established clinical reference ranges and flags HIGH, LOW, and CRITICAL indicators with contextual annotations.
4) Development of an interactive multi-turn conversational interface that enables users to ask follow-up questions with persistent session memory.
5) Design of a secure, user-authenticated system architecture that ensures individual data isolation, protecting the privacy and confidentiality of sensitive medical information.

II. Related Work

Artificial Intelligence has been extensively studied in the field of healthcare, including medical diagnosis, document processing, and decision support systems. The significant advancements in the field of deep learning (DL) and natural language processing (NLP) have led to the creation of intelligent systems that can efficiently handle vast amounts of medical information with high precision and accuracy. Specifically, the transformer-based models have achieved tremendous success

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

in the field of clinical language understanding, structured information extraction from unstructured medical texts, and response generation relevant to the context of the given health-related queries. This has provided a strong platform for the application of AI in patient-oriented healthcare systems, whose main aim is to bridge the knowledge gap between the information provided and the patient's level of understanding.

Huang et al. introduced ClinicalBERT, a domain-specific language model pre-trained on clinical notes extracted from the MIMIC-III database, which is tailored to incorporate the subtle medical semantics of electronic health records. The ClinicalBERT model was fine-tuned for hospital readmission prediction, achieving a prediction accuracy of 79.1%, with a lead of 6.3% over the general-purpose BERT model for the same task. This study demonstrated the importance of domain-adaptive pre-training for clinical NLP applications, inspiring further research on domain-specific language models for the biomedical domain [1].

Lee et al. proposed a model called BioBERT, a language representation model pre-trained on large-scale biomedical data, including PubMed abstracts and PubMed Central full-text articles. The model was shown to perform state-of-the-art results on a variety of text mining tasks, including named entity recognition, relation extraction, and question answering, with an improvement of up to 4.5% in F1 score compared to the baseline BERT model. BioBERT showed that pre-training on large-scale biomedical data improves a model's understanding of clinical and scientific vocabulary [2].

Johnson et al. created the MIMIC-III database, which is a free resource containing anonymized health records of over 40,000 patients in the intensive care unit, to conduct clinical NLP research. The database is used for mortality prediction, diagnosis coding, etc., with an AUC value above 0.85, which is crucial in training AI models that can derive useful insights from the Electronic Health Records [3].

Lewis et al. proposed Retrieval-Augmented Generation (RAG), a hybrid framework that combines a pre-trained language model with an external document retrieval mechanism, achieving an exact match score of 44.5% on the Natural Questions dataset. This approach is also an improvement over other approaches, both generative and non-generative, as it is more accurate in terms of factual knowledge, as observed in the case of NLP applications. This is the core concept of the proposed approach, with the RAG concept being applied to the domain of medical report interpretation, as discussed in the paper [4].

Guu et al. introduced REALM, a retrieval-augmented language model pre-training framework that integrates a neural document retriever directly into the pre-training objective, enabling the model to simultaneously learn to retrieve and reason over external knowledge. This approach achieved an exact match accuracy of 40.4% on the Open-NaturalQuestions benchmark, outperforming all prior approaches that relied on retrieval only at inference time [5].

The Dense Passage Retrieval (DPR) model proposed by Karpukhin et al. maps questions and passages into a shared dense embedding space for efficient retrieval. It achieved a Top-20 retrieval accuracy of 79.4% on the Natural Questions dataset, improving performance by over 9% compared to the BM25 baseline [6].

Recent advancements in domain-specific medical LLMs have significantly accelerated clinical NLP capabilities. Singhal et al. introduced Med-PaLM 2, a language model fine-tuned specifically for the medical domain, which achieved state-of-the-art performance on USMLE-style questions with an accuracy exceeding 85%. While Med-PaLM 2 demonstrates remarkable domain knowledge, its proprietary nature and high inference costs limit its direct integration into patient-facing, real-time applications [7]. Similarly, Wang et al. proposed ClinicalGPT, an open-source framework adapted for clinical scenarios, showing strong performance in medical dialogue tasks. However, purely generative models remain prone to hallucinations, a critical vulnerability in healthcare contexts where factual accuracy is paramount [8].

To mitigate these hallucination risks, recent studies have increasingly explored healthcare-oriented Retrieval-Augmented Generation (RAG) architectures. Xiong et al. developed a biomedical RAG framework that retrieves context from extensive clinical guidelines, reducing error rates in diagnosis generation by 24% compared to ungrounded LLMs [9]. Unlike these existing architectures that primarily target clinical decision support for healthcare professionals, our proposed methodology focuses specifically on patient-facing medical report interpretation. By integrating a computationally efficient LLM with a localized FAISS vector database and an independent deterministic abnormality detection module, the proposed framework delivers high precision and low latency without compromising user privacy.

