Topic - 7: QA platform

Dhruv Shrimali - 2021A7PS0008P , Nikhil Agarwal - 2020B2A71611P , Parivesh Bajpai - 2020B3A70752P

Abstract

In this report, we present the development and implementation of a web-based medical assistant application that utilizes advanced natural language processing (NLP) techniques to facilitate the extraction and interpretation of medical documents. The application, built using Flask, integrates a Large Language Model (LLM) fine-tuned for medical contexts, specifically the Medical-Llama3-8B model (Ruslanmv, 2023). By leveraging Hugging Face's transformers library and EasyOCR for text extraction, the system can process both image and PDF formats, extracting relevant information to answer user queries. The primary goal of this project is to provide a reliable and efficient tool for medical professionals and students to interact with medical documents and obtain precise, contextually relevant answers to their questions. The application features secure authentication, robust error handling, and a user-friendly interface, ensuring a seamless experience for end-users. This report details the underlying architecture of the application, the methods used for text extraction and processing, and the integration of the LLM for generating responses. Additionally, we discuss the challenges faced during development, including managing large model sizes and ensuring the accuracy of the extracted information, and the solutions implemented to address these issues. The results demonstrate the potential of using advanced IR models in enhancing the accessibility and utility of medical information, paving the way for future advancements in AI-assisted healthcare applications.

1 Motivation

Before diving into the details of our project, we'd like to share the personal stories that ignited our passion for developing this solution.

1.1 Dhruv:

A relative of mine recently faced a troubling health issue that brought to light the significant challenges in accessing and understanding medical information. They started experiencing a persistent and uncomfortable sensation in their left leg, described as a "buzzing sensation" that oscillated between their ankle and knee. This sensation particularly flared up after sitting for extended periods, often lasting for hours.

Concerned about these symptoms, they visited their healthcare provider, who conducted several consultations and diagnostic tests. The results were complex: the diagnosis included chronic denervation in the left S1 nerve root muscles and a mild fall out of motor units in the left gastrocnemius muscle. This technical medical language was overwhelming for all of us. We had to understand what it meant and what the implications were for their health and daily life. To manage the condition, the doctor prescribed new medications, including Gabapin 300, Mirator 0.5, and Myostaal liniment. Each medication came with detailed instructions and potential side effects, adding another layer of complexity. Additionally, the doctor noted that they had low vitamin levels, which required dietary adjustments and supplementation.

Navigating through this sea of information was daunting. We had piles of medical documents: test results, clinical notes, medication instructions, and therapeutic guidelines. Each document was filled with specialized medical terminology and detailed descriptions that were hard to decipher without a medical background. Feeling overwhelmed, I turned to an AI assistant to help understand these complex terms and conditions. With chatGPT's assistance, I was able to break down the medical jargon into more understandable language. For example, chatGPT explained that "chronic denervation" referred to long-term damage or dysfunction in the nerves supplying specific muscles, and "mild fall out of motor units" indicated a decrease in the number or function of motor units in the affected muscle. ChatGPT also helped us understand the im-

plications of the prescribed medications. Gabapin 300 (Gabapentin) was explained as a medication used to manage neuropathic pain, and Mirator 0.5 (Pregabalin) as another neuropathic pain treatment. The AI provided insights into how these medications work, their potential side effects, and why they were chosen for this condition.

Moreover, chatGPT helped us comprehend the significance of low vitamin levels and how proper supplementation and dietary adjustments could support overall health and possibly alleviate some symptoms. This guidance was invaluable in helping us navigate through the complexity of the treatment plan. Despite these efforts, the sensation they described as an "electric current" traveling through their leg remained distressing and affected daily activities. It hindered their ability to concentrate and perform routine tasks comfortably. They found some relief when lying down, but this was not a practical solution for their busy lifestyle. This experience underscored the critical need for a solution that can bridge the gap between complex medical information and patient comprehension. It highlighted how difficult it is for patients to access, understand, and utilize medical information effectively, which is essential for making informed decisions about their health and treatment

1.2 Nikhil:

In the heart of our family, my grandmother stands as a pillar of strength, having navigated through numerous health challenges with grace and resilience. Her medical journey has been far from easy—diabetes, high blood pressure, and surgeries for her heart and knee have made regular checkups and frequent tests a necessity. Each medical report she receives is a maze of complex medical jargon, leaving us bewildered about her health status. We constantly grapple with understanding whether her levels are within a healthy range, what immediate precautions we should take, and what her results truly signify.

