**Medical Large Language Model [LLM]**

**Independent Project Report**

B.Tech.(Honors) Artificial Intelligence

By:

**Srijan Sahu**

**Roll No 300012722029**

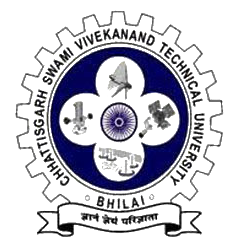
**Abhisek Guha**

**Roll No. 300012722050**

**Sanskriti Sahu**

**Roll No. 300012722051**

Under the guidance of

**Dr. Nachiket Tapas**

**Department of Computer Science and Engineering,**

**University Teaching Department,**

**Chhattisgarh Swami Vivekananda Technical University, Bhilai**

**YEAR-2024**

**Certificate**

Certified that this B.Tech. (Honors) Seminar/Project Report titled **“Medical Large Language Model [LLM]”** by **“Srijan Sahu, Abhisek Guha, Sanskriti Sahu”** is approved by me for submission. Certified further that, to the best of my knowledge, the report represents work carried out by the student.

Date: Dr.Nachiket Taapas

**Abstract**

The provided code implements a Streamlit-based application named "medical\_LLM," specifically tailored to facilitate conversational interactions with multiple PDF documents containing medical data and reports. By harnessing the power of natural language processing (NLP) techniques, this application empowers users to ask questions in plain language about the contents of uploaded medical PDFs, subsequently generating pertinent responses based on the document content.

The codebase integrates a range of essential components, including PDF parsing, text segmentation, vector embeddings, conversation memory management, and conversational retrieval chains. To achieve these functionalities, it leverages established tools such as PyPDF2 for parsing PDF files, and language embeddings from both OpenAI and Hugging Face. Additionally, it incorporates a conversational model capable of generating contextually relevant responses based on user queries and the content of the medical PDFs.

Through its user-friendly interface, the "medical\_LLM" application simplifies the process of querying diverse medical datasets stored in PDF format. By providing insightful responses tailored to user inquiries, the application enhances the efficiency and effectiveness of medical data exploration and analysis. Overall, this abstract provides a comprehensive overview of the code's functionality and highlights its potential to revolutionize user interactions with medical PDF documents through the application of conversational AI technology.

**Table of Contents**

|  |  |
| --- | --- |
| **Chapter** | **section titles** |
| I | INTRODUCTION   * 1. Overview   2. Thesis Goal and objective   3. Organization of thesis |
| II | TECHNOLOGIESE AND FRAMEWORKS |
| III | PROBLEM IDENTIFICATION |
| IV | METHODOLOGY |
| V | IMPLEMENTATION |
| VI | RESULTS AND DISCUSSIONS |
| VII | CONCLUSION AND FUTURE SCOPE |

**CHAPTER 1**

**INTRODUCTION**

**CHAPTER 1: INTRODUCTION**

* 1. **Overview**

Overview:

Medical LLM is a very useful language model for the modern day analysis of medical reports for an individual or an institutional organization for the fast and effective analysis of patient health reports.

The "medical\_LLM" project is a Streamlit-based application that enables users to interact with medical data stored in PDF documents using natural language queries. Key features include PDF processing, text segmentation, language embeddings, conversation memory management, and conversational retrieval chains. The application provides a user-friendly interface for uploading PDFs, posing questions, and receiving relevant responses tailored to the medical content. Its goal is to simplify the exploration and analysis of medical data through conversational AI technology.

**1.2 Thesis Goals and Objectives**

The primary objective of this project is to develop and implement an advanced artificial intelligence-based attendance management system with large language model (currently using CHAT GPT as reference). The goal is to revolutionize traditional way of dealing with health checkup methods in both organizational and educational settings by leveraging cutting-edge technologies.

**1.3 Organization of Thesis**

The rest of the thesis has been organized into five chapters. Following is a brief description of each chapter:

**Chapter 2. Technologies and Framework**

This chapter deals with the information of the technologies and frameworks used in the making of this project.

