AI in Healthcare: Simplifying Medical Reports for Enhanced Patient Comprehension

Rahulprasath S[1], Pranav Harshan S[2], Kabilash PV[3], Lakshith Raj A[4], Sreemathy J[5]
[1,2,3,4,5] *Department of Computer Science, Sri Eshwar College of Engineering,* Coimbatore - 642 202

***Abstract*—Medical reports can contain complex and detailed information, making them difficult for patients to understand and healthcare providers to summarize. This project presents a Medical Report Summariser that uses LLAMA models and Retrieval-Augmented Generation (RAG) to generate concise summaries of medical data from two types of inputs: unstructured text (e.g., patient symptoms and issues) and structured documents (e.g., medical reports from clinics or diagnostic centres). The proposed system first extracts relevant medical concepts and terminology from user inputs via text processing and Optical Character Recognition (OCR) for document inputs. Next, the LLAMA model, reinforced by RAG, creates a contextually appropriate summary, obtaining relevant medical knowledge as needed to improve the summarisation. Experimental results show that our method effectively generates coherent and short summaries that help patients understand and increase documentation efficiency for healthcare providers. This paper discusses the methodology, implementation obstacles, and evaluation metrics used to analyse the system's performance before concluding with thoughts on the possible implications of automated medical report summarisation in clinical settings.**

I.INTRODUCTION

The growing digitisation of healthcare has resulted in an increase in the volume of medical paperwork produced everyday, ranging from clinical notes to diagnostic reports. These documents are critical for expressing a patient's health state, documenting treatment progress, and helping healthcare practitioners make educated decisions. However, the material in medical reports is sometimes highly technical, making it difficult for patients to grasp and time-consuming for healthcare professionals to review. The requirement for automated solutions that can swiftly and accurately summarize medical documents is becoming increasingly important. Such solutions can improve patient comprehension by providing medical information in an understandable way, allowing for more efficient clinical procedures and supporting telemedicine applications. The capacity to summarize both structured documents and unstructured inputs, such as patient-reported symptoms, will significantly increase the systems' flexibility and usefulness. Recent advances in Natural Language Processing (NLP), notably transformer-based models like LLAMA, have shown considerable promise for text summarisation tasks. LLAMA, with its strong language understanding and generation skills, is ideal for producing succinct summaries of extensive and difficult documents. However, medical terminology and domain-specific expertise present additional hurdles, frequently necessitating extra information beyond the immediate context. Here is where Retrieval-Augmented Generation (RAG) comes into play. RAG blends retrieval-based and generative language models, allowing the system to retrieve and incorporate essential medical knowledge during the summarisation process, resulting in more accurate and informative outputs.

In this study, we offer a Medical Report Summariser that uses LLAMA and RAG to handle two sorts of inputs: unstructured text, such as patient symptoms, and structured medical reports from clinics or diagnostic centers. The system uses Optical Character Recognition (OCR) to extract text from scanned documents, allowing for the processing of a variety of input formats. This project attempts to bridge the gap between raw medical data and relevant insights for patients and healthcare professionals by producing concise and contextually appropriate summaries.

Section II presents the study's research objectives. Section III provides a comprehensive assessment of relevant literature, focusing on major studies and discoveries from the subject. Section IV describes in detail the proposed system's design, which includes the integration of important components and the system architecture. Section V describes our evaluation measures, conclusions, and a performance analysis of the proposed system. Section VI of the study concludes with a summary of the major findings and recommendations for future system research and development.

II.RESEARCH OBJECTIVE

The goal of this study is to create an automated summary system capable of accurately and simply summarizing medical information from two separate input types: unstructured patient-reported symptoms and structured medical records, such as diagnostic reports. To do this, the proposed system uses advanced Natural Language Processing (NLP) techniques, notably LLAMA models combined with Retrieval-Augmented Generation (RAG), to improve the accuracy and context-awareness of the generated summaries [7]. The system is designed to handle a variety of input formats using a configurable input processing pipeline that incorporates Optical Character Recognition (OCR) for transforming scanned documents into text. The system uses transformer-based summarizing via LLAMA to generate coherent and contextually relevant summaries, even when dealing with sophisticated medical language. Furthermore, RAG is used to extract relevant information from external medical knowledge sources, which improves the summarizing process. The system's performance will be assessed across both text and document inputs using metrics such as ROUGE, BLEU, and domain-specific evaluations to determine its flexibility, accuracy, and robustness. Finally, this study intends to make a contribution to the field of medical text summarization by creating a system that improves patient comprehension and assists healthcare workers in expediting clinical recording.

