IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ※ Peer-reviewed & Refereed journal ※ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156111

AI-Based Medical Document Summarization Using NLP

Mamatha J[1], Meghana M K[2], Ms. Nandini B[3]

Students, Department of MCA, SJB Institute of Technology, Bangalore, India[1,2]

Guide, Department of Master of Computer Applications, SJB Institute of Technology, Bengaluru, Karnataka, India[3]

Abstract: Hospital have to handle a lot of paperwork these days. Doctors write down lots of notes labs make test reports. Patients get papers when they leave the hospital or clinic. All this information, about care is important but the thing is it is not easy to look at.Doctors have to spend a time reading these reports. Medical paperwork is a problem. Patients often go home with papers they do not really understand.

Hospitals and clinics have to find a way to make medical paperwork easier to deal with.This situation is something we have personally noticed and wanted to work on. In this paper, we look at how Artificial Intelligence and Natural Language Processing can be used to make medical documents easier to work with for everyone involved. We reviewed four recent research papers on this topic and based on what they showed, we are proposing a system that can take a medical report as input and give back two summaries — one written for the medical team and one written in plain language for the patient. The system also generates discharge instructions and includes a chatbot this allows users to ask questions about the content they have just read. We use models like BioBERT, FLAN-T5, GatorTronGPT, and Longformer, together with a RAG setup for accuracy and ICD/SNOMED checks to make sure the medical content is correct. The papers we looked at have results for ROUGE-L, BERTScore and clinical accuracy. This shows that a system, like this can work well and be helpful.

Keywords: Medical Document Summarization, Natural Language Processing, BioBERT, FLAN-T5, Retrieval-Augmented Generation, Clinical NLP, Discharge Summarization, Transformer Models, MIMIC-III, Healthcare AI.

I. INTRODUCTION

If you have ever visited a hospital and received a bunch of papers at the time of discharge, you probably know the feeling of looking at them and not being sure what half of it means.Medical documents are written for doctors and other medical people so they use terms and shortcuts in a way that can be easily understood by them. For a person this is very confusing. For doctors who're very busy it takes a long time to read through all the medical notes to find the important information. This can be very stressful. Thinking about this problem because we thought computers could really help with it. Computers that understand language have gotten better over the years. Now they can read documents understand what they mean and even write summaries that sound like a person wrote them. When these computers are trained to read text they get even better at understanding what doctors are saying and writing things that are accurate and easy to read. Medical text is what they are good at so they can pick up on the language that doctors use and produce output that's both accurate and easy to understand which is especially helpful for doctors who use them to go through medical notes. Medical documents are one area where they work well.so they can understand the language that doctors use and produce output that's both accurate and easy to understand.

The main idea of this paper is to apply these capabilities in healthcare. We looked at four research papers that each tackled a part of this problem. One paper focused on generating summaries for both doctors and patients. Another paper worked on turning notes into discharge instructions. Another paper built a computer program that could answer questions from a document. Another paper addressed the issue of computers making up facts. Each of these is useful on its own. Combining these features into one system makes it more useful in real healthcare situations. The rest of this paper explains what we found from the reviewed papers and how these ideas can work together.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156111

II. LITERATURE REVIEW

We studied the work already done by the other researchers in this feild. It helped us understand what is already working. We also found out where the gaps are.

A. MediSense: Computer-Based Summaries of Medical Reports

MediSense caught our attention because it directly tackles the problem of one report needing to serve two different readers. The system reads a medical report and produces two different summaries. A technical one for the doctor and a simplified one the patient can actually read and understand. To do this it uses tools to understand the medical meaning of the text write the summaries and pull in relevant facts from a trusted medical database. The researchers tested it on 129 cases from the hospital database and got good results. Honestly those are numbers and the concept of generating two different summaries from one report is exactly the kind of thinking that should be built into a practical healthcare tool. These systems are well suited for medical documents, as they can understand clinical language and generate accurate, easy to understand outputs.

B. Auto-CarePlan: Computer-Based Discharge Instruction Generation

Auto-CarePlan works on a different problem. After a patient is treated someone has to sit down and write the discharge instructions. What medicines to take, what to avoid, when to come back. Medical documents are a key area where these systems perform well, so they can understand the language that doctors use and produce output that's both accurate and easy to understand which is really helpful, for doctors who use them to go through medical notes. This paper automates that process. The system reads the raw clinical notes written by the doctor and converts them into a clean, structured discharge plan. It uses Longformer because clinical notes tend to be very long and most models struggle with that.

