MedAId: A Multilingual, Multimodal, and Retrieval-Augmented Framework for Medical Query Resolution

Jay Sawant[1], **Mayukh Sharma**[1], **Maria Tavares**[1]
[1] **University of California San Diego, CA**

Abstract

Health inequities and barriers, such as language, literacy, and accessibility, remain major challenges in healthcare systems, contributing to disparities in health outcomes. MedAId, a multilingual and multimodal Retrieval-Augmented Generation (RAG) system, is designed to address these issues by integrating speech, text, and medical images to provide accurate, contextually relevant medical responses. By leveraging advanced datasets like MIMIC-IV, and BioMedBERT embeddings, MedAId enhances the accessibility of medical information for diverse populations. This project explores the potential of combining cutting-edge AI models, medical knowledge bases, and user-centered design to create a tool that empowers patients, fostering more equitable healthcare delivery and improved decision-making.

Introduction

Barriers that lead to health disparities and inequity are pressing issues that impact patients and the healthcare ecosystem as a whole. Many of these barriers stem from patients either not trusting or not fully understanding the information provided by their healthcare providers, which can undermine effective communication and shared decision-making. Research on health disparities has identified several critical obstacles to equitable care, with language barriers emerging as one of the most significant.[1] Language gaps in healthcare can result in miscommunication between medical professionals and patients, reduce patient satisfaction, and compromise the quality of care.[2] These issues often lead to unequal access to healthcare and disparate health outcomes, particularly for linguistically diverse populations. MedAId, aims to bridge these gaps through a multilingual and multi-modal Medical Question-Answering (QA) system. Unlike traditional QA systems, which are typically text-based, MedAId incorporates speech-to-text functionality, multiple language capabilities, and the ability to process medical images such as X-rays and MRI scans. By leveraging advanced Retrieval-Augmented Generation (RAG) techniques and embedding models, MedAId provides personalized, accurate medical information to patients, ensuring inclusivity and accessibility across diverse user groups.

Background and Related Work

Healthcare delivery often requires an integrated understanding of patient history, medical imaging, and clinical documentation. Despite ongoing efforts to address health inequities and achieve optimal healthcare delivery, the tools currently available often fail to meet the complex needs of diverse populations. Although Question and Answer(QA) models offer great potential in bridging these gaps, most existing QA systems rely solely on text inputs, limiting their effectiveness in real-world medical scenarios. Recent mandates, such as the 2014 US Department of Health and Human Services regulation granting patients direct access to their medical records and laboratory results, have sought to empower patients, but access alone is not sufficient. Patients often struggle to interpret the information, particularly when it involves complex data such as imaging or laboratory test results. Without proper contextualization, these tools may not bridge the understanding gap, leaving patients unable to participate effectively in shared decision-making with their providers.

Studies have found a significant inverse relationship between health literacy and numeracy and the ability to make sense of imaging and laboratory test results.[3] Patients with limited health literacy are more likely to misinterpret or misunderstand their results (either overestimating or underestimating their results), which in turn may delay them seeking critical medical attention. These challenges underscore the importance of providing not just access but also meaningful contextualization of medical data, enabling patients to act upon it in an informed and timely manner. The integration of imaging analysis into medical QA systems is particularly promising, as it allows patients and healthcare providers to benefit from a deeper understanding of diagnostic findings. By combining medical images with textual

and patient history contextual data, such systems can provide comprehensive and tailored responses that improve both comprehension and clinical outcomes.

The multilingual aspect of healthcare communication further complicates this landscape. Language barriers are a well-documented impediment to effective healthcare delivery, often resulting in miscommunication, reduced patient satisfaction, and diminished health outcomes.[2] Addressing this issue requires tools that can seamlessly translate and interpret queries in multiple languages, ensuring that linguistically diverse populations can access accurate and relevant medical information. Multilingual capabilities not only improve accessibility for patients, but also improve the efficiency of providers, allowing them to provide equitable care regardless of language barriers.

Recent studies evaluating the performance of current LLMs in answering medical questions highlight their potential, but also reveal significant limitations.[4] For example, it is often unclear what sources these models use to generate their answers, which raises concerns about transparency and reliability. Additionally, LLM responses are typically not tailored to individual patient contexts, and they often fail to ask clarifying questions when crucial details are missing from the input. We aimed to enhance the quality and applicability of our model by training it with robust medical datasets, such as MIMIC-IV, that offers a comprehensive repository of clinical notes and radiology reports. By leveraging these resources, we created a system that delivers accurate, transparent, and contextually relevant medical responses.