III.LITERATURE REVIEW

However, existing Medical report summarisers often have drawbacks like Zero-Shot learning like LLAMA 2, even with retrieval-augmented generation (RAG), are prone to generating inaccurate information i.e.,hallucinations from ambiguous data, posing a reliability risk in clinical applications [10].

Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records(2024) [1]: This paper investigates recent advances in artificial intelligence (AI) for summarizing electronic health records (EHRs) in aged care institutions show remarkable accuracy, notably in extracting data on malnutrition. The study found an overall accuracy of 94.5%, with summarization utilizing retrieval-augmented generation (RAG) obtaining an astonishing 99.25% accuracy and extraction tasks reaching 90% [6]. Despite these advancements, limitations remain, particularly with zero-shot learning models like as LLAMA 2, which, even when combined with RAG, are still prone to producing erroneous information, or "hallucinations," as a result of ambiguous or implicit data. This emphasizes the importance of carefully applying AI in therapeutic situations to ensure reliability.

Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models(2023) [2]: Recent clinical text summary research has investigated approaches for automatically summarizing Brief Hospital Course (BHC) sections from electronic health records (EHRs) using extractive, abstractive, or ensemble models. Extractive summarization uses sentence ranking models like TextRank and Bi-LSTM to choose the top-k salient sentences, which are then embedded using approaches like GloVe and Sentence-BERT (S-BERT) [8]. Abstractive summarization uses fine-tuned transformer models (T5, BERT-to-BERT, and BART) to provide coherent, contextually accurate summaries of BHC sections[3]. Clinically-guided techniques combine domain-specific knowledge, with tools such as MedCAT extracting clinical ideas to guide the process, and a modified BART architecture using this guidance for encoding and decoding. Ensemble models combine extractive selection and guided abstractive generation, with the goal of leveraging both approaches. These solutions have the potential to reduce physician burden, improve discharge summary quality, and emphasize critical information [11]. Factual errors, the necessity for human oversight, and dealing with complex situations are some of the challenges. Accuracy evaluations reveal ROUGE-L scores of 34.9 on the MIMIC-III and 26.6 on the KCH datasets [12], with human-level performance estimated at 70-80%. While clinically directed models showed modest improvement, 30-40% of summaries remained problematic. These findings show promise, but they highlight the need for additional research before clinical implementation.

IV.PROPOSED SYSTEMS

The proposed Medical Report Summarizer system is intended to handle user-submitted medical reports in a systematic manner, using Retrieval-Augmented

Generation (RAG) and a finely calibrated Fine-tuned LLAMA-3 8b model to provide accurate, contextually relevant summaries. The system begins by receiving medical reports, which might be free-form text or structured documents like scanned clinical reports. For document inputs, Optical Character Recognition (OCR) is used to extract text, allowing the system to handle a wide range of formats. Following preprocessing, the data is routed to the RAG component, which gets relevant information from external medical knowledge sources to provide additional context.This retrieval is crucial for dealing with complex medical terminology and ensuring that the summary include not just the major content but also contextual information relevant to the specific medical conditions being discussed [5]. The RAG-enhanced input is then fed into a Fine-tuned LLAMA-3 8b (MoE) model that has been trained on medical data to handle domain-specific language and provide logical, succinct summaries. Fine-tuned LLAMA-3 8b creates a summary that represents the important points of the original report while smoothly incorporating the contextual data provided by RAG, resulting in output that is both useful and understandable.

Our MoE model is a merged model of two pre-trained models in the medical domain. *medllama3-v20* is the base model which is trained on Medical databases with average accuracy around 90%, *Llama3-Aloe-8B-Alpha* is the other model which is merged with the base one. This model is also trained with a different set of medical databases, and this model has an average accuracy of 73%. These two models are merged using Mergekit.

