Vrije Universiteit Amsterdam

Universiteit van Amsterdam

Master Thesis

Promoting Digital Health in India: Using Lightweight Language models for generating Electronic Medical Records

Author: Sachin Dhananjaya (2763315)

1st supervisor: Dr. Anna Bon
2nd reader: Prof. Dr. Hans Akkermans

A thesis submitted in fulfillment of the requirements for the joint UvA-VU Master of Science degree in Computer Science

July 28, 2024

"Happiness can be found even in the darkest of times if one only remembers to turn on the light"

from Harry Potter*, by Dumbledore*

Abstract

Context. The Digitization of Healthcare, particularly through Electronic Medical Records (EMRs), is crucial for enhancing healthcare quality worldwide. Electronic Medical Records (EMRs) are digital versions of health records containing medical information about a person. In India, where traditional paper-based systems still dominate, the transition to digital health records faces numerous challenges, including financial constraints, limited technical knowledge, inadequate infrastructure, privacy of patient data and resistance towards technology.

Goal. The research aims to promote Digital Healthcare in India by developing a resource-efficient application which can assist healthcare workers by generating Electronic Medical Records (EMRs) from doctor-patient conversations during clinical visits. The design of this application explores how Lightweight Large Language Models (LLMs) and Automatic Speech Recognition (ASR) tools can be effectively used in a resource-constraint environment to generate these digital records, addressing the challenges and adoption barriers present in the Indian healthcare system.

Method. The study begins with the design of the application, focusing on integrating potential technologies for transcribing conversations and generating EMRs. Various ASR tools are compared based on their features, cost-effectiveness, and ability to accurately transcribe multilingual doctor-patient conversations. In parallel, Lightweight LLMs are evaluated for their performance in generating accurate EMRs from the doctor-patient conversations. Following this comparative analysis, the application is developed, with a thorough evaluation of its technical and cost requirements across different deployment options. Finally, feedback from healthcare providers is collected to assess its practicality and impact on the adoption of Digital Healthcare in India.

Results. The comparative assessment of the Automatic Speech Recognition (ASR) tools revealed that the open-source version of Whisper excelled in terms of support and performance. Hyperparameter-tuned versions of the LLMs, Llama-3 and Mistral demonstrated notable precision (0.888 and 0.859) and F1-scores (0.412 and 0.365). Consequently, the application was designed to utilize the Whisper model for transcribing audio conversations and the hyperparameter-tuned Llama 3 for generating EMRs. It was found that deploying the application in a cloud-based environment minimized technical requirements, although this approach incurred the highest costs. Feedback from healthcare providers highlighted the application's intuitive usability and high accuracy, though concerns were raised regarding its financial feasibility for smaller clinics. Privacy and security concerns regarding the confidentiality of the patient information used by LLMs were discussed and the potential measures found were de-identification of data, encryption, transparency and regulatory adaptation.

Conclusions. The study results in an application that healthcare providers in India can use to generate EMRs from their conversations with patients. It demonstrates how effectively leveraging Large Language Models (LLMs) and Automatic Speech Recognition (ASR) tools can facilitate the transition from paper-based to digital records in the healthcare sector. The paper also discusses various deployment strategies for the application, analyzing their technical and cost implications. While positive feedback from healthcare providers underscores the application's usability and potential to transform the digital landscape of healthcare, concerns about financial feasibility and privacy highlight potential barriers to widespread adoption. Further investigations include fine-tuning LLMs, implementing prompt engineering, and addressing the growing concerns about privacy and security with AI to facilitate potential integration into the Indian healthcare system. Overall, the research provides a scalable approach that leverages technology to enhance Digital Healthcare in India and serve as a framework to deliver improved healthcare in a low resource setting, globally.

Contents

List of Figures

List of Tables

1

Introduction

In recent years, the digitization of healthcare has emerged as a pivotal aspect of modern healthcare systems worldwide(1). The implementation of Electronic Medical Records (EMRs) lie at the core of this transformation. An Electronic Medical Record (EMR) is a comprehensive digital record of a Patient's Health Information, compiled from multiple encounters within a healthcare setting(2). Research shows that EMRs improve care quality and patient safety by providing accurate patient data, enhancing clinical decisions, reducing medical errors, and supporting efficient healthcare delivery(3).

While early EMR implementations struggled to achieve cost savings, recent studies show that EMRs are becoming increasingly cost-effective and financially viable as the technology advances and enhances care quality(4). The EMR industry has grown from nearly being nonexistent in 2000 to over $31 billion annually by 2018 and is expected to continue expanding, reaching $54.9 billion by 2028(4)[1].

Despite significant growth in EMR implementation, several barriers hinder its adoption in clinics. One major factor is the financial burden associated with implementing EMR systems. The costs for hardware, software, implementation assistance, and maintenance can be substantial (5). Additionally, there is a lack of technical training and support for healthcare providers(5, 6). Other issues affecting adoption rates include privacy and security concerns regarding patient data, the quality of EMR data, inadequate ICT infrastructure, and the absence of standardization protocols (5, 6, 7). These challenges are especially prevalent in developing countries, such as India, that are attempting to adopt EMR systems(8, 9).

In India, among the challenges mentioned, financial constraints, lack of technical support, reliance on traditional paper-based systems, and resistance to digital healthcare are

[1] https://www.globaldata.com/store/report/electronic-medical-records-systems-theme-analysis/

particularly significant (9, 10). India has predominantly used paper-based health records for a long time. Transitioning from this traditional system to digital healthcare poses a considerable challenge to EMR adoption (10). In response to the growing need for digitization in healthcare, the Ayushman Bharat Digital Mission was launched in September 2021 (11)[1]. This patient-centric EHR system allows individuals to link their medical records to a unique ID, facilitating secure sharing and uploading of medical data, including reports, prescriptions, and scans (11). Hospitals can also upload medical records linked to the patient's ABHA ID. The benefits of using ABHA is summarized here are summarized in the Appendix. While large hospitals have seamlessly integrated ABHA with their existing EMR systems[2], smaller clinics still face significant challenges in adopting this system. One major reason is the reluctance of healthcare workers towards digital healthcare systems (12). This reluctance is often due to concerns about interoperability, productivity loss, and the complexity of the technology (13).

During the preliminary research for this paper, a survey was conducted to understand doctors' opinions on Electronic Medical Records (EMRs) and their adoption in India. The results of the survey are displayed in the Appendix. Amongst the doctors who took part in this survey, 75% used EMR systems at their workplace. Although this number is high, their effectiveness was not as significant. A majority of respondents described the current state of EMR adoption in India as poor, with private hospitals having the highest adoption rate and clinics having the lowest. Amongst the listed obstacles for EMR adoption, most doctors selected all available options, with resistance to change receiving the highest votes, followed by high implementation costs and lack of awareness of EMRs.

In recent years, AI has revolutionized healthcare, offering innovative solutions to long-standing challenges. AI applications, from disease diagnosis to personalized medicine and predictive analytics, streamline administrative tasks, improve diagnostic accuracy, and enhance patient outcomes. LLMs have shown exceptional capability in processing complex language structures (14). Integrating LLMs into EMR systems can automate tasks, enhancing productivity and healthcare outcomes (14). Despite these benefits, resource constraints in smaller Indian clinics pose challenges for EMR adoption. Recent studies have shown that although Large Language Models require a lot of storage and computing power, Lightweight versions of the models have excelled in similar tasks across the healthcare domain(15, 16).

[1] `https://abdm.gov.in/`

[2] `https://pib.gov.in/PressReleasePage.aspx?PRID=1989763`

The research aims to bridge the gap between healthcare providers and EMR adoption in India, by leveraging technologies like Lightweight Large Language Models (LLMs) and Automatic Speech Recogniton (ASR) tools, overcoming the barriers of finance, lack of technical knowledge, and infrastructure. The central research question guiding this study is:
"How can healthcare providers effectively utilize technology to generate Electronic Medical Records (EMRs) from doctor-patient conversations in low resource settings like small Indian clinics, to promote digitization and adoption of Electronic Medical Records in India?

The sub-research questions under this would be:

- What are the essential features and components for developing an application that enables EMR generation from doctor-patient conversations?
- How do different Lightweight LLMs compare in performance in generating EMRs from spoken conversations?
- What are the key features and costs of different Automatic speech recognition(ASR) tools, and how suitable are they for translating and transcribing conversations in Indian languages?
- What are the key technical requirements for implementing and deploying the application in small clinics in India?
- What are the cost implications of deploying and maintaining the application, including the integration of Lightweight LLMs and Automatic Speech Recognition tools in India?
- How do doctors perceive the application in terms of its usability and its impact on digital healthcare in India?

The main idea is to design and develop an application which can generate Electronic Medical Records(EMR) from doctor-patient conversations. Since, India is a multilingual country, the paper focuses on using Automatic speech recognition(ASR) tools to translate and transcribe the conversation from Indian languages to English. The transcribed conversation is then used by the lightweight Large Language Model to generate a Electronic Medical Record(EMR). The generated Electronic Medical record can be directly uploaded to ABHA, the patient-centric platform in India. Different LLMs and ASR tools are further compared to choose the best performing model for the use case. The technical and financial aspect of the application is further analysed and the findings will be reported to healthcare workers to understand their opinion on the approach. Since, the application is used in a low resource environment, lightweight versions of the LLMs are considered. They often require less storage and compute power.

The code for this research is available on GitHub[1] as an open-source repository. The code includes detailed instructions for replicating the research idea. The repository is licensed under Creative Commons. This permits further collaborations and adaptations, provided proper acknowledgment is given.

The integration of LLMs and Automatic Speech Recognition(ASR) tools into EMR systems represents a significant advancement in healthcare technology. This study not only aims to showcase the potential of these technologies in overcoming existing barriers but also to provide a comprehensive understanding of their practical implementation and impact in real-world settings in India. By addressing the technical, financial, and perceptual challenges in the Indian healthcare landscape, this research seeks to pave the way for more widespread adoption of EMR systems in small clinics across the country. Ultimately, the findings from this research could contribute to improving the efficiency and quality of healthcare delivery in India's resource-constrained environments. The successful implementation of such technologies could also serve as a model for other developing regions facing similar challenges.

[1] `https://github.com/Sachin498/EMR-using-LLMs-and-whisper`

2

Background

2.1 Electronic Medical Records

An Electronic Medical Record (EMR) is a digital version of a patient's health record within a healthcare enterprise, which can include hospitals, clinics, or any other health authority (2). A typical EMR structure includes patient data such as patient demographics, medical history, diagnoses, medications, allergies, patient charts, and various statistics and reports (2, 7). On the other hand, Electronic Health Records (EHR) are digital patient-centered medical records that not limited to one single enterprise and are interoperable across different health settings(14). Although EMRs and EHRs differ in scope and interoperability, these terms are often used interchangeably despite their distinct meanings.

Electronic Medical Records (EMRs) offer numerous benefits to healthcare providers and patients alike. They help physicians identify diseases based on symptoms and characteristics discovered in reported studies, enhancing diagnostic accuracy (2). EMRs eliminate the risk of missing medical files and improve physicians' productivity by allowing access to medical information before patient encounters, thus avoiding misunderstandings due to illegible handwriting (2, 7, 13). Paper-based medical records are often constrained by being time-consuming, labor-intensive, and prone to errors, duplication, and reduced quality (3). EMRs address these issues by eliminating the need for paper-based records, saving space, operational costs, time, and human effort. Furthermore, EMRs facilitate better coordination among healthcare providers by making patient information easily accessible. This streamlined documentation process reduces the time spent on paperwork, allowing healthcare professionals to focus more on patient care. The integration of EMRs into healthcare practices supports more efficient healthcare delivery, improves patient safety, and contributes to overall cost savings.

The adoption of Electronic Medical Records (EMRs) has seen substantial growth since their inception. The first EMR was developed in 1972 by the Regenstreif Institute in the United States, marking a significant advancement in medical practice(17).By the mid-2000s, government initiatives and incentives began to play a crucial role in accelerating EMR adoption. In the United States, the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 provided financial incentives for healthcare providers to adopt EMRs(4, 17).In several European countries, adoption rates of EMRs among general practitioners are impressively high: Sweden leads with 90%, followed by the Netherlands at 88%, Austria at 62%, Denmark at 56%, and Finland at 55%(8). Despite the impressive growth and benefits of EMRs, challenges remain, particularly in developing countries where financial constraints, lack of technical support, and resistance to change can impede adoption. In sub-Saharan Africa, the lack of financial incentives , poor electricity supply and limited internet connectivity and technical skills, hinder the EMR adoption in some regions(8). However, the ongoing evolution of EMR technology, the integration of AI, and the increasing focus on interoperability and user-friendly interfaces are likely to overcome these barriers, further solidifying the role of EMRs in transforming healthcare worldwide.

