Ollama-Driven Medical Insights Using LLMs with a Federated Learning Approach

Journal:	*IEEE Internet of Things Journal*
Manuscript ID	IoT-49395-2025
Manuscript Type:	Regular Article
Date Submitted by the Author:	05-Apr-2025
Complete List of Authors:	Lal, Gurbaksh; Netaji Subhas University of Technology Rathee, Geetanjali; Netaji Subhas University of Technology Kerrache, Chaker; Universite Amar Telidji Laghouat Song, Houbing ; University of Maryland Baltimore County
Keywords:	eHealth and mHealth < Sub-Area 3: Services, Applications, and Other Topics for IoT, Smart Cities < Sub-Area 3: Services, Applications, and Other Topics for IoT

Ollama-Driven Medical Insights Using LLMs with a Federated Learning Approach

Gurbaksh Lal, Geetanjali Rathee, Chaker Abdelaziz Kerrache, and Houbing Herbert Song

***Abstract*—Traditional medical diagnostics often suffer from delays and inconsistencies due to the manual interpretation of unstructured patient data. To overcome these challenges, we introduce Our Model (given name as 'AI Doctor')—a novel diagnostic system built on the Ollama platform that integrates multiple pre-trained large language models (meditron, medllama2, wizardlm2, and mistral) through an innovative prompt filtering mechanism. AI Doctor accurately interprets patient-reported symptoms to deliver precise diagnoses and personalized treatment recommendations, while its design supports robust local deployment and includes a theoretical framework for federated learning. This federated approach facilitates decentralized, privacy-preserving model updates across healthcare institutions. Performance evaluations using BLEU scores, structured output analysis, and inference speed measurements demonstrate that AI Doctor consistently outperforms individual models, ensuring high diagnostic accuracy and realtime clinical applicability.**

***Index Terms*—Ollama, Large Language Model (LLM), Federated Learning, Meditron, MedLLaMA2, WizardLM2, Mistral, Performance Metrics, BLEU Score, Inference Speed, Output Consistency, Medical AI.**

I. Introduction

The advent of Large Language Models (LLMs) has significantly transformed the landscape of Medical AI, offering new avenues for automated clinical decision-making, diagnosis, and patient care [1].Traditional diagnostic methodologies often suffer from inconsistencies, delays, and the challenge of interpreting complex medical narratives. In response, this research introduces AI Doctor, a cutting-edge diagnostic system built upon Ollama, a platform that facilitates seamless access to multiple LLMs tailored for healthcare applications. At the core of AI Doctor is the strategic integration of specialized models, including Meditron, MedLLaMA2, WizardLM2, and Mistral, each contributing unique strengths to the diagnostic pipeline. Through an innovative prompt filtering mechanism, the system synthesizes insights from these models to generate accurate disease diagnoses, personalized treatment recommendations, and structured medical insights [2], [3]. This multi-model approach ensures that AI Doctor harnesses the domain-specific expertise embedded in each LLM, thereby enhancing overall diagnostic reliability. Recognizing the critical importance of privacy and scalability in medical AI, this study also explores the potential of Federated Learning as a means of enabling decentralized model updates across health care institutions. By allowing institutions to collaboratively train AI models without sharing sensitive patient data, federated learning enhances data security while continuously refining diagnostic capabilities. To rigorously evaluate the system's effectiveness, AI Doctor is assessed using key performance metrics such as BLEU Score (to measure diagnostic accuracy), inference speed (to ensure real-time application feasibility), and output consistency (to validate the reliability of model-generated diagnoses across multiple runs) [4], [5]. These metrics provide a comprehensive evaluation of the system's strengths and imitations, offering insights into its real-world applicability in clinical settings. By integrating state-of-the-art LLMs, federated learning, and robust performance assessment, AI Doctor represents a significant advancement in Medical AI. This research not only demonstrates the feasibility of multi-model diagnostic systems but also lays the groundwork for scalable, privacy-preserving, and high-accuracy medical decision support tools that have the potential to revolutionize modern healthcare.

A. Motivation

The rapid evolution of artificial intelligence is transforming medical diagnostics, which have traditionally relied on manual interpretation of patient symptoms—a process often prone to inconsistencies and delays. The increasing complexity of clinical data, characterized by unstructured narratives and variable patient descriptions, further exacerbates these challenges [6]. Recent advancements in Large Language Models (LLMs) offer a promising alternative [7]. By processing vast amounts of clinical text, LLMs can extract actionable insights, ensuring datadriven decision-making and consistent diagnostic outcomes. Models such as Meditron, MedLLaMA2, WizardLM2, and Mistral demonstrate how specialized LLMs can be leveraged to enhance diagnostic precision and generate detailed treatment recommendations. Yet, while centralized AI models can achieve impressive performance, they typically require the aggregation of large volumes of sensitive patient data, raising significant privacy and regulatory concerns. Federated Learning (FL) addresses this issue by enabling decentralized model training. In an FL frame work, healthcare institutions can collaboratively update a shared model without the need to transfer raw patient data, thereby preserving privacy while still benefiting from collective intelligence [8], [9].