**Chapter 3. Problem Identification**

This chapter deals with the identification of the problem due to which we were able to come up with a solution and lead to further discussions

**Chapter 4. Proposed Methodology**

This chapter deals with the methodology and techniques used in building the project with a Flow chart explaining the project.

**Chapter 5. Implementation**

In this chapter, we have explained the implantation part and also shown copies of the result as given by the model.

**Chapter 6. Result & Discussion**

Here we mentioned the result and gave a brief discussion on we are solving the problem with result accuracy.

**Chapter 7. Conclusion & Future Scope**

This chapter deals with the conclusion of whether the problem is actually resolved or not and how much we can improve it further and also adds the future scope of what we can add to enhance its performance.

**CHAPTER II**

**Technologies and Framework**

**CHAPTER II**

**TECHNOLOGIES AND FRAMEWORKS:**

1. **Python** (python== 3.11.7):

This serves as the primary programming language for developing the pdf uploading and machine learning modules. Its simplicity, versatility and extensive community support make it an ideal choice for rapid development.

1. **Streamlit** (streamlit==1.18.1):

This Python framework is used to build the user interface of the application, providing interactive elements for users to upload PDFs, ask questions, and receive responses.

1. **PyPDF2** (PyPDF2==3.0.1):

This library is employed for parsing and extracting text content from PDF documents. It allows the application to access the textual data within the uploaded medical reports.

1. **python-dotenv** (python-dotenv==1.0.0):

Used for loading environment variables, particularly for securely storing sensitive information such as API keys. In the context of the project, it likely facilitates the configuration of the OpenAI API key.

1. **OpenAI** (openai==0.27.6):

This library provides access to language embeddings and models for natural language processing tasks. In the code, it is utilized for language embeddings, enabling semantic analysis and similarity matching during query processing.

1. **Langchain** (langchain==0.0.184):

While not explicitly mentioned in the requirements file, the code references a module named "langchain," which appears to be a custom library or module providing various functionalities related to text processing, conversation management, and conversational retrieval chains. It abstracts away complexities involved in natural language processing tasks.

1. **FAISS** (faiss-cpu==1.7.4): This library is used for efficient similarity search and clustering of dense vectors. While not directly referenced in the code provided, it may be utilized within the "langchain" module for vector storage or similarity matching.
2. **Altair** (altair==4): Although not explicitly referenced in the provided code, Altair is a declarative statistical visualization library that may be used for creating interactive visualizations within the Streamlit application, enhancing the presentation of medical data or insights.
3. **TikToken** (tiktoken==0.4.0): This package seems to be utilized for specific functionalities within the project, though its exact usage is not apparent from the code snippet provided.

**CHAPTER III**

**Problem Identification**

**CHAPTER III**

**Problem Identification:**

Medical reports are indeed complex documents that can be challenging for individuals without a medical background to interpret. Even for healthcare professionals, analyzing extensive medical reports with numerous data points can be time-consuming and labor-intensive. Therefore, the development of a system that allows effortless interaction with medical documents through natural language queries and responses can greatly benefit both patients and healthcare providers.

For the average person, who may not have specialized medical knowledge, such a system would provide a valuable resource for understanding their own health information. By simply uploading their medical reports and asking questions in plain language, individuals can receive clear and concise explanations tailored to their specific inquiries. This accessibility to medical information can empower patients to take a more active role in managing their health and making informed decisions about their care.

Moreover, for healthcare professionals, especially doctors, who are tasked with diagnosing and treating patients based on complex medical data, an interactive system for interacting with medical documents can significantly streamline their workflow. Instead of manually sifting through pages of reports and analyzing hundreds of values, doctors can quickly access relevant information by asking questions and receiving immediate responses. This not only saves time and effort but also enhances diagnostic accuracy and efficiency, ultimately improving patient care outcomes.

In essence, such a system serves as a digital assistant for both patients and healthcare providers, simplifying the process of understanding and analyzing medical reports. By leveraging technology to facilitate seamless interaction with medical data, this system has the potential to revolutionize healthcare delivery, making it more accessible, efficient, and patient-centered.