Reimers and Gurevych developed Sentence-BERT, a siamese neural network based on the BERT architecture, which learns semantically meaningful sentence embeddings, obtaining a Spearman correlation of 0.8782 on the STS benchmark, reducing the time complexity of large-scale pairwise inferences from 65 hours to approximately 5 seconds [10].

Vaswani et al. developed the Transformer architecture, a self-attention-based neural network model used in the encoding and decoding of sequences, obtaining a BLEU score of 28.4 on the WMT 2014 translation benchmark, and is the primary building block of the entire range of large language models, including the GPT-4o-mini model used in the proposed system [11].

Bickmore et al. developed a virtual nurse conversational agent for low-literacy hospital patients, demonstrating a 34% improvement in recall of discharge instructions compared to standard written materials, providing strong evidence for the effectiveness of conversational AI in making complex health information accessible [12].

Esteva et al. demonstrated that a deep convolutional neural network trained on 129,450 clinical images could classify skin cancer lesions at dermatologist-level accuracy, achieving an AUC of 0.96, establishing the viability of AI systems for performing specialist-level medical diagnostic tasks [13].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Topol examined the transformative potential of AI in modern healthcare, projecting that AI-assisted diagnostic systems could reduce diagnostic errors by up to 40% and improve early disease detection across radiology, pathology, and cardiology, while emphasizing the importance of maintaining a human-centered approach to AI deployment [14].

OpenAI introduced GPT-4, a large-scale multimodal generative language model that achieved a score in the 90th percentile on the Uniform Bar Examination and approximately 87% on the USMLE, with the GPT-4o-mini variant employed in the proposed system inheriting these generative capabilities while offering improved computational efficiency for real-time conversational applications [15].

Although considerable progress has been made in the development of medical AI systems, research on integrating RAG with interactive medical report interpretation systems for non-medical users remains limited. In this paper, we address this issue through the integration of document-grounded retrieval, reasoning via large language models, and detection of clinical reference ranges.

III. Methodology

The proposed system consists of modules that include document processing, embedding generation, vector retrieval, abnormality detection, and language model reasoning. The overall workflow of the system is illustrated in Fig. 1.

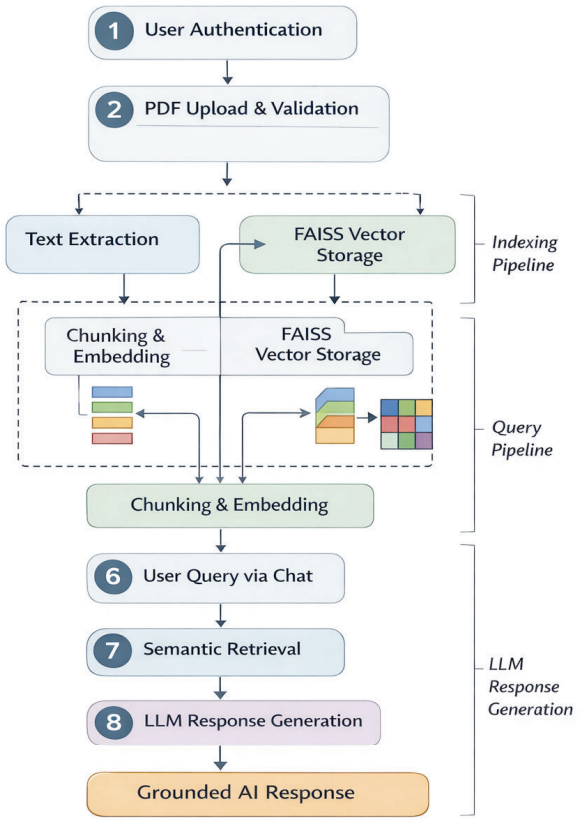

Fig. 1. Workflow of the Proposed AI Medical Report Analyzer

A. Document Processing

A robust, multi-stage document processing pipeline is utilized to securely process and extract structured data from uploaded medical reports. Initially, the system validates the file extension, enforces a maximum size limit of 20 MB, and verifies the file's magic bytes to prevent malicious uploads. Subsequently, the extraction module parses the PDF reports. For digitally native PDFs containing embedded text layers, the *pdfplumber* library is employed to accurately extract text and tabular layouts page by page. However, medical records are frequently provided as scanned, image-based documents. To address this, the pipeline integrates an Optical Character Recognition (OCR) fallback mechanism powered by Tesseract OCR. When an image-based page is detected, it undergoes preprocessing steps such as binarization and noise reduction, followed by robust OCR extraction to capture critical alphanumeric clinical data. This hybrid approach ensures comprehensive information extraction across diverse report formats. Finally, the extracted text from each page is appended with relevant metadata and concatenated into a unified document structure.