MedAId aims to improve the functionality and reliability of medical QA systems by training the model on high-quality, domain-specific datasets while incorporating advanced capabilities such as speech-to-text processing, multilingual support, and medical image analysis. These enhancements ensure that the system can generate individualized, accurate, and contextually relevant responses for diverse users, paving the way for a more inclusive and effective healthcare support platform.

Datasets

This study leverages two distinct datasets, labeled examples from PubMedQA[5] and MIMIC-IV-Note,[6] to develop and evaluate our RAG-based medical Q&A system. The datasets play complementary roles in embedding evaluation, knowledge base construction, and contextual retrieval. Below, we provide an overview of these datasets and the pre-processing techniques applied.

1. PubMedQA
The PubMedQA (labeled) dataset[5] contains 1,000 expert-annotated examples designed for medical question-answering tasks. Each example includes a natural language query, a related context derived from biomedical literature, and an expert-provided answer. This dataset serves a specific role in our study: evaluating the performance of various embedding models for the RAG-based system. By benchmarking embeddings against these examples, we ensure the retrieval mechanism identifies and retrieves the most contextually relevant documents, thereby enhancing the accuracy and reliability of downstream question answering tasks.

2. MIMIC-Note-IV
The MIMIC-IV-Note dataset,[6] developed by Johnson et al. (2021), provides a comprehensive collection of de-identified clinical notes, including radiology reports and discharge summaries. This dataset serves as the foundation for constructing the knowledge base in our RAG system. Specifically, it includes approximately 2.3 million radiology reports from 237,427 patients and 331,794 discharge summaries from 145,915 patients. Due to the large size and length of these notes, pre-processing was necessary to align with computational constraints. Radiology reports, averaging 1,000 characters each, were truncated to the first 700 characters for embedding generation, while retaining the full text for retrieval during inference. For discharge summaries, which average 10,000 characters per note, we used GPT-4o-mini (OpenAI, 2024) to create concise summaries of approximately 700 characters. Resource limitations led us to sample 50,000 radiology reports and re-summarize 3,000 discharge summaries. These processed records were embedded using BiomedBERT[7] (Gu et al., 2021) and combined to form a compact yet clinically rich knowledge base for the retrieval system.

Methods

System Architecture
Our system, **MedAid**, is designed to handle multimodal user input through an intuitive user interface. The workflow begins with preprocessing the input using Google's Cloud APIs.

1. Input Processing

- **Speech**: Converted into text using the Translate API.
- **Images**: Processed via the Vertex AI API to generate a detailed textual summary, which is integrated with the input pipeline.
- **Text**: Directly enters the pipeline.

2. RAG Pipeline

- The processed text is sent to the **Retrieval-Augmented Generation (RAG)** pipeline, which combines **BioMed BERT** and a **FAISS datastore**.
- BioMed BERT generates embeddings for the input text, which are compared against the FAISS datastore to retrieve the top 5 most relevant contexts.

3. Response Generation

- The extracted contexts, along with the original query and conversation history, are provided to the **Gemini model** for generating a response.
- The response is returned to the user via the interface.

This architecture ensures an efficient and accurate system for answering user queries by leveraging Google's advanced AI tools and custom RAG pipelines.