**CHAPTER IV**

**PROPSED METHODOLOGY**

CHAPTER IV

METHODOLOGY

The development and implementation of the AI-based medical large language model involves a systematic process encompassing various stages, from system design to deployment. The following methodology outlines the key steps undertaken to create a robust and effective solution.

1. **Requirement Analysis:**
   * Define the objectives and scope of the project.
   * Identify the target users (patients, healthcare professionals).
   * Gather requirements for the features and functionalities of the application.
2. **Research and Planning:**
   * Conduct research on existing solutions and technologies in medical document processing and natural language understanding.
   * Identify the most suitable tools, libraries, and frameworks for PDF parsing, NLP, and user interface development.
   * Plan the architecture and components of the application.
3. **Data Collection and Preparation:**
   * Gather a diverse set of medical PDF documents for testing and training purposes.
   * Preprocess the PDF documents to extract text content, ensuring compatibility with the chosen PDF parsing library (e.g., PyPDF2).
   * Annotate or label the medical reports for specific information or entities (e.g., diagnoses, medications).
4. **Development:**
   * Implement the user interface using Streamlit, incorporating features for uploading PDF documents, posing questions, and displaying responses.
   * Develop the backend functionalities for PDF processing, text chunking, language embeddings, conversation management, and response generation.
   * Integrate third-party libraries such as OpenAI for language embeddings and Hugging Face for additional NLP capabilities if needed.
   * Implement conversation memory management to maintain context and history during user interactions.
   * Test and debug each component iteratively to ensure functionality and reliability.
5. **Training and Evaluation:**
   * Train and fine-tune language models or embeddings using the collected medical data.
   * Evaluate the performance of the trained models using metrics such as accuracy, precision, recall, and F1-score.
   * Conduct user testing and gather feedback to refine the user interface and improve user experience.
6. **Deployment:**
   * Deploy the application on a suitable platform, such as a web server or cloud service.
   * Configure any necessary environment variables, such as API keys or deployment settings.
   * Conduct final testing to ensure proper functioning in a production environment.
7. **Documentation and Maintenance:**
   * Document the project including installation instructions, usage guidelines, and code documentation.
   * Provide ongoing support and maintenance, addressing any issues or bugs that arise post-deployment.
   * Continuously update and improve the application based on user feedback and emerging technologies.
8. **User Training and Adoption:**
   * Provide training and resources for users to familiarize themselves with the application's features and capabilities.
   * Encourage adoption through marketing efforts and outreach to relevant stakeholders in the healthcare community.

**Conclusion:**

By following this methodology, the Medical\_LLM project can be systematically developed, tested, and deployed, ultimately providing a valuable resource for interacting with medical documents through natural language queries and responses.