Our primary doctor is exceptional but always in high demand. Each time we get my grandmother's medical reports, scheduling an appointment to decipher them becomes a necessity. This often translates to waiting hours for a brief, rushed conversation that leaves our questions partially answered and us feeling even more perplexed. There have been times when we've had to travel to another city or state to see specialist doctors, only to wait an entire day due to delayed reports or the doctor's unavailability to review them promptly. The brief discussions we manage to have often feel rushed, leaving us with more questions than answers and a lingering sense of frustration.

This recurring challenge inspired me to develop a solution to transform how patients access and comprehend their medical information. Thus, we created a Q&A platform designed to interpret and respond to queries related to medical prescriptions, lab tests, and diagnostic reports. Through this innovation, we aim to empower patients like my grandmother, ensuring they receive timely, accurate information and enhancing their overall healthcare experience.

1.3 Parivesh:

Motivated by my parents' dedication to dentistry, I have developed a simple Q&A chatbot that leverages the Medical Llama LLM to handle patient reports and queries. This innovative tool aims to streamline their practice by providing accurate, timely responses to common dental questions and concerns, thereby enhancing patient engagement and improving efficiency. Not only designed for my parents' practice, this chatbot is also intended to benefit other doctors by offering a reliable, AI-driven solution to support their clinical operations and elevate the standard of patient care. Through this project, I hope to contribute meaningfully to the healthcare community by integrating advanced AI technology into everyday medical practice.

2 Introduction

The integration of Artificial Intelligence (AI) in the healthcare sector has shown immense potential to revolutionize medical practices, offering innovative solutions to improve patient care, streamline administrative processes, and enhance clinical outcomes. Among the various AI technologies, Natural Language Processing (NLP) has emerged as a crucial tool in the interpretation and utilization of medical information. This report outlines the development and implementation of a web-based medical assistant application designed to leverage advanced NLP techniques for extracting and interpreting medical documents, facilitating efficient and accurate information retrieval for healthcare professionals and students. Our application is built using Flask, a lightweight web framework for Python, and incorporates the Medical-Llama3-8B model (Ruslanmv, 2023), a fine-tuned Large Language

Model (LLM) tailored specifically for medical contexts. This model, integrated via the Hugging Face transformers library, enables the system to understand and respond to complex medical queries with a high degree of accuracy and relevance. The motivation behind this project stems from the need for a reliable tool that can assist in navigating the vast and often complex information contained in medical documents. Traditional methods of information retrieval in the medical field can be time-consuming and require significant manual effort. By automating this process with an AI-driven application, we aim to enhance the efficiency and effectiveness of medical professionals in their daily tasks.

Key Features and Components

- Text Extraction: The application supports both image and PDF formats for document input. EasyOCR is used for extracting text from images, while PyMuPDF handles text extraction from PDF files. These tools were chosen for their effectiveness and compatibility with our system's resource constraints.
- Language Model Integration: The core of our application is the Medical-Llama3-8B model (Ruslanmv, 2023). This LLM has been fine-tuned on a large dataset of medical information, enabling it to generate accurate and contextually relevant responses to user queries. The model is loaded using quantization techniques to manage its substantial size and computational requirements efficiently.
- User Interface: A user-friendly web interface was developed to facilitate interaction with the system. Users can upload medical documents and submit queries through the interface, receiving detailed and relevant responses generated by the LLM.
- Authentication and Security: To protect user data, the model runs locally, preventing any medical data from being leaked. The model is also instructed to avoid disclosing any personal information in its responses, ensuring privacy and confidentiality.

3 Task Description

The task involves the development of a Question and Answer (Q&A) platform designed to assist users in interpreting and understanding their medical documents. The system accepts various types of medical documents as input, including:

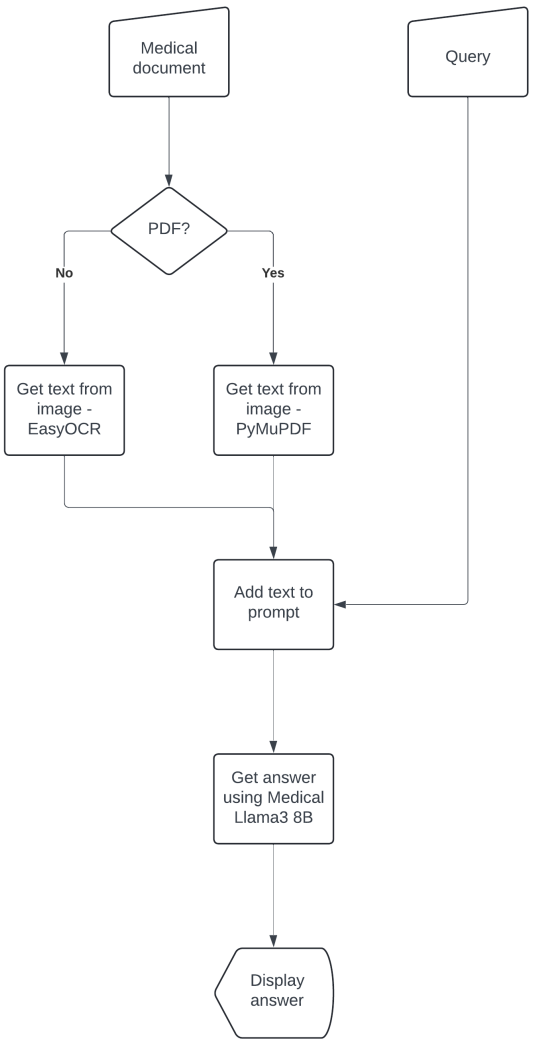

Figure 1: Block diagram of the system

- **Medical Prescriptions:** Documents detailing prescribed medications and treatment plans.
- **Lab Test Reports:** Reports containing the results of medical tests and diagnostics.
- **Diagnostic Reports:** Comprehensive reports from medical imaging and other diagnostic procedures.

Users can upload these documents and submit questions related to their contents. The system processes the documents, extracts relevant information, and provides detailed answers to the user queries. This task requires the integration of advanced Optical Character Recognition (OCR) technology to accurately interpret and respond to medical inquiries.

Why Focus on This Project Healthcare accessibility and literacy are significant challenges, particularly in regions with limited access to medical professionals. In India, the doctor-to-patient ratio is alarmingly low, especially in rural areas, leading to delays in appropriate treatment and increased stress for patients. Many individuals struggle to understand their medical reports due to the complex language used. This project aims to address these issues by:

1. **Enhanced Patient Understanding:** Medical documents are often complex and filled with jargon that can be difficult for patients to understand. By providing a platform that interprets these documents and answers related questions, we can help patients gain a clearer understanding of their medical conditions and treatment plans.
2. **Improved Healthcare Communication:** Effective communication between healthcare providers and patients is essential for successful treatment outcomes. This platform can bridge the gap by ensuring that patients have access to accurate and understandable information about their health, facilitating better-informed discussions with their healthcare providers.
3. **Accessibility and Empowerment:** Empowering patients with knowledge about their health can lead to more proactive and engaged healthcare management. By making medical information more accessible, we enable patients to take a more active role in their healthcare decisions.
4. **Efficiency for Healthcare Providers:** Healthcare providers often face time constraints and high workloads. This platform can assist by providing patients with accurate information, potentially reducing the number of routine inquiries and allowing providers to focus on more critical aspects of care.
5. **Integration of Advanced Technologies:** The task leverages cutting-edge NLP and OCR technologies, showcasing the potential of AI in transforming healthcare. By focusing on this task, we contribute to the broader field of AI in healthcare, demonstrating practical applications and benefits.
6. **Data Privacy and Security:** Running the model locally ensures that sensitive medical data is kept secure, addressing privacy concerns and building trust with users.

By leveraging technology to simplify and clarify medical information, this project aims to make healthcare more accessible, understandable, and efficient, ultimately contributing to better health outcomes and a more informed society.

4 Related Work

Several existing applications and systems aim to assist patients in understanding their medical documents and providing answers to their health-related questions. These systems leverage various technologies, including Optical Character Recognition (OCR), Natural Language Processing (NLP), and machine learning. Below are a few notable examples, along with their advantages and limitations:

1. MyChart by Epic Systems: MyChart is a patient portal that provides users with access to their health records, lab results, and prescriptions. It allows patients to ask questions directly to their healthcare providers and view their medical history.

Cost: Starting from $500,000 for smaller healthcare systems

Advantages:

- Integration with Healthcare Providers: Directly connected with healthcare providers, ensuring accurate and personalized responses.
- Comprehensive Access: Provides a comprehensive view of the patient's medical history and current health data.

