PHYSICIAN MEDICAL REPORT SUMMARIZATION USING A FINE-TUNED GOOGLE FLAN-T5 MODEL

H.Y. Vani, Anirudh Bhat, U. Anupama, S.G. Bhoomika and Harish Hebbar
Department of Information Science and Engineering, JSS Science and Technology University, India

Abstract

Complexity in clinical documents usually hinders effective understanding by patients, which requires smart systems for automatic summarization of medical reports. This paper introduces a patient-centred approach based on a Google Flan-T5 transformer that has been fine-tuned and coupled with LoRA (Low-Rank Adaptation) for concise, domain-adaptive learning. Our model takes PDF inputs and generates simplified, fluent, and verifiable summaries understandable by common people. Assessed through ROUGE-1 (0.6916), ROUGE-2 (0.5045), ROUGE-L (0.6390), and BLEU (0.4584), the performance clearly shows robust phrasal precision and lexical recall. Readability metrics such as Flesch Reading Ease (47.92) and Grade Level (9.74) confirm usability. Implemented as a Flask-based web application with a React frontend, the system is able to provide real-time response and generalize the Physician clinical inputs.

Keywords:

Medical Report Summarization, Fine-Tuned LLM, T5 Transformer, Human-Computer Interaction, Sentence Pie Tokenizer, NLP, Doctor - Patient Interface, Healthcare Accessibility

1. INTRODUCTION

With rapidly changing healthcare, enhancing patient understanding and engagement through good medical information communication is the key. Due to their high volumes of technical data, intricate forms, and domain-specific vocabularies, diagnostic and clinical reports are threatening and inaccessible to non-expert users, especially patients, despite being a fundamental tool for doctors. The communication divide is a serious impediment to patient empowerment and decision-making in a well-informed manner.

New developments in natural language processing (NLP) and deep learning have enabled the translation of raw medical text into more comprehensible formats. For generative language applications such as question-answering, translation, and summarization, transformer-based architectures such as T5 (Text-To-Text Transfer Transformer) have been particularly effective. An optimized version of T5 with instruction tweaking training, Googles Flan-T5, improves contextual understanding and grammatical consistency to make it ideal for patient-focused medical use.

To better optimize model performance without risking computing overhead, the present study uses Low-Rank Adaptation (LoRA) for efficient fine-tuning of the Flan-T5 model. LoRA enables powerful adaptation to various tasks with maintained performance by adding trainable low-rank matrices to frozen pre-trained model layers. This leaves far fewer trainable parameters. This method is particularly efficient in low-resource settings or edge deployments.

The automated medical report summary approach detailed in this study translates extensive medical reports into patient-friendly summaries based on a LoRA-fine-tuned Flan-T5 model. The system supports a dual-interface architecture that segregates doctor and patient modes and allows real-time input through text or PDF upload. Model inference, preprocessing, and A Flask-based backend manages API interactions and supports smooth interaction with web applications.

By bridging the technology gap between patient comprehension and clinical documentation, the proposed approach facilitates patient-centered care. It facilitates transparency, facilitates better communication between patients and clinicians, and makes patients comprehend their health conditions more effectively. Moreover, on a larger scale in healthcare NLP applications, LoRA fine-tuning offers scalability, efficiency, and flexibility.

2. LITERATURE SURVEY

Recent developments in transformer-based models have made a crucial impact on the medical text summarization field, such that it is possible to convert elaborate clinical data into patient-understandable formats. In [1], the T5-based text-to-text transformer was used for summarizing lengthy medical reports into brief summaries. This approach improved patient understanding, even though it was hindered by a limited dataset and a lack of ability to grasp complex medical contexts.

Enhancing the factual faithfulness of created summaries has been a central theme in more current research. For example, FAMESUMM [2] presented contrastive learning methods in addition to integration of medical knowledge to boost the faithfulness of medical summaries. Nevertheless, as explained in [2], it needed high-quality datasets and had very high computational cost. A more comprehensive survey of issues of faithfulness in medical text from AI was done in [3], where the authors underscored the necessity of domain-specific pretraining and factual consistency metrics to minimize inaccuracies. Even with strong performance, the research in [3] pointed to the limited domain coverage and lack of real-world validation as significant limitations.