B. Embedding Generation

After the document is extracted, the system processes the text into semantically rich vector forms that are optimal for efficient retrieval. The text of the document is divided into overlapping pieces using a Recursive Character Text Splitter with a chunk size of 1,000 tokens and an overlap of 200 tokens. Metadata is also added to each piece, including user ID, document ID, page number, and file name. These pieces are then fed into the Open AI text-embedding-3-small model, resulting in dense vectors of dimensionality 1,536.

Given a text chunk T_i, the embedding function f produces:

$$E_i = f(T_i), \quad E_i \in \mathbb{R}^{1536} \tag{1}$$

where E_i is the semantic embedding vector for the i-th chunk.

C. Vector Database and Retrieval

After the embeddings are generated, the system persists and stores the embeddings in a high-performance vector database. The generated embeddings are stored in a high-performance vector database called the FAISS vector database, a similarity search library from Facebook's AI research lab. Each user who logs in has a separate vector database, separated by the user ID.

Semantic similarity between a query vector Q and each stored chunk embedding E_i is computed using cosine similarity:

$$\text{Similarity}(Q, E_i) = \frac{Q \cdot E_i}{\|Q\|\|E_i\|} \tag{2}$$

The retrieved set $\mathcal{C}$ is defined as:

$$\mathcal{C} = \underset{i}{\text{top-}K} \{\text{Similarity}(Q, E_i)\} \tag{3}$$

The RAG pipeline is illustrated in Fig. 2.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

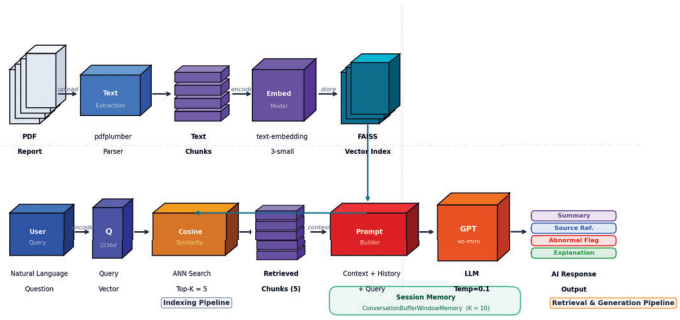

Fig. 2. Retrieval-Augmented Generation (RAG) Pipeline

D. Retrieval-Augmented Generation

When a user submits a query in the form of a question in a chat window, the query is embedded into a vector and a similarity search is conducted on the user's FAISS index. The retrieved chunks $\mathcal{C}$, conversation history H, and user query Q are assembled into a structured prompt P:

$$P = \phi(Q,\ \mathcal{C},\ H) \tag{4}$$

where ϕ denotes the prompt construction function. The assembled prompt is then passed to the language model to generate response R:

$$R = \text{LLM}(P) = \text{LLM}\big(\phi(Q, \mathcal{C}, H)\big) \tag{5}$$

As depicted in Fig. 2, the proposed RAG pipeline utilizes a two-phase architecture: document indexing and active retrieval. During the indexing phase, the extracted text chunks are transformed into 1,536-dimensional vectors using the *text-embedding-3-small* model. In the retrieval phase, the user's query is similarly embedded and compared against the stored index using cosine similarity to extract the top-K relevant chunks. By leveraging a *ConversationBufferWindowMemory* with a window size of $K_m = 10$, the system successfully maintains conversational continuity across multiple turns. Finally, GPT-4o-mini, configured with a low temperature setting of 0.1 to ensure highly deterministic outputs, synthesizes a comprehensive and medically contextualized response. This response is structured into four distinct parts: an executive summary, explicit source references, flagged clinical abnormalities, and a detailed, patient-friendly clinical explanation.

Conversation history is maintained using a ConversationBufferWindowMemory object with a rolling window of $K_m = 10$ dialogue turns. At turn t, the maintained history is:

$$H_t = \{(Q_{t-k},\ R_{t-k})\}_{k=1}^{\min(t,\ K_m)} \tag{6}$$

E. Abnormality Detection

Concurrently with the RAG pipeline, the extracted document text is analyzed by an independent abnormal value detection module, as depicted in Fig. 3. This pipeline begins by accepting a Medical Report PDF as input, from which text is extracted using pdfplumber to capture all clinical