This final summary condenses complex medical information into an understandable manner, making it accessible to patients seeking clarity on their health concerns and valuable to healthcare professionals who require efficient summarizing of clinical data. By merging RAG and Fine-tuned LLAMA-3 8b, the system represents an innovative method to medical text summarizing, with the goal of increasing comprehension and facilitating communication in healthcare settings.

Tokenizing it, and guaranteeing conformity with the summarization model. Preprocessing removes any extraneous elements or noise, standardizes the format, and prepares the text for analysis[4]. This stage ensures that the inputs are optimized for the summarizing task, allowing the model to focus on relevant content while being consistent across different types of medical reports.The Medical Report Summarizer system converts user inputs into brief and easily understandable summaries using a structured, multi-step process. The process starts with the Input step, when users are given two alternatives for submission: write a prompt explaining their symptoms, health concerns, or pertinent facts, or upload a Medical Report document directly from a healthcare provider or diagnostic center. The system is designed to accept a variety of document types, making it responsive to both organized and unstructured inputs. For document uploads, the system uses Optical Character Recognition (OCR) to extract text from scanned reports, allowing it to process both typed and handwritten documents.Once the raw data is acquired, it goes through a preparation phase. This stage entails cleaning the text,

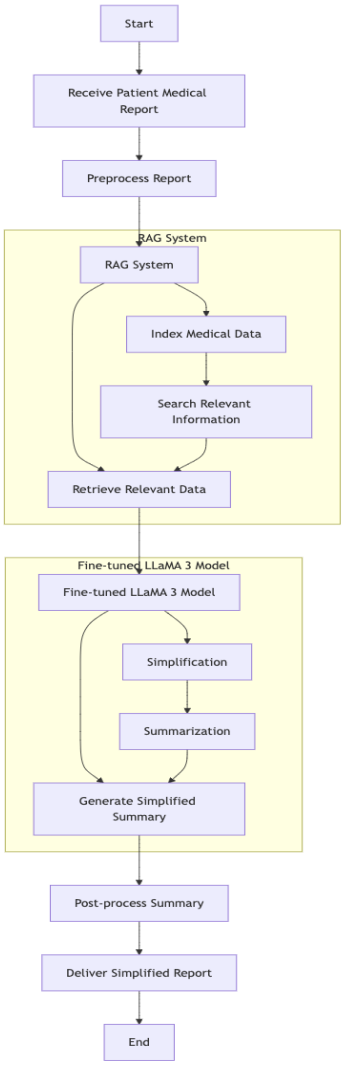

Fig 4.1 -Proposed Architecture Diagram of Fine tuned LLAMA model for Medical Report summarizing

Fig 4.2 – Proposed Fine-tuned LLAMA-3 Model

A Fine-Tuned LLAMA-3 8b Model serves as the system's foundation. This model has been specifically trained and fine-tuned on medical datasets to deal with domain-specific language and context. Using advanced Natural Language Processing (NLP) techniques, the LLAMA model interprets the combined inputs—both user prompts and specific medical report content—to provide a cohesive, contextually appropriate summary [13]. The fine-tuning procedure entails training the model on massive volumes of medical data, allowing it to comprehend complex terminology, identify medical conditions, and accurately interpret clinical jargon. The LLAMA-3 8b model's sophisticated language generating skills ensure that the generated summaries are not only brief, but also capture the important aspects of the original report.The system then creates a Simplified Summarized Medical Report, which condenses important data into an understandable format. This output is designed to be both comprehensive and accessible, summarizing the main points of the original report while leaving out unnecessary details. The summary report aims to overcome the communication gap between healthcare providers and patients by making complex medical information more understandable. This improved accessibility promotes informed decision-making by allowing patients to better understand their health state and doctors to communicate critical information more effectively.

Using this methodology, the Medical Report Summarizer creates a dependable, efficient, and user-friendly tool for converting extensive and detailed medical reports into useful summaries. The use of RAG for contextual retrieval and the fine-tuned LLAMA model guarantees that the final output is accurate and relevant, highlighting the potential of NLP technology in clinical applications.