(Corresponding Authors: Chaker Abdelaziz Kerrache)

Gurbaksh Lal is with the Department of Computer Science and Engineering, Netaji Subhas University of Technology, Dwarka Sector-3, New Delhi-110078, India. (e-mail: geetanjali.rathee123@gmail.com)

Geetanjali Rathee is with the Department of Computer Science and Engineering, Netaji Subhas University of Technology, Dwarka Sector-3, New Delhi-110078, India. (e-mail: gurbakshlal999@gmail.com)

Chaker Abdelaziz Kerrache is with the Laboratoire d'Informatique et de Mathématiques, Université Amar Telidji de Laghouat, Laghouat, Algeria. (e-mail: ch.kerrache@lagh-univ.dz)

Houbing Herbert Song is with the Department of Information Systems, University of Maryland, Baltimore County (UMBC), Baltimore, USA. (e-mail: songh@umbc.edu)

II. Literature Survey

This section reviews the current state of research relevant to our work. We discuss key concepts including Large Language Models (LLMs), Federated Learning (FL), the Ollama platform, Medical AI, and the integration of Federated Learning with LLMs. The figures (Figures 1, Figure 2, and Figure 3) provide architectural overviews from recent studies, and Table I summarizes recent contributions arranged by publication year.

A. Large Language Models (LLMs)

Large Language Models (LLMs) are deep neural networks—typically based on the transformer architecture—that are pretrained on massive text corpora. They capture complex linguistic patterns and semantic relationships, enabling a wide range of natural language processing tasks such as text generation, summarization, translation, and question answering. Key components include:

- Embedding Layers: Convert words or tokens into dense vector representations.
- Self-Attention Mechanisms: Enable the model to weigh the importance of different words in a sentence.
- Feed-forward Networks: Process the aggregated information to produce contextualized outputs.

Recent models like GPT-3 and BERT have set new performance benchmarks and are increasingly fine-tuned for specialized tasks across various domains. As shown in Figure 1, a typical LLM architecture integrates these components in a transformer-based design.

B. Federated Learning

Federated Learning (FL) is a decentralized machine learning paradigm that enables multiple participants—such as hospitals or edge devices—to collaboratively train a model without sharing raw data. Key aspects include:

- Decentralization: Each participant trains a local model, and only model updates are shared with a central server.
- Privacy Preservation: Techniques such as secure aggregation and differential privacy protect sensitive data.
- Heterogeneity Handling: Various methods have been designed to handle non-IID data distributed among different participants.

FL is particularly beneficial in healthcare, where data privacy and regulatory compliance are critical. Figure 2 represents an overview of the FL architecture, showing how local models update a central server without compromising individual data privacy.

C. Ollama overview

Ollama is a platform designed to simplify access to and customization of pretrained LLMs for specific applications. It offers:

- Unified Interface: A single platform to query, fine-tune, and deploy LLMs.
- Modularity: Support for rapid prototyping and iterative development.
- Scalability: Reduced computational overhead by leveraging pretrained models rather than training from scratch.

Ollama accelerates the integration of LLMs into domain-specific systems such as medical diagnostics.

D. Medical AI

Medical AI involves applying artificial intelligence techniques to address healthcare challenges. Its goals include:

- Enhanced Diagnostics: Improving the accuracy and speed of disease diagnosis.
- Personalized Treatment: Tailoring therapy recommendations based on patient-specific data.
- Operational Efficiency: Streamlining administrative and clinical workflows.

Advanced deep learning models in Medical AI aim to transform patient care by integrating diverse data sources while addressing issues of data privacy and interoperability.

E. Federated Learning with LLMs

Integrating Federated Learning with LLMs leverages the powerful natural language understanding of LLMs and the decentralized, privacy-preserving benefits of FL. This approach is particularly promising for Medical AI, as it:

- Ensures Data Privacy: Allows institutions to update models collaboratively without sharing sensitive patient data.
- Enhances Model Robustness: Aggregates diverse data from multiple sources, leading to more generalized and robust models.
- Facilitates Real-time Diagnostics: Supports scalable, real-time diagnostic tools that can op erate across various clinical settings.

Figure 3 illustrates an integrated architecture where federated updates are applied to specialized LLM components, ensuring both high performance and data privacy.

F. Summary of Related Work

As summarized in Table I, recent research has focused on a range of topics from enhancing LLM capabilities with few-shot learning to integrating retrieval mechanisms and applying FL for secure, decentralized model updates in healthcare settings. The table below, arranged by publication year, details the research focus, methodology, and key contributions of each work. Our work builds upon these concepts by integrating multiple pretrained LLMs via the Ollama platform and exploring federated learning as a strategy for future scalability and enhanced privacy preservation.

III. Methodology

This study's methodology is organized into three main sections: (I) System Architecture and Functional Mapping (II) Code Implementation and Explanation, (III) Experimental Outputs and Analysis.

A. System Architecture and Functional Mapping

This section describes the architecture and functional workflow of the AI Doctor system under three different configurations: (1) AI Doctor with LLM only, (2) AI Doctor with FL only, and (3) AI Doctor integrating both LLM and FL.