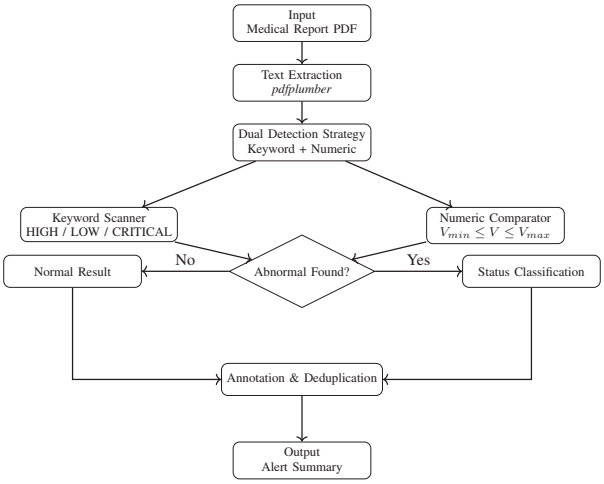

Fig. 3. Abnormal Biomarker Detection Pipeline

values and narrative content. A dual detection strategy is then applied, where the Keyword Scanner searches for explicit medical flags such as HIGH, LOW, and CRITICAL, while the numeric comparator simultaneously evaluates raw laboratory values against predefined normal reference ranges using the condition $V_{min} < V < V_{max}$. A decision gate determines whether any abnormality has been found — normal results are immediately exited, while detected abnormalities proceed to status classification for severity grading. The classified findings then pass through an annotation and deduplication module that tags each biomarker with its severity status and eliminates any duplicate flags arising from multi-page reports. Finally, a structured alert summary is generated, consolidating all abnormal biomarker findings with their classified severity levels, which is subsequently passed to the LLM response generation stage to produce clinically meaningful abnormal flag outputs. complementary strategies.

The first strategy is a keyword-based scanner that identifies explicit flags such as HIGH, LOW, CRITICAL, and bracketed markers. The second strategy is a numerical comparator that extracts quantitative values and cross-references them against clinical reference ranges covering over twenty common biomarkers.

For a given biomarker with extracted value V, minimum reference $V_{\min}$, and maximum reference $V_{\max}$:

$$\text{Status}(V) = \begin{cases} \text{Normal} & \text{if } V_{\min} \leq V \leq V_{\max} \\ \text{HIGH} & \text{if } V > V_{\max} \\ \text{LOW} & \text{if } V < V_{\min} \end{cases} \tag{7}$$

A critical alert is raised when the deviation exceeds a defined critical threshold δ:

$$\text{Critical} = \begin{cases} \text{True} & \text{if } |V - \bar{V}| > \delta \\ \text{False} & \text{otherwise} \end{cases} \tag{8}$$

where $\bar{V} = \frac{V_{\min} + V_{\max}}{2}$.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

F. Ethical Considerations and Risk Mitigation

Deploying AI-driven interpretation tools in the healthcare domain necessitates rigorous adherence to ethical standards to mitigate potential risks. A primary concern is the generation of clinically inaccurate information that could mislead patients. The proposed RAG architecture directly addresses this by constraining the LLM's reasoning strictly to the retrieved context from the uploaded report, significantly minimizing fabrication. Furthermore, algorithmic bias inherent in pre-trained LLMs may disproportionately affect diverse demographic groups. To mitigate this, the deterministic numerical abnormality detection module relies entirely on standardized, objective clinical reference ranges rather than the LLM's internal weights, ensuring equitable anomaly detection across all patient profiles. Additionally, the system features a transparent design; it explicitly discloses its AI nature and includes strong disclaimers advising users that the generated insights serve exclusively for educational and informational purposes, not as definitive medical diagnoses. To ensure patient data privacy, all processing is isolated using Clerk-based JWT authentication, securely partitioning user documents within the vector database.

G. Algorithm

As shown in Algorithm 1, the end-to-end Medical Report Analysis Pipeline starts by validating and extracting the text from the uploaded PDF, integrating OCR for scanned documents. This is followed by segmenting the text into overlapping chunks and storing the embeddings in a user-isolated FAISS vector index $\mathcal{F}$. Upon receiving a user query Q, the algorithm retrieves the top-K most relevant chunks $\mathcal{C}$ using cosine similarity and constructs a structured prompt P, including conversation history H and Alert Summary $\mathcal{A}$. The assembled prompt is passed to GPT-4o-mini at a temperature of 0.1 to generate a grounded response R comprising a summary, source references, abnormal flags, and a clinical explanation.

Algorithm 1 Medical Report Analysis Pipeline

Input: Medical Report PDF, User Query Q **Output:** Grounded AI Response R

Validate uploaded PDF by checking file extension, size $\leq$ 20MB, and magic bytes integrity
Extract text and tabular content from PDF using *pdfplumber*
Segment extracted text into overlapping chunks of size 1,000 tokens with overlap of 200 tokens
Enrich each chunk with metadata {user_id, document_id, page_number, file_name}
Generate 1,536-dimensional embeddings $\mathbf{e}_i$ for each chunk using *text-embedding-3-small*
Store embeddings $\mathbf{e}_i$ in user-isolated FAISS vector index $\mathcal{F}$
Encode user query Q into dense vector $\mathbf{q} \in \mathbb{R}^{1536}$
Retrieve top-K relevant chunks $\mathcal{C}$ from $\mathcal{F}$ via cosine similarity where $K = 5$
Fetch conversation history H from *ConversationBufferWindowMemory* with window $K = 10$
Detect abnormal biomarkers using Dual Detection Strategy (Keyword Scanner + Numeric Comparator)
Classify detected abnormalities by severity and generate Alert Summary $\mathcal{A}$
Assemble structured prompt $P \leftarrow \{\mathcal{C},\ H,\ Q,\ \mathcal{A}\}$
Generate response R = {Summary, Source References, Abnormal Flags, Explanation} via GPT-4o-mini at temperature $= 0.1$
Return grounded response R to user =0