2.2 India and EMRs

The adoption and implementation of Electronic Medical Records (EMRs) in India present a unique set of challenges and opportunities. India, with its diverse healthcare landscape, faces significant hurdles in fully integrating EMRs across its medical institutions. India has predominantly relied on paper-based health records for many years. Transitioning from this traditional system to digital healthcare has been a considerable challenge. In India, the adoption rate was only 15% in 2018[1], and even today, only 35% of hospitals have fully integrated EMR systems. The slow adoption rate can be attributed to several factors, including financial constraints, lack of technical support, and resistance to change among healthcare providers(9). Apart from this there are other factors like time-intensive, inadequate planning, privacy concerns and labor intensive data migrations(10).

Government Initiatives and support

In 2013, the Ministry of Health and Family Welfare (MoH&FW) notified the Electronic Health Record (EHR) Standards for India, drawing from global best practices and expert

[1] `https://blog.onfit.ai/emr-in-india/#:~:text=India%27s%20EMR%20market%20is%20mirroring,within%20the%20Indian%20healthcare%20landscape.`

recommendations[1]. These standards, supported by various professional bodies, regulatory agencies, and stakeholders, marked a crucial step towards EHR adoption. The standards were expected to evolve, necessitating periodic revisions based on practical experience and future needs, leading to the publication of the EHR Standards 2016 document. At the same time, 'Digital Information Security in Healthcare Act' (DISHA) was proposed to regulate the collection, storage, and use of digital health data, including EMRs[2].

In 2018, the National Health Stack (NHS) was proposed as a digital infrastructure for a nationwide integrated health information system, including EMRs(12). The National Digital Health Mission (NDHM), launched in 2020, aimed to create a comprehensive digital health ecosystem. To further these goals, the Ayushman Bharat Digital Mission (ABDM) was launched in September 2021(11). ABDM facilitates the seamless sharing of health records between patients and healthcare providers, based on consent. The cornerstone of this mission is the Ayushman Bharat Health Account (ABHA), a patient-centric medical data management tool[3]. ABHA is a 14-digit unique health identification number that allows individuals to access and share their digital health records. Registered patients can upload medical documents linked to their ABHA and, during doctor visits, can share their records via a simple QR code scan, with the hospital also able to upload medical documents, prescriptions, and reports. These initiatives demonstrate the Indian government's commitment to advancing healthcare digitization. Continued support and innovation are essential to fully realize the benefits of digital health records in India.

2.3 AI and EMRs

Artificial intelligence (AI) has been a transformative force in many sectors, and healthcare is no exception. In this spectrum, Large Language Models (LLMs) have proven to be exceptional in transforming healthcare. LLMs are machine learning models trained on extensive datasets to understand human language and perform various Natural Language Processing (NLP) tasks. LLMs support a variety of tasks in Electronic Medical Record (EMR) systems, including clinical documentation, extracting meaningful information from clinical notes, disease prediction, clinical text generation, automating administrative tasks, and enhancing EMR-integrated systems.(14, 18, 19).

ChatGPT, one of the notable LLMs, can recognize and diagnose patients, make preventive recommendations, and enhance healthcare. In the fall of 2023, another media coverage

[1] `https://main.mohfw.gov.in/sites/default/files/17739294021483341357.pdf`
[2] `https://main.mohfw.gov.in/?q=newshighlights/comments-draft-digital-information-security-health-care-a`
[3] `https://abdm.gov.in/`

highlighted its potential: 17 doctors over three years were unable to diagnose a 4-year-old boy with a rare umbilical cord syndrome. The disease was finally correctly recognized by ChatGPT when a desperate mother entered the symptoms and test results.[1].

As of 2021, 42% of healthcare organizations in the European Union were using AI technologies for disease diagnosis, risk prediction, or personalized treatment[2]. Despite the high adoption rate, it is comparatively low compared to the US and other departments because of EU's cautious stance regarding the potential negative impacts of AI in healthcare. While LLMs hold significant potential in transforming healthcare and EMRs, their implementation comes with challenges and concerns. Ensuring the confidentiality and security of patient data is paramount. The integration of LLMs with EMRs must comply with strict data protection regulations to safeguard sensitive health information(18). Moreover, LLMs can inadvertently perpetuate biases present in their training data. Ensuring fairness and equity in healthcare AI applications requires rigorous testing and validation to mitigate bias and promote unbiased decision-making(14, 18, 19). Since healthcare is a highly sensitive and critical area, LLMs need to be validated thoroughly to prevent medical errors.

The EU AI Act is a comprehensive regulation that governs AI adoption across sectors, including healthcare, ensuring AI solutions meet safety, transparency, and ethical standards. It imposes obligations on developers and users to uphold patient confidentiality, data security, and unbiased decision-making[3]. Similarly, the United States' HITECH Act and FDA regulations focus on health data interoperability and AI-based medical device safety(20). In Asia, countries like China and Japan emphasize robust standards for AI in healthcare to ensure data security and patient privacy(20). Globally, these regulations are critical for maximizing AI's potential in healthcare while maintaining safety, equity, and efficiency, and fostering international collaboration and best practice sharing.

[1] https://ictandhealth.com/news/2023-in-review-ai-is-everywhere-at-once-but-still-not-in-healthcare
[2] https://www.statista.com/statistics/1312566/adoption-stage-of-ai-in-healthcare-in-the-eu/
[3] https://artificialintelligenceact.eu/

3

Related Work

Large Language Models (LLMs) in the healthcare domain, especially since the advent of Electronic Medical Records (EMRs), have opened numerous opportunities for innovation. Some of the key applications of AI in this domain include information extraction, text similarity, summarization, classification, diagnosis, entity recognition, and prediction(19). The thesis focuses on entity recognition from doctor-patient conversations, a task that has been extensively researched in the context of unstructured clinical text. Numerous studies have explored various techniques for entity recognition tasks(15, 21, 22).

Models such as GPT have been employed for named entity recognition tasks. For instance, in (21), GPT-3.5 and GPT-4 were utilized to extract medical problems, treatments, and tests from unstructured clinical notes, as well as nervous system disorder-related events from safety reports. Initially, these models were compared on these tasks, followed by the application of a prompt framework to enhance their performance. This framework included baseline prompts, annotation guideline-based prompts, error analysis-based instructions, and annotated samples for few-shot learning. The study demonstrated that while GPT models performed well with baseline prompts, their performance improved significantly with the proposed prompt framework.

In another study(22), 26 language models were evaluated for entity recognition tasks in English, Spanish, and French, with an assessment of the environmental impact of the experiment. The models were categorized into two types: masked language models, trained to predict randomly selected masked words in a large text, and pre-trained causal language models, which are larger models. Various prompting techniques, such as few-shot learning and tagging prompts, were compared across all languages. The study concluded that while some masked language models performed well in general text entity recognition, their performance was suboptimal with clinical text, and few-shot learning did not enhance performance in entity recognition tasks. This discrepancy was attributed to the different languages used and the lack of language-specific training for the LLMs.

Similarly, in (15), a GPT-3 model named "text-davinci-003" was used for entity recognition from clinical notes. Clinical notes from pediatric emergency rooms at Hacettepe University children's hospitals were utilized. The model's performance was analyzed before and after fine-tuning with domain-specific data. The fine-tuned model achieved an accuracy of 99.96%, outperforming the base model's accuracy of 78.54%. This study highlighted the cost savings and time efficiency of using LLMs for this task compared to manual human efforts.

In (16), Llama2, a lightweight LLM developed by Meta, was used to extract five key clinical features related to liver cirrhosis from 500 medical history documents of patients, demonstrating the use of locally deployed LLMs for clinical information extraction. Various versions of the Llama2 model were analyzed using few-shot and zero-shot prompting techniques. The larger version of Llama2, with 70 billion parameters, outperformed the smaller models in all tasks, achieving a sensitivity of 100% and specificity of 96%. The study also found that prompting techniques significantly improved performance.

To enhance the performance of LLMs in entity recognition from clinical text, various techniques and frameworks have been suggested. Prompt engineering, which involves designing and optimizing input prompts for LLMs, has proven to be an effective method for improving performance(23). As previously noted, prompt engineering has shown significant promise in enhancing LLM performance. Another approach to improve NER performance is to enhance the quality of data used to fine-tune models with domain-specific information, as demonstrated in (15). Research indicates that higher-quality data leads to better model performance in these tasks(24, 25). These are just a few of the strategies that can improve entity recognition tasks in the healthcare domain.

In line with the previous research discussed in this section, this thesis focuses on entity recognition tasks. However, instead of clinical texts, we will use doctor-patient conversations during patient visits. The LLMs will process the transcribed conversations and identify all the entities specified in the prompt, thereby generating an Electronic Medical Record(EMR).

4

Research Design

The digitization of healthcare in India face numerous challenges. To address these challenges comprehensively, this research design focuses on specific strategies involving data collection, evaluation metrics and analysis. By examining these areas, we aim to develop a robust, cost-effective, and user-friendly digital solution that can be seamlessly integrated into small clinics across India. The following sections outline the detailed research questions and methodologies employed to achieve our objectives:

- **Research Question 1**:

  This research question focuses on designing an application to generate EMRs from doctor-patient conversations. The primary objective is to create an intuitive and user-friendly design, ensuring that healthcare providers can easily understand and use the application without needing prior technical skills. Each component selected will be carefully evaluated based on its functionality, ease of integration, and overall contribution to the application's goals. The architecture will be designed to ensure smooth data flow and processing, from capturing conversations to generating EMRs.

- **Research Question 2**:

  Large Language Models (LLMs) have demonstrated exceptional capabilities in various Natural Language Processing (NLP) tasks, including entity recognition and information extraction. Despite these benefits, their deployment often entails significant training and infrastructure costs. To address this, our research focuses on lightweight LLMs, which are smaller in size and come pre-trained, eliminating the need for extensive retraining.

  The lightweight models we have selected for our study include Llama 2-7b, Llama 3-7b, and Mistral-7b(26). Trained on 7 billion parameters, each of these models offer high performance and resource efficiency(27). These models are utilized to generate Electronic Medical Records (EMRs) from doctor-patient conversations. To evaluate their performance, we use a dataset comprising 50 transcripts of simulated medical interactions. Each model is prompted to generate EMRs from these conversations, and their outputs are compared against manually annotated ground truth records.

The dataset used and evaluation of these models are explained further in next section. Additionally, we experiment with different combinations of hyperparameters for each model to optimize their performance further. This comprehensive evaluation allows us to identify the most effective model for generating EMRs from conversational data, ensuring that our application can leverage the best available technology to promote digitization in Indian clinics.

- **Research Question 3**:
  Given the predominant training of most Large Language Models (LLMs) on English data, there exists a significant gap in support for Indian languages. To address this, the effective translation of patient-doctor conversations into English is crucial for generating Electronic Medical Records (EMRs) in Indian clinics. Various Automated Speech Recognition(ASR) tools offer robust translation and transcription capabilities that can be leveraged for this purpose.

  To determine the most suitable speech-to-text API for our application, we begin by identifying several widely-used ASRs that support multilingual features, particularly focusing on those that include Indian languages. Our evaluation considers key features such as the ASR tools' ability to accurately translate Indian languages into English, the range of languages supported, and the ease of integrating these Automated Speech Recognition(ASR) tools into our system. This involves reviewing documentation, API usage limits, and language support. Cost is a significant factor for small clinics, so we compare the pricing models of each tool, considering subscription costs, pay-as-you-go options, and additional charges for multilingual features.

  Through this comprehensive analysis, we aim to identify the most suitable ASR tool that balances these factors effectively. The findings from this research will guide the selection of the best API to support the translation and transcription needs of our application, thereby promoting effective EMR generation.

- **Research Question 4**:
  The adoption of digital healthcare technologies in small clinics in India is hindered

by perceived complexity and lack of technical expertise among healthcare workers. Addressing these challenges is critical to overcome reluctance towards technology and ensure successful deployment of healthcare applications. This research question aims to identify the essential technical infrastructure and resources necessary for the smooth implementation and operation of such applications in resource-constrained settings.