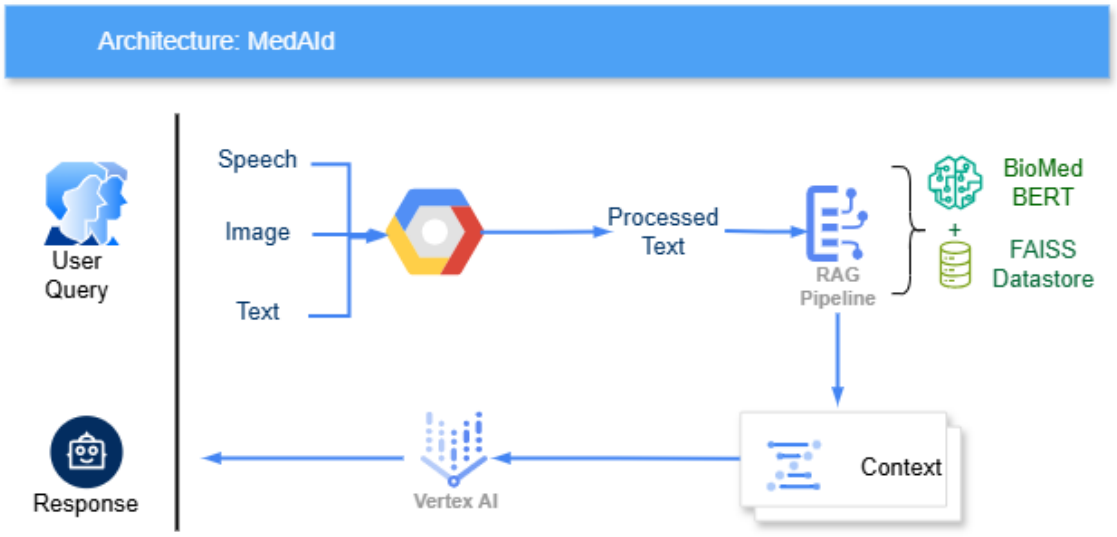

Figure 1: System architecture for MedAId application.

Machine Translation Pipeline
A key component of our pipeline is machine translation, which enables the system to process text or speech in any supported language. Using the Translate API, we convert preprocessed text from the source language into English. This step is crucial since most medical datasets used in RAG pipelines are primarily available in English.

After generating an optimal response through the RAG pipeline and Gemini API, the response is translated back into the source language. This ensures that language barriers are effectively eliminated, allowing users from diverse linguistic and ethnic backgrounds to interact seamlessly with our application.

Vector Embedding and Context Retrieval
To facilitate efficient context retrieval, we leverage the BiomedBERT model, a domain-specific transformer with 110 million parameters. This model processes input text with a context window of 512 tokens, generating 768-dimensional embeddings from its last layer. These embeddings encapsulate the semantic meaning of clinical documents and queries, enabling precise similarity-based retrieval.

We precompute embeddings for a knowledge base comprising 3,000 resummarized discharge summaries and 50,000 radiology reports from the MIMIC-IV-Note datset. For efficient retrieval, we utilize LangChain in conjunction with the FAISS[8] vector store, which indexes the embeddings. Retrieval is based on the L2 distance similarity metric, ensuring the selection of the most relevant contexts for each query. This process ensures that our RAG pipeline integrates the most pertinent clinical knowledge into downstream generative tasks.

Retrieval Augmented Generation
Our Retrieval-Augmented Generation (RAG)[9] pipeline leverages a combination of clinical data and contextual retrieval to enhance query resolution. The pipeline begins by analyzing input data, such as medical images and patient medical histories, using a Gemini 1.5 Pro large multimodal model to generate preliminary findings. The user query is then combined with these findings to construct a query for context search, which retrieves the top five most relevant contexts from a curated knowledge base comprising discharge summaries and radiology reports. These contexts are further combined with the user query and conversation history to form the final query, ensuring rich contextual relevance. The final query is processed by the Gemini 1.5 Pro model, producing the output that integrates both retrieved knowledge and generative reasoning to address the user query effectively.

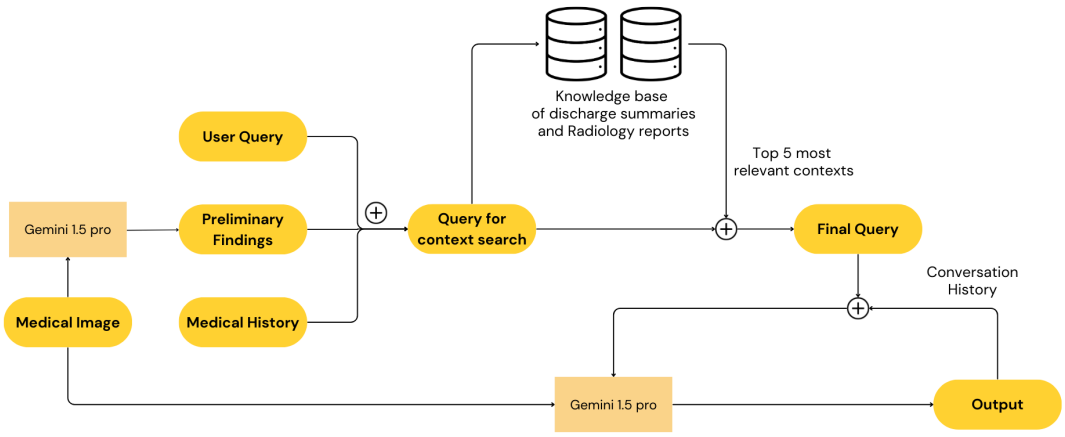

Figure 2: Overview of the Retrieval-Augmented Generation (RAG) pipeline integrating medical imaging, patient history, and knowledge base retrieval for contextualized clinical query resolution

