Devansh, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

GEN-AI Driven Automated Medical Report Generation and Discharge Summary

Devansh, Darshit Chauhan, Aditya Singh, Assistant Professor Mr. Anurag Chandana
Department of Computer Science and Engineering (AIML)
Quantum University, Roorkee

Abstract- The healthcare industry generates enormous volumes of clinical data daily, including patient histories, diagnostic results, lab reports, and physician notes. Converting this unstructured data into structured, readable medical reports and discharge summaries is a time-consuming and error-prone process when done manually. This research presents a GEN-AI driven system for automated medical report generation and discharge summary creation using large language models (LLMs), machine learning (ML) algorithms, and natural language processing (NLP) techniques. The system accepts raw clinical inputs including patient demographics, symptoms, diagnosis codes, prescribed medications, and lab results and automatically generates structured medical reports and professional discharge summaries. The project utilizes a web-based interface developed with modern frontend technologies and integrates backend ML pipelines for intelligent content generation. The results demonstrate significant improvements in documentation speed, accuracy, and consistency compared to manual approaches.

Keywords— Generative AI, Medical Report Generation, Discharge Summary, Large Language Models (LLMs), Natural Language Processing, Healthcare Automation, Machine Learning, Clinical Documentation

I. INTRODUCTION

Background of the Study

Healthcare documentation is one of the most critical and resource-intensive activities in modern medical practice. Physicians, nurses, and administrative staff spend a considerable portion of their working hours generating reports, recording observations, and preparing discharge summaries for patients. According to various healthcare industry studies, clinicians spend nearly 35 to 40 percent of their work time on documentation tasks rather than direct patient care.

The introduction of Generative Artificial Intelligence (GEN-AI) and Large Language Models (LLMs) presents a transformative opportunity to automate and streamline clinical documentation. Systems powered by these technologies can analyze structured and unstructured medical data, extract meaningful insights, and produce coherent, professional-grade medical documents with minimal human intervention.

This project was developed by final year B.Tech CSE (AIML) students at Quantum University, Roorkee, with the goal of building an end-to-end intelligent system that automates medical report generation and discharge summary creation using machine learning and generative AI technologies.

2. Problem Statement

Despite advancements in hospital management systems, clinical documentation remains largely manual. Key challenges include significant time spent by healthcare professionals on paperwork instead of patient care, inconsistencies and errors in manually prepared medical reports and discharge summaries, difficulty in maintaining standardized formats across hospitals and departments, lack of accessible intelligent tools that can process raw clinical inputs and produce formatted output, and high operational costs associated with documentation personnel and administrative overhead.

3. Objectives of the Study

The primary objectives of this research and development project are outlined in Table 1 below.

Table 1: Objectives of the Study

No.	Objective
1	To design and develop a GEN-AI powered system for automated medical report generation.
2	To automate the creation of discharge summaries from structured patient clinical data.
3	To integrate machine learning models and NLP pipelines for intelligent document generation.
4	To develop a user-friendly web-based interface for healthcare professionals to interact with the system.
5	To evaluate system performance in terms of report accuracy, generation speed, and usability.

4. Scope of the Study

This project focuses on automating two primary healthcare documentation workflows: generating structured medical reports from patient intake and diagnostic data, and creating discharge summaries upon patient release. The system is designed for use in general hospitals and outpatient departments, with the web-based interface supporting English-language clinical data. The research leverages secondary references from academic literature on NLP in healthcare, as well as primary development data collected during system testing.

II. LITERATURE REVIEW

1. Overview

The intersection of artificial intelligence and healthcare documentation has been an active area of research. Several studies have explored the use of NLP and deep learning for clinical text mining, information extraction, and automated report generation. This section reviews key theoretical frameworks and prior work relevant to the project.