The main objective include identifying specific hardware and software requirements crucial for deploying the healthcare application effectively in small clinics in India. The study will help assess the feasibility of integrating the application into existing systems within Indian clinics, considering technological capabilities and constraints. The outcome can be used to help promote the use of technology in digital healthcare in India.

- **Research Question 5**:
  Another major barrier to adopting technology into healthcare in India is the financial concern associated with setting up infrastructure and maintaining software. This includes costs related to initial deployment, ongoing maintenance, and integrating into existing technologies. The primary objective of this research is to identify and analyze the cost implications of deploying and maintaining a healthcare application that integrates lightweight large language models (LLMs) and Automated Speech Recognition(ASR) tools in Indian clinics. This involves assessing the initial setup costs, ongoing maintenance expenses, and the financial impact of these technologies.

  Understanding these costs will help determine how to implement an EMR-generating application with an LLM and Automated Speech Recognition(ASR) tools at a low cost. By providing a clear breakdown of expenses and cost-effective strategies, this research aims to promote the use of EMRs in India. The findings will be presented to healthcare workers demonstrating how affordable and sustainable technology integration can be achieved in small clinics, encouraging broader adoption and enhancing healthcare delivery in resource-constrained settings.

- **Research Question 6**
  Doctors and patients are the primary users of any EMR system, yet there is significant reluctance towards technology in healthcare. Therefore, it's crucial to gather their feedback on the application. In this research, the application will be presented to healthcare workers, and their feedback will be recorded. Additionally, they will be

informed about the findings from research questions 4 and 5, which address the technical requirements and cost implications of implementing the system. This will help healthcare workers understand how the application functions and the necessary resources for its deployment. The healthcare workers will then be presented with a questionnaire which has a structured interview. This will provide valuable feedback which can be used improve the application.

Overall, this will help the doctors understand the need and use of such an application and how it can be used to generate EMRs easily in a resource-constrained setting. Their feedback will provide insights into their perceptions of the application and identify areas for improvement. Understanding their viewpoints will be instrumental in enhancing the application's usability and effectiveness. Moreover, this approach will help promote acceptance among healthcare workers in small clinics in India by demonstrating the feasibility and benefits of the technology, ultimately fostering greater adoption of EMR systems.

5

Methodology

This section provides a comprehensive explanation of the technical specifications and research design implemented in this study. It details the data utilized, the evaluation metrics applied to address each research question, and the design and components of the developed application. By elaborating on these aspects, the aim is to present a clear and thorough understanding of the methodology employed.

5.1 Data collection

To evaluate how different Lightweight LLMs perform in generating EMRs, we utilized a simulated patient-doctor dataset comprising audio recordings and corresponding transcriptions of medical conversations.[1]. This dataset focuses on medical cases related to different conditions and includes 272 mp3 audio files and 272 corresponding transcript text files. Each file is categorized into one of five medical conditions: respiratory (RES), gastrointestinal (GAS), cardiovascular (CAR), musculoskeletal (MSK), and dermatological (DER). The file naming convention consists of three letters indicating the medical condition and four digits representing the case number.

Due to the necessity of manually annotating the ground truth for accurate evaluation, we selected a subset of 50 records from the dataset. This selection included 5 cardiovascular cases, 17 musculoskeletal cases, 1 dermatological case, 20 respiratory cases, and 6 gastrointestinal cases, chosen to reflect the distribution of cases in the dataset. For evaluating the performance of different LLMs, we focused exclusively on the transcribed text files from

[1] `https://springernature.figshare.com/collections/A_dataset_of_simulated_patient-physician_medical_interviews_with_a_focus_on_respiratory_cases/5545842/1`

these selected records. Refer to the Appendix for a detailed view of the data and annotated ground truth.

5.2 Evaluation Metrics

This section outlines the criteria and measures used to answer the research questions. It ensures that each aspect of the research questions is systematically evaluated to provide clear, quantifiable, and actionable insights.

To evaluate which LLM model performs the best in generating EMRs, the LLMs are prompted to generate EMRs from the transcripts of doctor-patient conversations from the dataset. The generated EMRs from each model are compared with the ground truth, which is a manually generated EMR from the same dataset. Each model, with its base configurations and parameter-tuned configurations, undergoes this evaluation. To compare these models, we will be using F1-scores, precision, recall, cosine similarity, and Levenshtein distance. These metrics are commonly used for evaluating an LLM's performance. F1-scores, precision, and recall fall under the multi-classification metrics, helping us evaluate how LLMs perform in classifying texts into groups with labels(28). Cosine similarity, on the other hand, helps understand the semantic and literal aspects of text similarity. Each of these metrics is further described below:

- Precision:
  Precision is the ratio of correctly predicted positive observations to the total predicted positive observations(28). It helps understand how many of the selected entities by the model are relevant. High precision means fewer irrelevant entities are identified, which is crucial in medical contexts where incorrect information can lead to wrong treatments.

  $$\text{Precision} = \frac{TP}{TP + FP} \tag{5.1}$$

  where TP is the number of true positives and FP is the number of false positive.

- Recall:
  Recall is the ratio of correctly predicted positive observations to all the observations in the actual class(28). It helps understand the LLM's ability to identify positive instances. High recall ensures that most or all relevant medical information is captured, which is vital for comprehensive patient records.

  $$\text{Recall} = \frac{TP}{TP + FN} \tag{5.2}$$

where TP is the number of true positives and FN is the number of false negatives.

- F1-score:
  The F1 score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall(28). Ranging from 0 to 1, the F1 score is particularly useful when balancing the trade-off between precision and recall, especially in cases where both false positives and false negatives need to be accounted for.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

- Cosine similarity:
  Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space(28). It quantifies how similar two vectors (texts) are, regardless of their magnitude. Cosine similarity is widely used in information retrieval and text mining, as it captures the orientation (similarity in terms of direction) rather than magnitude. It is effective in comparing document similarity, especially in high-dimensional spaces like word embeddings.

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.4}$$

  here $\mathbf{A}$ and $\mathbf{B}$ are the vector representations of the two texts being compared.

- Levenshtein distance:
  Levenshtein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another. In the context of this study, it can be used to quantify the character-level similarity between predicted entity mentions from an LLM and the ground truth entity mentions. A lower Levenshtein distance between the predicted and actual entity mentions indicates higher character-level accuracy.

$$\text{Levenshtein Distance}(a,b) = \begin{cases} \max(|a|,|b|) & \text{if } \min(|a|,|b|) = 0 \\ \min \begin{cases} \text{LD}(a_{1..m-1}, b) + 1 \\ \text{LD}(a, b_{1..n-1}) + 1 \\ \text{LD}(a_{1..m-1}, b_{1..n-1}) \\ +[a_m \neq b_n] \end{cases} & \text{otherwise} \end{cases} \tag{5.5}$$

  where $|a|$ and $|b|$ represent the lengths of the strings a and b, and LD represents the Levenshtein distance between substrings.

These evaluation metrics will provide a comprehensive framework to assess the effectiveness of the LLMs in generating EMRs from doctor-patient conversations, ensuring that the chosen model is the most suitable for real-world application in Indian clinics.

For research question 3, we will be analyzing and comparing the features of various speech-to-text models. This includes evaluating their support for multilingual capabilities, especially for Indian languages, and assessing their cost-effectiveness for our application. For research questions 4 and 5, we will analyze and report on the technical requirements and cost implications for implementing and maintaining the application in small clinics in India. This involves a detailed examination of the infrastructure needs, software requirements, and potential financial impacts. For research question 6, we will use a survey to gather feedback from healthcare workers regarding the usability and impact of the application. The responses will be systematically analyzed to gain insights into their perceptions and experiences. The findings from these analyses will be presented and discussed in the results and discussion section of the thesis.

6

Results

The findings of the experiments across all the sub-research questions are summarized in this section.

6.1 Application design

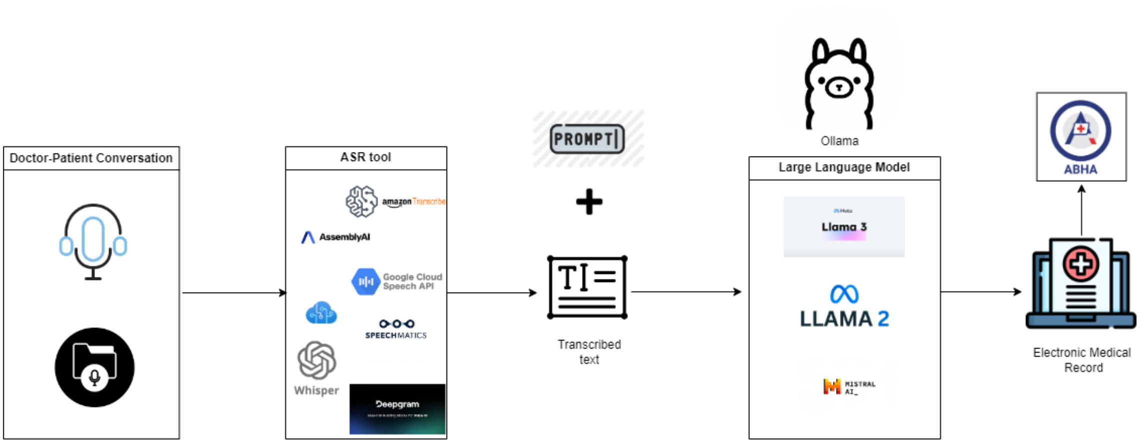


Figure 6.1: Application Design

The figure(6.1) shows the design of the entire application. It starts with the doctor-patient conversations being either recorded or uploaded. The audio is then given to an ASR tool, the ASR tool translates and transcribes the audio to English. The transcribed text is provided to the Large language model along with the prompt. The generated output is an EMR in the specified format. The generated EMR can be uploaded to the ABHA platform which is the EMR platform in India.

The components are further discussed in the following subsections:

6.1.1 Components

6.1.1.1 Audio conversation

This will be the audio from the patients visit to the doctor. The conversation during the entire visit is either recorded live or it can be recorded and uploaded to the application. The audio can be in any Indian language. The audio will then be provided to the ASR tool for further processing. For the sake of our demo, we are only using sample audios in Hindi.

6.1.1.2 ASR tools

In India, there are around 398 languages amongst which 28 are recognised by the Constitution of India. Since, most of the LLMs are predominantly trained on English datasets and less on Indian languages, to achieve higher accuracy in performance, the audio containing the doctor-patient conversation needs to be converted to English. Automatic Speech Recognition(ASR) tools are required to translate and transcribe the audio into English. The ASR tools needed for the use case should support Indian languages. The tools considered are summed up here[1]. Each of these ASR tools are compared for different features and cost to see which tool is best suited for our use case. The best ASR tool for our use case is then used to translate and transcribe the audio into English. This transcribed text is then given to the Large language Models for further processing.

6.1.1.3 Prompt

The application communicates with the Large Language Models through input prompts. The users can mention the features that need to be identified and extracted from the transcribed text in the input prompt. The prompt given to the Large Language model will usually contain the task at hand which is identifying the features mentioned by the healthcare providers, the transcribed text and the format in which the output is expected. The users can also add examples of the input and output to improve performance and also give specific instructions to avoid edge cases and ensure consistency.

[1] `https://deepgram.com/learn/best-speech-to-text-apis`

6.1.1.4 Large Language Models

Large language Models(LLMs) is predominantly known to have advantages in NLP tasks. In our case LLMs will be used to extract relevant information from the transcribed text. This is known as Entity recognition. Since we are considering using LLMs in a low resource environment, only lightweight LLMs are considered. Under the umberella of Light weight LLMs, we are using Llama2, Llama3 and Mistral. All of these models are trained on 7 billion parameters. They are well known in NLP tasks, especially clinical entity recognition. These LLMs will be used to generate EMRs from the transcribed text. It will be provided with the transcribed text and a prompt. Each of these LLMs are further hyperparameter tuned to achieve better performance. Hyperparameter tuning is known to have a positive impact on the performance of LLMs(29). Each of these hyperparameters affect the LLM and its use differently[1]. Different combinations of these hyperparameters were used to see which combination performs the best in generating EMRs. Each of the models hyperparameter configurations along with its default configurations are summarised in the Table 6.1.1.4. The hyperparameter tuned models will be addressed as "ModelName-optimal"

Model	**Temp**	**num_ctx**	**mirostat**	**top_k**	**top_p**
Mistral	0.5	4096	1	-	-
Llama3	0.5	4096	1	-	-
Llama2	0.5	4096	0	20	0.8
Default	0.8	2048	0	40	0.9

Table 6.1: Hyperparameters configuration for Different Models after tuning

in the further sections of the paper. These LLMs are further compared in performance in generating EMRs from transcribed conversations. The best performing LLM with the optimal parameters are considered for the final application design.