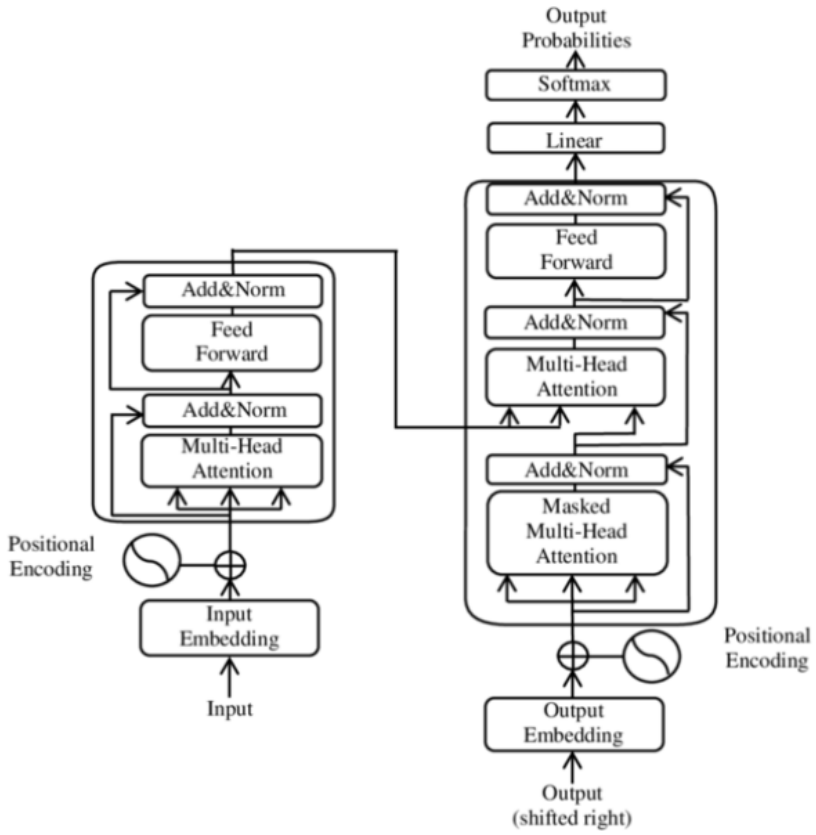

Fig. 1. Overview of a typical Large Language Model architecture, illustrating key components such as embedding layers, multi-head attention, and feed-forward networks

B. AI Doctor Architecture with LLM Only

The AI Doctor system utilizing only a Large Language Model (LLM) follows a structured pipeline where patient data is processed and used to generate diagnostic reports. The architecture comprises the following key components:

- Input Module: Captures patient symptoms and medical history using an interactive interface (e.g., Gradio).
- Preprocessing Module: Tokenizes, standardizes, and extracts relevant features from input text.
- Model Integration Module: Employs multiple LLMs, such as Ollama-based models, to interpret queries and generate responses.
- Aggregation Module: Uses Retrieval-Augmented Generation (RAG) to merge responses from different models.
- Evaluation Module: Measures response accuracy, completeness, and efficiency.
- Output Module: Produces diagnostic reports with detailed explanations.
- Feedback Loop: Captures user interactions and corrections to refine future predictions.

The system includes a feedback loop to improve model performance iteratively. Figure 4 illustrates this architecture.

C. AI Doctor Architecture with FL Only

In this architecture, the AI Doctor system leverages Federated Learning (FL) to train models across distributed client nodes (e.g., hospitals, clinics) while preserving data privacy. The key components include:

- Client Nodes: Hospitals or medical institutions perform local training on private patient data.
- Preprocessing Module: Standardizes and extracts features from patient records before training.
- Federated Aggregation Module: Securely aggregates model updates from local nodes without exposing raw data.
- Global Model Update: Combines distributed model updates to improve overall accuracy.
- Evaluation Module: Assesses model performance across different medical cases.
- Output Module: Provides AI-generated diagnostic assistance based on FL-trained models.
- Privacy Protection Mechanisms: Implements encryption and differential privacy techniques to ensure secure data handling.

This privacy-preserving approach ensures compliance with medical data regulations. Figure 5 illustrates the AI Doctor FL-based architecture.

D. AI Doctor Architecture with LLM and FL

The hybrid architecture combines the natural language capabilities of LLMs with the privacy advantages of Federated Learning. The components include:

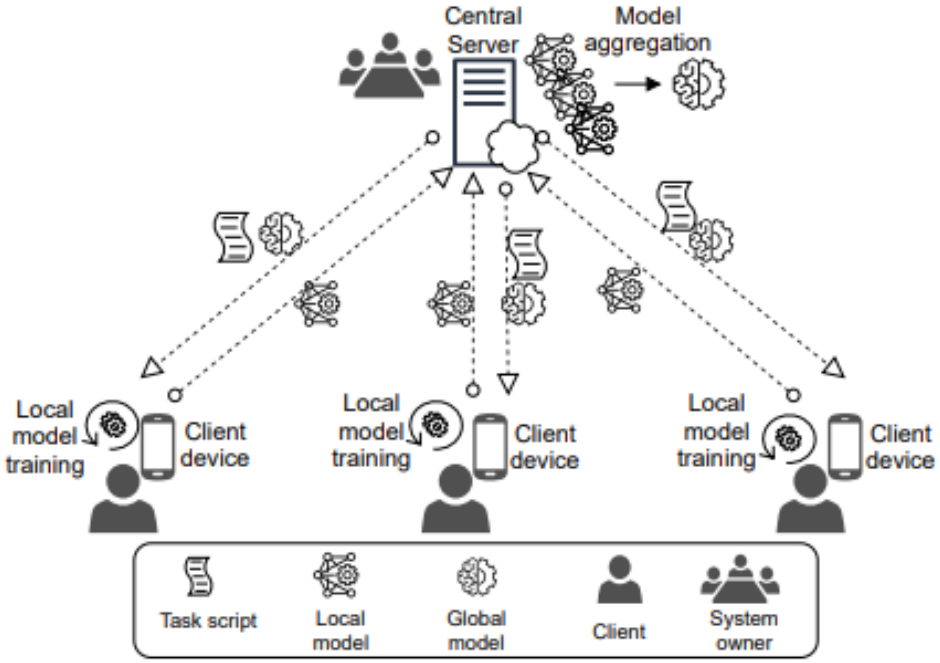

Fig. 2. Overview of Federated Learning architecture, illustrating decentralized training with local models updating a central server while ensuring data privacy