Implementation Details

Prompt Structure
The generative model is guided using a structured prompt template that combines various components of the available information. The template is shown below:

```
Based on the following information, provide a clear and accurate response
to the last question in the conversation.
    {preliminary_findings}
    {medical_history_section}
    {contexts_section}
Conversation History:
    {conversation}
```

Here:

- {preliminary_findings}: Represents the outputs generated from the analysis of medical images and patient history.
- {medical_history_section}: Incorporates the relevant details from the patient's medical history.
- {contexts_section}: Includes the top retrieved contexts from the knowledge base (discharge summaries and radiology reports).
- {conversation}: Captures the ongoing conversational history for continuity.

This structured approach ensures that the model synthesizes information effectively, maintaining consistency and accuracy in the generated responses.

User Interface

Our interface is designed to handle both multilingual and multimodal inputs, allowing users to interact with the system using speech, text, or images. For instance, Figure 3 illustrates the interface processing an image of a chest X-ray, extracting relevant features, and generating an initial diagnosis. Meanwhile, Figure 4 highlights the interface's ability to accept multilingual input, process it efficiently, and provide a response in the user's preferred language.

These capabilities are made possible by integrating Google Cloud APIs, which facilitate accurate text translation and sophisticated image understanding. This ensures that users from diverse linguistic backgrounds and varying input preferences can interact seamlessly with the system, making it both inclusive and versatile.

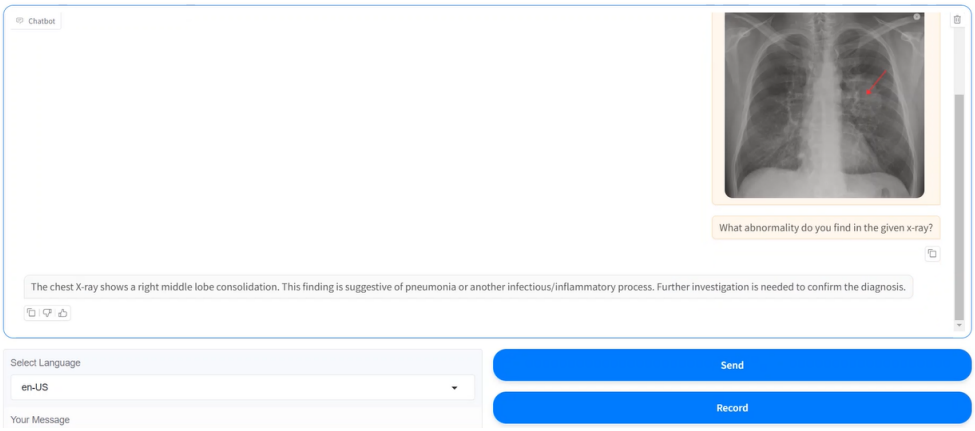

Figure 3: Interface with multimodal capability.

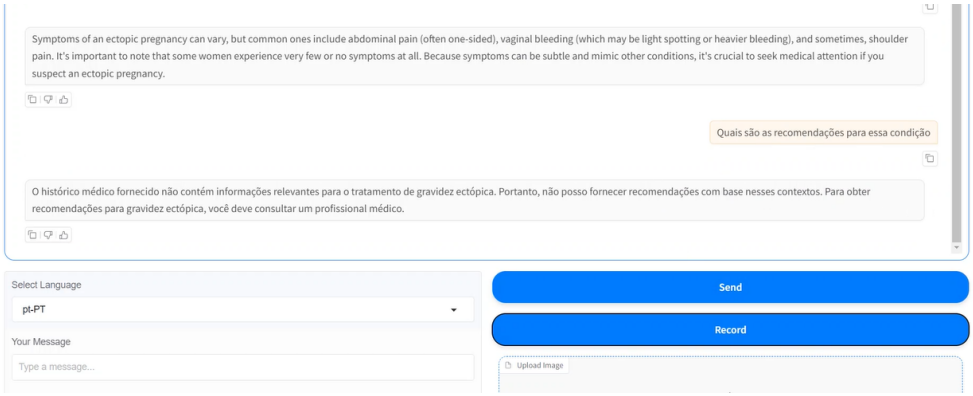

Figure 4: Interface with multilingual capability.