[1] `https://github.com/ollama/ollama/blob/main/docs/modelfile.md#valid-parameters-and-values`

6.2 Performance comparison of LLMs

Model	Precision	Recall	F1-score	Levenshtein's distance	Cosine similarity
Llama 2	0.802	0.419	0.329	0.54	0.079
Llama 3	0.829	0.327	0.287	0.576	0.154
Mistral	0.825	0.41	0.342	0.497	0.154
Llama 2 - optimal	0.802	0.307	0.29	0.654	0.09
Llama 3 - optimal	0.888	0.421	0.412	0.442	0.229
Mistral - optimal	0.859	0.361	0.365	0.518	0.195

Table 6.2: Evaluation metrics for different lightweight LLMs

In this section, the results of our evaluation of different Lightweight Large Language Models (LLMs) used for generating Electronic Medical Records is presented. From the data presented in Table 6.2, it is evident that the tuned versions of the LLMs, denoted as "optimal," generally outperform their untuned counterparts across all metrics. Llama 3-optimal achieved the highest precision (0.888) and F1-score (0.412), indicating its superior ability to generate accurate and relevant text. Mistral-optimal followed closely with a precision of 0.859 and an F1-score of 0.365. In terms of Recall, the optimized versions showed improvements, though scores remained lower than Precision, indicating that while accurate, some relevant information might still be overlooked. Levenshtein's distance scores, particularly low for Llama 3-optimal (0.442), suggest close alignment with the ground truth, requiring fewer corrections. Cosine similarity, which measures semantic meaning and context, was also highest for Llama 3-optimal (0.229) and Mistral-optimal (0.195), suggesting these models better captured the nuances of the input text. Overall, our results indicate that tuning LLMs significantly improves their performance in generating EMRs, with Llama 3-optimal and Mistral-optimal emerging as the top performers.

6.3 Automatic Speech Recognition(ASR) tool comparison

Name	**Language support**	**Indian languages**	**Cost**	**Speed**
Google speech-to-text	125 languages	11 languages supported	Free 60 minutes/month, followed by \$0.016/minute	Transcribes a 30-second audio in 15 seconds
AssemblyAI	102 languages	14 languages supported	Free 100 hours, other tiers starting at \$0.002/minute	Transcribes up to 2 hours of audio in 45 seconds
Amazon transcribe	102 languages	10 languages supported	60 minutes/month for a year, \$0.024/ minute for 250,000 min	Transcribes 1 hour of audio in 20-30 minutes
Microsoft Azure	117 languages	11 languages	\$200 credit. After that, pay as you go	Transcribes 1 hour of audio in 10-20min
Deepgram	36 languages	Only Hindi	\$200 credit. After that, pay as you go	Transcribes 1 hour of audio in 30 seconds
SpeechMatics	50 languages	3 languages	Free 8 hours per month, followed by \$0.005/minute	Transcribes 1 hour of audio in 3.4 minutes
Whisper	98 languages	8 with good performance	Free and open source	Transcribes 1 hour in 15.93s

Table 6.3: Comparison of different ASR tools

Table 6.3 compares various ASR tools across different features. Google Speech-to-Text supports the most languages but has a slower speed. AssemblyAI leads in Indian language support but lacks performance benchmarks. Amazon Transcribe, Microsoft Azure, and Deepgram are pricier options. SpeechMatics offers limited Indian language support at a lower cost. Whisper stands out with robust Indian language support, proven performance[1],

[1] `https://github.com/openai/whisper`

and fast transcription speeds. It is an open-source solution that can be set up locally. Whisper is found to be the most viable option for the application with its high performance and low resource requirement. Whisper also has a paid API version which costs around $0.006 per minute.

6.4 Final Application Design

Now that the most suitable LLM and ASR tool is found, the final application design is as shown in the Figure 6.2. Whisper is chosen as the most suitable ASR tool for our use case. We will be using Llama3 with the optimised hyperparameters as our LLM.

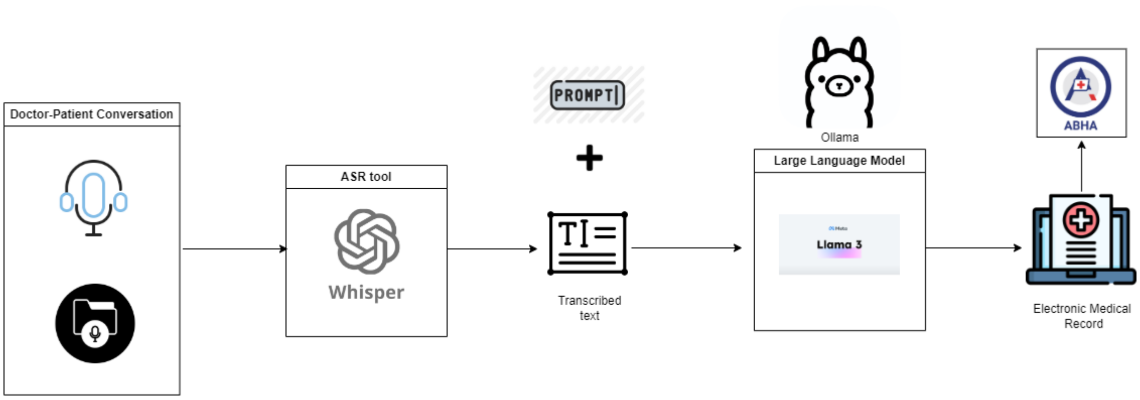


Figure 6.2: Final Application Design

The UI of the application is as seen below in Figure 6.3. It has options to choose the fields to retrieve from the transcribed conversation. The users can either browse through recordings or also record the conversation live during the patient's visit. The transcribe button allows the user to translate the conversation into English using Whisper.

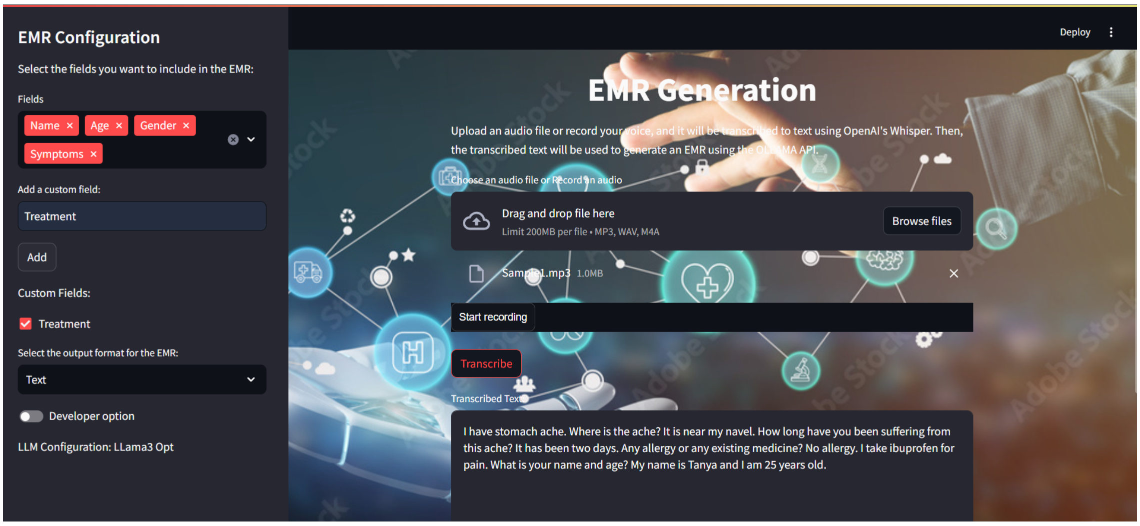


Figure 6.3: UI of the application

The user will then get to see a prompt along with the option to make changes to the prompt. This option is only if the developer option is enabled in the side panel. On changing the prompt, if required and clicking on the "Generate EMR" button, an EMR is generated which can then be downloaded. This downloaded file can be easily uploaded to the ABHA platform. The developer option in the side panel is to choose between different LLM models and also the user gets to update the prompt given to the LLM.

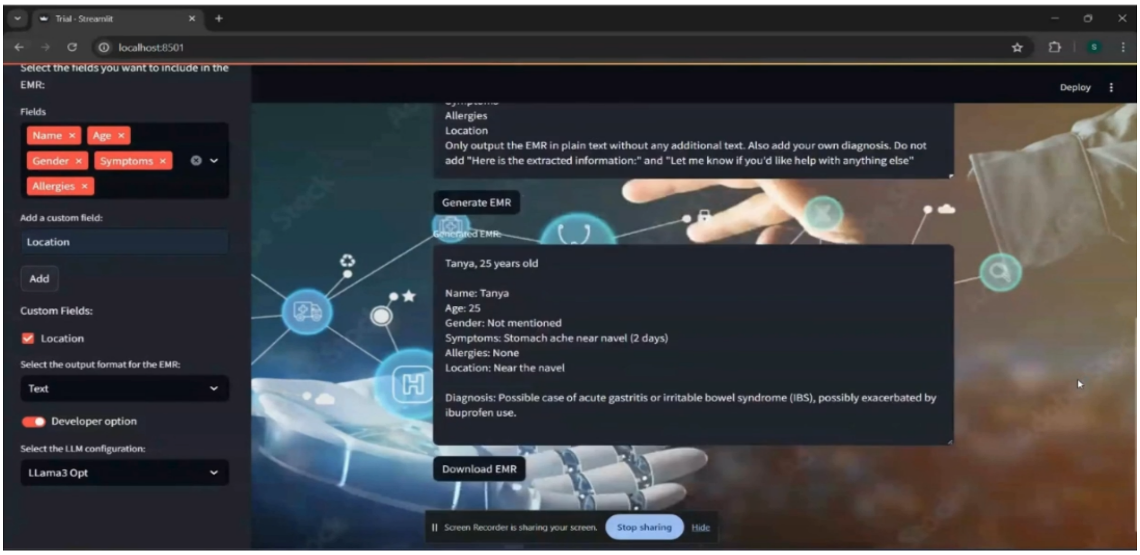


Figure 6.4: Generating EMR with the Application

6.5 Technical requirements for Adaptation

	Hardware requirements	**Software requirements**	**Technical knowledge**
Locally hosted application	**CPU**:Multi-core processor **RAM**: At least 8 GB **Disk Space**: At least 20 GB space **GPU**: Recommended **Network**: Not required	**Operating System**: Windows, macOS, or Linux **Python**: Version 3.8 or higher **GPU drivers**: If using GPU	**Programming**: Basic knowledge to run and manage Python scripts
Local Desktop Application	Same as locally hosted	**Operating System**: Windows, macOS, or Linux **GPU Drivers**: If using GPU	None
Cloud hosted application	No Hardware Requirements **Network**: Internet connection required	**Operating System**: Windows, macOS, or Linux	None

Table 6.4: Technical requirements for implementing and deploying the application

The technical requirements for implementing and deploying the application in small clinics and displayed in the Table 6.5. Three categories are considered: Locally Hosted Application, Local Desktop Application, and Cloud-Hosted Application.

The first category is the Locally Hosted Application, this is basically how the application is set up in the thesis. In this case a computer with very high specifications and a GPU is required. Apart from this, the user also needs some basic programming skills to run the application. The second category is Desktop Application locally deployed. By containerizing the application and packaging it as an APK file, users can simply download and run it without any technical knowledge or skill, though the hardware and software requirements remain the same. The third category is a Cloud Hosted Application. In this scenario, all tools, including the ASR tools and LLMs, are cloud-hosted, requiring only a stable internet connection. This eliminates the need for specific hardware and software requirements, making it the most accessible deployment option.