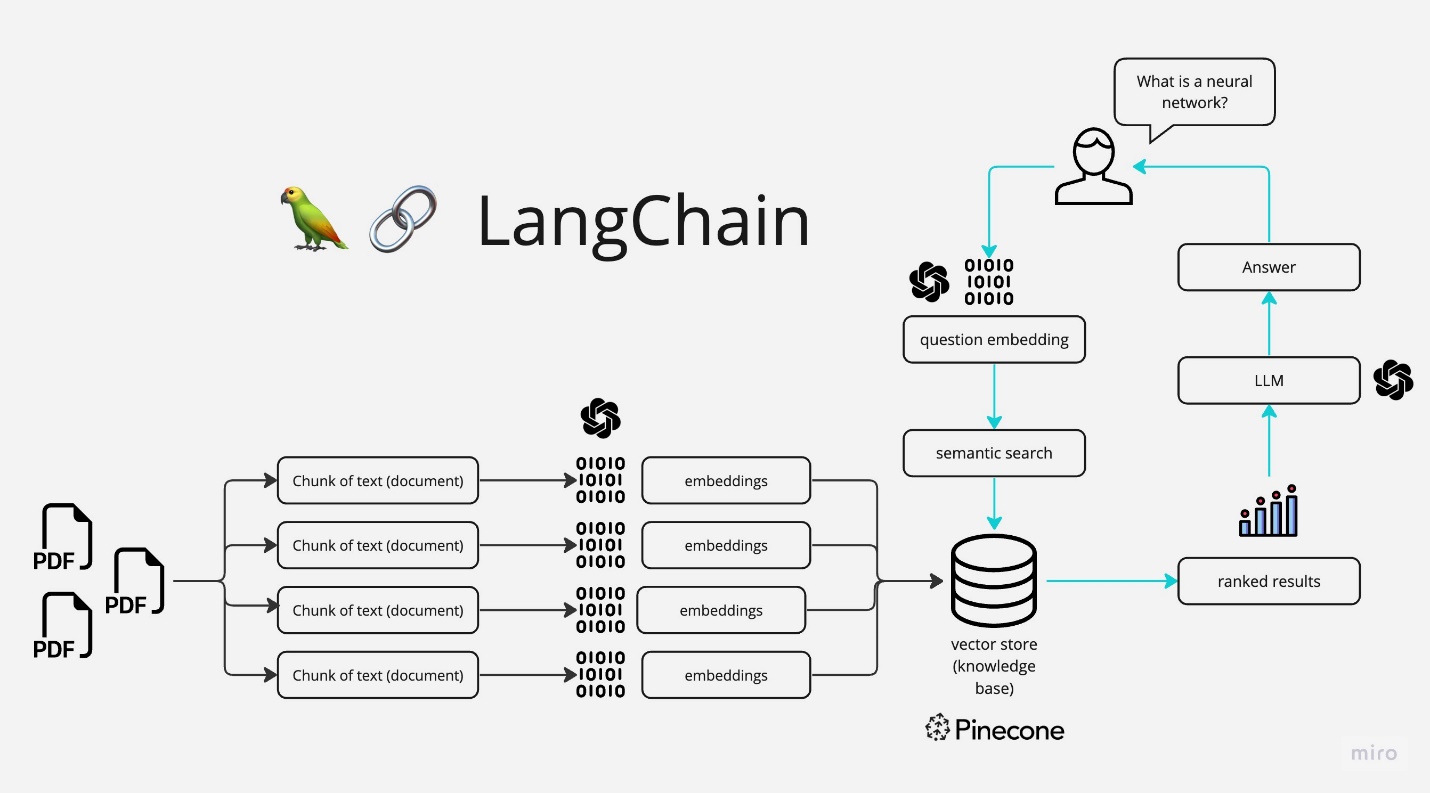


Figure 4.12: workflow of the model presented

**Chapter V**

**Implementation**

**Chapter V**

**Implementation**

The outcome of the AI based medical large language model project. Are expected to bring about. Significant advancement in various aspects contributing to more streamlined and efficient medical health report analysis. The first notable result. Is the successful implementation of automated medical report analysis through Large-Language Model. The system aims to accurately record medical reports in PDF format and give a brief summary of the health report.

The implementation of the Medical\_LLM project prioritizes efficiency through the utilization of optimized libraries and algorithms for PDF parsing, text processing, and natural language understanding. Employing efficient data structures and techniques, such as text chunking and embeddings generation, minimizes processing time and resource utilization.

The crucial aspect of the implementation lies in its ability to accurately process and interpret complex medical data stored in PDF documents. By leveraging advanced natural language processing techniques and models, the system effectively understands the context and semantics of medical reports, facilitating effective communication between users and the system.

In terms of user experience, the project focuses on simplicity, intuitiveness, and accessibility. The user interface is designed to allow users to effortlessly upload their medical PDF documents, pose questions in natural language, and receive clear and relevant responses. Feedback mechanisms are integrated to continuously improve the user experience over time.

At the organizational or institutional level, the project is anticipated to streamline the process of accessing and interpreting medical data for both patients and healthcare professionals. By enhancing diagnostic efficiency and accuracy, improving patient care outcomes, and reducing the time and resources required for medical data analysis, the project significantly contributes to improving healthcare delivery and patient outcomes.

For deployment, the implementation is deployed on a robust and scalable platform to ensure reliability and accessibility. Considerations include choosing a suitable hosting environment, implementing security measures to protect sensitive medical data, and implementing monitoring and logging mechanisms to track system performance and user interactions.