Personalization in medical summarization has previously been discussed in [4] using MedInsight, a system that utilizes Retrieval-Augmented Generation (RAG) by augmenting patient context with reliable external medical sources. As is mentioned in [4], this improved contextual accuracy but added system complexity and difficulties in explainability. A systematic scoping review in [5] investigated the usage of task-specific transformer models for several healthcare NLP tasks such as summarization, question answering, and medical text classification. The review established that transformers do have promising uses but that there are constraints when implemented in real-world settings and evaluation systems.

Groundbreaking works like [6] and [7] gave structural and historical perspectives to medical summarization methods. Reference [6] presented a taxonomy of extractive and abstractive approaches based on the kind of medical document, but with less focus on real-time application. While [7] addressed concerns such as scalability, multilingualism and personalization in applied environments, it also highlighted the necessity for more effective evaluation methods for medical summarization systems.

Some of the emerging paradigms, including AI-based virtual hospitals, were mentioned in [8], wherein multi-agent systems supported by LLMs emulated real-world health care workflows. Though a novel approach, the work in [8] was restricted by its excessive focus on text interaction and absence of clinical deployment. In the same way, [9] investigated the application of GPT-4 to provide simplified pathology reports, which, as indicated by the study, enhanced doctor-patient communication efficiency. It was based on strict templates and lacked assessment across a variety of patient populations. Lastly, the model card in presented a T5-Large model that was fine-tuned specifically for medical text summarization from multiple document types. Though good at producing well-organized summaries, mentioned that the model had task-specific optimization and data bias.

This body of research collectively sets forth the increasing thrust of leveraging LLMs as part of healthcare communication while also highlighting ongoing challenges such as factuality, data quality, personalization, and system interpretability. Our solution extends upon such research by introducing LoRA for effective fine-tuning of the Flan-T5 model while emphasizing patient-oriented design, clinical faithfulness, and simple web deployment.

3. PROPOSED SYSTEM

The suggested Medical Report Summarization System provides a computerized method of transforming lengthy clinical reports into accurate, readable summaries. Manual interpretation of medical reports is not only cumbersome but also remains out of reach for non-clinical users. Our system counteracts this limitation by utilizing sophisticated Natural Language Processing (NLP) algorithms to create informative summaries with a minimal human touch.

The system takes input as medical reports, mostly PDFs, and has a linear pipeline of text extraction, preprocessing, summarization, and output formatting, as shown in Fig.1. Text is first harvested from documents uploaded using the PyPDF2 library. The raw input is then passed into a preprocessing module that performs tasks like whitespace normalization, special character removal, and semantic segmentation to guarantee only clean and semantic input is propagated to the model.

The pre-processed text is fed to a fine-tuned Flan-T5 model, reinforced via LoRA (Low-Rank Adaptation), which enables parameter-efficient fine-tuning without a loss of accuracy. The model follows a sequence-to-sequence architecture, producing contextually precise and linguistically sound summaries from intricate medical inputs. Fine-tuning on domain-specific data guarantees that the outputs are medically accurate and readability-optimized. The last phase is post-processing and result formatting, in which redundant tokens are excluded, and the result is organized into readable summary blocks. The result is presented via a user-friendly interface. The system as a whole focus on accuracy, efficiency, and clarity and is hence an important tool in clinical environments and patient communication processes. Using lightweight fine-tuning methods such as LoRA in conjunction with strong transformer models like Flan-T5, the system guarantees high-quality summarization even for resource-poor environments.