Limitations:

- Dependency on Providers: Patients need their healthcare providers to be part of the Epic Systems network to access their data.
- Limited AI Capabilities: The platform relies more on direct provider interaction rather than AI-driven insights, which may delay responses.

2. Ada Health: Ada Health is a health assessment and symptom-checking app powered by AI. It asks users a series of questions about their symptoms and provides possible conditions and advice.

Cost: Free

Advantages:

- User-Friendly Interface: Easy to use with a conversational interface that guides users through symptom assessment.
- Broad Symptom Database: Covers a wide range of symptoms and conditions, providing valuable insights to users.

Limitations:

- Generic Advice: The app provides general health advice and possible conditions but may not be as accurate for specific medical documents and reports.
- No Direct Document Analysis: Focuses on symptom checking rather than interpreting detailed medical documents.

3. Google Health: Google Health provides various tools and services aimed at improving health outcomes, including AI-powered diagnostics and health record management.

Cost: Free

Advantages:

- Advanced AI Integration: Utilizes state-of-the-art AI and machine learning models for diagnostics and health insights.
- Strong Data Analytics: Provides robust data analytics and health tracking features.

Limitations:

- Privacy Concerns: Handling and storage of sensitive health data by a large tech company may raise privacy concerns among users.
- Limited Direct Interaction: While powerful in analytics, it may not provide direct Q&A capabilities for detailed medical documents.

4. IBM Watson Health: IBM Watson Health uses AI to analyze vast amounts of health data, providing insights and recommendations for both healthcare providers and patients.

Cost: $140 per month

Advantages:

- Powerful AI and Data Analysis: Leverages IBM's advanced AI capabilities to provide deep insights and recommendations.
- Clinical Decision Support: Assists healthcare providers with clinical decision-making by analyzing patient data.

Limitations:

- Complexity and Cost: Can be complex to implement and may be cost-prohibitive for smaller healthcare providers or individual users.
- Primarily Provider-Focused: More focused on aiding healthcare providers rather than direct patient interaction and Q&A.

5. Apple Health Records: Apple Health Records allows users to store and view their health records on their iPhones. It integrates data from various healthcare providers into one accessible platform.

Cost: Free

Advantages:

- Convenient Access: Provides easy access to health records on a user's mobile device.
- Secure and Private: Emphasizes data security and privacy, leveraging Apple's security infrastructure.

Limitations:

- Limited AI Functionality: Does not provide AI-driven insights or Q&A functionality.
- Provider Dependency: Requires healthcare providers to participate in Apple's Health Records program for data integration.

4.1 Related Papers

In the realm of medical document analysis and question answering, several notable research papers have contributed to the development and understanding of this field. These papers explore various methodologies, advancements, and applications of large language models (LLMs) and other AI technologies to improve the accuracy and reliability of medical question-answering systems.

1. MedPaLM: Towards Expert-Level Medical Question Answering with Large Language Models (Singhal et al., 2023)

Conference/Journal: arXiv

Authors: Google Research and DeepMind Team

Publication Year: 2023

Summary: This paper presents MedPaLM, a model designed to achieve expert-level performance in medical question answering. It builds on the Pathways Language Model (PaLM) framework and introduces medical domain-specific fine-tuning and prompting strategies.

Advantages:

- **High Accuracy**: Achieved state-of-the-art accuracy in medical question answering.
- **Clinical Relevance**: Preferred by human evaluators over physician answers in several clinical scenarios.

Limitations:

- **Resource Intensive**: Requires significant computational resources for training and inference.
- **Access**: Primarily available to research institutions and not widely accessible for individual use.

2. JMLR: Joint Medical LLM and Retrieval Training for Enhancing Reasoning and Professional Question Answering Capability (Wang et al., 2024)

Conference/Journal: arXiv

Publication Year: 2024

Summary: The paper introduces Joint Medical LLM and Retrieval Training (JMLR), a novel approach that integrates retrieval systems with LLMs during fine-tuning to reduce "hallucination" and improve accuracy in medical question answering.

Advantages:

- **Improved Accuracy**: Demonstrated significant accuracy improvements over state-of-the-art models.
- **Efficient Training**: Reduced training time and resource requirements.

Limitations:

- **Preprint Status**: Findings are preliminary and require further peer review and validation.
- **Complex Integration**: Combining retrieval and LLM training can be complex and challenging to implement.