TABLE I
SUMMARY OF RECENT WORK IN FEDERATED LEARNING AND LARGE LANGUAGE MODELS

Name of Author	Method/Scheme	Result Metrics	Limitation
Moon et al. [10]	Few-shot learning with GPT models	Analysis of GPT's few-shot capabilities	Demonstrates GPT's potential to perform tasks with minimal examples, emphasizing efficiency in medical NLP.
Gupta et al. [11]	Retrieval Augmented Generation in LLMs	Integration of retrieval mechanisms with LLMs	Enhances knowledge intensive tasks by incorporating external information, improving contextual understanding and output quality.
Win et al. [12]	Multi-institutional FL collaboration	Federated strategy for MRI-based classification	Proposes a method for multi-institutional collaboration with out raw data exchange, enhancing diagnostic accuracy.
Jin et al. [13]	Advances in Federated Learning	Comprehensive survey	Identifies challenges in scalability, communication efficiency, and fairness in federated systems, guiding future research
Nguyen et al. [14]	Wellness detection with FL	Clustered federated learning	Improves detection accuracy by clustering clients based on data similarity and reducing communication overhead.
Fan et al. [15]	Privacy-preserving FL in healthcare	Secure federated framework	Demonstrates robust privacy mechanisms enabling model updates without exposing patient data.
Oh et al. [16]	Time series fore casting via LLMs	Reprogramming pretrained LLMs	Repurposes LLMs for time series analysis, demonstrating superior forecasting accuracy
Greenspan et al. [17]	FL integration for medical diagnostics	Combining FL with transformer-based models	Provides insights on integrating decentralized learning with LLM architectures for real-time diagnostic support.
Poulain et al. [18]	Scalable FL for clinical decision support	Large-scale federated learning	Presents a scalable approach for training LLMs on distributed medical data while preserving patient privacy.
Cheng et al. [19]	Multimodal clinical prediction using LLMs	Unified prompt-based integration	Enhances clinical prediction by leveraging text, image, and sensor data for improved diagnostic precision.

- Input Module: Captures patient queries and medical history.
- Data Augmentation: Enhances input records with structured metadata.
- Preprocessing Module: Tokenizes and extracts relevant features from the input.
- Prompt Filtering: Ensures queries are structured optimally for LLMs.

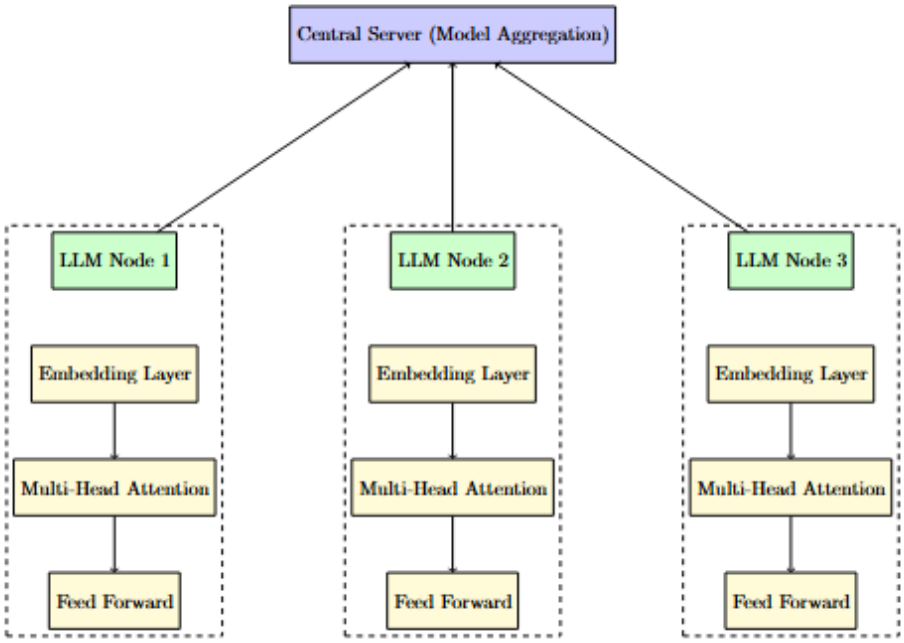

Fig. 3. Integrated architecture demonstrating how Federated Learning is combined with LLMs to enable collaborative, privacy-preserving training for medical applications

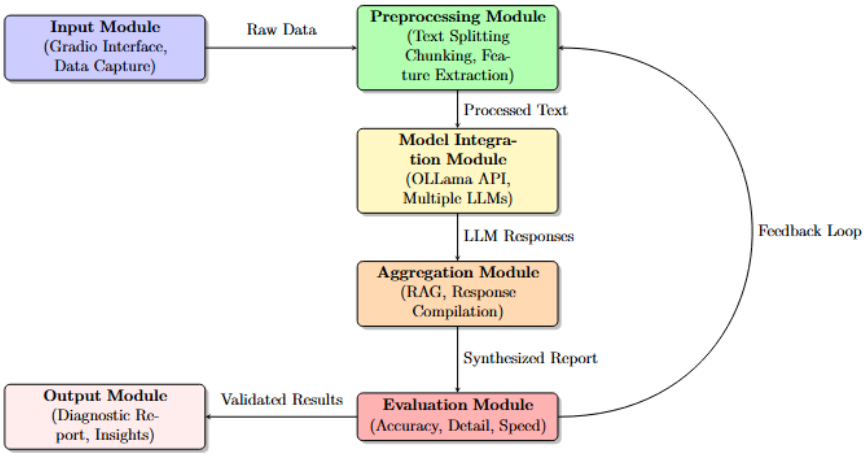

Fig. 4. AI Doctor system architecture using LLM only

- Model Integration Module: Utilizes multiple LLMs (e.g., Meditron, MedLLaMA2) to generate medical insights.
- Aggregation Module: Combines outputs from different LLMs for improved accuracy.
- Federated Learning Module: Trains models locally at hospitals and aggregates updates without sharing patient data.
- Evaluation Module: Measures BLEU scores, response quality, and clinical relevance.
- Output Module: Provides validated AI-generated diagnoses.
- Security Mechanisms: Implements end-to-end encryption, federated averaging, and model pruning to enhance security and efficiency.
- Scalability Considerations: Supports multi-institutional collaboration, ensuring the system remains effective in diverse medical environments.