Experiments

Comparison of BiomedBERT and OpenAI embeddings

To evaluate the effectiveness of different embedding strategies for context retrieval, we conducted a comparative experiment using BiomedBERT embeddings and OpenAI embeddings. The experiment utilized the Gemini 1.5 Flash model for output generation, configured with a constant temperature of 0.1 and a fixed retrieval of the top 5 most relevant contexts. The only variable in the setup was the embedding model used for vectorizing the knowledge base and queries.

The PubMedQA dataset, which contains 1,000 expert-labeled query-answer pairs along with their relevant contexts, was employed for this evaluation. Each query was matched with retrieved contexts using embeddings generated either by BiomedBERT (768-dimensional vectors from its last layer) or OpenAI embeddings.

This experiment aimed to assess the impact of domain-specific embeddings (BiomedBERT) versus general-purpose embeddings (OpenAI) on retrieval quality and the resulting output accuracy for clinical queries. Results are discussed in the subsequent sections.

Evaluation and Results

ROUGE and BLEU Scores

The performance of the BiomedBERT and OpenAI embeddings was evaluated using ROUGE[10] and BLEU[11] scores to assess the quality of generated responses. ROUGE scores (**ROUGE-1**, **ROUGE-2**, **ROUGE-L**) measure the overlap between the generated output and reference answers, evaluating unigram, bigram, and longest common subsequence matches, respectively. BLEU scores assess n-gram precision, capturing how closely the generated text matches the reference text.

Across all metrics, OpenAI embeddings consistently outperformed BiomedBERT embeddings, suggesting a higher degree of overlap and contextual relevance in retrieved contexts when OpenAI embeddings were used. While BLEU scores were relatively low for both embeddings, OpenAI embeddings exhibited marginally better n-gram precision.

Metric Formulas

- **ROUGE-1 (Unigram Overlap):**

$$\text{ROUGE-1} = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in the reference}}$$

- **ROUGE-2 (Bigram Overlap):**

$$\text{ROUGE-2} = \frac{\text{Number of overlapping bigrams}}{\text{Total bigrams in the reference}}$$

- **ROUGE-L (Longest Common Subsequence):**

$$\text{ROUGE-L} = \frac{\text{LCS length}}{\text{Reference length}}$$

- **BLEU (Bilingual Evaluation Understudy):**

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

where:

  - p_n = $\frac{\text{Number of matching n-grams}}{\text{Total n-grams in the generated text}}$
  - w_n = Weight for n-gram precision (commonly $w_n = \frac{1}{N}$).
  - BP = $\exp\left(1 - \frac{\text{Reference Length}}{\text{Generated Length}}\right)$, if Generated Length $<$ Reference Length, otherwise 1.

Table 1: Comparison of ROUGE and BLEU scores for BiomedBERT and OpenAI embeddings using the PubMedQA dataset.

Metric	BiomedBERT Embeddings	OpenAI Embeddings
ROUGE-1	0.2758	**0.3213**
ROUGE-2	0.0774	**0.0939**
ROUGE-L	0.1848	**0.2061**
BLEU	0.0256	**0.0294**

Assessment Criteria and Scoring Framework for MedAId

To evaluate the performance of MedAId, English-speaking and Portuguese-speaking medical professionals independently graded its responses on a scale from 0 to 10 across four metrics: Medical Accuracy, Comprehensiveness, Clarity and Readability, and Ethics and Tone. These metrics were carefully selected to assess not only the technical reliability of the system but also its alignment with ethical considerations, especially given the importance of ensuring that the LLM operates as a support tool and not as a replacement of medical diagnosis. The Medical Accuracy metric focuses on the precision and correctness of the medical content provided. Comprehensiveness evaluates the system's ability to provide complete responses to all relevant aspects of the query. Meanwhile, Clarity and Readability measure how well the response communicates complex information in a user-friendly manner, and Ethics and Tone assess adherence to ethical guidelines, ensuring responses are supportive and culturally sensitive. Additionally, we added gradings by another LLM(GPT-4) and calculated the mean scores for each metric for both English and Portuguese responses.

For the response in English shown in Figure 4, the following average scores were recorded:

Medical Professional 1: 9

Medical Professional 2: 9.12

Medical Professional 3: 8.75

LLM: 9.25

For the response in Portuguese:

Medical Professional 1: 8

Medical Professional 2: 7.5

Medical Professional 3: 7.2

LLM: 8.5

These results indicate that MedAId excels in delivering accurate, clear, and ethically sound responses.