3. Large Language Models Encode Clinical Knowledge (Singhal et al., 2022)

Conference/Journal: Nature

Authors: Google Research and DeepMind Team

Publication Year: 2023

Summary: This study evaluates the clinical knowledge encoded by large language models using the MultiMedQA benchmark, which combines multiple medical question-answering datasets.

Advantages:

- **Comprehensive Evaluation**: Utilizes a diverse benchmark for thorough evaluation.
- **High Performance**: Flan-PaLM model achieved state-of-the-art results across several datasets.

Limitations:

- **Gap in Consumer Questions**: Highlighted limitations in generating responses to consumer medical questions.
- **Instruction Prompt Tuning Required**: Additional tuning needed to align responses with expert knowledge.

4. RJUA-MedDQA: A Multimodal Benchmark for Medical Document Question Answering and Clinical Reasoning (Jin et al., 2024)

Conference/Journal: arXiv

Publication Year: 2024

Summary: This paper introduces RJUA-MedDQA, a benchmark designed to evaluate LLMs and large multimodal models on medical document understanding and clinical reasoning tasks.

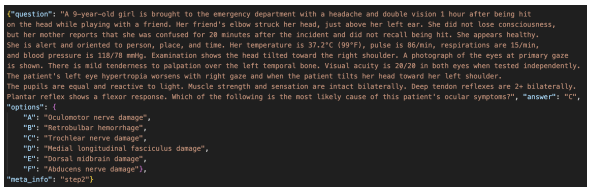

Figure 2: Sample of MEDQA dataset

Advantages:

- **Real-World Data**: Uses real-world medical report images to challenge AI models.
- **Enhanced Annotation Efficiency**: Introduces the Efficient Structural Restoration Annotation (ESRA) method.

Limitations:

- **Performance Limitations**: Current models still show limited performance on this challenging benchmark.
- **Focus on Chinese Reports**: Benchmark primarily based on Chinese medical reports, which may limit generalizability.

5. LingYi: Medical Conversational Question Answering System based on Multi-modal Knowledge Graphs (Xia et al., 2022a)

Conference/Journal: arXiv

Publication Year: 2022

Summary: LingYi integrates a Chinese Medical Multi-Modal Knowledge Graph and a large-scale Chinese Medical CQA dataset to facilitate dynamic, knowledge-based medical dialogue.

Advantages:

- **Multimodal Integration**: Combines text and image data for comprehensive understanding.
- **Dynamic Dialogue**: Supports dynamic and context-aware medical conversations.

Limitations:

- **Resource Intensive**: Requires substantial computational resources for multimodal integration.
- **Language Limitation**: Primarily focused on Chinese, limiting applicability in other languages.

6. MedConQA: A Medical Conversational Question Answering System (Xia et al., 2022b)

Authors: Research Team (EMNLP 2022)

Publication Year: 2022

Summary: MedConQA is designed to improve the quality and reliability of medical conversational systems, integrating entity disambiguation, central records memory, and symptom selection algorithms.

Advantages:

- **Structured Approach**: Handles different phases of medical consultation effectively.
- **Personalized Responses**: Generates contextually appropriate and medically sound responses.

Limitations:

- **Complex Architecture**: The system's sophisticated architecture may be complex to implement and maintain.
- **Limited Real-World Testing**: Requires extensive real-world validation to ensure reliability.

4.1.1 Summary

The existing research highlights significant advancements in the application of large language models and AI technologies to medical question answering. These papers present various innovative approaches and benchmarks, each contributing to the field's growth. However, they also underscore the challenges, such as resource requirements, integration complexities, and the need for further validation. Our proposed system aims to build on these advancements, addressing some of the limitations by focusing on secure, efficient, and accessible solutions for analyzing medical documents and providing accurate user queries.