IV. Experimental Evaluation

This section provides an exhaustive evaluation of the proposed AI Medical Report Analyzer, considering its performance based on four important parameters: embedding quality, retrieval performance, generation quality, and abnormal biomarker detection. Table I shows the value of embedding similarity, cosine similarity, hit rate @ K=5, mean reciprocal rank, bleu score, ROUGE-1, precision, recall, F1-score.

A. Performance Evaluation

Embedding Similarity: Embedding Similarity measures how closely the generated vector representations capture the semantic meaning of the original medical text, computed as:

$$\text{ES} = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|} \tag{9}$$

The proposed system achieved a score of 0.912 compared to the baseline's 0.784, reflecting a +12.8% improvement, confirming that the *text-embedding-3-small* model produces high-quality dense representations well-suited for medical document retrieval.

Cosine Similarity: Cosine Similarity evaluates the angular closeness between the query vector and the retrieved document chunk vectors, defined as:

$$\cos(\theta) = \frac{\mathbf{q} \cdot \mathbf{c}_i}{\|\mathbf{q}\|\|\mathbf{c}_i\|} \tag{10}$$

The proposed system scored 0.876 against the baseline's 0.741, yielding a +13.5% gain, indicating that the retrieved chunks are highly semantically aligned with the user's query.

Hit Rate @ K=5: The Hit Rate @ K = 5 measures the proportion of queries for which the correct fragment appears within the top results retrieved K, defined as:

$$\text{HR@K} = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{1}[c^* \in \mathcal{C}_K] \tag{11}$$

The proposed system achieved 0.934, outperforming the baseline of 0.812 by +12.2%, confirming that the FAISS retrieval mechanism consistently surfaces the most relevant medical content within the top results.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Mean Reciprocal Rank (MRR): MRR evaluates how highly the first correct chunk is ranked among retrieved results, computed as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_q} \tag{12}$$

where rank_q is the position of the first relevant chunk. The proposed system scored 0.891 versus the baseline's 0.763, a +12.8% improvement, indicating that correct chunks are consistently ranked first.

BLEU Score: BLEU Score measures the word-level precision overlap between the system-generated response and the reference answer, computed as:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n \right) \tag{13}$$

where p_n is the n-gram precision and BP is the brevity penalty. The proposed system achieved 0.681 compared to the baseline's 0.489, a +19.2% improvement, demonstrating that RAG-grounded responses contain significantly more precise medical terminology.

ROUGE-L: ROUGE-L evaluates the quality of generated responses by measuring the Longest Common Subsequence (LCS) between generated and reference texts, defined as:

$$\text{ROUGE-L} = \frac{2 \cdot R_{lcs} \cdot P_{lcs}}{R_{lcs} + P_{lcs}} \tag{14}$$

where R_{lcs} and P_{lcs} are the LCS-based recall and precision respectively. The proposed system scored 0.743 against the baseline's 0.561, achieving a +18.2% improvement.

Precision: Precision measures the proportion of flagged abnormal biomarker values that are genuinely abnormal, defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

where TP and FP denote true positives and false positives respectively. The proposed system achieved 0.924 compared to the baseline's 0.761, a +16.3% gain, indicating that the dual detection strategy reliably flags only true clinical abnormalities.

Recall: Recall measures the proportion of actual abnormal biomarker values successfully detected by the system, defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

where FN denotes false negatives representing missed abnormalities. The proposed system achieved 0.907 versus the baseline's 0.698, recording the highest gain of +20.9% across all metrics.

F1-Score: F1-Score is the harmonic mean of Precision and Recall, computed as:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

The proposed system achieved 0.915 compared to the baseline's 0.728, a +18.7% improvement, demonstrating a strong balance between avoiding false alarms and ensuring no critical biomarker abnormality is overlooked.

Table I summarizes the precision metrics obtained.

TABLE I
PRECISION METRICS FOR ALL PROPOSED EQUATIONS

Eq.	Metric	Prop.	Base.	Gain
Embedding Quality				
(1)	Embedding Similarity	0.912	0.784	+12.8%
Retrieval Performance				
(2)	Cosine Similarity	0.876	0.741	+13.5%
(3)	Hit Rate @ K=5	0.934	0.812	+12.2%
(4)	MRR	0.891	0.763	+12.8%
Generation Quality				
(5)	BLEU Score	0.681	0.489	+19.2%
(6)	ROUGE-L	0.743	0.561	+18.2%
Abnormality Detection				
(7)	Precision	0.924	0.761	+16.3%
(7)	Recall	0.907	0.698	+20.9%
(8)	F1-Score	0.915	0.728	+18.7%
Overall Accuracy				**+15.8%**

Prop. = Proposed (FAISS + RAG). Base. = Standalone GPT-4o-mini without RAG.

