Vinay Jukanti
June 13th, 2025

Large Language Model-Driven Retrieval-Augmented Generation Medical Report Analysis System

Introduction

Healthcare generates vast amounts of data and reports, most of which are in unstructured formats like medical reports. Despite the critical information they hold, such documents are challenging to parse and analyze efficiently. Patients often struggle to interpret their lab results, while healthcare providers face difficulties in tracking trends over time. With the rapid advancement in AI, there is an opportunity to transform this data into actionable insights.

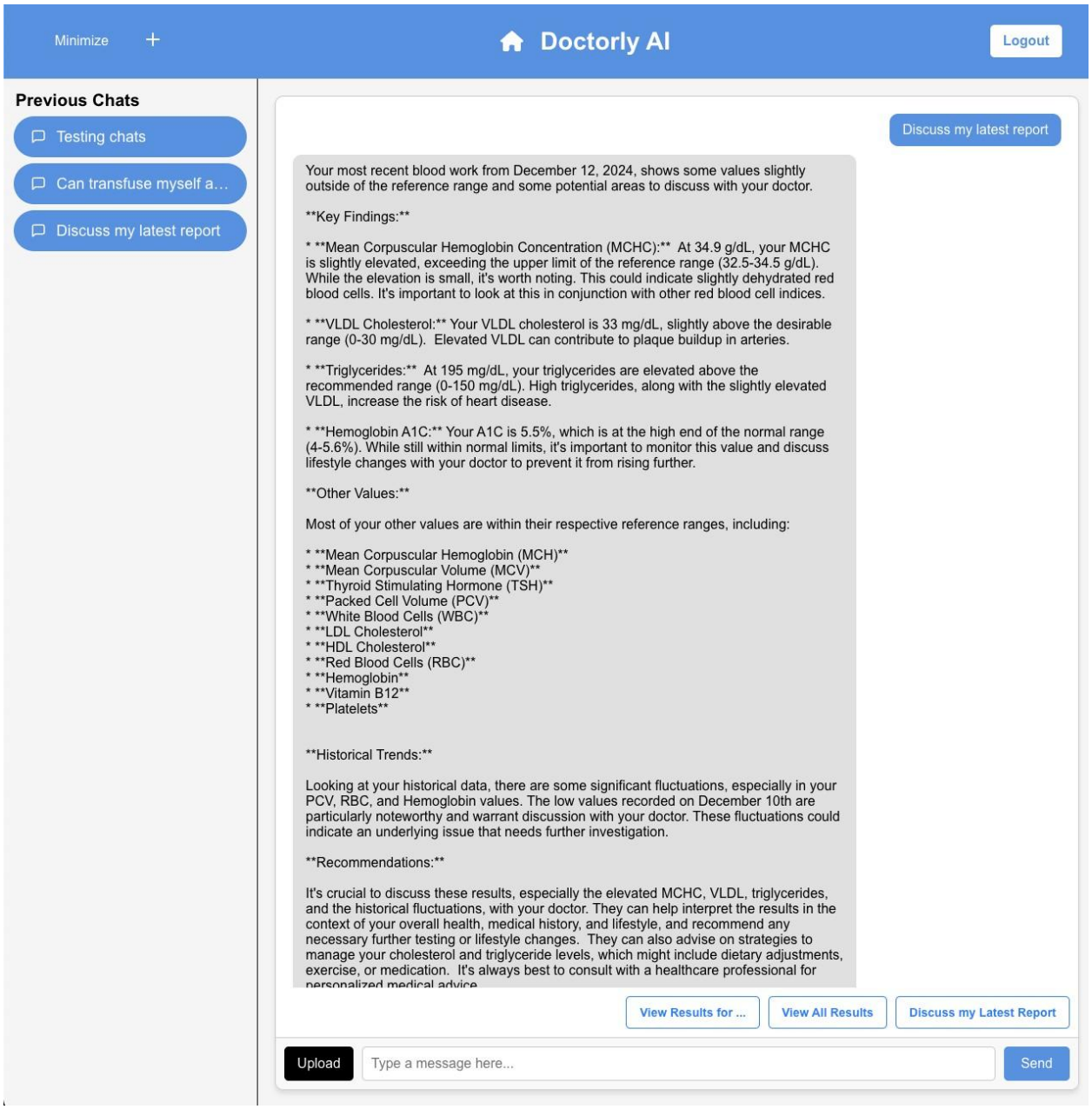
To address this challenge, I developed a Retrieval-Augmented Generation (RAG) Medical Report Analysis System. This system combines the power of a fine-tuned Large Language Model (LLM) and natural language processing (NLP) to extract, analyze, and visualize critical data from medical reports. The system enables users to upload medical reports, parse relevant biomarkers, store and track historical data, and leverage AI to get personalized insights and recommendations.