2. Natural Language Processing in Healthcare

Natural Language Processing (NLP) has been widely applied in the healthcare domain for tasks such as named entity recognition (NER), clinical text summarization, ICD code assignment, and information extraction from electronic health records (EHRs). Early NLP-based systems for medical documentation relied on rule-based methods and template filling. Modern approaches leverage transformer-based models such as BERT, BioBERT, and GPT variants, which have demonstrated superior performance in understanding and generating clinical language.

3. Generative AI and Large Language Models

Generative AI refers to a class of machine learning models capable of producing new content such as text, images, or structured data based on learned patterns. Large Language Models (LLMs), including the GPT series by OpenAI, LLaMA by Meta, and Gemini by Google, have shown remarkable capabilities in text generation tasks. In healthcare, these models have been fine-tuned on clinical datasets to generate radiology reports, clinical summaries, and patient discharge notes with high coherence and factual accuracy.

4. Automated Discharge Summary Generation

Discharge summaries are critical clinical documents that communicate the patient's hospital course, diagnosis, treatment received, and follow-up instructions to outpatient providers and the patient. Automating this process reduces clinician burden and improves care continuity. Prior research has demonstrated that seq2seq models and abstractive summarization techniques can generate discharge summaries comparable in quality to human-written ones when trained on sufficiently large clinical datasets such as MIMIC-III.

5. Machine Learning in Clinical Decision Support

Machine learning models including Random Forests, Support Vector Machines, and neural networks have been extensively used in clinical decision support systems (CDSS) for diagnosis prediction, risk stratification, and treatment recommendation. In the context of report generation, ML models help classify patient conditions, extract structured features from raw inputs, and predict relevant diagnostic codes, which are then fed into the generative pipeline.

III. RESEARCH METHODOLOGY

1. Introduction to Research Methodology

This section describes the system design, technology stack, data processing pipeline, and development methodology adopted for the GEN-AI driven medical report generation system. The project follows an agile software development approach, with iterative design, development, testing, and refinement cycles.

2. System Architecture

The system is organized into three primary layers. The Frontend or Presentation Layer is the web-based user interface for data input and report visualization. The Backend or Processing Layer is the REST API server hosting ML models and NLP processing pipelines. The AI/ML Engine or Intelligence Layer consists of GEN-AI models responsible for report and summary generation.

3. Technology Stack

The technology components used in the project are listed in Table 2 below.

Table 2: Technology Stack

Component	Technology Used
Frontend	HTML5, CSS3, JavaScript, React.js / Vanilla JS
Backend	Python (Flask / FastAPI)
ML Libraries	Scikit-learn, Pandas, NumPy, NLTK, SpaCy
Generative AI	OpenAI GPT API / Hugging Face Transformers (BioBERT, GPT-2 fine-tuned)
Database	SQLite / MySQL for patient records
Report Export	ReportLab (PDF), python-docx (Word)
Deployment	Localhost / Cloud deployment (AWS / Heroku)

4. Data Collection and Sources

The system was developed and tested using the following data sources: MIMIC-III (Medical Information Mart for Intensive Care), a de-identified clinical database from Beth Israel Deaconess Medical Center used for model training and testing; synthetically generated patient records curated to simulate realistic patient demographics, diagnoses, and treatments for system validation; academic literature covering published papers on clinical NLP, discharge summary generation, and healthcare AI; and publicly available ICD-10 diagnostic code datasets for structured classification.

5. System Workflow

The end-to-end workflow of the system follows six stages. First, the user inputs patient details via the web interface including name, age, gender, symptoms, diagnosis, medications, and lab results. Second, the backend preprocesses the data using NLP pipelines for tokenization, entity extraction, and normalization. Third, ML classification models identify diagnostic categories and relevant clinical entities. Fourth, the GEN-AI engine receives structured input and generates the medical report or discharge summary. Fifth, post-processing refines the output for formatting, completeness, and clinical accuracy. Sixth, the final document is presented on screen and can be exported as PDF or Word.

6. ML Models Used

The machine learning models used in the system and their respective purposes are summarized in Table 3 below.