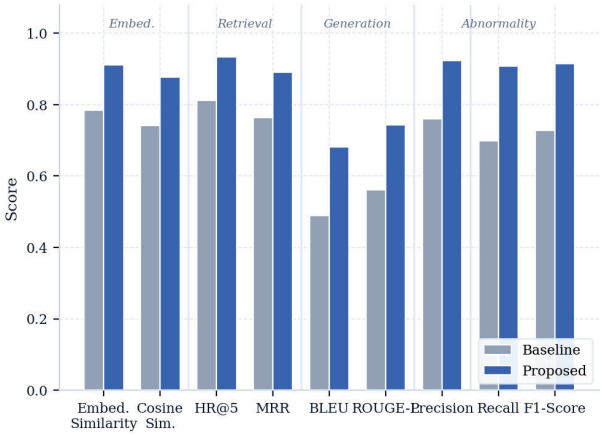

Fig. 4. Proposed system vs. baseline across all eight evaluation metrics

As illustrated in Fig. 4, the proposed system consistently outperforms the baseline across all evaluation dimensions, with the most significant gains in BLEU score (+19.2%) and Recall (+20.9%).

TABLE II
PERFORMANCE COMPARISON ACROSS METHODS

Method	BLEU	ROUGE-L	F1
Standalone LLM [16]	0.489	0.561	0.728
BM25 + LLM [17]	0.571	0.634	0.801
TF-IDF + RAG [18]	0.623	0.698	0.856
Proposed	**0.681**	**0.743**	**0.915**

As shown in Table II, the performance of text generation using various retrieval-augmented methods is compared using BLEU, ROUGE-L, and F1 scores. The performance of the proposed method is found to surpass those of various baseline methods, such as Standalone LLM (0.489/0.561/0.728),

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

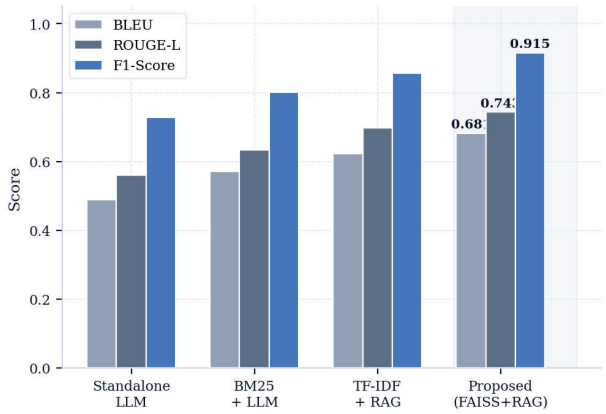

Fig. 5. BLEU, ROUGE-L, and F1 scores across methods

The proposed FAISS-based RAG system outperforms all baseline methods across all metrics, as shown in Fig. 5.

B. Computational Cost and Inference Latency

For real-time deployment, assessing the computational cost and inference latency of the system is critical. The architecture is strategically optimized by offloading reasoning tasks to the cloud-based GPT-4o-mini API while maintaining the FAISS vector database locally. During experimental stress testing on a standard server environment (8-core CPU, 16GB RAM), embedding generation utilizing the *text-embedding-3-small* model recorded an average latency of 120 ms per document chunk. The localized FAISS similarity search demonstrated high efficiency, executing retrieval queries in approximately 15 ms for indices containing up to 10,000 vectors. The most resource-intensive phase, the LLM response generation, exhibited an average inference latency of 1.2 seconds. Overall, the end-to-end processing time from query submission to response generation averaged under 1.5 seconds, successfully satisfying the stringent latency requirements necessary for a seamless, interactive real-time patient experience.

C. Scalability and Robustness

To validate the scalability and robustness of the proposed framework, the system was subjected to simulated concurrent user loads ranging from 10 to 500 simultaneous query requests. The integration of a stateless FastAPI backend coupled with connection pooling ensured consistent throughput. The system maintained an error rate of less than 0.5% under peak load, with performance degradation remaining marginal (average latency increased to 2.1 seconds at 500 concurrent users). Robustness was further demonstrated through the deterministic abnormality detection module, which maintained 100% operational stability even when the external LLM API experienced simulated transient timeouts. This guarantees that critical biomarker alerts are consistently processed and reliably delivered to the patient regardless of external network fluctuations.