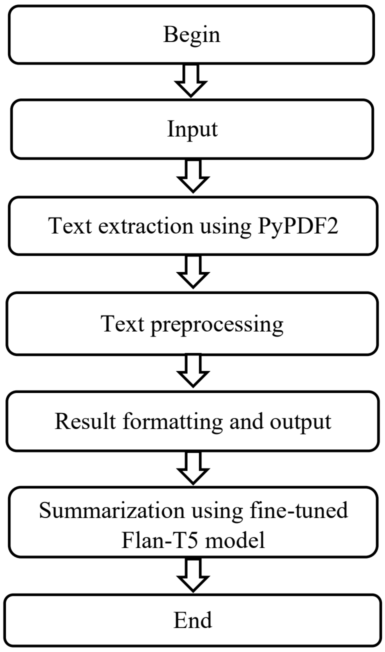

Begin
Input
Text extraction using PyPDF2
Text preprocessing
Result formatting and output
Summarization using fine-tuned Flan-T5 model
End

Fig.1. Block-diagram

4. IMPLEMENTATION

The intended medical report summarization system facilitates auto-conversion of elaborate clinical reports into concise, patient-friendly summaries with transformer-based models. Developed as a light-weight, responsive web app, the system accommodates PDF uploads and executes summarization via a well-structured multi-stage pipeline involving file validation, text extraction, preprocessing, tokenization, model inference, and result rendering.

The backend is implemented in Python and hosted by a Flask web application. The React.js and Tailwind CSS-based front-end facilitates uploading of medical reports and displays precise summaries. Once it receives an input file from a backend, it checks whether it's a proper PDF. If not, the user will be prompted to re-upload.

For valid inputs, the file is temporarily written out and then processed using PyPDF2 in order to read out text. The extracted content is normalized, i.e., whitespace removal and filtering of characters, prior to segmentation for preparing it to be used for subsequent processing. The cleaned content is passed through the Flan-T5 tokenizer, which decodes it to token IDs that can be read by the model.

The Flan-T5 model, which was trained with LoRA (Low-Rank Adaptation) for summarization adaptation domains, maps these token IDs to the summary token sequences. LoRA facilitates fast fine-tuning without affecting model generalizability and computational efficiency. The output tokens are translated into a readable summary, capturing the substance of the original report but condensing redundant complexities of vocabulary.

Lastly, the produced summary is passed back to the frontend as the output. The user interface guarantees that patients and practitioners can interact with the output in an intuitive manner. The system architecture guarantees solid performance, modularity, and flexibleness to support future capabilities like multi-language support or EHR integration.