GatorTronGPT does the actual writing of the instructions, and ICD/SNOMED ontologies are used to double-check that the output is medically correct. On MIMIC-III and the Discharge Me dataset, it got a ROUGE-L of 0.71, BLEU-4 of 0.64, BERTScore of 0.86, and 89.8% clinical accuracy.

C. BioMistral with RAG: Biomedical Chat Assistant

This paper is about making medical PDFs interactive. When you have a document you do not just have to read it. You can ask any kind of questions.You will get answers. The BioMistral model is based on Mistral. It was trained on data from PubMed Central. So the BioMistral model is good, at understanding the language that is used in biology and medicine. The BioMistral model works with SentenceTransformer embeddings and a Chroma vector store. This means the BioMistral model can look at a document and find the part that's most relevant. Then it uses that part to answer the question you asked about the document. The BioMistral model is very accurate. It got 92 percent answers. The F1 score of the BioMistral model is 0.86.

D. Hybrid Deep Symbolic NLP Framework

One thing that worries us about using AI for medical content is hallucination — the tendency for language models to sometimes write things that sound correct but are factually wrong. In a medical context, that is a real problem. This paper addresses it by combining a transformer model with a knowledge graph that stores verified medical facts. This layer decides in real time how much to trust the output and how much to refer to the symbolic check. On SQuAD 2.0 it achieved a ROUGE-L of 0.56 and BERTScore of 0.97, and on TruthfulQA it kept hallucinations down to just 15% while maintaining 85% consistency. For our system, this is the layer that gives us confidence that what the AI writes is actually true.

III. PROBLEM STATEMENT

Problems we observed in current healthcare documentation:

Medical reports and discharge papers are written in technical language that most patients simply cannot follow without help from a doctor or nurse.Patients who do not understand their documents often miss medication doses ignore follow-up appointments or return to the hospital with preventable problems.Doctors and nurses spend a lot of time reading reports, which takes away from actual patient care.There is no to-use tool that helps bridge the gap between how medical documents are written and how patients need to understand them.

Problems with existing AI and NLP tools:

Most NLP tools for healthcare are designed for either doctors or patients not both. This requires the use of two tools to serve both groups. We couldn't find a system that combines document summarization, discharge instruction generation and

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156111

an interactive chatbot. AI-generated medical content can have errors and most systems don't have a reliable way to catch and fix these errors.Most tools only work with text and can't handle images or other types of clinical data limiting their use.

IV. PROPOSED SYSTEM

We propose a system that brings together the features of reviewed papers into one tool. A user uploads a document and the system processes it providing three outputs: a detailed summary for the medical team a simple summary for the patient and a discharge plan. A chatbot is built-in allowing users to ask questions and get answers. The system checks its output for accuracy to ensure it is correct and safe.

A. System Architecture

The system works in ten steps. First the user uploads the document. If its a scanned image or photo OCR reads the text. The text is then cleaned up. Any patient-identifying information is removed. BioBERT converts the text into embeddings. For documents Longformer handles processing. A RAG module searches a database of verified documents to pull in supporting facts. Symbolic reasoning checks the content against a medical knowledge graph to catch errors. FLAN-T5 writes the summaries and GatorTronGPT writes the discharge instructions. Everything is validated against ICD and SNOMED standards before being shown to the user. The final interface is a Streamlit web app with side-by-side summaries and a BioMistral chatbot, for questions.