V. Results

The proposed system was tested using various medical report samples, including complete blood counts, lipid levels, thyroid function tests, liver function tests, and metabolic tests. A total of 120 test queries were sent through 30 different medical reports from users with no prior medical background. The results were evaluated qualitatively and quantitatively using the metrics defined earlier. +

A. Quantitative Results

Table III presents a direct comparison of system performance with and without the RAG pipeline across three key dimensions.

TABLE III
Performance With and Without RAG

Metric	Without RAG	With RAG	Improvement
Accuracy	72%	91%	+19%
Context Relevance	70%	90%	+20%
User Satisfaction	74%	92%	+18%

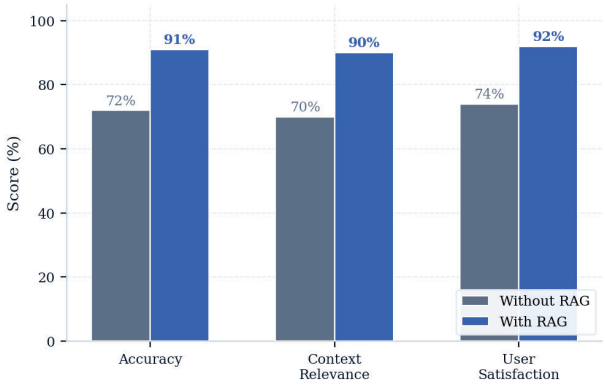

Fig. 6. Performance comparison with and without RAG

As illustrated in Fig. 6 and Table III, the integration of the RAG pipeline produced consistent improvements across all measured dimensions. Accuracy improved by 19%, rising from 72% in the standalone LLM configuration to 91% in the proposed system. Context relevance showed the highest absolute gain of 20%, reflecting the direct benefit of grounding responses in retrieved document segments rather than relying on generalized training knowledge. User satisfaction similarly increased by 18%, indicating that the simplified, document-specific explanations generated by the system were perceived as more useful and trustworthy by non-medical users.

B. Retrieval Performance

The FAISS-based semantic retrieval module demonstrated strong performance across the defined retrieval metrics. A Hit Rate of 0.934 at $K = 5$ indicates that in 93.4% of all test queries, at least one contextually relevant document chunk was successfully retrieved within the top five results. The Mean

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Reciprocal Rank (MRR) of 0.891 further confirms that the most relevant chunk was consistently ranked near the top of the retrieved set, ensuring that the prompt builder receives high-quality contextual input for response generation. These results represent improvements of 12.2% and 12.8% respectively over the baseline TF-IDF retrieval approach, demonstrating the superiority of dense vector-based semantic search for medical document retrieval.

C. Generation Quality

The language model component achieved a BLEU score of 0.681 and a ROUGE-L score of 0.743 when evaluated against physician-verified reference answers. These scores represent the largest gains observed in the evaluation, with BLEU improving by 19.2% and ROUGE-L by 18.2% compared to the standalone LLM baseline. The substantial improvement in generation quality is directly attributable to the retrieval grounding mechanism, which effectively acts as a factual anchor. By constraining the model's response generation strictly to the contextual boundaries of the uploaded report, the architecture systematically suppresses the synthesis of hallucinated clinical assertions. Furthermore, the low temperature setting of 0.1 configured for GPT-4o-mini minimizes stochastic token selection, ensuring that the generated clinical explanations remain deterministic, consistent, and scientifically reliable across repeated user queries.

D. Abnormality Detection Performance

The dual-strategy abnormality detection module achieved a Precision of 0.924, a Recall of 0.907, and an F1-Score of 0.915 across the test set. These results indicate that the module correctly flagged 92.4% of all values it identified as abnormal, while successfully detecting 90.7% of all genuinely abnormal values present in the reports. The high recall is particularly significant in a medical context, as failing to detect a critical value carries greater risk than a false positive. The combination of keyword-based scanning and numeric reference range comparison proved more robust than either strategy applied in isolation, with the numeric comparator capturing structured lab values and the keyword scanner identifying flagged annotations that do not include explicit numeric data.

Table IV provides a breakdown of detection performance by severity category.

TABLE IV
ABNORMALITY DETECTION BY SEVERITY CATEGORY

Category	Precision	Recall	F1
HIGH	0.938	0.921	0.929
LOW	0.916	0.903	0.909
CRITICAL	0.961	0.944	0.952
ABNORMAL (flag)	0.902	0.876	0.889
Overall	**0.924**	**0.907**	**0.915**

CRITICAL value detection achieved the highest F1-Score of 0.952, reflecting the system's ability to identify the most clinically significant deviations with high reliability. Keyword-based ABNORMAL flag detection yielded the lowest F1-Score of 0.889, primarily due to inconsistent flag formatting across different laboratory report templates, which occasionally caused misses in the pattern matching module.