A. Input Acquisition: The algorithm begins by receiving user input via two primary channels: a text prompt and a medical report document. Users can give a textual description of their symptoms or upload a document from a healthcare provider or diagnostic center. The system extracts text from uploaded documents using Optical Character Recognition (OCR), which can handle both typed and handwritten content. This adaptability means that the summarizer can handle a wide range of input formats, meeting the different needs of its users.

B. Preprocessing: Once the input is acquired, it goes through a thorough preprocessing stage. This includes text cleaning, which removes any unnecessary letters, symbols, or formatting flaws that could disrupt the analysis. Following that, the text is tokenized into smaller, more manageable units and normalized by standardizing it, which includes converting to lowercase, deleting stop words, and using stemming or lemmatization to reduce words to root forms. Furthermore, the approach uses a specialized medical lexicon to identify and categorize key medical terms, which improves the model's knowledge of the content and the accuracy of the following summary.

C. Contextual Retrieval (RAG): To enrich the input data, the system includes a Retrieval-Augmented Generation (RAG) component. This stage entails extracting pertinent contextual information from external knowledge bases, such as medical databases that contain words, conditions, and treatment regimens. The RAG process provides additional context to the input, which is especially valuable in cases involving complex or rare medical issues. By incorporating external knowledge, the system improves the completeness of the input data, guaranteeing that the summarization model has a broad comprehension of the medical content.

D. Summarization with Fine-Tuned LLAMA-3 8b Model: The Fine-Tuned LLAMA-3 8b model, which was specifically trained on medical datasets to handle specialized language and domain-specific terms, serves as the algorithm's foundation. The increased input from the RAG process is given into the model, which uses sequence analysis to comprehend sentence relationships and find essential information. The LLAMA model uses abstractive summarization techniques to provide a cohesive summary of the medical report's key elements. This fine-tuning process allows the model to efficiently rewrite and compress material, resulting in a compact output that appropriately reflects the report's key substance.

E. Post-Processing: After the summary is prepared, it goes through a post-processing phase to improve readability and accuracy. This process comprises grammatical fixes, sentence restructuring, and terminological accuracy checks. Additionally, a validation process is carried out to ensure that the summary is consistent with the original text. If differences are discovered, changes are performed, or the system may request further contextual information to fill any gaps. This careful post-processing guarantees that the summary is both accurate and easy to understand.

F. Output Generation: The system's ultimate result is a Simplified Summarized Medical Report, which condenses the key information from the original document into an understandable format. This report is intended to be understandable to both patients and healthcare practitioners, enabling successful communication. The Medical Report Summarizer bridges the gap between complex medical paperwork and patient-friendly information by delivering simple and short summaries, allowing clinicians to make more informed decisions.

V. RESULT

We assessed our model using Medical Benchmarking, which included MedMCQA,MedQA_4options, anatomy, clinical knowledge, college_biology, college_medicine, medical_genetics, professional_medicine, and PubMedQA. Using our advanced model, the MedNarra-X1, the system produced noteworthy results. The MedMCQA model attained an accuracy of 75.19%. MedQA_4options returned an accuracy of 87.27%. The model performed well in anatomy, with an accuracy of 96.30%, and clinical knowledge, with a score of 96.98%. College biology and medicine tasks yielded accuracy scores of 97.22% and 94.80%, respectively. Professional medicine was predicted exactly, with an accuracy of 100%, while PubMedQA's accuracy was 77.00%.

Tasks	Version	Filter	n-shot	Metric		Value		Stderr
medmcqa	Yaml	none	0	acc	↑	0.7519	±	0.0067
		none	0	acc_norm	↑	0.7519	±	0.0067
medqa_4options	Yaml	none	0	acc	↑	0.8727	±	0.0093
		none	0	acc_norm	↑	0.8727	±	0.0093
anatomy	1	none	0	acc	↑	0.9630	±	0.0163
clinical_knowledge	1	none	0	acc	↑	0.9698	±	0.0105
college_biology	1	none	0	acc	↑	0.9722	±	0.0137
college_medicine	1	none	0	acc	↑	0.9480	±	0.0169
medical_genetics	1	none	0	acc	↑	1.0000	±	0.0000
professional_medicine	1	none	0	acc	↑	0.9853	±	0.0073
pubmedqa	1	none	0	acc	↑	0.7700	±	0.0188