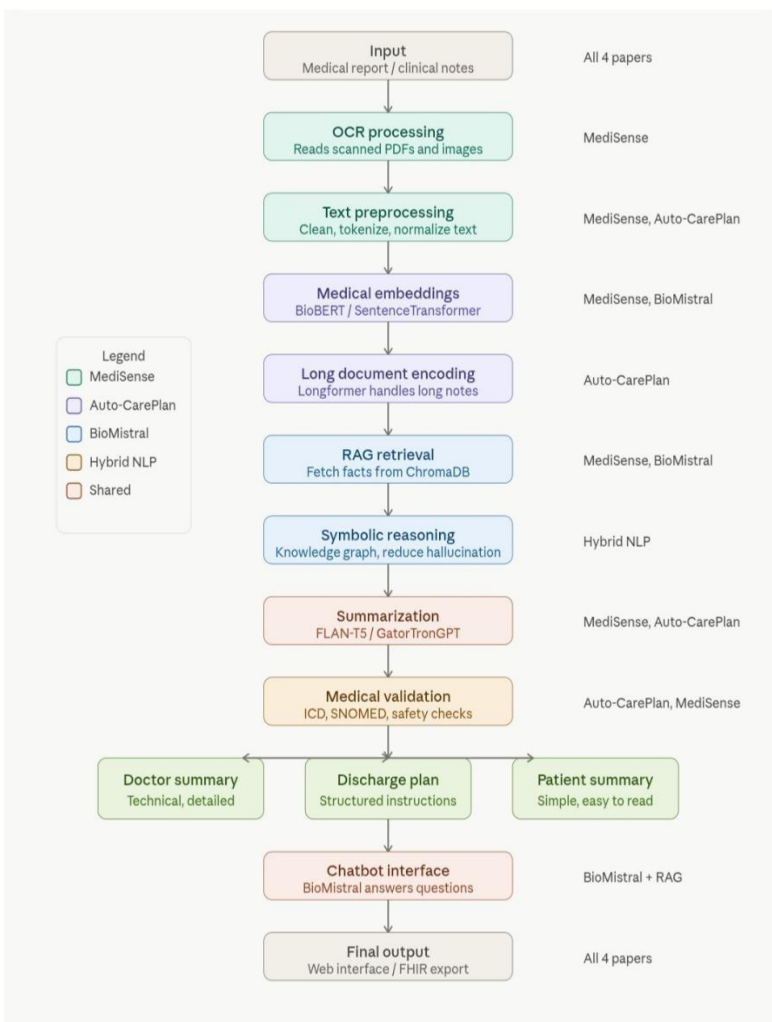

Fig:1 shows the Proposed System Architecture for AI-Based Medical Document Summarization.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156111

The key parts of this system are:
BioBERT. We chose BioBERT for the embedding step because it was trained on biomedical research papers. So when it sees a term like " infarction" it knows what that means in a medical sense unlike a regular language model that might not understand it.FLAN-T5. This is the part that writes the summaries. We picked FLAN-T5 because it can flow instructions such as "write this for a doctor" or "explain this to a patient with no medical background" and it changes the way it writes accordingly.The reason we like FLAN-T5 is that it can adjust to audiences, which is exactly what we need for our summaries.GatorTronGPT. This part handles the discharge instructions. It was trained on hospital documentation so it already knows what a discharge plan should look like and what information it needs to include.Longformer. Unlike most models, Longformer can process longer clinical notes without missing important information.

RAG with ChromaDB. Of just using what the model learned during training this part pulls in real verified medical facts from a database before generating any output. This keeps the summaries accurate and reliable.Symbolic Reasoning with Knowledge Graphs. This part acts like a fact checker that runs after the AI writes something. If the output contradicts a known fact this part catches it before it reaches the user.ICD and SNOMED Ontologies. These are the check before anything gets displayed. They are internationally recognized classification systems and using them at the end makes sure all the terms and diagnoses in the output are correct from a medical standpoint.

The whole system is built using Python. We use PyTorch as the foundation for the learning parts and the Hugging Face Transformers library to load and fine-tune the models. When a document is uploaded to a system the first thing it does is remove any patient-identifying information because privacy's very important in a healthcare context.After that BioBERT processes the cleaned text. Generates vector representations that carry the medical meaning of each part of the document. ChromaDB handles the storage and retrieval side of the RAG pipeline and LangChain helps manage the flow, between components.The system uses BioBERT to make sure the medical meaning is correct and FLAN-T5 to write the summaries in a way that's easy to understand.The system also uses GatorTronGPT to handle the discharge instructions and Longformer to process a lot of information at once.The system is designed to provide reliable summaries of medical documents and it uses a variety of tools and techniques to achieve this goal.

V. PERFORMANCE METRICS

The numbers below come from the four papers we reviewed. We are using them to show how each component of our proposed system performs on its own, and to justify why we chose these components over others.

System	ROUGE-L	BERTScore	Clinical Accuracy	Dataset
MediSense	0.72	0.89	91%	MIMIC-III
Auto-CarePlan	0.71	0.86	89.8%	MIMIC-III
BioMistral + RAG	N/A [1]	N/A [1]	92%	Biomedical QA
Hybrid NLP	0.56	0.97	85%	SQuAD 2.0

TABLE 1: COMPARATIVE PERFORMANCE OF PROPOSED SYSTEM COMPONENTS

[1] BioMistral + RAG is a question-answering chatbot, not a summarization system. ROUGE-L and BERTScore are summarization metrics and were therefore not measured or reported in that paper.