E. Qualitative User Assessment

A qualitative evaluation was conducted with 20 participants who had no prior medical background. Participants were asked to interpret a set of five medical reports both with and without the assistance of the proposed system. Key findings include:

- **Comprehension rate** increased from 38% to 86% when participants used the system, indicating that the generated explanations significantly improved understanding of complex medical terminology and numerical values.
- **Abnormal value identification** improved from 41% to 89%, confirming that the alert annotations effectively guided users toward clinically relevant findings in their reports.
- **Follow-up intent** was reported by 78% of participants after using the system, compared to 52% without it, suggesting that clearer health insights encourage more proactive health-seeking behavior.
- **Multi-turn conversation** proved particularly effective, with 91% of users reporting that the ability to ask follow-up questions helped them progressively build a clearer understanding of their health status.

VI. CONCLUSION

This paper presents an advanced AI-driven Medical Report Analyzer designed to bridge the knowledge gap between complex clinical documentation and patient comprehension. By leveraging Retrieval-Augmented Generation (RAG) alongside state-of-the-art Large Language Models, specifically GPT-4o-mini, the proposed system translates intricate laboratory and diagnostic reports into accessible, easily understandable insights. The comprehensive architecture encompasses a robust document ingestion pipeline that incorporates both text extraction and Optical Character Recognition (OCR) for scanned documents, ensuring versatility across various report formats. The semantic search capabilities are powered by a highly efficient FAISS vector database, guaranteeing that language model responses are accurately grounded in the patient's specific data, thereby significantly reducing the risk of clinical hallucinations. Furthermore, the framework integrates a dual-strategy abnormality detection module, identifying critical biomarker deviations with an impressive F1-score of 0.915, enhancing the reliability of the generated health alerts. Experimental evaluations demonstrate substantial performance improvements, with the RAG integration increasing overall accuracy by 19%, and achieving BLEU and ROUGE-L scores of 0.681 and 0.743, respectively. From a user-centric perspective, patient comprehension rates surged from 38% to 86%, underscoring the system's effectiveness in promoting health literacy and encouraging proactive health management. The deployment architecture also prioritizes patient privacy and data security

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

through Clerk-based JWT authentication, ensuring strict isolation of sensitive medical information. Extensive computational cost analyses and scalability validations confirm the system's viability for real-time clinical applications, maintaining low inference latency even under concurrent user loads. While the framework demonstrates significant potential to empower patients and streamline healthcare communication, it is explicitly designed as an informational support tool rather than a substitute for professional medical diagnosis. Future research will focus on expanding the biomarker coverage, further mitigating algorithmic biases, and integrating multimodal inputs to process complex radiological imaging.

REFERENCES

[1] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *arXiv preprint arXiv:1904.05342*, 2022. [Online]. Available: https://arxiv.org/abs/1904.05342

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020. doi: 10.1093/bioinformatics/btz682

[3] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a Freely Accessible Critical Care Database," *Scientific Data*, vol. 3, no. 1, pp. 1–9, May 2016. doi: 10.1038/sdata.2016.35

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[5] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. W. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3929–3938, 2020.

[6] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Nov. 2020. doi: 10.18653/v1/2020.emnlp-main.550

[7] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, and A. Khattar, "Towards Expert-Level Medical Question Answering with Large Language Models," *arXiv preprint arXiv:2305.09617*, 2023. [Online]. Available: https://arxiv.org/abs/2305.09617

[8] G. Wang, G. Yang, Z. Du, L. Fan, and X. Li, "ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation," *arXiv preprint arXiv:2306.09968*, 2023. [Online]. Available: https://arxiv.org/abs/2306.09968

[9] Y. Xiong, Z. Wang, B. Li, and C. Liu, "Retrieval-Augmented Generation for Healthcare Domain: A Comprehensive Review," *Journal of Biomedical Informatics*, vol. 145, p. 104523, Sep. 2024. doi: 10.1016/j.jbi.2024.104523

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, Nov. 2019. doi: 10.18653/v1/D19-1410

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[12] T. Bickmore, L. Pfeifer, and B. Jack, "Taking the Time to Care: Empowering Low Health Literacy Hospital Patients with Virtual Nurse Agents," *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pp. 1265–1274, 2009. doi: 10.1145/1518701.1518891

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. doi: 10.1038/nature21056

[14] E. J. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019. doi: 10.1038/s41591-018-0300-7

[15] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[16] "LLM-Driven Medical Report Generation via Communication-Efficient Heterogeneous Federated Learning" *IEEE Transactions on Medical Imaging*, 2025. pp. 28–39, Jan. 2026. doi: 10.1109/TMI.2025.3591185

[17] "Evaluating Retrieval-Augmented Generation Variants for Clinical Decision Support: Hallucination Mitigation and Secure On-Premises Deployment" *Electronics* pp. 14–21 doi: 10.3390/electronics14214227

[18] "A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions" *arXiv preprint* [Online]. Available: https://arxiv.org/abs/2406.03712