This hybrid model enhances both accuracy and data security while maintaining real-time medical insights. Figure 6 illustrates this architecture.

This integrated approach leverages the strengths of both LLMs and FL, ensuring an optimal balance between intelligence, security, and efficiency for AI-driven medical diagnostics. Additionally, the combination of models improves generalization across different patient demographics while preserving sensitive medical information through decentralized learning.

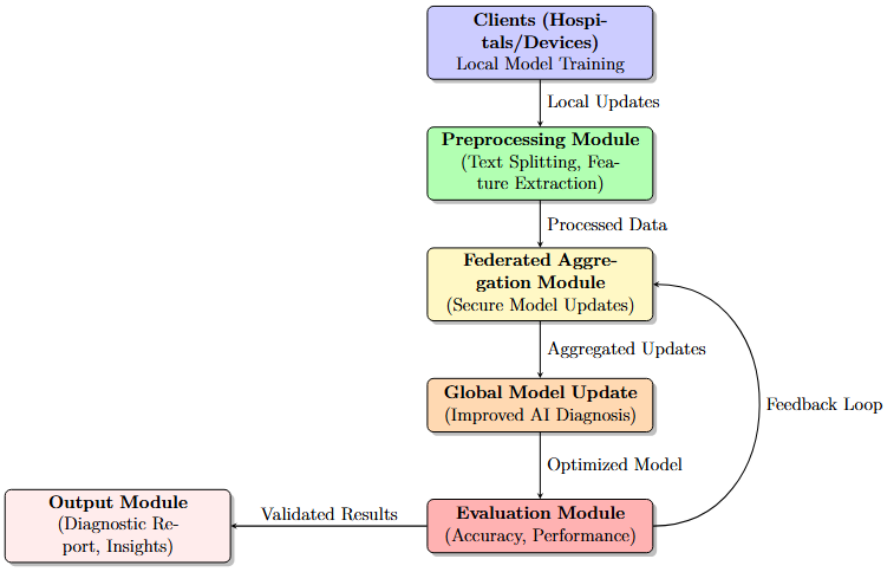

Fig. 5. AI Doctor system architecture using LLM only

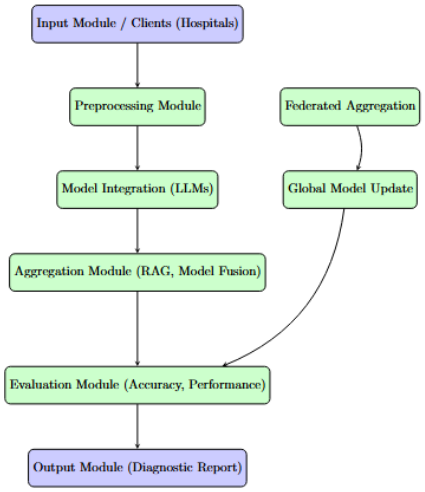

Fig. 6. AI Doctor system architecture combining LLM and FL

IV. Code Implementation and Explanation

The AI Doctor system is a sophisticated AI-driven medical diagnostic framework designed to leverage large language models (LLMs) for efficient and accurate diagnosis. The implementation follows a struc tured and modular approach, ensuring robustness, scalability, and efficiency. The execution process is outlined in Algorithm.

A. Algorithm Overview

Algorithm 1 provides a structured workflow of the AI Doctor system, consisting of six key stages:

- System Initialization: The system loads the AI Doctor model and pre-trained LLMs, including Meditron, MedLlama2, WizardLM2, and Mistral. System parameters such as temperature settings are initialized for optimal performance.
- Data Processing: Input patient symptoms undergo cleaning, normalization, and conversion into a structured format suitable for model interpretation.
- Model Querying: The processed input is sent to the AI Doctor model, where multiple LLMs generate responses, which are then stored.
- Federated Learning Implementation: A federated learning approach aggregates model updates across decentralized nodes, ensuring improved model learning without sharing raw patient data.
- Diagnosis Generation: The system compiles aggregated responses from different LLMs into a structured diagnosis and treatment recommendations.
- Performance Optimization: The final step involves optimizing inference speed, reducing latency, and improving diagnostic accuracy. Additionally, the system continuously refines model parameters based on user feedback and real-world testing.

B. Algorithm Explanation

The AI Doctor system follows a modular approach to ensure efficiency and accuracy in medical diagnostics. The key functional modules are defined in Algorithm 1: Each function in the algorithm plays a critical role in ensuring a structured diagnostic process. The federated learning approach enhances the AI Doctor system's ability to learn from diverse sources while maintaining patient data privacy.

C. Experiment Output and Analysis

To demonstrate the functionality and effectiveness of the AI Doctor system, we present experiment outputs below.