5 Proposed approach

5.0.1 Overview

Our proposed approach aims to develop a secure and efficient Q&A platform that can analyze medical documents and provide accurate answers to user queries. The system processes various types of medical reports, including prescriptions, lab test

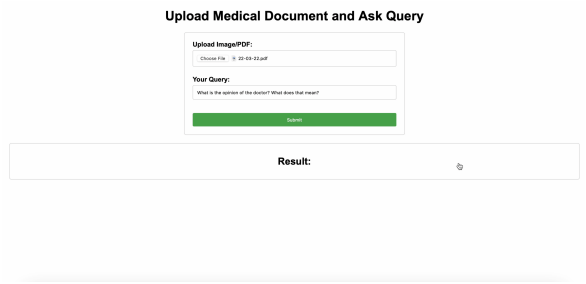

Figure 3: System interface

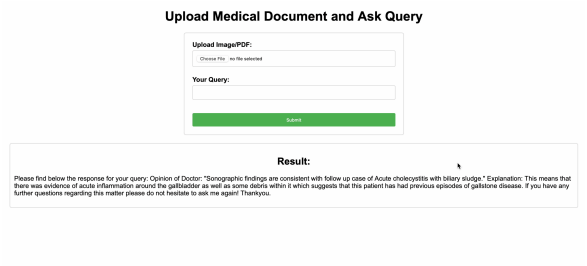

Figure 4: Result of the query

reports, and diagnostic reports, to respond to related user questions. The primary components of our approach include leveraging state-of-the-art language models, effective OCR tools, and a user-friendly interface. Below, we detail each aspect of our proposed approach.

5.1 System Architecture

5.1.1 User Interface:

Framework: The web-based interface is developed using Flask, a micro web framework written in Python. Flask is chosen for its simplicity and flexibility, allowing for rapid development and easy integration with other components of the system. The interface provides a platform for users to upload medical documents (PDFs or images) and input their queries.

Features: The interface includes multiple features to enhance user experience:

- Document Uploads: Users can upload their medical documents in various formats such as PDFs, PNG, JPG, and JPEG.
- Query Input Fields: Users can enter their medical-related questions in a dedicated input field.
- Response Display: The system displays the answers generated by the language model in a clear and concise manner.
- Error Handling: The interface includes mechanisms to handle errors gracefully, providing users with feedback in case of issues with file uploads or query processing.

OCR and Text Extraction:

- Image Files: EasyOCR is used to extract text from image files (PNG, JPG, JPEG). EasyOCR is selected for its high accuracy and efficiency in handling medical documents, which often contain complex terminology and handwriting. EasyOCR leverages deep learning models to recognize text, making it suitable for the intricate nature of medical documents.
- PDF Files: PyMuPDF, also known as Fitz, is utilized to extract text from PDF files due to its reliability and speed in processing such documents. PyMuPDF supports text extraction from both simple and complex PDFs, including those with embedded fonts and images. This makes it a robust choice for handling diverse types of medical reports.

Large Language Model (LLM):

- Model Selection: We opted to utilize the Medical LLaMA model(Ruslanmv, 2023) from Hugging Face for generating responses to user queries based on the extracted text from medical documents. This repository offers a fine-tuned version of the potent LLaMA3 8B model, specifically tailored for answering medical questions in an informative manner. Leveraging the extensive knowledge encapsulated within the AI Medical Chatbot dataset, the Medical LLaMA model (Ruslanmv, 2023) is adept at comprehensively addressing medical inquiries with its robust understanding of medical terminologies and contexts. This choice aligns with our objective of providing accurate and informative responses to users seeking medical guidance and information.
- Quantization: The model is loaded with quantization using BitsAndBytesConfig to optimize resource usage. Quantization reduces the model's memory footprint and computational requirements, allowing the system to run efficiently on available hardware without compromising performance.

Model Integration:

- Integration: The LLM is integrated into the Flask application to process user queries and provide responses based on the content of the uploaded medical documents. This involves setting up endpoints in the Flask application that handle file uploads, text extraction, and query processing.

- Prompt Design: The system generates responses through a series of prompts designed to ensure the relevance and accuracy of the answers. Prompts are crafted to guide the language model in generating precise and contextually appropriate responses. This includes including relevant extracted text from the medical documents and the user's query in the prompt.

Security and Privacy:

- Local Execution: The model runs locally to prevent any leakage of sensitive medical data. This ensures that all processing happens on the user's machine or a secure local server, enhancing data security and user privacy.

- Response Filtering: The system is programmed to avoid disclosing personal information in its responses. This involves implementing filters that detect and remove any sensitive data from the generated answers before they are displayed to the user.

5.1.2 Detailed Steps

User Upload:

- File Upload: Users upload their medical documents through the web interface. The interface supports drag-and-drop functionality and standard file selection dialogs.

- Supported Formats: The system supports multiple formats including images (PNG, JPG, JPEG) and PDF files. This flexibility ensures that users can upload documents in the format they have available.