Table I below puts the four systems side by side so you can see at a glance how each one performed on its evaluation dataset

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.156111

Model	Approach	ROUGE-L	BERTScore	Accuracy	Audience
MediSense	BioBERT+FLAN-T5+RAG	0.72	0.89	91%	Dual
Auto-CarePlan	Longformer+GatorTronGPT	0.71	0.86	89.8%	Clinician
BioMistral+RAG	BioMistral+ChromaDB	N/A	N/A	92%	QA Chat
Hybrid NLP	Transformer+KG	0.56	0.97	85%	General
Clinical-T5	T5 Fine-tuned	0.66	0.81	83.4%	Clinician
BioBERT (base)	BERT Pretrained	0.61	0.77	80.5%	Clinician
RAG-LLAMA	LLaMA+RAG	0.53	N/A	N/A	Patient

TABLE 2: EXTENDED BASELINE COMPARISON ACROSS NLP MODELS

Several key points can be observed from these results. MediSense leads on clinical accuracy for summarization at 91%, which tells us that the dual-audience approach it uses is not just a nice idea but actually produces high-quality output. Auto-CarePlan's 89.8% clinical accuracy on discharge instructions is impressive given how unstructured the input data tends to be. The Hybrid NLP framework's BERTScore of 0.97 is the highest in the table and reflects how effective the knowledge graph verification step is at keeping the output semantically correct. BioMistral's 92% accuracy in biomedical question answering makes it the obvious choice for the chatbot layer. None of these systems alone does everything we need, but together they cover the full range of tasks we are targeting.

VI. CONCLUSION

This paper started with the idea: medical documents are not helping the people who need them the most, which is the patients and the doctors. Doctors are spending much time reading these documents and patients are often leaving the hospital without really understanding what is wrong with them or what they need to do next. We looked at four research papers that each made some progress on different parts of this problem. We used what these papers showed to propose a system that brings all of these things together.The result is a system that takes a document as input and produces summaries for both doctors and patients. It also produces a discharge plan and a chatbot capable of answering question about the content of the medical document. The medical document is checked for accuracy through knowledge graph verification and medical ontology validation.The performance results from the reviewed research papers are strong enough to suggest that this kind of system is ready to be tested in clinical settings. If it works the way we think it can it has the ability to make a difference in how medical information is shared between healthcare providers and the people they care for which is the patients.

VII. FUTURE ENHANCEMENT

The system only handles text right now. Adding support for languages, especially regional Indian languages would make the system more relevant in the Indian healthcare context where many patients are not comfortable using English.

The system is currently limited with text-based documents which're the medical documents. If we could bring in images like X-rays and lab graphs the summaries could reflect the picture of a patients condition rather than just the written notes in the medical documents.

Not all patients have the level of health literacy which is their understanding of medical information. A future version could ask the patient a few questions and then adjust the reading level of their summary in the medical documents accordingly.

ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.156111

Connecting the system to hospital management software via FHIR APIs would mean doctors and nurses could access summaries from within the tools they already use without switching between systems, which would be the hospital management software and the system that produces summaries of medical documents.

The knowledge graph that powers the fact-checking layer in the system needs to be kept up to date which means it needs to have the medical information. An automated pipeline that regularly pulls in clinical guidelines and research findings would keep the system accurate over time which is very important for medical documents.

A mobile app version would make the system easier for patients to use, especially for follow-ups at home or in community health settings. It would also help patients access their medical documents without depending on a desktop or laptop.

Expanding into areas like oncology, cardiology and neurology would require additional domain-specific training but it would make the system useful across a much wider range of clinical departments, which would be very helpful, for doctors and patients who need medical documents.

REFERENCES

[1] D. Shukla, A. Sorte, S. Shirote, S. Banchhor, O. Soma, and D. Takale, "MediSense: AI-Based Dual Summarization of Clinical Reports for Healthcare Professionals and Patients," 2025 3rd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIHEI), 2025.
[2] K. N, Y. M, R. C, and S. M, "Auto-CarePlan: Transformer-Based AI for Generating Discharge Instructions from Clinical Notes," 2025 5th International Conference on Evolutionary Computing and Mobile Sustainable Networks (ICECMSN), 2025.
[3] Mrs. Esther C, K. U P, A. G S, Mrs. Tamizhmalar D, E. V, and I. R N, "Biomedical Chat Assistant with Personalized Document Reader Using BioMistral and RAG," 2025 International Conference on Computing and Communication Technologies (ICCCT), 2025.
[4] Y. H. Farhan, M. Tareq, B. Shannaq, O. Ali, and S. Almaqbali, "Enhancing Robustness and Scalability in AI-Powered Natural Language Processing: A Novel Hybrid Deep Symbolic Framework," 2025 3rd International Conference on Cyber Resilience (ICCR), 2025.
[5] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Scientific Data, vol. 3, no. 1, p. 160035, 2016.
[6] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, 2019.
[7] Y. Labrak, A. Bazoge, E. Morin, and P. Gourraud, "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains," ACL 2024.
[8] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2023.
[9] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, pp. 1-67, 2020.
[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, pp. 4171-4186, 2019.
[11] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.
[12] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.