The first image (Figure 7) shows the AI Doctor system responding to a query about its identity. It provides an overview

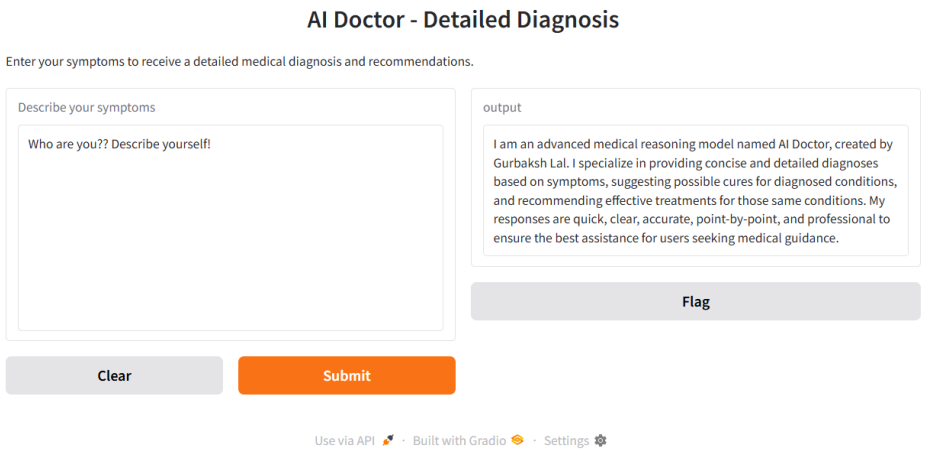

Fig. 7. AI Doctor system introducing itself and its creator

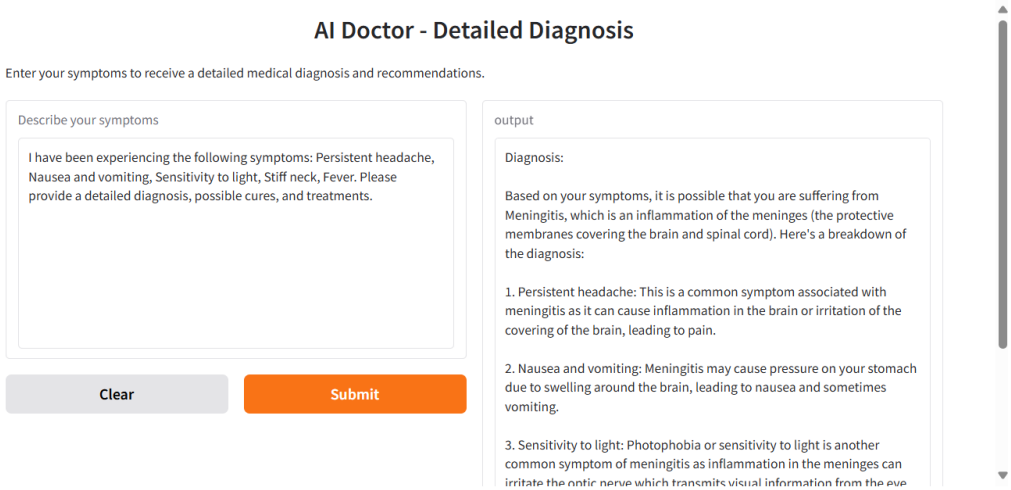

Fig. 8. AI Doctor providing diagnosis, cure, and treatment recommendations based on symptoms

of its capabilities, including diagnosing diseases, suggesting possible cures, and recommending treatments.

Figure 8 demonstrates the diagnostic process of the AI Doctor system. The system analyzes input symptoms and provides the following:

- A detailed diagnosis explaining the potential medical condition.
- Possible cures and treatments based on medical reasoning.
- Structured recommendations that can assist in further medical evaluation.

These results validate the efficiency and accuracy of the AI Doctor system, making it a viable tool for AI-driven medical diagnosis.

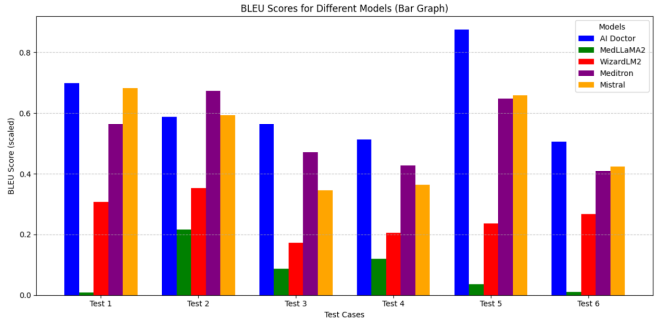

Fig. 9. Bar chart representation of accuracy results

D. Accuracy Analysis

To further analyze the accuracy of the AI Doctor system, we present three visual representations of accuracy testing results: bar chart, line chart, and pie chart.

Figure 9 illustrates the BLEU Score accuracy for different test cases. The variation in BLEU Score highlights cases where the model achieves higher accuracy and where improvements may be required. Figure 10 presents a line chart that visualizes the trends in BLEU Score accuracy measurement over multiple tests. A consistently high BLEU Score indicates reliable diagnostic accuracy, while fluctuations suggest potential inconsistencies in model predictions.