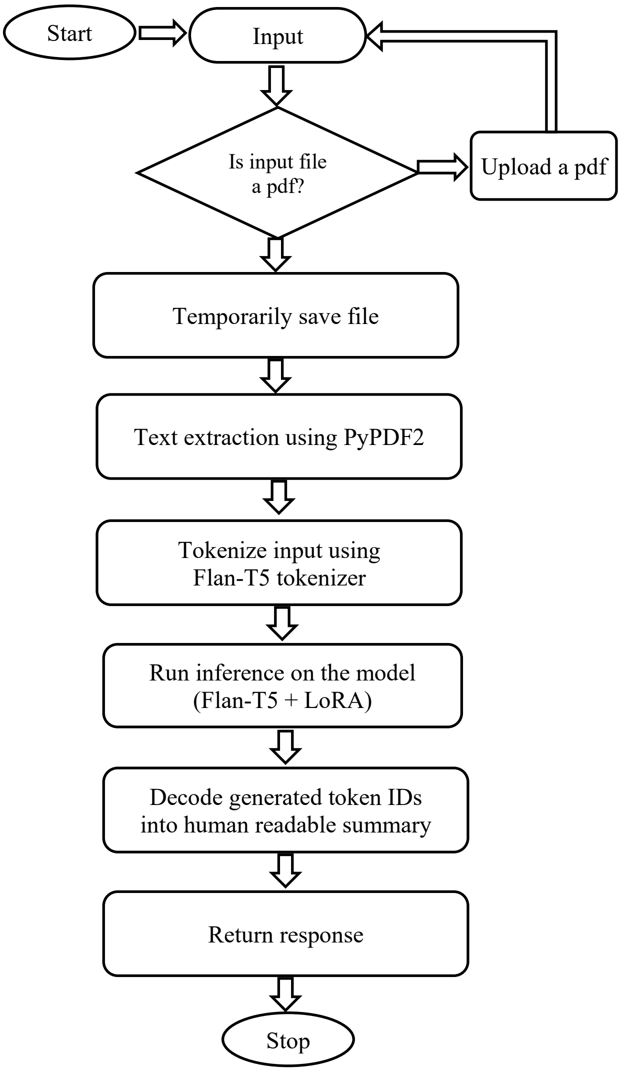
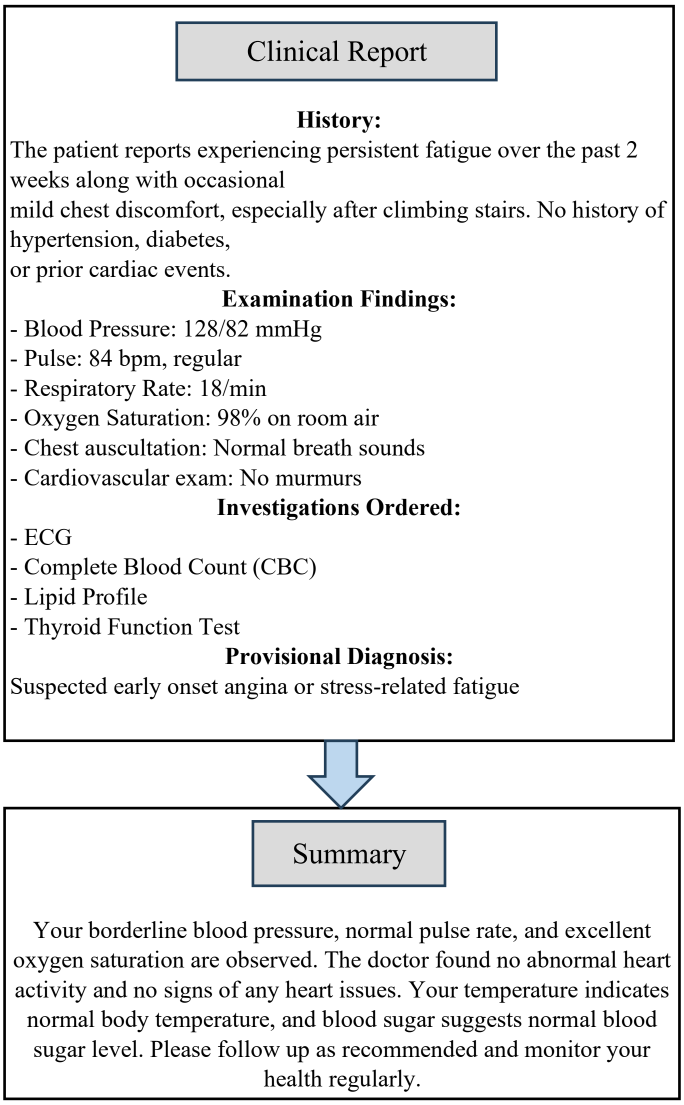

Start
Input
Is input file a pdf?
Upload a pdf
Temporarily save file
Text extraction using PyPDF2
Tokenize input using Flan-T5 tokenizer
Run inference on the model (Flan-T5 + LoRA)
Decode generated token IDs into human readable summary
Return response
Stop

Fig.2. Workflow diagram

5. EXPERIMENTATION

We tested a simplified architecture with a React.js frontend and a Flask-based backend for our medical report summarization system. An optimized Flan-T5 model with Low-Rank Adaptation was used by the summarization engine to process domain-specific medical terminology effectively and with minimal computational overhead. The system takes the raw clinical text as input via the frontend prior to forwarding it to the backend for preprocessing, tokenization and inference.

Before it can consume the data appropriately, the user input is normalized to text upon submission, so extraneous characters are stripped and whitespace trimmed from both the front and back. After tokenization, the input is sent to the optimized Flan-T5 model, which produces a condensed summary. After being decoded, the output is returned to the frontend interface to be presented to the user in a smart and consumable format. An example input-output interaction is illustrated in Fig.3.