Table 3: ML Models Used in the System

Model	Purpose	Library
Logistic Regression	Diagnosis category classification	Scikit-learn
Random Forest	Severity prediction	Scikit-learn

BioBERT	Clinical NER and entity extraction	Hugging Face
GPT-based LLM	Medical report and summary generation	OpenAI / HF
TF-IDF + SVM	Symptom-to-diagnosis mapping	Scikit-learn

7. Website Design

A fully functional website was designed and developed as the primary interface for the system. Key design principles include responsive design compatible with desktop and tablet browsers, patient data input forms with field validation to ensure data integrity, real-time report generation display with a loading indicator during AI processing, report preview and download options in PDF and Word format, and a clean professional UI with healthcare-appropriate color schemes and typography.

8. Limitations

The project has certain limitations that should be acknowledged. The system has been primarily tested on synthetic and publicly available datasets, and may require additional fine-tuning before deployment in live clinical environments. The system currently supports English-language inputs only. Additionally, LLM-generated content may occasionally require human review to ensure clinical accuracy, as hallucination remains an inherent risk in generative models.

IV. DATA ANALYSIS AND RESULTS

1. System Performance Metrics

The system was evaluated on a test set of 200 synthetic patient records. Key performance metrics assessed include report generation accuracy, time taken, and user satisfaction scores collected from a group of medical students and healthcare professionals who reviewed the outputs. The results are presented in Table 4 below.

Table 4: System Performance Comparison

Metric	Manual Process	AI-Generated System
Average Report Generation Time	25 to 40 minutes	Less than 30 seconds
Content Accuracy (Clinician Rating)	~88%	~85 to 90%
Format Consistency	Variable	100% standardized
Discharge Summary Completeness	~82%	~91%
User Satisfaction Score (1 to 5)	N/A	4.3 / 5.0

2. NLP Model Accuracy Comparison

Different NLP and AI approaches were evaluated to compare their accuracy in generating medical documents. The results are summarized in Table 5.

Table 5: NLP Approach Accuracy Comparison

NLP / AI Approach	Task	Accuracy / Score
Rule-based Template System	Report Generation	72% (BLEU Score)
TF-IDF + ML Classifier	Symptom Classification	81%
BioBERT (NER)	Entity Extraction	89% F1-Score
GPT-based LLM Fine-tuned	Full Report and Discharge Summary	91% ROUGE-L

The results clearly demonstrate that advanced generative AI models fine-tuned on clinical data significantly outperform traditional rule-based and simpler ML approaches in generating clinically meaningful and coherent medical reports. BioBERT performs strongly for structured entity extraction, while the end-to-end LLM pipeline delivers the best overall results for full document generation.

3. Expert Reviewer Evaluation

Sample medical reports and discharge summaries generated by the system were reviewed by three healthcare professionals including two MBBS graduates and one medical faculty member. The reviewers rated the outputs on four criteria as shown in Table 6.

Table 6: Expert Reviewer Evaluation Scores

Evaluation Criterion	Reviewer 1	Reviewer 2	Reviewer 3
Clinical Accuracy	4.5 / 5	4.2 / 5	4.4 / 5
Language and Readability	4.7 / 5	4.6 / 5	4.5 / 5
Format and Structure	5.0 / 5	4.8 / 5	4.9 / 5
Completeness	4.3 / 5	4.1 / 5	4.4 / 5

4. Summary of Results

The data analysis reveals several important findings. The AI-driven system reduces report generation time from 25 to 40 minutes to under 30 seconds. Fine-tuned GPT-based models achieve 91% ROUGE-L score on discharge summary generation. BioBERT achieves an 89% F1-Score on clinical named entity recognition tasks. Expert reviewers rated the system outputs 4.3 out of 5 on average across all criteria. The web interface received a user satisfaction score of 4.3 out of 5. Standardized formatting was achieved consistently across 100% of generated reports.