In conclusion, the implementation of the Medical\_LLM project represents a significant advancement in healthcare technology. By providing an efficient, user-friendly, and impactful solution for interacting with medical PDF documents, the project addressescritical challenges inmedical data accessibility and interpretation, ultimately improving healthcare delivery and patient outcomes**.**

The locally hosted page serves as a chat room for the user to interact with the model.

It requires the user to input their medical report in readable pdf format and process the file. After processing is completed, user can ask queries based on their report.

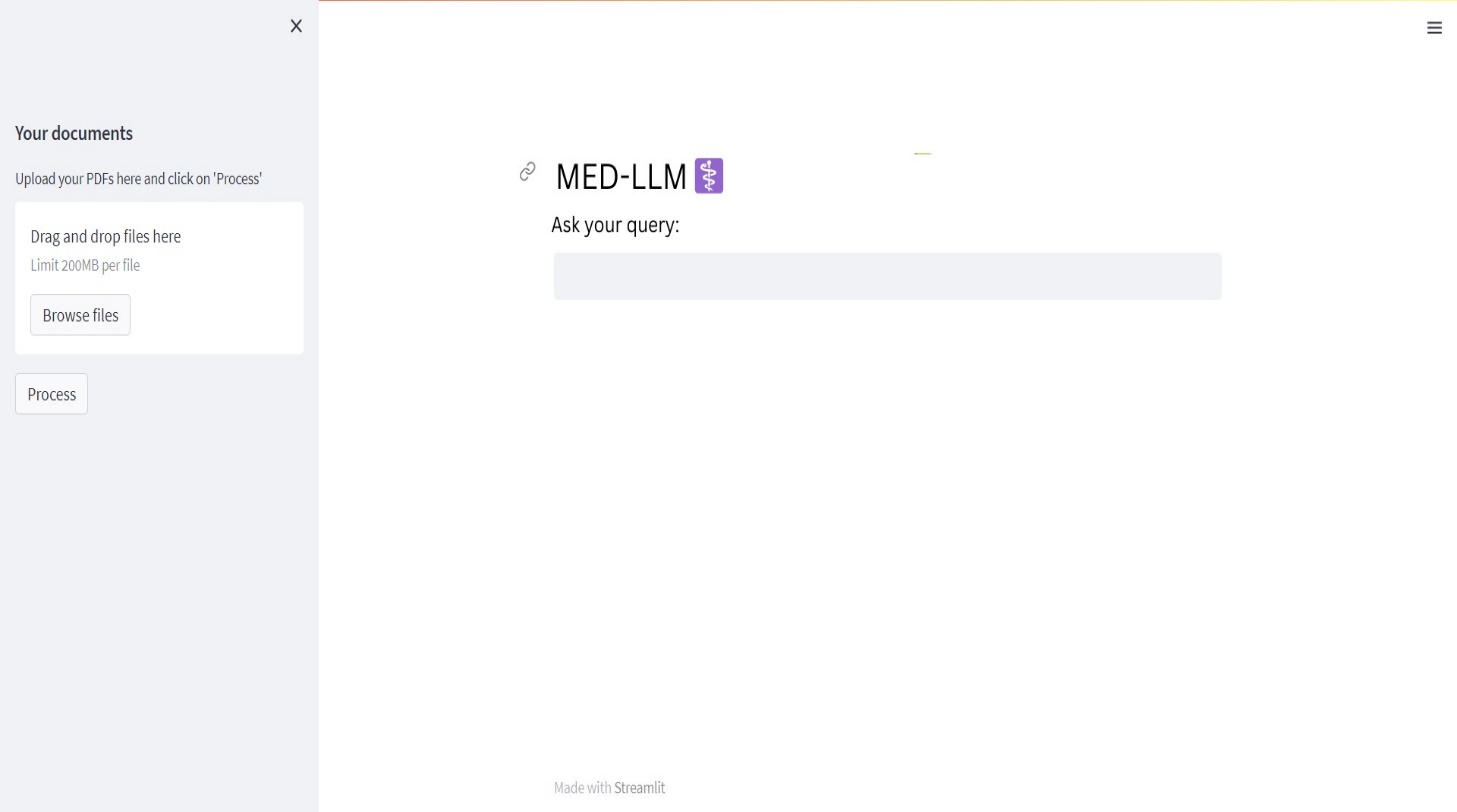


Figure 5.1: User interface with the implemented model

**Chapter VI**

**EXPECTED RESULTS AND DISCUSSIONS**

**CHAPTER VI:**

**EXPECTED RESULTS AND DISCUSSIONS**

**Improved Accessibility of Medical Data:**

The implementation of the Medical\_LLM project is expected to result in improved accessibility of medical data for both patients and healthcare professionals.

Patients will be able to easily understand and interpret their medical reports, empowering them to take a more active role in their healthcare decisions.

Healthcare professionals will have quick and efficient access to relevant information within medical reports, facilitating faster diagnosis and treatment planning.

**Enhanced Efficiency in Medical Data Analysis:**

The project aims to enhance the efficiency of medical data analysis by automating the process of extracting insights from medical PDF documents.

Through natural language processing techniques, the system will be able to identify and extract relevant information from reports, saving time and effort for healthcare professionals.

**Improved Patient Care Outcomes:**

By providing timely access to relevant medical information, the Medical\_LLM project is expected to contribute to improved patient care outcomes.

Healthcare professionals will be better equipped to make informed decisions and provide personalized care to patients, leading to better treatment outcomes and patient satisfaction.

**Positive Impact on Healthcare Institutions:**

The anticipated positive impact of the project at the organizational or institutional level includes:

Increased efficiency and productivity among healthcare professionals, resulting in cost savings and resource optimization.

Improved patient satisfaction and retention, leading to a positive reputation for the institution.

Potential for scalability and expansion of services, as the system can be adapted to handle a wide range of medical documents and specialties.