Discussion

This project showcases the transformative potential of integrating multimodal and multilingual inputs to address critical challenges in medical question-answering systems. By incorporating speech, text, and medical image processing, the system demonstrates a significant enhancement in usability and flexibility for patients from diverse backgrounds. This multimodal approach ensures that users receive tailored, context-rich responses that are not only relevant but also actionable, thus improving the overall healthcare experience. Additionally, the system's multilingual capability bridges a critical accessibility gap, enabling linguistically diverse populations to access high-quality healthcare information and services that were previously out of reach.

One of the key strengths of this system lies in its multimodal capabilities, which enable it to analyze and generate responses based on diverse inputs such as textual queries, spoken questions, and medical imaging. This adaptability allows the system to address a wide range of user needs, making it suitable for various clinical and patient-facing scenarios. Furthermore, by leveraging curated medical datasets, the system provides responses that are rich in context, offering users information that is both accurate and relevant. These strengths are particularly valuable in improving healthcare accessibility and outcomes, especially for non-English speaking populations and patients with limited health literacy.

However, the project also faced certain limitations. Budget and compute constraints restricted the ability to fully explore advanced models and large-scale datasets, which could have further enhanced the system's capabilities. Additionally, computational limitations impacted the scale and efficiency of the training and testing processes, potentially limiting the system's overall performance. Despite these constraints, the system demonstrates substantial potential, paving the way for future enhancements.

For future directions, we hope to incorporate advanced embeddings, such as OpenAI embeddings, that could significantly enhance the relevance and semantic understanding of responses. Additionally, integrating specialized vision-language models like LlavaMed[12] would enable the system to more accurately interpret and utilize medical imaging, further improving diagnostic support. Fully leveraging the MIMIC-IV dataset with extended contextual capabilities could allow the system to address a broader range of clinical scenarios with greater precision. Lastly, expanding the knowledge base by incorporating additional high-quality datasets would further strengthen the model's understanding of diverse medical queries. Despite limitations, the strengths of the system and the outlined future directions provide a clear pathway for transforming medical information delivery, ultimately improving patient outcomes, enhancing clinical efficiency, and addressing critical healthcare accessibility challenges.

Conclusion

MedAId marks a significant advancement in the field of medical QA systems, offering a robust, accessible, and contextually aware platform for patients and healthcare providers. By addressing linguistic and technical barriers, MedAId has the potential to transform healthcare delivery, enabling equitable access to accurate medical information and empowering users to make informed decisions. As the system continues to evolve, it promises to set a new standard for multimodal and multilingual AI applications in healthcare, fostering improved outcomes and bridging gaps in medical knowledge.

References

1. Quadri NS, Wilkins S, Krohn K, Mann EM, Stauffer WM, Walker PF. Language Justice: Addressing Linguistic Disparities Begins with Language Data Collection; 2023. Available from: https://doi.org/10.4269/ajtmh.23-0237.
2. Shamsi HA, Almutairi AG, Mashrafi SA, Kalbani TA. Implications of Language Barriers for Healthcare: A Systematic Review; 2020. Available from: https://doi.org/10.5001/omj.2020.40.
3. Zikmund-Fisher BJ, Exe NL, Witteman HO. Numeracy and Literacy Independently Predict Patients' Ability to Identify Out-of-Range Test Results; 2014. Available from: https://doi.org/10.2196/jmir.3241.
4. He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, et al.. Quality of Answers of Generative Large Language Models Versus Peer Users for Interpreting Laboratory Test Results for Lay Patients: Evaluation Study; 2024. Available from: https://doi.org/10.2196/56655.

5. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 2567-77.
6. Johnson A, Pollard TJ, Mark RG, Shen L, Moody BE. MIMIC-IV-Note: Deidentified free-text clinical notes; 2023. PhysioNet. Available from: https://doi.org/10.13026/1n74-ne17.
7. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al.. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing; 2020.
8. Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al.. The Faiss library; 2024. Available from: https://arxiv.org/abs/2401.08281.
9. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al.. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks; 2021. Available from: https://arxiv.org/abs/2005.11401.
10. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74-81. Available from: https://aclanthology.org/W04-1013.
11. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics; 2002. p. 311-8. Available from: https://aclanthology.org/P02-1040.
12. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:230600890. 2023.