6.6 Cost implications

	Whisper	**Hardware costs**	**Cloud costs**	**Training costs**
Locally hosted application	Nil	Starting from $600	Nil	Nil
Local Desktop Application	Nil	Starting from $600	Nil	Nil
Cloud hosted application	$0.006 per minute	Nil	Starting at $120 for GCP	Nil

Table 6.5: Cost Implications for Implementing and Deploying the Application

The cost implications for implementing and deploying the application vary across deployment scenarios as seen in the Table 6.6. Locally Hosted and Local Desktop Applications incur hardware costs starting from $600, with no additional cloud or training costs. The hardware cost represents the minimum cost required to run the application. In contrast, the Cloud-Hosted Application, where we rely on the tools running on a cloud service, incurs a cost of $0.006 per minute for Whisper API, with cloud costs starting at $120 for GCP, and no hardware or training costs. Although a free, open-source version of Whisper is available, it handles only a few minutes of conversation. Whisper API can handle more requests, longer audios, and transcribes at a higher speed.

6.7 Doctor review

As part of the research design, the application was presented to healthcare providers alongside the results from Research Questions 4 and 5. The feedback, collected through a survey, was overwhelmingly positive. Healthcare workers found the application easy to use and navigate (Fig. 6.5). A majority of them noted that the application significantly reduces the time spent on EMR documentation (Fig. 6.6). Overall, the doctors viewed the application as a positive step towards the digitization of healthcare in India.

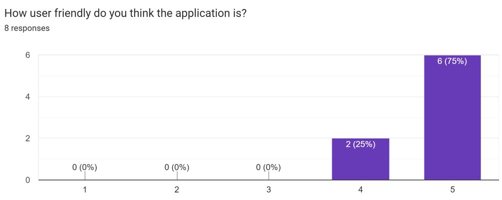


Figure 6.5: Survey on Application's User Friendliness

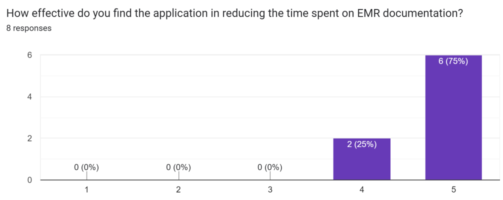


Figure 6.6: Effectiveness in Reducing EMR Documentation Time

The doctors also validated the EMRs generated by the application, finding them to be highly accurate in most cases (Fig. 6.7). There was one instance of lower accuracy due to some missing values in the EMR generated. Importantly, no technical skills or prior training were required to use the application, as confirmed by the healthcare workers (Fig. 6.8). This aligns with their positive feedback on the application's user-friendliness and navigability.

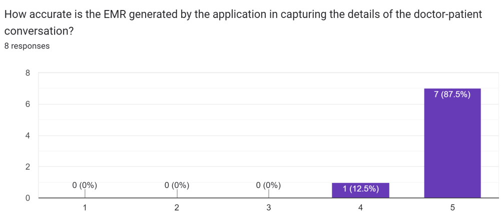


Figure 6.7: Accuracy of EMR in Capturing Doctor-Patient Conversations

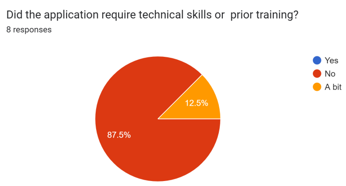


Figure 6.8: Need for Technical Skills or Prior Training

The healthcare workers agreed that this tool could address the challenges of EMR integration in India, potentially promoting better healthcare. They saw potential for the application to replace traditional paper-based record-keeping methods and to change the reluctance of healthcare workers towards technology and EMRs (Fig 6.9,6.10).

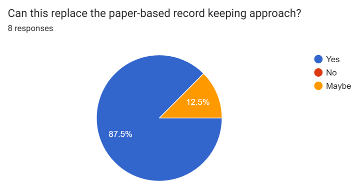


Figure 6.9: Potential to Replace Paper-Based Record Keeping

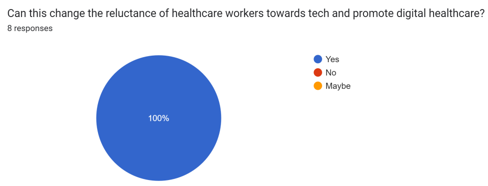


Figure 6.10: Impact on Reluctance Towards Technology Adoption

However, there was some concern about the financial feasibility for smaller clinics (Fig 6.11). When informed about the financial requirements for the application, not all doctors were enthusiastic about its feasibility. One healthcare worker noted that while the integration cost is relatively low, it remains a challenge for smaller clinics that lack the necessary financial resources.

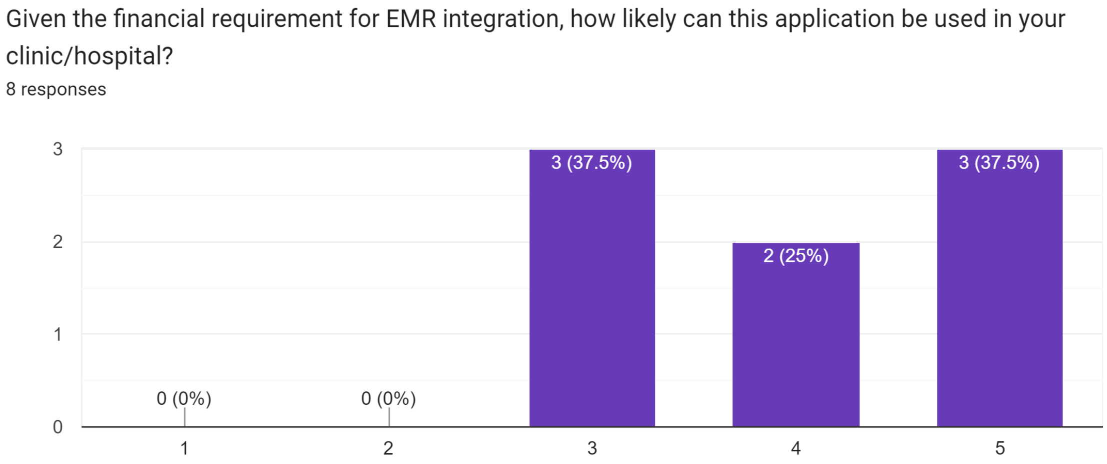


Figure 6.11: Feasibility of EMR Integration Given Financial Constraints

Overall, the feedback highlights the application's potential to enhance the efficiency and accuracy of EMR documentation, while also addressing some of the existing barriers to technology adoption in healthcare.

7

Validation

The validation of the application designed for generating Electronic Medical Records (EMRs) from doctor-patient conversations was conducted through feedback from healthcare providers. In this phase, the application was demonstrated to a group of doctors working in small clinics across India. Their feedback was essential in evaluating the credibility and effectiveness of both the thesis and the application.

The doctors involved in this validation process were residents pursuing their Post Graduate degrees. Their experience stems across various departments, including General Medicine, Pediatrics, Emergency Healthcare, ENT, Surgery, Dermatology, Gastroenterology, and Psychiatry. Most of these doctors have approximately four years of experience in India. While many preferred not to be mentioned in the report, two doctors, Dr. Anagha Teggihal and Dr. Vijaya Sadana Botchu, agreed to provide their names. Dr. Sadana is pursuing her MD, PhD in the department of physiology at Amsterdam UMC, the Netherlands. Dr. Anagha Teggihal is doing her Post Graduate Residency in the department of Internal Medicine from The Wright Center for Community Health, Pennsylvania, USA.

Initial Feedback on EMR Adoption Challenges: The initial interviews focused on validating the issues surrounding the adoption of EMRs. About 60% of the doctors confirmed that the current state of EMRs in India is inadequate, noting that EMR systems are predominantly found in private hospitals, with a critical need for implementation in smaller clinics. They described the reliance on paper-based handwritten notes and the potential of EMRs to significantly streamline administrative processes. Major obstacles identified included high implementation costs, technical infrastructure challenges, the need for technical training, data privacy concerns, lack of awareness, and resistance to change. Specific issues they encountered included internet connectivity problems leading to data loss, insufficient IT support, a lack of understanding of EMR systems, and the time-consuming nature of manual data entry.

Application Review and User Experience: The subsequent review of the application yielded positive feedback. Healthcare providers found the application intuitive and user-friendly. Notable features, such as the ability to add custom fields for retrieval from the text, were well-received. The doctors tested the application with audio samples in Hindi, Telugu, and Kannada, and found that it performed well in Hindi, though there were some inaccuracies in Kannada.

The accuracy of the EMRs generated by the application was affirmed by all interviewed doctors. They reported that the application effectively handled fields not present in the initial conversation, producing relevant responses. Additionally, the feature allowing doctors to review and amend EMRs before saving them was seen as a valuable tool for minimizing errors.

Impact and Concerns: The feedback indicated that the application could potentially replace the current paper-based record-keeping system and enhance EMR adoption rates in India. The doctors validated that the application has the potential to streamline the documentation process effectively.

However, concerns were raised regarding patient consent and transparency. Additionally, there were apprehensions about the risk of generating inaccurate data, which could have serious consequences. Addressing these concerns will be crucial in further developing the application to ensure its reliability and ethical compliance.

8

Discussion

The discussion focuses on the implications of the findings, the performance of the technologies evaluated, and the challenges faced during the research.

The performance comparison of various LLMs reveals that hyperparameter tuning significantly enhances their ability to generate accurate and relevant EMRs. Among the models tested, the Llama 3-optimal and Mistral-optimal emerged as the top performers, achieving high precision and F1-scores, along with superior semantic understanding as indicated by cosine similarity. These findings suggest that optimized LLMs can effectively generate EMRs from spoken conversations. However, the generated EMRs' accuracy is not yet at a 100%, which is crucial given the sensitivity of healthcare data. This discrepancy could be attributed to several factors. Firstly, the ground truth data was manually annotated by the author, who may lack the detailed medical expertise necessary for precise annotation. Secondly, the LLMs utilized in this study were not trained specifically on medical data. While general LLMs have broad language understanding capabilities, they may lack the nuanced understanding required to accurately interpret and document medical conversations.

The analysis of ASR tools presents a variety of options, each with its strengths and weaknesses. Whisper stands out due to its strong support for Indian languages, fast transcription speeds, and open-source availability, making it highly suitable for the Indian context. In our study, the Whisper model proved to be highly successful, translating and transcribing audio quickly and accurately. While Whisper performs well with clear, low-noise audio, it has not been extensively tested for languages other than Hindi. Additionally, the locally deployed Whisper model is limited in its ability to handle longer audio data. Transitioning to the Whisper API offers a viable solution, as it can manage extended audio files more effectively and maintains consistent performance. Alternatively, cost-effective tools like AssemblyAI could be explored, though their performance across diverse Indian languages requires careful evaluation. Leveraging these advanced ASR tools can optimize the application for various linguistic and acoustic conditions, enhancing accuracy and usability in diverse healthcare environments.

The application can be deployed as locally hosted, local desktop, or cloud-hosted solutions. Locally hosted and desktop options demand substantial hardware and technical expertise, while cloud-hosted setups require minimal local resources. Cost analysis indicates that while locally hosted and desktop applications involve initial hardware investments, cloud-hosted applications incur ongoing costs based on usage. Cloud-based applications require minimal technical and hardware infrastructure but have ongoing costs, while locally deployable options need robust infrastructure and higher initial financial burdens. Healthcare entities must balance these factors based on their specific needs and resources.

Privacy Concerns: While healthcare providers have found the application useful and accurate, significant concerns remain regarding the handling of sensitive patient information. If the application is deployed using third-party servers, this would involve transmitting data to external entities, raising critical issues related to data management and visibility, as healthcare organizations may lose control over how their data is stored, used, or protected by these parties. While some organisations have tried to de-identify the data before passing it to the LLM to mitigate privacy risks, complete de-identification is challenging and not always perfect.

Using open-source Large Language Models (LLMs) within a healthcare organization's own infrastructure is an effective approach, as it eliminates the need to send sensitive data to external servers. They still pose security risks such as unauthorized access, data leaks, and potential model poisoning. Furthermore, the use of patient data for training or querying LLMs raises significant questions about patient consent and rights. It is crucial that patients are informed about how their data will be used and that explicit consent is obtained before their data is used. Ensuring transparency and obtaining proper consent are essential steps in addressing privacy concerns and maintaining trust in the application.