Text Extraction:

- OCR for Images: EasyOCR processes image files to extract text. The extracted text is then cleaned and prepared for further processing.

- PDF Text Extraction: PyMuPDF processes PDF files to extract text. PyMuPDF handles various PDF structures, ensuring that text is accurately extracted from both simple and complex documents.

Context Formation: The extracted text serves as the context for the language model to generate answers. This context is crucial for ensuring that the model understands the content of the medical documents and can generate relevant responses.

Question Processing:

- User Input: Users input their medical-related questions through the interface. The system ensures that the input is received correctly and is free from any formatting issues.

- Prompt Creation: The system combines the extracted text and the user question to form a prompt for the LLM. This involves structuring the prompt in a way that maximizes the relevance and accuracy of the generated response.

Answer Generation:

- Response Generation: The LLM processes the prompt and generates a detailed response. The model uses its extensive training on diverse text data to provide accurate and relevant answers to the user's queries.

- Filtering: The system ensures the response is relevant to the query and does not include personal information. This involves post-processing the generated text to remove any sensitive data and ensure the answer is coherent and contextually appropriate.

Response Delivery:

- Display: The generated response is displayed on the web interface, providing the user with the information they requested. The interface is designed to present the answers clearly and concisely, making it easy for users to understand the response.

5.1.3 Implementation of Suggestions and Comments

Difference between KerasOCR and Pytesseract

- Resource Constraints: We initially considered KerasOCR, but it was not feasible due to GPU memory constraints. KerasOCR's deep learning model could not be loaded simultaneously

with the LLM, which occupied most of the GPU memory. This led us to seek alternative OCR solutions that were more memory-efficient.

- Pytesseract Issues: Pytesseract required the installation of the tesseract module, but we lacked the necessary sudo privileges to install it.

- Chosen Solutions: Instead, we used EasyOCR for image text extraction due to its efficiency and accuracy, and PyMuPDF for PDF text extraction for its reliability and speed. EasyOCR's lightweight model and high accuracy made it an ideal choice for our system.

Pre-processing

- **Techniques Tested:** We experimented with various pre-processing techniques including spell correction, medically augmented spell correction, stemming, and lemmatization. These techniques were evaluated for their impact on the quality of the input text and the subsequent performance of the LLM.

- **Observations:**

  - No Pre-processing: This approach yielded the best results as the LLM could understand the semantics of the medical document. The raw text, despite potential spelling errors or irregularities, provided the most contextually rich input for the model.

  - Spell Correction: This often led to incorrect modifications of medical terms, degrading the quality of the input document. Medical terminology is highly specific, and spell correction algorithms frequently made erroneous changes.
  - Stemming and Lemmatization: These techniques were less effective due to the complexity of medical terms and the LLM's better handling of raw text. The nuanced nature of medical language made these pre-processing steps less beneficial.

Implementation

System Development: The system was developed using Flask for the web interface, integrated with the Medical LLaMA model (Ruslanmv, 2023) for question answering. The detailed implementation code has been provided. The Flask application includes routes for file uploads, query input, and response generation, all seamlessly integrated with the LLM.

Interface Development

User-Friendly Design: We developed a user-friendly interface using Flask to facilitate document uploads and query input. The design focused on being intuitive and accessible for users. Features such as drag-and-drop uploads, clear input fields, and responsive design were prioritized to enhance user experience.

LangChain for Conversation Management

- Consideration: LangChain was considered for maintaining conversation context, which would allow the system to handle follow-up questions and maintain the flow of a conversation.

- Challenges: Documentation for using LangChain with a locally running LLM was insufficient, leading us to exclude it from the current implementation. Despite its potential benefits, the lack of robust support and examples for our specific setup posed significant challenges.

Quantization of LLM

- Quantization: The LLaMA model (Ruslanmv, 2023) was successfully quantized using BitsAndBytesConfig to reduce resource usage, enabling efficient model loading and inference on available hardware. Quantization involves reducing the precision of the model's weights, which significantly lowers memory and computational requirements without a substantial loss in performance.

- Implementation Check: Quantization was implemented in the provided code to optimize the model's performance. This involved configuring the model loading process to apply quantization, ensuring that the system runs efficiently even on limited hardware.

By incorporating these suggestions, we aimed to create a robust and efficient system that can accurately process and respond to user queries based on medical documents while ensuring data security and privacy. This comprehensive approach ensures that the platform is both effective in its functionality and secure in handling sensitive medical information.