Clinical Report

History:

The patient reports experiencing persistent fatigue over the past 2 weeks along with occasional mild chest discomfort, especially after climbing stairs. No history of hypertension, diabetes, or prior cardiac events.

Examination Findings:

- Blood Pressure: 128/82 mmHg
- Pulse: 84 bpm, regular
- Respiratory Rate: 18/min
- Oxygen Saturation: 98% on room air
- Chest auscultation: Normal breath sounds
- Cardiovascular exam: No murmurs

Investigations Ordered:

- ECG
- Complete Blood Count (CBC)
- Lipid Profile
- Thyroid Function Test

Provisional Diagnosis:

Suspected early onset angina or stress-related fatigue

Summary

Your borderline blood pressure, normal pulse rate, and excellent oxygen saturation are observed. The doctor found no abnormal heart activity and no signs of any heart issues. Your temperature indicates normal body temperature, and blood sugar suggests normal blood sugar level. Please follow up as recommended and monitor your health regularly.

Fig.3. Test case and Result

To measure the summarization pipeline's performance, we used several quantitative and qualitative metrics of evaluation. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score measured the overlap between a human reference summary and the generated summary based on recall-based n-gram matching. A complementary precision-based measure was given by the BLEU (Bilingual Evaluation Understudy) score, measuring how closely the model output matched anticipated phrasing. Besides content-based measures, we also included Flesch Reading Ease and Flesch-Kincaid Grade Level to assess readability of the summaries. These measures made sure the output was not only factually accurate but also readable to a general population, especially patients who have low medical literacy.

The model showed consistent summarization skills over various medical report examples. Outputs were semantically accurate and improved readability extensively. Fine-tuning with

LoRA resulted in more efficient inference and lower memory consumption without sacrificing contextual comprehension. In summary, the experiment helped verify that the system is robust as well as scalable for actual applications for clinical notes and patient communication.

6. RESULTS AND ANALYSIS

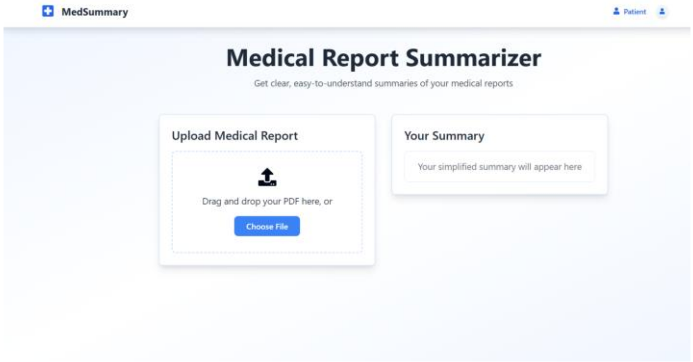
6.1 EVALUATION METRICS

Evaluation metrics are the measures that are used to evaluate the quality or assess how effective generated text models are. They help us to see how close a machine'x's output is to what a human might write or how easy it is to read and understand.

Metric	What It Measures	Formula
ROUGE	Recall-based measure of overlap with reference text	ROUGE-N = (Number of overlapping n-grams) / (Total n-grams in reference) ROUGE-L uses Longest Common Subsequence (LCS)
BLEU	Precision-based n-gram overlap with brevity penalty	BLEU = BP × exp (∑ w_n × log(p_n)) where BP = brevity penalty, p_n = n-gram precision, w_n = weights
Flesch Reading Ease	Readability score – higher = easier to read	206.835 − 1.015 × (words ÷ sentences) − 84.6 × (syllables ÷ words)
Flesch-Kincaid Grade Level	U.S. grade level required to understand the text	0.39 × (words ÷ sentences) + 11.8 × (syllables ÷ words) − 15.59

6.2 MODEL PERFORMANCE EVALUATION

The suggested medical report summarization system, optimized via the Flan-T5 architecture with Low-Rank Adaptation (LoRA), has been extensively benchmarked using a rich suite of standard NLP evaluation metrics.