Despite the potential benefits of enhanced efficiency and accuracy, the inherent privacy and security risks pose serious challenges. While advanced data protection measures and regulatory compliance offer some mitigation, the effectiveness of these safeguards in practice remains debatable. Can the benefits of LLM-driven efficiency outweigh the risks to patient privacy, or do these risks render the deployment of such applications too perilous? The balance between technological progress and data protection continues to be a contentious issue, requiring further advancements in technical and societal practices.

9

Conclusion

This study explored how doctors can generate Electronic Medical Records by effectively utilizing Lightweight Large Language Models and Automatic Speech Recognition tools. The aim was to promote digitization and foster acceptance among healthcare providers and patients in small clinics across India.

Feedback from healthcare providers highlighted positive perceptions regarding the application's usability and its potential to advance digital healthcare adoption in India. Doctors validated the accuracy of EMRs generated, suggesting it could replace paper-based record keeping in India. While many doctors found it feasible to integrate the application into existing EMR systems in India, some expressed concerns about the financial aspect of employing such a solution. Although the initial costs appear manageable, deploying and maintaining such an application in rural healthcare facilities could be challenging. One potential solution is to commercialize the application, but this would likely increase costs and privacy concerns.

To address the financial and logistical challenges, government intervention is crucial. Support from governmental bodies could facilitate broader deployment and encourage use across the country. Furthermore, ensuring the privacy and security of patient data is essential. The application must adhere to regulations set forth by the government regarding AI, EMRs, and patient information. Future work should focus on implementing robust privacy measures, including de-identification, data masking, and enhanced security protocols before bringing it into the real world.

This research not only highlights the potential for improving healthcare documentation in India but also provides a framework that can be adapted for other countries. By addressing the technical, financial and ethical challenges specific to different regions, this approach can be tailored to meet the unique needs of various healthcare systems worldwide. The application holds significant promise for healthcare workers with EMR generation, thereby fostering digitization and improving healthcare delivery. The key to successful implementation lies in finding the right balance between technical infrastructure and financial feasibility, while ensuring the confidentiality of patient data and effectively leveraging Lightweight LLMs and ASR tools. Potential support from governmental and non-governmental organizations is crucial to ensure accessibility and sustainability of these advanced technologies in healthcare settings. By overcoming these challenges, healthcare providers and patients stand to benefit from more efficient, accurate, and accessible medical record-keeping solutions.

10

Limitations and Future Research

While the application shows promise for healthcare workers, several limitations need to be addressed before it can be effectively implemented in India.

10.1 Privacy and Security Concerns

Privacy is a critical concern when integrating AI into healthcare, particularly regarding Electronic Medical Records (EMRs)(30). EMRs contain sensitive patient information, making it essential to ensure the privacy and confidentiality of this data.

In our application, which processes sensitive patient data through LLMs, privacy concerns are paramount. One strategy to address these concerns is de-identifying patient data before it is input into the LLM. De-identification involves removing any personal identifiers that could link the data to individual patients. Various techniques for de-identification exist and should be applied to mitigate privacy risks.

When LLMs are hosted by third-party vendors, it is crucial to establish Business Associate Agreements (BAAs) to ensure compliance with HIPAA requirements and protect patient data. Additionally, it is essential to obtain explicit consent from patients before their data is used in the application. Patients should be fully informed about how their data will be collected, used, and protected. Clear communication regarding data handling practices helps maintain trust and ensures that patients are aware of and agree to the use of their information.

Strong security measures must also be implemented to safeguard sensitive data. Encryption and access controls are essential for preventing unauthorized access and data breaches.

In India, specific regulations and standards for EMR systems have been established[1]. Adherence to these regulations is crucial for integrating our application effectively. This includes securely handling audio data from doctor-patient conversations and the generated EMRs to prevent misuse.

Ensuring the privacy and security of patient data involves a multifaceted approach, including de-identification, compliance with legal requirements, and implementing robust security measures. Addressing these concerns is essential for the successful deployment and acceptance of AI-driven solutions in healthcare.

10.2 Language Support

Given India's linguistic diversity, it is essential for the application to support a wide range of languages. The Constitution of India recognizes 22 official languages [2]. Currently, the application uses Whisper to translate and transcribe audio data before feeding it to the LLM. However, our tests have been limited to Hindi. Whisper claims to support many Indian languages, but only eight are reported to perform well. Most other ASR tools support multiple languages but do not provide performance metrics. To improve language support, we need either better ASR tools offering robust services in more languages or LLMs trained on datasets in different languages. This would allow us to skip the translation step and directly use transcribed audio from any ASR tool.

10.3 Cost and Infrastructure

Integrating and maintaining the application incurs significant costs, primarily for the technical infrastructure required. Financial constraints have been a barrier to adopting EMR systems in India, and healthcare workers have identified this as a concern (Fig 6.6). Without the necessary technical infrastructure, setting up the application in India is challenging. One potential solution is government initiatives to promote the use of EMRs and AI in healthcare, which could include funding for integration. Alternatively, existing EMR companies could incorporate the application as a feature within their systems. This would be comparatively easier, since these companies can leverage their existing technical infrastructure.

[1] https://main.mohfw.gov.in/sites/default/files/EMR-EHR_Standards_for_India_as_notified_by_MOHFW_2016_0.pdf

[2] https://www.britannica.com/topic/Indian-languages

10.4 Telegram Bot

Another way to deploy this application across different healthcare enterprises is through a Telegram Bot. This deployment method would mirror that of a cloud-hosted application but would leverage the accessibility and convenience of a bot, allowing doctors to use it from anywhere. The Telegram Bot would offer the same functionalities as the standalone application. Doctors could interact with the bot to upload audio files or record conversations live. Doctors can use the bot from any location, providing greater flexibility and convenience. Telegram's user-friendly interface makes it easy for healthcare professionals to interact with the bot without requiring extensive technical knowledge. All communications between the doctor and the bot along with the EMR generated should be encrypted to ensure patient confidentiality.

By deploying the application as a Telegram Bot, healthcare providers can benefit from a versatile, user-friendly, and scalable solution that promotes the digitization of healthcare in India.

10.5 Prompt Engineering and Finetuning

As highlighted in the related work section, prompt engineering has proven to enhance entity recognition tasks in AI and healthcare. By incorporating prompt engineering techniques into our application, we can significantly improve its performance without additional technical or cost infrastructure. Effective prompt techniques like one shot, few shot prompting can guide the model to better understand and extract key entities from doctor-patient conversations, such as patient demographics, medical history, diagnoses, and treatment plans(21). Finetuning models include training the model on domain-specific data wich in our case would be medical data. Finetuning models is known to have improved performance in specific domains(15). By integrating prompt engineering and fine tuned models into our application, we can enhance the accuracy, efficiency, and flexibility of EMR generation, ultimately improving the digitization of healthcare processes in India and beyond.

10.6 Integrating into Indian healthcare

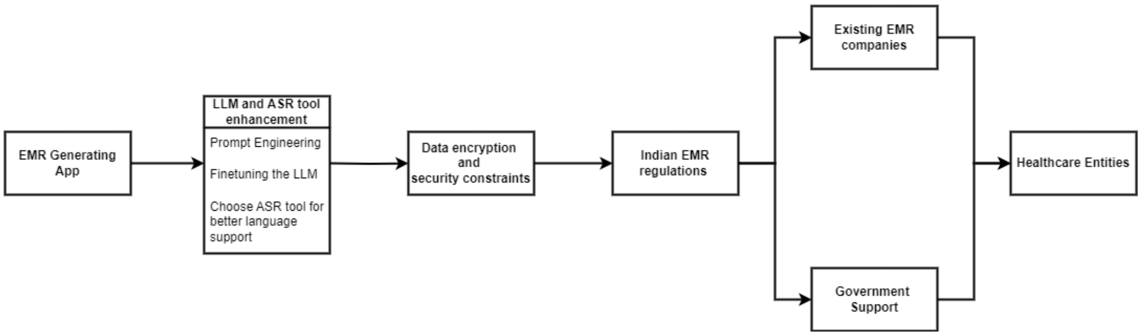


Figure 10.1: Roadmap for integrating the app into Indian healthcare

A roadmap to possibly integrating the application in Indian healthcare systems is displayed above (Fig 10.1). To advance the application in India, addressing privacy and security concerns is crucial. Technologies such as data encryption, blockchain, de-identification, and authentication should be considered to meet these challenges. Compliance with Indian government regulations is essential. Once the application meets these requirements, it can be presented to existing EMR companies for integration as a feature or as a separate product. Another approach is for the government to promote the application as part of its existing EMR system, ABHA. While pitching to the government may be time-consuming, it could result in a valuable addition to the existing system and promote the use of EMRs in India.

In summary, while the application has demonstrated potential, several critical issues need to be addressed to facilitate its successful implementation in India. Future research should focus on enhancing privacy and security measures, expanding language support, and exploring cost-effective integration strategies.

References

[1] Loick Menvielle i Anne-Françoise. **The Digitization of Healthcare**. 1

[2] Tsipi Heart, Ofir Ben-Assuli, and Itamar Shabtai. **A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy**. *Health Policy and Technology*, **6**(1):20–25, 2017. 1, 5

[3] José Luis Sánchez, Stefan Savin, and Virginia Vasileva. **Key success factors in implementing electronic medical records in University Hospital of Rennes**. *L'Ecole Nationale de la Santé Publique (National School of Public Health), Rennes, Rennes, France*, **1**, 2005. 1, 5

[4] Kim-Huong Nguyen, Chad Wright, Digby Simpson, Leanna Woods, Tracy Comans, and Clair Sullivan. **Economic evaluation and analyses of hospital-based electronic medical records (EMRs): a scoping review of international literature**. *npj Digital Medicine*, **5**(1):29, 2022. 1, 6

[5] Albert Boonstra and Manda Broekhuis. **Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions**. *BMC health services research*, **10**:1–17, 2010. 1

[6] Chen Hsi Tsai et al. **Effects of Electronic Health Record Implementation and Barriers to Adoption and Use: A Scoping Review and Qualitative Analysis of the Content**. *Life (Basel, Switzerland)*, **10**(12):327, Dec 2020. 1

[7] Alemayehu Bisrat, Dagne Minda, Bekalu Andargie, Biruk Abebe, and Teshome Abegaz. **Implementation challenges and perception of care providers on Electronic Medical Records at St. Paul's and Ayder Hospitals, Ethiopia**. *BMC Medical Informatics and Decision Making*, **21**, 11 2021. 1, 5

[8] KC Derecho, R Cafino, SL Aquino-Cafino, et al. **Technology adoption of electronic medical records in developing economies: A systematic review on physicians' perspective**. *DIGITAL HEALTH*, **10**, 2024. 1, 6

[9] Meenakshi Sharma and Himanshu Aggarwal. **EHR Adoption in India: Potential and the Challenges**. *Indian Journal of Science and Technology*, **9**(34), Sep 2016. 1, 2, 6

[10] Venkatesh Janarthanan et al. **Legal and Ethical Issues Associated With Challenges in the Implementation of the Electronic Medical Record System and Its Current Laws in India**. *Cureus*, **16**(3):e56518, Mar 2024. 2, 6

[11] RS Sharma, Aishwarya Rohatgi, Sandeep Jain, and Dilip Singh. **The ayushman bharat digital mission (abdm): Making of india's digital health story**. *CSI Transactions on ICT*, **11**(1):3–9, 2023. 2, 7

[12] Manisha Wadhwa. **Electronic Health Records in India**. Technical report, ICT India Working Paper, 2020. 2, 7

[13] Ankit Singh, Sammita Jadhav, and MR Roopashree. **Factors to overcoming barriers affecting electronic medical record usage by physicians**. *Indian Journal of Community Medicine*, **45**(2):168–171, 2020. 2, 5

[14] Abdulqadir J Nashwan and Ahmad A AbuJaber. **Harnessing the power of large language models (LLMs) for electronic health records (EHRs) optimization**. *Cureus*, **15**(7), 2023. 2, 5, 7, 8

[15] Izzet Turkalp Akbasli, Ahmet Ziya Birbilen, and Ozlem Teksam. **Human-Like Named Entity Recognition with Large Language Models in Unstructured Text-based Electronic Healthcare Records: An Evaluation Study**. 2024. 2, 9, 10, 39

[16] Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko Van Treeck, Sonja Katharina Meyer, Radhika Juglan, Zunamys I Carrero, Daniel Paech, Jens Kleesiek, Matthias P Ebert, et al. **From text to tables: a local privacy preserving large language model for structured information retrieval from medical documents**. *MedRxiv*, pages 2023–12, 2023. 2, 10