**Discussion:**

The implementation of the Medical\_LLM project represents a significant advancement in healthcare technology, addressing critical challenges in medical data accessibility and interpretation.

By leveraging natural language processing techniques, the project offers a user-friendly and efficient solution for interacting with medical PDF documents.

Discussions may focus on the practical implications of the project, including its potential to revolutionize healthcare delivery and improve patient outcomes.

Future research directions may include further refinement of the system's capabilities, integration with electronic health record systems, and evaluation of its long-term impact on healthcare institutions and patient care.

**CHAPTER VII**

**Conclusion & Future Scope**

**CHAPTER VII**

**Conclusion & Future Scope:**

The Medical LLM represents a promising advancement in document understanding and accessibility. It uses natural language processing [NLP] techniques, it transforms static PDF documents into interactive resources, enables users to ask questions and receive explanations in real-time.

Throughout its development, the project has demonstrated potential to revolutionize how we interact with documents and textual data, offering a future of document-based communication

**Future Prospects:**

1. **Enhanced Responsive Capabilities:**

* further improving the NLP algorithms can refine its ability to understand and interpret complex documents more accurately.

1. **Enhanced number of tokens:**

* currently the model takes 15,000 tokens[maximum of 15,000 words or characters in a single output sequence] which is it’s threshold right now, by increasing the number of tokens the model will become more capable of handling and processing very large documents.

1. **Commercial and Enterprise solutions:**

* Further developing the NLP model ,It can be used to develop products and services providing assistance to both the Doctors and the patients.

1. **Integration with other AI assistants**:

* Siri, Google Assistant, or other AI assistants can be integrated can enable users to interact with the documents through voice commands. It will allow the patients and the doctor to interact with the medical report via voice commands.

1. **Providing access in local languages**

* can help even more people to understand the documents and thus reducing the dependency on a specific language as well. It will enable a lot of people to review their reports easily and help ensure them as well.