Figure 11 provides a pie chart displaying the distribution of BLEU Score accuracy. A balanced distribution indicates a stable and consistent model, whereas any disproportionately

Algorithm 1 AI Doctor System Workflow

Step 1: Initialize AI Doctor System
Step 2: Load AI Doctor model from modelfile.py
Step 3: Load pre-trained LLMs: Meditron, MedLlama2, WizardLM2, Mistral
Step 4: Set system parameters (temperature = 1.0)
Step 5: Data Processing
Step 6: function $PROCESS_{INPUT}$ (symptom description)
Step 7: Clean and normalize input
Step 8: Convert input into structured format
Step 9: return $processed_{input}$
Step 10: end function
Step 11: Model Querying
Step 12: function $QUERY_{AIDOCTOR}$ (processed input)
Step 13: Retrieve responses from AI Doctor
Step 14: Store model outputs
Step 15: return $model_{responses}$
Step 16: end function
Step 17: Federated Learning Implementation
Step 18: function $APPLY_{FL}$ (model responses)
Step 19: Distribute learning across decentralized nodes
Step 20: Aggregate model updates without raw data sharing
Step 21: Optimize AI Doctor's reasoning and diagnostic capabilities
Step 22: return $optimized_{AIDoctor}$
Step 23: end function
Step 24: Diagnosis Generation
Step 25: function $GENERATE_{DIAGNOSIS}$ $(optimized_{AIDoctor})$
Step 26: Aggregate responses from multiple LLMs
Step 27: Structure final diagnosis and treatment recommendations
Step 28: return final diagnosis
Step 29: end function
Step 30: Performance Optimization
Step 31: function $OPTIMIZE_{INFERENCE}$
Step 32: Reduce latency and enhance efficiency
Step 33: Ensure federated learning robustness
Step 34: Improve AI-driven diagnostic accuracy
Step 35: Implement real-time feedback integration for continuous model improvement
Step 36: return optimization status
Step 37: end function

low BLEU Scores suggest areas for improvement in diagnostic accuracy.

By analyzing these visual outputs, we can conclude that the AI Doctor system demonstrates high ac curacy in its diagnoses, supporting its potential for real-world medical applications. Further refinements can focus on improving areas with inconsistencies, ensuring a more reliable diagnostic process.

E. Computation Speed Analysis

To assess the computation efficiency of the AI Doctor system, we analyze the system's execution time through bar, line, and pie charts. Figure 12 illustrates the time taken for different test

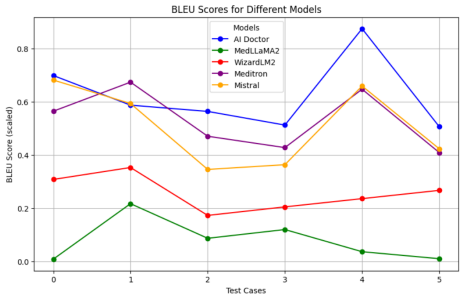

Fig. 10. Line chart representation of accuracy trends over multiple tests

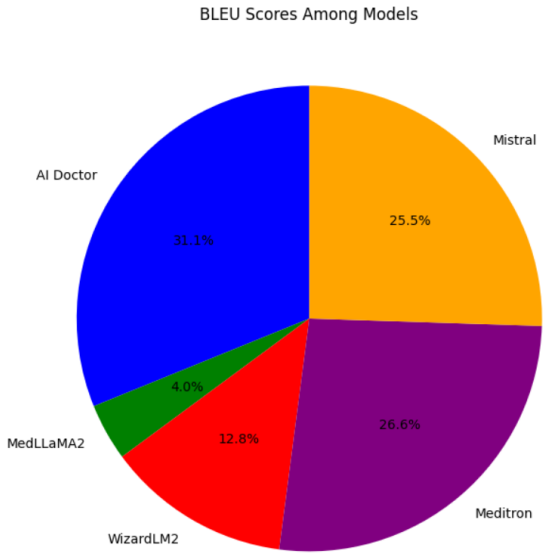

Fig. 11. Pie chart representation of accuracy distribution

cases. The variation in computation time highlights cases where the model is faster and where optimization may be required.

Figure 13 presents a line chart that visualizes the trends in computation speed over multiple tests. A steady computation time indicates efficient performance, while fluctuations suggest potential bottlenecks.

Figure 14 provides a pie chart displaying the distribution of execution times. A balanced distribution indicates a stable system, whereas any disproportionately high computation times suggest areas for optimization.

V. Conclusion

From the analysis of computation speed and accuracy, it is evident that AI Doctor performs exceptionally well compared to other models. Among the compared models—MedLLaMA2, WizardLM2, Meditron, Mistral, and AI Doctor—MedLLaMA2 has the least execution time but the worst accuracy. AI Doctor, while taking slightly more time than MedLLaMA2, achieves the highest accuracy among all models. These results highlight the efficiency of AI Doctor, as it balances computation time and accuracy effectively. It completes execution significantly faster than other high-accuracy models while maintaining superior diagnostic precision. The AI Doctor system demonstrates a robust diagnostic capability with a reliable execution speed,