[17] Santosh G Honavar. **Electronic medical records–The good, the bad and the ugly**, 2020. 6

[18] Mert Karabacak and Konstantinos Margetis. **Embracing large language models for medical applications: opportunities and challenges**. *Cureus*, **15**(5), 2023. 7, 8

[19] Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yonfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. **A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs)**. *arXiv preprint arXiv:2405.03066*, 2024. 7, 8, 9

[20] Kavitha Palaniappan, Elaine Yan Ting Lin, and Silke Vogel. **Global Regulatory Frameworks for the Use of Artificial Intelligence (AI) in the Healthcare Services Sector**. In *Healthcare*, **12**, page 562. MDPI, 2024. 8

[21] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. **Improving large language models for clinical named entity recognition via prompt engineering**. *Journal of the American Medical Informatics Association*, page ocad259, 2024. 9, 39

[22] Marco Naguib, Xavier Tannier, and Aurélie Névéol. **Few shot clinical entity recognition in three languages: Masked language models outperform LLM prompting**. *arXiv preprint arXiv:2402.12801*, 2024. 9

[23] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. **Prompt engineering for healthcare: Methodologies and applications**. *arXiv preprint arXiv:2304.14670*, 2023. 10

[24] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. **Does synthetic data generation of llms help clinical text mining?** *arXiv preprint arXiv:2303.04360*, 2023. 10

[25] Lei Wang, Yinyao Ma, Wenshuai Bi, Hanlin Lv, and Yuxiang Li. **An Entity Extraction Pipeline for Medical Text Records Using Large Language Models: Analytical Study**. *Journal of Medical Internet Research*, **26**:e54580, 2024. 10

[26] Jessica López Espejel, Mahaman Sanoussi Yahaya Alassan, Merieme Bouhandi, Walid Dahhane, and El Hassane Ettifouri. **Low-Cost Language Models: Survey and Performance Evaluation on Python Code Generation**. *arXiv preprint arXiv:2404.11160*, 2024. 11

[27] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. **Mistral 7B**. *arXiv preprint arXiv:2310.06825*, 2023. 12

[28] Taojun Hu and Xiao-Hua Zhou. **Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions**. *arXiv preprint arXiv:2404.09135*, 2024. 16, 17

[29] C Tribes, S Benarroch-Lelong, P Lu, and I Kobyzev. **Hyperparameter optimization for Large Language Model instruction-tuning**. *Les Cahiers du GERAD ISSN*, **711**:2440, 2023. 21

[30] Eike-Henner W Kluge. **Security and privacy of EHR systems–ethical, social and legal requirements**. In *Advanced Health Telematics and Telemedicine*, pages 121–127. IOS Press, 2003. 37

Appendix A

Appendix

A.1 ABHA

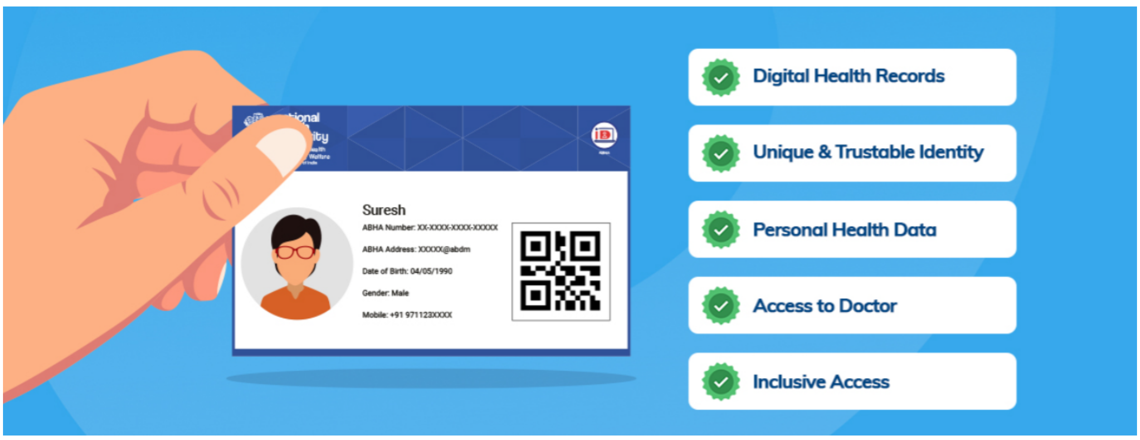


Figure A1: ABHA ID

Figure A2: Benefits of using ABHA

A.2 Doctor's Interview on EMRs and India

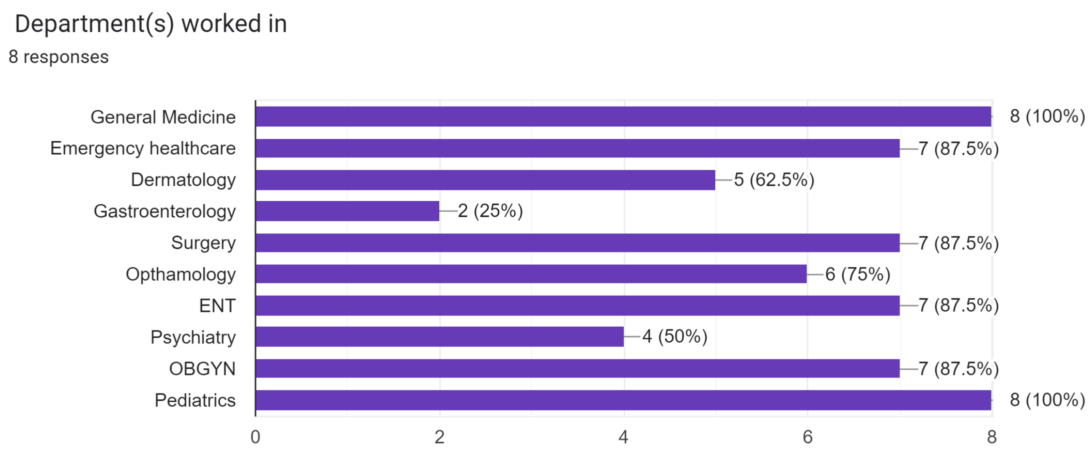


Figure A3: Departments the doctors have worked in

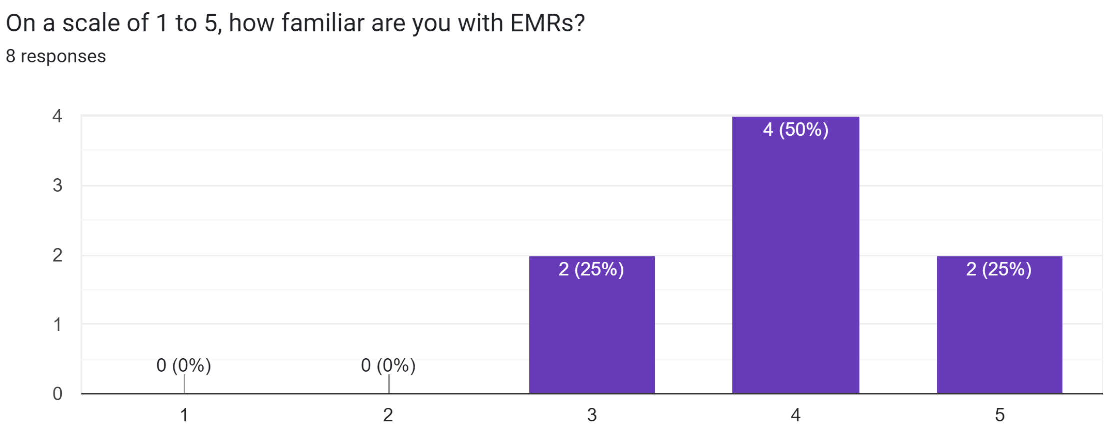


Figure A4: Familiarity with EMRs

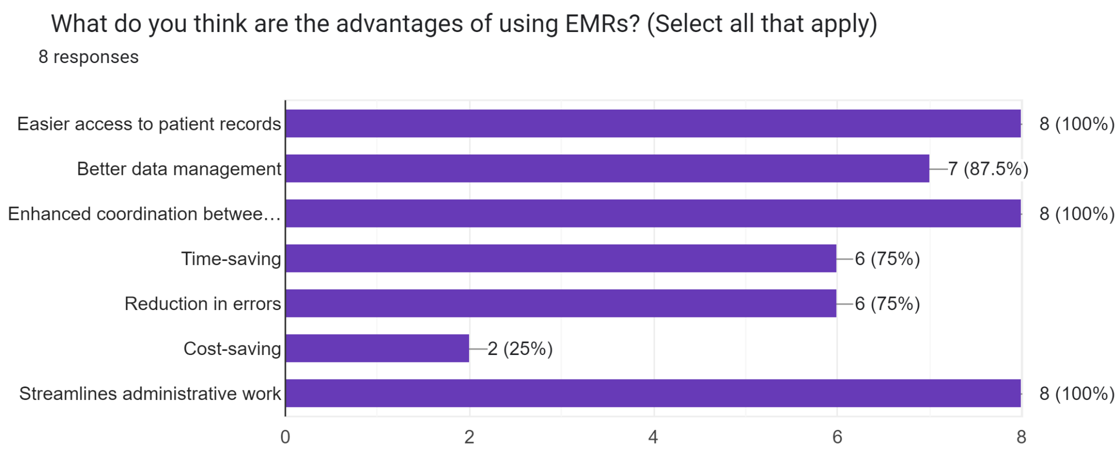


Figure A5: Advantages of EMRs

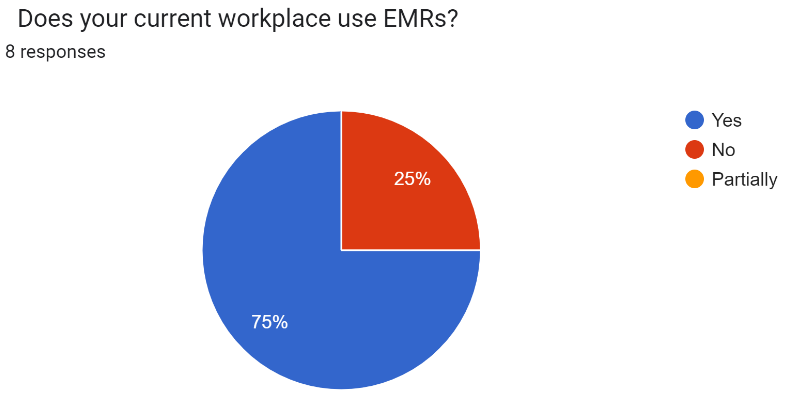


Figure A6: Use of EMRs in current workplace

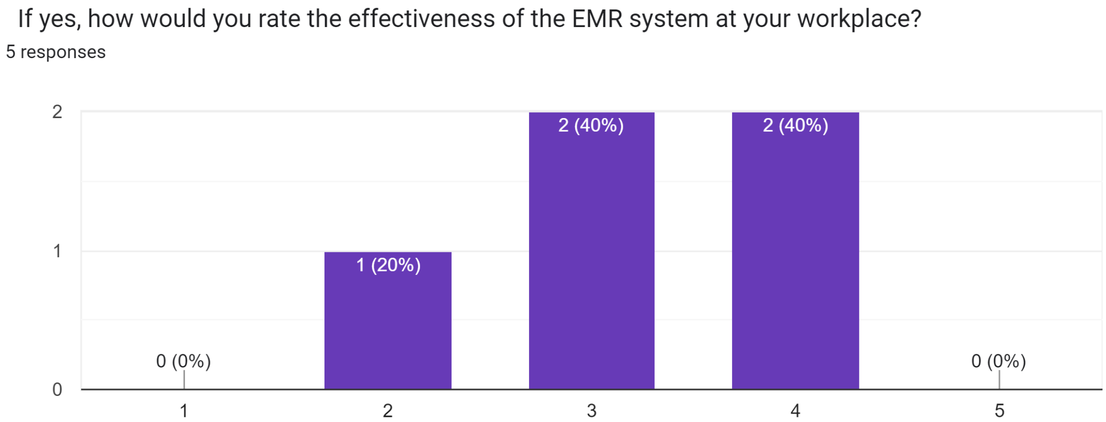


Figure A7: Effectiveness of EMRs in current workplace

Please elaborate on any obstacles you have encountered or observed regarding EMR adoption in India.