6 Results and Discussion

6.1 Result

The implementation of the Q&A platform using a large language model (LLM) demonstrated promising capabilities in accurately answering user queries based on medical documents such as prescriptions, lab reports, and diagnostic reports. The system effectively processed text extracted from images and PDF files using EasyOCR and PyMuPDF libraries, respectively. It was able to generate relevant and informative answers to a wide range of medical questions posed by users.

Key observations from the results include:

Accuracy and Relevance The system provided accurate and contextually relevant answers to user queries based on the content of the provided medical documents. The LLM's ability to understand and interpret medical terminology and context played a significant role in the quality of the responses.

Performance with Different Document Types The system handled both image-based and PDF-based documents efficiently. Text extraction from images using EasyOCR and from PDFs using PyMuPDF was generally successful, though the quality of text recognition from handwritten notes remained a limitation.

Hyper-parameter Tuning Extensive tuning of hyper-parameters such as temperature, repetition penalty, and token limits was essential to optimize the model's performance. Adjustments to these parameters helped mitigate issues like the generation of excessive or repetitive text.

Quantization Implementing quantization allowed the deployment of a large model like LLaMA on available hardware resources, improving performance without compromising response quality significantly.

6.2 Discussion

The development of the Q&A platform using the Medical-Llama3-8B model (Ruslanmv, 2023) has demonstrated the potential of LLMs in enhancing medical information retrieval and patient support systems. The approach of leveraging advanced text extraction techniques coupled with a robust LLM enables accurate and context-aware responses to user queries.

Despite the positive results, the project highlighted several areas for improvement and future work:

Handwritten Document Processing Current OCR solutions struggle with handwritten notes, which are common in medical documents. Future iterations could explore more advanced or specialized OCR technologies capable of handling handwritten text.

Further Hyper-parameter Optimization While significant progress was made in tuning the model, continuous refinement and experimentation with hyper-parameters could further enhance performance and reduce the occurrence of irrelevant text generation.

User Interface Enhancements Developing a more user-friendly and interactive interface would improve the overall user experience, making the system more accessible and efficient for end-users.

Security and Privacy Although the model runs locally to prevent data leakage, further security measures and privacy protocols should be implemented to ensure the confidentiality of sensitive medical information.

7 Conclusion

In this project, we developed an AI-powered Question and Answer (QA) platform designed to process medical documents such as prescriptions, lab test reports, and diagnostic reports. The system leverages advanced OCR techniques and the Llama3-8B language model (Ruslanmv, 2023), fine-tuned specifically for medical inquiries. By combining the capabilities of PyMuPDF and EasyOCR, our system effectively extracts text from both PDF and image files, ensuring versatility in handling various types of medical documents.

The implementation of this platform demonstrated its potential to accurately answer user queries related to medical reports. However, the project faced several limitations. Firstly, the OCR libraries struggled with handwritten notes, leading to inaccuracies in context extraction. Secondly, the language model occasionally generated excessive text beyond the user's query response. Despite these challenges, fine-tuning the model's hyper-parameters helped mitigate some issues, resulting in more precise and relevant answers.

The results underscore the importance of using sophisticated language models for medical Q&A applications. The Llama3-8B model's ability to understand and generate medically relevant information highlights its value in providing informative responses to users. Nonetheless, the system's

performance could be further enhanced by addressing the limitations encountered, such as improving OCR accuracy for handwritten notes and optimizing text generation parameters.

Overall, this project represents a significant step toward creating an intelligent medical assistant capable of aiding users in understanding their medical reports. Future work will focus on refining the system's accuracy and expanding its capabilities to ensure it can handle a broader range of medical documents and inquiries effectively.

References

Congyun Jin, Ming Zhang, Xiaowei Ma, Li Yujiao, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, Jinjie Gu, Chenfei Chi, Xiangguo Lv, Fangzhou Li, Wei Xue, and Yiran Huang. 2024. Rjua-meddqa: A multimodal benchmark for medical document question answering and clinical reasoning.

Ruslanmv. 2023. Medical-llama3-8b. Hugging Face. License: Apache-2.0. Finetuned from model: meta-llama/Meta-Llama-3-8B.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022a. Lingyi: Medical conversational question answering system based on multi-modal knowledge graphs.

Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022b. Medconqa: Medical conversational question answering system based on knowledge graphs. pages 148–158.