System Overview

The system, Doctorly AI, automates the analysis of medical reports end-to-end. Users upload their medical reports through an intuitive web interface, which processes these documents using natural language processing by extracting biomarker data, such as hemoglobin, white blood cell count, platelet count, etc, and validates the data against a predefined list of biomarkers reported within blood and lab reports. Thi extracted biomarker data is stored in a MongoDB database, allowing for efficient retrieval and analysis of historical trends, and the uploaded reports are securely stored in AWS S3. The system is structured to provide a secure report management workflow and a comprehensive view of a user’s health over time.

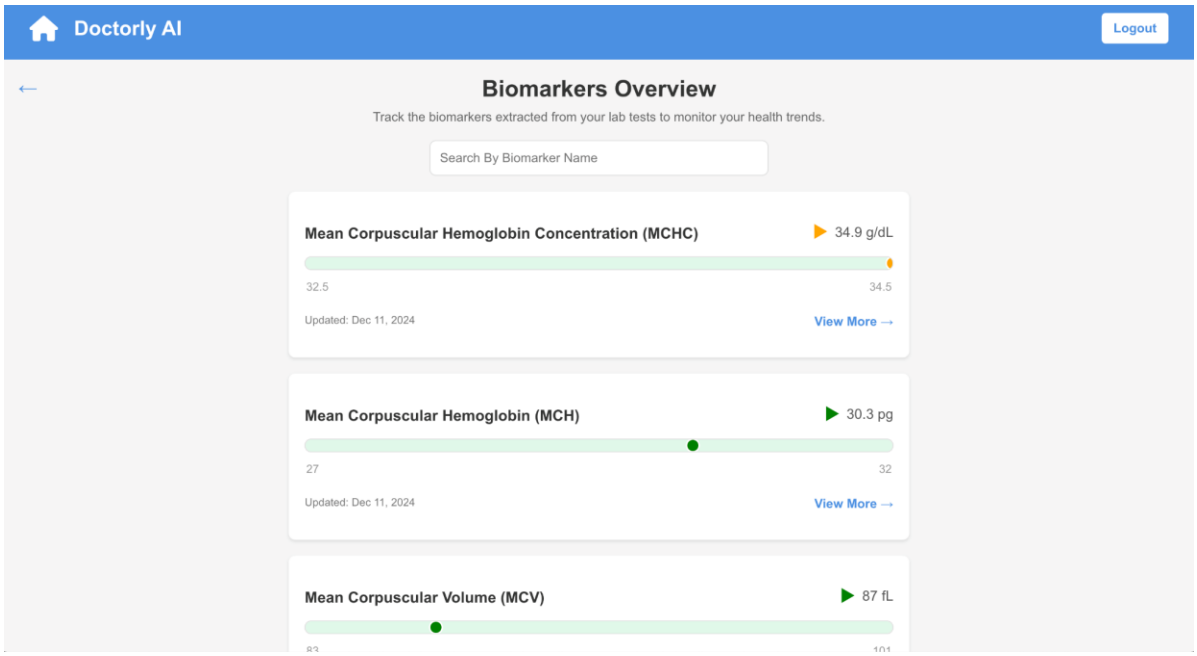
The platform also integrates a fine-tuned Meta LLaMA 8B model, which was trained on medical reports, doctor-patient conversations, and clinical datasets using Google Colab. After fine-tuning,

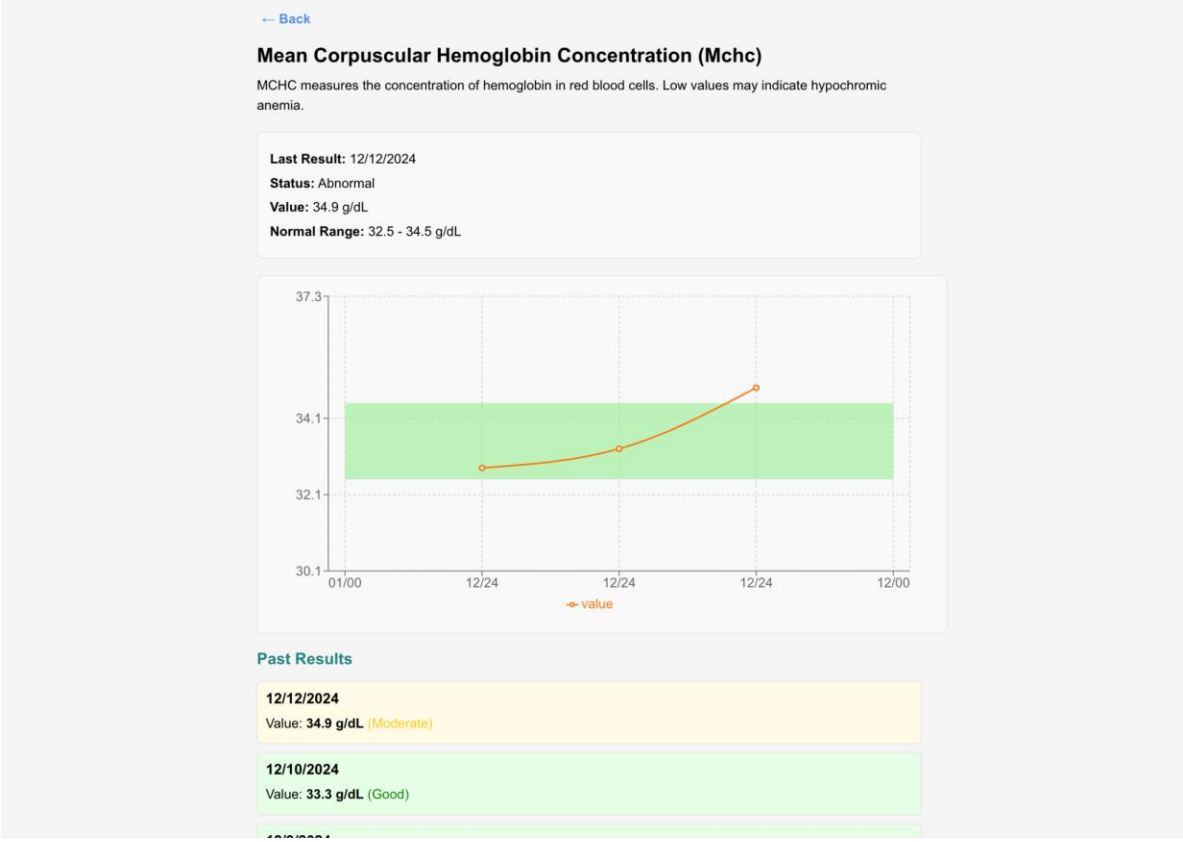
the model was quantized and deployed locally using Ollama to ensure efficient performance. This integration enables the system to combine historical data retrieval with generative AI capabilities, providing personalized health insights, recommendations, and alerts about potential risks. The chat functionality is displayed below:



Users can view their biomarker trends and insights through an interactive dashboard. Biomarker trends are visualized using Recharts, featuring dynamic line charts and color-coded indicators for quick interpretation. The dashboard allows users to explore their most recent test results, drill

into specific biomarkers, and monitor historical trends with ease. The images below display this functionality.





Vinay Jukanti
June 13th, 2025

Technologies Used

This project integrates several different technologies to achieve the overall objective. On the backend, Node.js and Express.js were utilized for building REST APIs, while MongoDB is used for storing extracted biomarker data. The uploaded reports are stored securely in AWS S3 and are available for user access at any time through the platform using single-use access links, generated with AWS. The text extraction and processing steps are accomplished using pdf-parse and custom created regular expressions to identify and structure biomarker data from PDF reports.

On the frontend, React.js provides a dynamic and responsive interface for the user to interact with, and the Recharts library helps power the visualizations, offering a clear representation of biomarker trends.

The system's AI components rely on Meta's LLaMA 8B model, which were fine-tuned on medical datasets to provide contextual and actionable health insights. The fine-tuning process was carried out using Google Colab, leveraging high-performance GPU resources to train the model on a curated dataset that included medical reports, doctor-patient conversations, and clinical datasets obtained from various sources.

This involved preprocessing the data, training the model over multiple epochs, and validating its accuracy to ensure it could handle domain-specific queries effectively. After fine-tuning, the model was quantized to optimize performance and deployed locally through Ollama, ensuring a balance between inference speed and model precision. There was a significant amount of time devoted to optimizing the model prompt in order to receive specific and curated responses to user questions. This architecture ensures that the AI outputs are accurate, context-aware, and reliable.

How Components Connect

Doctorly AI is designed to provide a seamless user experience by ensuring that all components work together efficiently. When users upload a medical report, the backend processes the document, extracting biomarker data using pdf-parse and custom parsing logic. The extracted

data is stored in MongoDB, organized by user ID and timestamp to enable easy retrieval and historical tracking.

When users access their health dashboard, the backend retrieves the most recent and historical biomarker data and feeds it into the fine-tuned LLaMA model. The AI interprets trends, generates personalized insights, and offers actionable recommendations based on the user's health data. The frontend displays these insights through interactive charts and detailed views, making it easy for users to understand their health metrics.

The system's integration of retrieval-augmented generation ensures that the AI combines factual historical data with generative capabilities, delivering insights that are both accurate and meaningful.

Key Outcomes

Doctorly AI fully automates the parsing and analysis of medical reports, significantly reducing manual effort and improving efficiency. Users receive actionable health recommendations tailored to their unique data, enabling them to make informed decisions about their health. The system tracks comprehensive historical data, allowing users to monitor long-term trends and identify patterns that may require attention.

Interactive visualizations make it easy to interpret complex health metrics, with color-coded indicators providing a quick assessment of whether values fall within normal ranges. The integration of the fine-tuned LLaMA model adds a layer of intelligence to the system, making it more insightful and user-friendly.

The system has the potential to be a transformative tool in healthcare, combining the latest advancements in AI with natural language processing and data visualization. By empowering users to understand their health data and providing personalized recommendations, this system bridges the gap between raw medical data and actionable insights. Future expansions could include integration with wearable devices, support for additional document formats, and

predictive modeling for long-term health outcomes. This project sets the stage for more intelligent and accessible healthcare solutions.