V. SYSTEM ENHANCEMENT STRATEGIES

1. Prompt Engineering

Carefully crafted prompts are used to guide the LLM in producing clinically appropriate output. Prompt templates are structured to include patient context, required output format, and relevant clinical constraints, reducing hallucination and improving report relevance.

2. Model Fine-Tuning on Clinical Data

The base LLM is fine-tuned on MIMIC-III clinical notes and discharge summaries to adapt its language patterns to the medical domain. This significantly improves the clinical accuracy and terminology used in generated outputs.

3. Retrieval-Augmented Generation

A Retrieval-Augmented Generation (RAG) approach is incorporated to allow the model to reference patient history records and similar case documents stored in the database before generating reports. This grounds the generation in factual clinical context and reduces the risk of fabricated content.

4. Human-in-the-Loop Validation

The system includes a review and edit interface where healthcare professionals can review AI-generated content, make corrections, and approve the final document before it is saved or printed. This ensures clinical responsibility remains with qualified practitioners.

Findings and Discussion

The study and system development produced the following key findings. GEN-AI models, particularly fine-tuned LLMs, are highly effective for automated medical report generation and produce outputs comparable in quality to human-written documents. Machine learning classifiers improve the accuracy of structured data extraction, which directly enhances the quality of AI-generated reports. The integration of BioBERT for NER significantly improves the system's ability to correctly identify and utilize clinical entities such as diagnoses, medications, and procedures.

The web-based interface provides an accessible, user-friendly platform that healthcare professionals can adopt with minimal training. Automated report generation dramatically reduces documentation time, allowing clinicians to focus more on patient care. Standardized output format ensures consistency and reduces the risk of documentation-related medical errors. Expert reviews confirm that AI-generated discharge summaries are clinically meaningful, well-structured, and largely accurate, with minor corrections needed only in edge cases.

VI. CONCLUSION

This project demonstrates that Generative AI, combined with machine learning and NLP, can effectively automate clinical documentation tasks such as medical report generation and discharge summary creation. The developed system achieves high accuracy, significant time savings, and strong expert approval ratings, making it a viable tool for healthcare settings.

The project was successfully implemented by Devansh, Darshit Chauhan, and Aditya Singh, fourth-year B.Tech CSE (AIML) students at Quantum University, Roorkee, under the mentorship of Asst. Prof. Mr. Anurag Chandana. The system delivers a complete end-to-end solution from patient data intake to professional document output, with a user-friendly web interface and robust AI backend.

Future enhancements to the system include expanding multilingual support, integrating with hospital EHR systems via HL7/FHIR APIs, incorporating voice-to-text input for real-time physician dictation, and exploring blockchain-based audit trails for generated medical documents to enhance data integrity and compliance. As generative AI continues to evolve, its integration into clinical workflows holds immense promise for transforming healthcare delivery, reducing administrative burden, and improving patient outcomes through more timely and accurate clinical documentation.

REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019.
2. Johnson, A. E. W., et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.
3. Lee, J., et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
4. Brown, T. B., et al. (2020). Language Models are Few-Shot Learners (GPT-3). Advances in Neural Information Processing Systems.
5. Sanger, P. C., et al. (2020). Automated generation of discharge summaries using natural language processing. JAMIA Open, 3(2), 258-265.
6. World Health Organization (WHO). (2020). Global Strategy on Digital Health 2020-2025. Geneva: WHO.
7. Rajpurkar, P., et al. (2022). AI in health and medicine. Nature Medicine, 28, 31-38.
8. Esteva, A., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.
9. Scikit-learn Documentation. (2024). Machine Learning in Python. Retrieved from https://scikit-learn.org
10. Hugging Face. (2024). Transformers Documentation. Retrieved from https://huggingface.co/docs/transformers
11. Khaleel Khan Mohammed. "A Survey on Digital Health Care Data Analysis Techniques for Developing Machine Learning Models