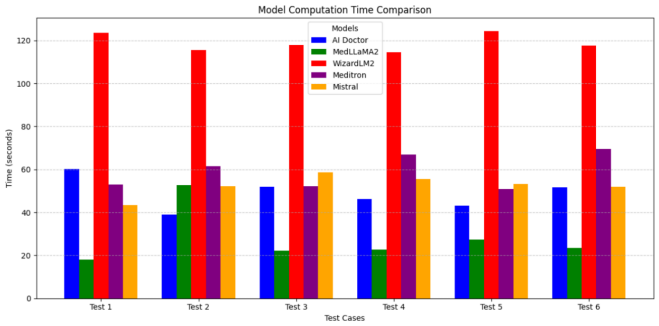

Fig. 12. Bar chart representation of computation time

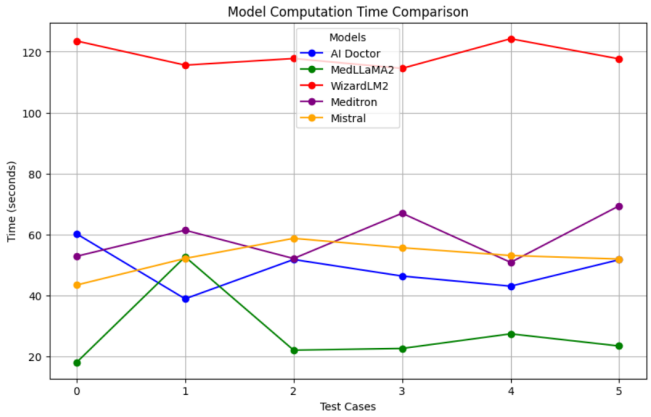

Fig. 13. Line chart tracking computation time across multiple test cases

making it a strong contender for AI-driven medical diagnosis. Further optimizations could focus on reducing computation time while maintaining its superior accuracy, ensuring it remains a competitive and effective solution for real-world medical applications.

ACKNOWLEDGMENTS

This work has been funded by R&D project SERB-SURE SUR/2022/001051.

REFERENCES

[1] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly," *High-Confidence Computing*, p. 100211, 2024.

[2] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," in *Forty-first International Conference on Machine Learning*, 2024.

[3] M. Fahim-Ul-Islam, A. Chakrabarty, M. G. R. Alam, and S. S. Maidin, "A resource-efficient federated learning framework for intrusion detection in iomt networks," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2025.

[4] M. Babar, B. Qureshi, and A. Koubaa, "Review on federated learning for digital transformation in healthcare through big data analytics," *Future Generation Computer Systems*, 2024.

[5] X. Li, Q. Lin, F. Khan, S. Kumari, M. J. Alenazi, and J. Yang, "Enhancing cancer detection capabilities in medical consumer electronics through split federated learning and deep learning optimization," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2025.

[6] P. Kanhegaonkar and S. Prakash, "Federated learning in healthcare applications," in *Data Fusion Techniques and Applications for Smart Healthcare*. Elsevier, 2024, pp. 157–196.

[7] X. Ma, G. Fang, and X. Wang, "Llm-pruner: On the structural pruning of large language models," *Advances in neural information processing systems*, vol. 36, pp. 21 702–21 720, 2023.

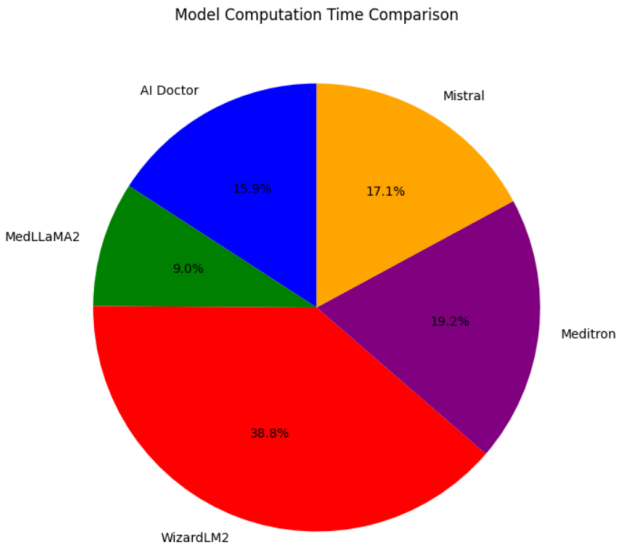

Fig. 14. Pie chart showing proportional distribution of computation time

[8] J. Wen, Z. Zhang, Y. Lan, Z. Cui, J. Cai, and W. Zhang, "A survey on federated learning: challenges and applications," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 2, pp. 513–535, 2023.

[9] F. S. Marcondes, A. Gala, R. Magalhães, F. Perez de Britto, D. Durães, and P. Novais, "Using ollama," in *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs*. Springer, 2025, pp. 23–35.

[10] S. Moon and W. H. Lee, "Privacy-preserving federated learning in healthcare," in *2023 International Conference on Electronics, Information, and Communication (ICEIC)*. IEEE, 2023, pp. 1–4.

[11] A. Gupta, C. Maurya, K. Dhere, and V. K. Chaurasiya, "Wellness detection using clustered federated learning," in *2022 IEEE 6th conference on information and communication technology (CICT)*. IEEE, 2022, pp. 1–5.

[12] C. Winston, C. Winston, C. Winston, C. Winston, and C. Winston, "Multimodal clinical prediction with unified prompts and pretrained large-language models," in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2024, pp. 679–683.

[13] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan *et al.*, "Time-llm: Time series forecasting by reprogramming large language models," *arXiv preprint arXiv:2310.01728*, 2023.

[14] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Computing Surveys (Csur)*, vol. 55, no. 3, pp. 1–37, 2022.

[15] T. Fan, Y. Kang, G. Ma, W. Chen, W. Wei, L. Fan, and Q. Yang, "Fate-llm: A industrial grade federated learning framework for large language models," *arXiv preprint arXiv:2310.10049*, 2023.

[16] W. Oh and G. N. Nadkarni, "Federated learning in health care using structured medical data," *Advances in kidney disease and health*, vol. 30, no. 1, pp. 4–16, 2023.

[17] H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor, *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part V*. Springer Nature, 2023, vol. 14224.

[18] R. Poulain, M. F. Bin Tarek, and R. Beheshti, "Improving fairness in ai models on electronic health records: The case for federated learning methods," in *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 2023, pp. 1599–1608.

[19] Y. Cheng, Y. Liu, T. Chen, and Q. Yang, "Federated learning for privacy-preserving ai," *Communications of the ACM*, vol. 63, no. 12, pp. 33–36, 2020.