Fig 5.1 – Performance analysis metric of Medical Report Summarizing model using Fine tuned LLAMA Model

A new approach to the problem of turning complicated medical data into clear, understandable summaries is provided by the Medical Report Summarizer. Utilizing sophisticated models such as Fine-Tuned LLAMA-3 8b and including the Retrieval-Augmented Generation (RAG) methodology, the system exhibits adaptability in managing both comprehensive medical data and text inputs based on symptoms [9]. The system's usefulness for patients and medical professionals is increased by its capacity to recover pertinent medical context and process domain-specific language, which guarantees accurate and contextually appropriate outputs. The technology is appropriate for real-world applications where clear communication is crucial because of the post-processing procedures, which further enhance the resulting summaries' readability and coherence [14].

But there are still a number of issues that need more research. Notwithstanding the system's resilience, problems like sporadic factual errors highlight the necessity of human supervision in clinical settings, especially in complicated or uncommon medical instances. Variability in training data quality can also effect summarization consistency, suggesting that future iterations will benefit from bigger, more consistent datasets. To guarantee accuracy without compromising completeness, it is also crucial to strike a balance between the summaries' conciseness and the inclusion of important medical facts. By resolving these issues, the system's full potential to lessen administrative workloads and enhance patient comprehension in medical settings may be realized.

VI. CONCLUSIONS

The Medical Report Summarizer represents a substantial development in the automated summarizing of complicated clinical papers, utilizing cutting-edge NLP methods such as Retrieval-Augmented Generation (RAG) and a fine-tuned LLAMA-3-8b model. This system provides flexibility and accommodates the various data sources often seen in healthcare settings by allowing users to submit either symptom descriptions or entire medical reports. The multi-step method, which includes contextual retrieval, domain-specific language modeling, and rigorous post-processing, ensures that the resulting summaries are accurate and

easily accessible [15].

This study demonstrates the potential of AI-driven solutions to alleviate healthcare professionals' cognitive and administrative burdens while also providing patients with understandable health information. Despite the difficulties of dealing with sophisticated medical language and ensuring contextual accuracy, the system shows potential for real-world clinical applications. Future enhancements, such as incorporating larger medical knowledge bases and fine-tuning the model's training with larger datasets, could increase the summarizer's reliability and performance even more. Finally, the Medical Report Summarizer is a useful tool for bridging the communication gap in healthcare, promoting informed decision-making, and enhancing patient-centered care.

REFERENCES

[1] Mohammad Alkhalaf, et a "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records," 2024.
[2] Thomas Searle, et al., "Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models," 2023.
[3] Akshara Ramprasad, et al., "Context-Aware Summarization for PDF Documents using Large Language Models," 2024.
[4] Sherif, Ahmed Abdelfattah Saleh., "Language Independent Text Summarizer and Deep Self-Organizing Cube," 2024.
[5] Demiao Lin, et al., "Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition," 2024.
[6] Yubing Ren, et al., "Retrieve-and-Sample: Document-level Event Argument Extraction via Hybrid Retrieval Augmentation," 2023.
[7] Deepak Dharrao, et al., "Patients' Medical History Summarizer using NLP," 2023.
[8] Jen-Yuan Yeh, et al., "Text summarization using a trainable summarizer and latent semantic analysis," 2005.
[9] Lin C., "Rouge: a package for automatic evaluation of summaries," 2004.
[10] Gupta, V., Lehal, G., "A survey of text summarization extractive techniques. J. Em. Technol. Web Intel. 2(3), 258–268" 2001.
[11] Cheung, J., "Comparing abstractive and extractive summarization of evaluative text: controversiality and content selection," 2008
[12] Nenkova, A., "Computational Linguistics" 2012.
[13] Shengjie Liu, et al., "Towards a Robust Retrieval-Based Summarization System," 2024.
[14] Darren Edge, et al., "From Local to Global: A Graph RAG Approach to Query-Focused Summarization," 2024.
[15] Sriram Veturi, et al., "RAG based Question-Answering for Contextual Response Prediction System," 2024.