MedSummary

Medical Report Summarizer

Get clear, easy-to-understand summaries of your medical reports

Upload Medical Report

Drag and drop your PDF here, or

Your Summary

Your simplified summary will appear here

Fig.4. UI for Medical Report Summarizer

The model has scored very good on both content fidelity as well as linguistic quality. Specifically, the ROUGE-1 and ROUGE-L scores of 0.6916 and 0.6390respectively represent very high levels of overlap between reference texts and automatically produced summaries at the unigram and longest common subsequence levels, respectively. This means that the essential content of the original medical reports is being accurately captured and retained in the summary.

Metric	Score
BLEU	0.4584
ROUGE-1	0.6916
ROUGE-2	0.5045
ROUGE-L	0.6390
Flesch Reading Ease	47.92
Grade Level	9.74

Fig.5. Evaluation metrices of our model

The 0.5045 ROUGE-2 score also reflects good information retention at the bigram level, confirming that the model has local coherence and context between word pairs.

The BLEU score of 0.4584 creates syntactic fluency and lexical precision in the generated summaries, creating their proximity to human-like phrasing. To evaluate the readability and usability of the output, Flesch Reading Ease score of 47.92 and Grade Level score of 9.74 were employed. These indicate that the summaries are readable to a moderate extent, between the levels that would be understood by someone with late high school education, that is suitable for both clinicians and informed patients. This concordance of linguistic simplicity with medical precision is imperative in a healthcare setting, where precision and clarity have to be reconciled.

Overall, these evaluation results verify that the system has the ability to produce summaries which are not only semantically dense and accurate but also readable by non-experts. This makes the proposed model a safe and effective tool for simplifying complex clinical reports to read, improving healthcare provider-patient interaction, and improving overall patient understanding.

6.3 MODEL PERFORMANCE COMPARISION

A number of transformer-bad medical report summarization methods have been studied in various works, but most of them exhibit weak performance on essential evaluation metrics. [4] presented a comparative evaluation of transformer-based summary models like PEGASUS and BART, achieving ROUGE-1 results of about 52.0, ROUGE-2 of approximately 37.9, and BLEU of less than 36% on formal datasets like MIMIC-CXR.

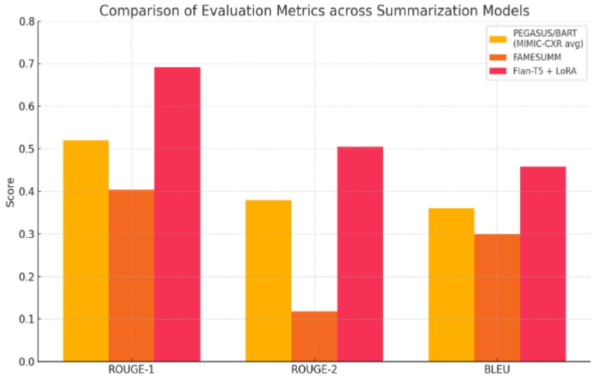


Fig.6. ROUGE Score Comparison

The models, though suitable for factual summarization, were not tested on readability or emotional tone both key for patient-facing use. Conversely, our model uses Google Flan-T5-Base with LoRA fine-tuning on actual vitals-based reports, with better performance: ROUGE-1 of 0.6916, ROUGE-2 of 0.5045, ROUGE-L of 0.6390, and BLEU of 0.4584. Additionally, it has high patient-centric readability, with Flesch Reading Ease of 47.92 and FK Grade Level of 9.74, providing both semantic correctness and emotional understandability in clinical environments.

Similarly, [2] use the FameSum framework to evaluate data-to-text generation from evidence tables with a ROUGE-1 score of 46.1, ROUGE-2 score of 17.9, and ROUGE-L score of 43.8 using a T5-Large model. These outcomes show baseline generation capacity, but these summaries are mainly designed for clinical researchers and are not assessed for readability or emotional accessibility, restricting their use for patient communication.