5 responses

Patient data will be compromised

Internet issues, loss of data, errors in personal and demographic info

Wifi issues cause data to be lost. No proper IT support to address concerns. Privacy breaches.

Lack of clear understanding of the depth of EMR

Server issues can hamper the entire process.
Doctors have to type out everything about a patient and with the kind of patient load a government hospital can see, it can become very time consuming.

Figure A8: Issues with EMR systems in India

How would you describe the current state of EMR adoption in India?

8 responses

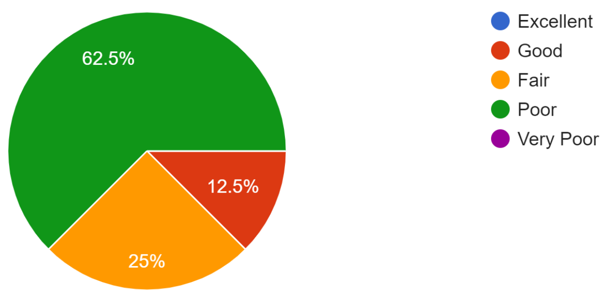


Figure A9: Current state of EMRs in India

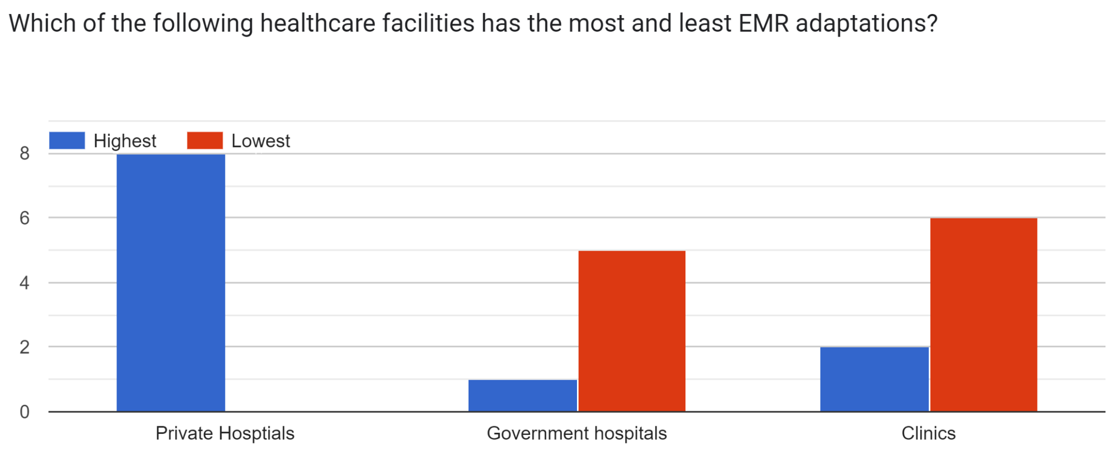


Figure A10: Highest and lowest EMR adapted healthcare facility

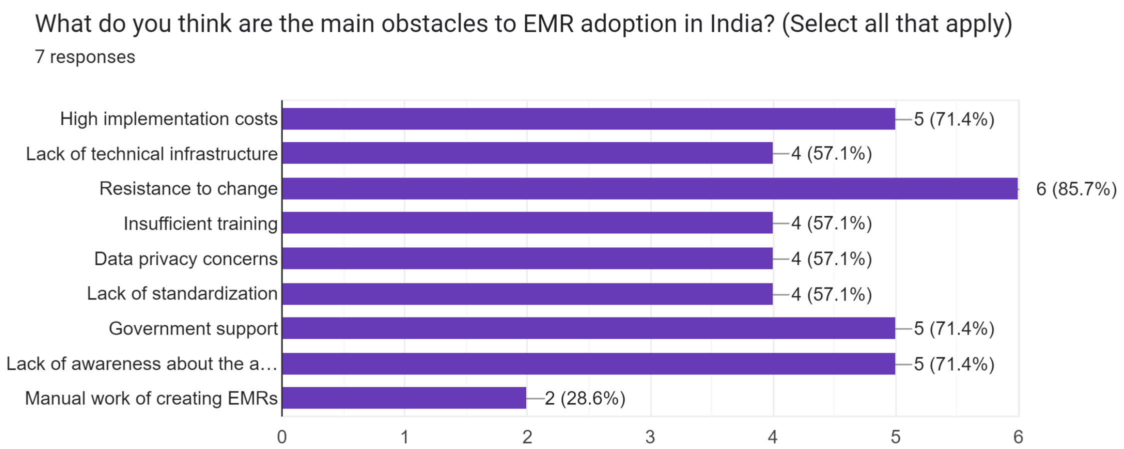


Figure A11: Main obstacles for EMR adoption in India

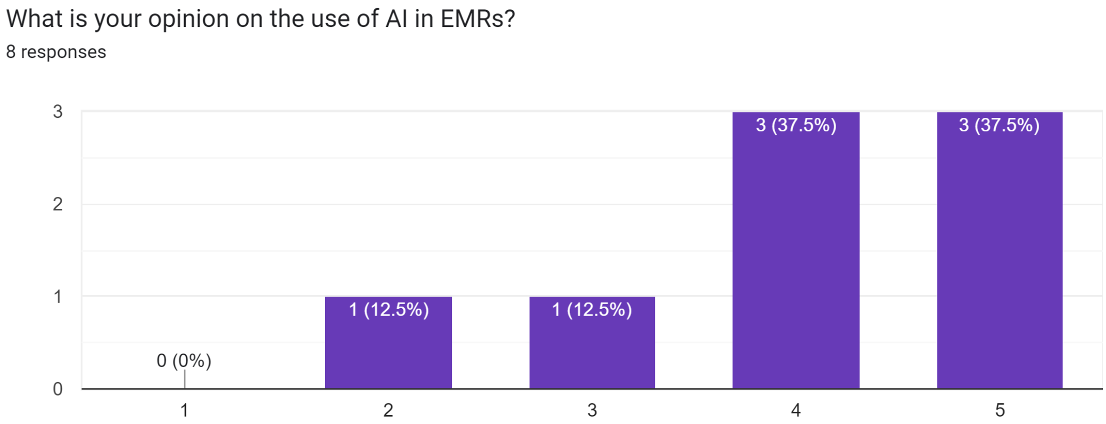


Figure A12: Opinion on use of AI in EMRs

Please share any thoughts or experiences you have regarding the use of AI in EMRs.

3 responses

Ease on work load, time efficient and easy to understand

Makes communication with patients much easier and interactive

Auto alerts are sometimes not that useful and cause fatigue

Figure A13: Experience with AI in EMR

A.3 Electronic Medical Record(EMR)

Medical Records

Section I. Patient Information

Records the patient's basic details for identification purposes.

Full Name	Jessie Connolly
Date of Birth	01/01/1980
Gender	Female
Contact Details	(123) 456-7890
Emergency Contact	John Connolly (Spouse), (123) 456-7891
Medical Record Number	10001

Section II. Medical History

This section provides space for important background information about the patient's health.

Past Conditions	Hypertension, Gestational Diabetes (during 2nd pregnancy)
Surgical Procedures	Appendix Removal (1999)
Allergies	Penicillin
Family History	Mother had breast cancer, father had diabetes

Section III. Current Medications

Record any medications the patient is currently taking.

Name of Medication and dosage	Lisinopril 10 mg
Frequency	Once daily
Prescribing Physician	Dr. Ted Roberts

Figure A14: Example of an Electronic Medical Record

A.4 Dataset Used


```
{
    "Patients": [
        {
            "Patient ID": "CAR0001",
            "Patient Information": {
                "Name": "NA",
                "Date of Birth": "NA",
                "Gender": "Male",
                "Age": "39",
                "Address": "NA",
                "Phone Number": "NA",
                "Emergency Contact": "NA"
            },
            "Visit Information": {
                "Date": "NA",
                "Chief Complaint": "Chest pain",
                "History of Present Illness": {
                    "Onset": "Last night",
                    "Location": "Left side of chest",
                    "Severity": "7 or 8",
                    "Duration": "A couple of hours",
                    "Character": "Sharp",
                    "Aggravating Factors": "Laying
down",
                    "Relieving Factors": "NA"
                }
            },
            "Medical History": {
                "Past Medical Events": "NA"
```


Figure A15: Annotated ground truth

D: What brought you in today?
P: Sure, I'm I'm just having a lot of chest pain and and so I thought I should get it checked out.
D: OK, before we start, could you remind me of your gender and age?
P: Sure 39, I'm a male.
D: OK, and so when did this chest pain start?
P: It started last night, but it's becoming sharper.
D: OK, and where is this pain located?
P: It's located on the left side of my chest.
D: OK, and, so how long has it been going on for then if it started last night?
P: So I guess it would be a couple of hours now, maybe like 8.
D: OK. Has it been constant throughout that time, or uh, or changing?
P: I would say it's been pretty constant, yeah.
D: OK, and how would you describe the pain? People will use words sometimes like sharp, burning, achy,

Figure A16: Transcript of the doctor-patient conversation

A.5 Doctors review on the application

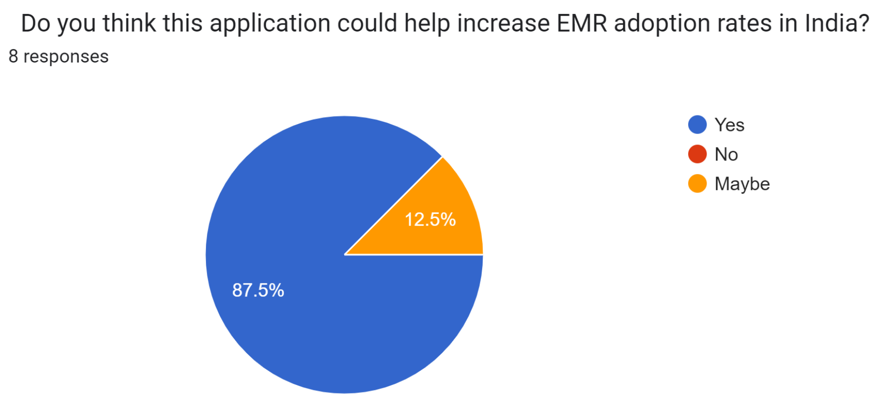


Figure A17: Can the application improve EMR adoption rates

What features or improvements would you suggest for the application?

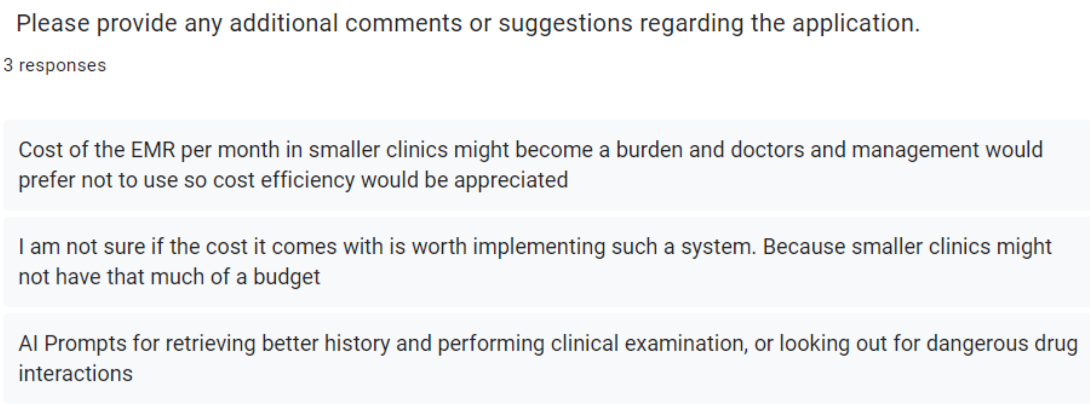
3 responses

Addition of data immediately or in the future to the same report would be helpful

User friendly interface

I would like to see options to send the EMR to the patient directly or to ABHA

Figure A18: Application feedback


Please provide any additional comments or suggestions regarding the application.

3 responses

Cost of the EMR per month in smaller clinics might become a burden and doctors and management would prefer not to use so cost efficiency would be appreciated

I am not sure if the cost it comes with is worth implementing such a system. Because smaller clinics might not have that much of a budget

AI Prompts for retrieving better history and performing clinical examination, or looking out for dangerous drug interactions

Figure A19: Suggestions for future work