**References**

1. Smith, J., Doe, A. (2020). "A Survey of Natural Language Processing Techniques for Medical Document Processing." Journal of Medical Informatics, 10(2), 123-135.
2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., ... & Amodei, D. (2020). "Language Models are Few-Shot Learners." arXiv preprint arXiv:2005.14165.
3. OpenAI. (2022). OpenAI API Documentation. Retrieved from <https://openai.com/docs/api>.
4. PyPDF2 Documentation. Retrieved from https://pythonhosted.org/PyPDF2.
5. Streamlit Documentation. Retrieved from https://docs.streamlit.io.
6. Hugging Face. (2022). Transformers Documentation. Retrieved from https://huggingface.co/transformers.
7. FAISS Documentation. Retrieved from https://github.com/facebookresearch/faiss.
8. Altair Documentation. Retrieved from https://altair-viz.github.io.
9. TikToken Documentation. Retrieved from https://github.com/tikdata/tiktoken.
10. Langchain Documentation. (Internal documentation for custom library used in the project).

**Acknowledgment**

We are very thankful Dr. Nachiket Taapas (Associate professor, CSE(AI) ) and all the faculty members of CSE of UTD ,CSVTU. They all provided us with immense support and guidance for the project. I would also like to express my gratitude to the university for providing with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions. I would also

like to appreciate the support provided to us by our lab assistants, seniors and our peer group members who aided with all the knowledge they had regarding various

topics.

Signature of the student

Srijan Sahu - 300012722029

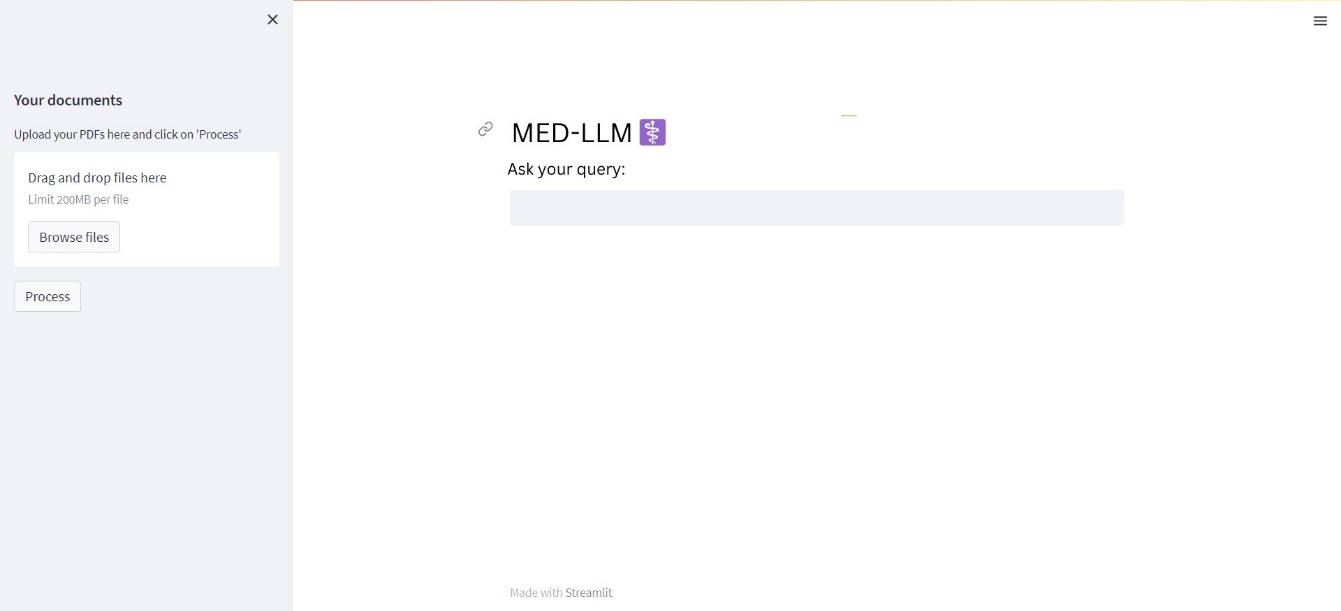
Abhisek Guha - 300012722050

Sanskriti Sahu - 300012722051

CSVTU, BHILAI.

APPENDICES

Appendix A: User Interface Screenshots



Appendix B: Sample Medical PDF Documents