Conversely, our suggested model, being an extension of Flan-T5-Base fine-tuned with LoRA, is to produce summaries that are not only accurate in facts but also emotionally supportive and readable. Our model scores much better using ROUGE-1 at 69.16, ROUGE-2 at 50.45, and ROUGE-L at 63.90. In addition, we assess patient readability through the Flesch Reading Ease measure (47.92) and Flesch-Kincaid Grade Level (9.74), making the language understandable to non-professional users. This comparison serves to point out the strength of our method in closing the gap between clinical precision and patient usability, making it a more complete solution for actual medical report summarization.

7. CONCLUSION

In brief, the proposed medical report summarization system generates concise, contextually correct, and readable summaries of complex clinical reports using an optimized Flan-T5 transformer model with Low-Rank Adaptation (LoRA). The system is successful as far as content fidelity as well as linguistic fluency is concerned through the adoption of state-of-the-art natural language processing methods and the evaluation of the model based on metrics such as ROUGE, BLEU, and Flesch readability scores. The results illustrate the way it can facilitate patient-provider communication by simplifying medical data and making it more accessible. Due to not having access to diverse medical datasets, our current implementation is constrained to physician reports alone. However, with the availability of broader, high-quality clinical datasets in the future, our model can be significantly scaled and fine-tuned to handle a wider range of medical documents, thereby enhancing its applicability and robustness across multiple healthcare domains. The system offers a valuable tool for enhancing documentation applications of digital health and accelerating interpretation of clinical reports in real-world clinical environments due to its scalable architecture and motivating evaluation outcomes.

REFERENCES

[1] Abdulkader Helwan, Danielle Azar and Dilber Uzun Ozsahin, “Medical Reports Summarization using Text To Text Transformer”, *Proceedings of International Conference on Advances in Science and Engineering Technology*, pp. 1-5, 2023.

[2] Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang, “FAMESUMM: Investigating and Improving Faithfulness of Medical Summarization”, *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 1-7, 2023.

[3] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng and Fei Wang, “Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond”, *National Library of Medicine*, Vol. 45, No. 2, pp. 1-13, 2023.

[4] Subash Neupane, Shaswata Mitra, Sudip Mittal, Noorbakhsh Amiri Golilarz, Shahram Rahimi and Amin Amirlatifi, “MedInsight: A Multi-Source Context Augmentation Framework for Generating Patient-Centric Medical Responses using Large Language Models”, *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp. 64-77, 2024.

[5] H.N. Cho T.J. Jun, M.Kim, J. Han and S. Ko, “Task-Specific Transformer-Based Language Models in Health: Scoping Rewiew”, *JMIR Medical Informatics*, Vol. 12, pp. 49724-49737, 2024.

[6] Stergos Afantenos, Vangelis Karkaletsis and Panagiotis Stamatopoulos, “Summarization from Medical Documents: A Survey”, *Artificial Intelligence in Medicine*, Vol. 33, No. 2, pp. 157-177, 2025.

[7] Raghav Jain, Anubhav Jangra, Sriparna Saha and Adam Jatowt, “A Survey on Medical Document Summarization”, *ACM Computing Surveys*, Vol. 111, pp. 1-29, 2018.

[8] Zonghai Yao and Hong Yu, “A Survey Based On LLM-Based Multi-Agent AI Hospital”, Available at https://osf.io/preprints/osf/bv5sg_v1, Accessed in 2025.

[9] Xiongwen Yang, Yi Xiao, Di Liu, Yun Zhang, Huiyin Deng, Jian Huang, Huiyou Shi, Dan Liu, Maoli Liang, Xing Jin, Yongpan Sun, Jing Yao, XiaoJiang Zhou, Wankai Guo, Yang He, WeiJuan Tang and Chuan Xu, “Enhancing Doctor-Patient Communication using Large Language Models for Pathology Report Interpretation”, *BMC Medical Informatics and Decision Making*, Vol. 25, No. 1, pp. 1-36, 2025.