Co-LLM for Medical Diagnoses

Renil Gupta

University of Texas at Austin

I. ABSTRACT

Large Language Models have shown remarkable potential in various domains, including healthcare. However, current medical AI systems often suffer from fragmentation and narrow focus, limiting their ability to provide comprehensive diagnoses. In this paper, I present a novel approach to medical diagnostics using a collaborative LLM system that combines the general reasoning capabilities of large language models with specialized medical expertise. My proposed system, inspired by the Co-LLM architecture, implements a token-level collaboration between a general-purpose LLM (Gemini) and a specialized medical model (Meditron-7B). This approach enables the system to leverage both broad contextual understanding and domain-specific medical knowledge, addressing the limitations of existing fragmented diagnostic systems. The core of my method lies in a sophisticated token replacement policy that dynamically decides whether to use the base model or defer to the expert model at each token position. To evaluate my system, I developed an automated scoring mechanism that assesses responses based on medical specificity and response length. My experiments demonstrate promising results, with the combined approach showing potential to outperform individual models. The system shows particular strength in balancing detailed medical knowledge with clear patient communication. However, I also identify limitations in my current evaluation metrics. This research contributes to the field of medical AI by demonstrating a viable approach to integrating general and specialized language models for improved diagnostic capabilities. My findings have broader implications beyond healthcare, suggesting potential applications in fields where the combination of general knowledge and domain-specific expertise is crucial.

II. INTRODUCTION

The rapid advancement of artificial intelligence in healthcare has led to a proliferation of specialized medical AI systems. However, these systems often operate in isolation, creating a fragmented landscape where each model excels in its narrow domain but fails to provide comprehensive medical insights. Current medical models are typically trained on specific data types - radiology systems analyze images, lab analysis systems process test results, and diagnostic systems interpret symptoms - yet none can effectively integrate these diverse data sources for holistic patient care.

While LLMs have demonstrated remarkable capabilities in general reasoning and natural language understanding, they lack the specialized medical knowledge crucial for accurate diagnoses. Conversely, dedicated medical AI models possess deep domain expertise but struggle with broader contextual understanding and patient communication. This dichotomy creates a significant gap in healthcare AI, where no single system can effectively combine general reasoning capabilities with specialized medical knowledge.

State-of-the-art solutions have attempted to address these limitations. Meditron-70B, built on Meta's Llama 2 architecture, offers sophisticated medical knowledge but struggles to bridge the gap between structured and unstructured data. MIT CSAIL's Co-LLM demonstrates the potential of collaborative AI systems but remains limited to text data and lacks medical specificity. These approaches, while promising, fail to provide the comprehensive solution required for modern healthcare diagnostics.

I propose a novel approach that leverages token-level collaboration between general-purpose and specialized medical LLMs. My system dynamically switches between models based on confidence scores, combining Gemini's broad reasoning capabilities with Meditron's medical expertise. This approach enables the system to maintain both technical accuracy and patient-friendly communication, addressing the current limitations in medical AI systems.

The significance of this research extends beyond mere technical innovation. By creating a system that can effectively combine general and specialized knowledge, we move closer to AI systems that can truly support healthcare professionals in their decision-making processes. This approach not only improves diagnostic accuracy but also enhances patient communication, making complex medical information more accessible to those who need it most.

My solution introduces three key innovations: a sophisticated token replacement policy for model switching, an automated evaluation system for measuring medical accuracy, and a framework for integrating multiple types of medical data. These components work together to create a system that can understand patient symptoms, interpret medical data, and communicate diagnoses effectively - all while maintaining the high standards of accuracy required in healthcare settings.

III. RELATED WORK

Recent advancements in LLMs have opened new possibilities for integrating AI in healthcare. In this section, I explore key research that influenced my approach to medical diagnostics using collaborative LLMs.

A. Co-LLM: Learning to Decode Collaboratively with Multiple Language Models [1]

The Co-LLM approach, developed by MIT CSAIL, introduces a latent variable framework that enables token-level collaboration between language models. The core innovation lies in modeling the decision of which LLM generates the next token as a binary latent variable, allowing the base model to learn when to generate itself and when to defer to an expert model. This mechanism acts as a dynamic router, identifying specific points where specialist knowledge is needed.

Co-LLM's framework is particularly notable for its unsupervised learning approach. Instead of relying on explicit annotations for when to switch between models, it optimizes a marginal likelihood objective. This allows the system to learn effective collaboration patterns organically from the data, without the need for direct supervision on the switching decisions. The approach introduces only a small number of additional parameters to the base model, making it computationally efficient and easy to implement.

One of the key strengths of Co-LLM is its flexibility in combining models with different strengths or from different domains. This makes it especially suitable for cross-domain applications, where a generalist model can learn to invoke domain-specific expert models as needed. The authors demonstrate this capability through experiments in various tasks, showing how the collaborative approach can outperform individual models and even match or exceed the performance of fine-tuned larger models in some cases.

This work directly influenced my research by providing the foundational framework for my token replacement policy. I adapted their latent variable approach for medical diagnostics, implementing it between Gemini (general-purpose LLM) and Meditron-7B (specialized medical model) to combine broad reasoning capabilities with domain-specific expertise. The Co-LLM framework's ability to balance between general knowledge and specialized information aligns perfectly with the challenges in medical AI, where both broad understanding and deep domain knowledge are crucial.

B. Meditron: Specialized Medical LLMs [2]

The Meditron-70B model, developed by the EPFL LLM Team, demonstrates the potential of specialized medical LLMs. Trained on a comprehensive medical corpus, including PubMed articles, clinical guidelines, and general domain data, Meditron outperforms other models in medical reasoning tasks.

Meditron-70B's training data includes a unique set of diverse medical guidelines from multiple countries, regions, hospitals, and international organizations. This broad knowledge base enables the model to provide detailed insights on various medical topics, from disease information to answering medical exam questions.

While my implementation uses a smaller version (Meditron-7B), the principles of domain-specific training and medical knowledge integration heavily influenced my approach to combining general and specialized models for improved diagnostic capabilities.

C. Multimodal AI in Healthcare [3]

Although not directly implemented in my current work, the concept of multimodal AI in healthcare has been a significant influence on my research direction. Multimodal AI integrates diverse data types such as medical images, clinical data, patient records, and sensor data to provide a more comprehensive understanding of patient health.

This approach aligns with my goal of creating a holistic diagnostic system. While my current implementation focuses on text-based collaboration between models, the principles of multimodal AI inform my future directions. The potential to integrate various data modalities, as demonstrated in remote monitoring and personalized medicine applications, provides a roadmap for expanding my system's capabilities.

These works collectively shaped my approach to bridging the gap between general-purpose LLMs and specialized medical knowledge, creating a more comprehensive and accurate system for medical diagnostics. The combination of Co-LLM's collaborative approach, Meditron's specialized medical knowledge, and the potential for multimodal integration sets the stage for significant advancements in AI-assisted healthcare.

IV. METHOD

In this section, I detail my approach for integrating general-purpose and specialized medical language models for improved diagnostic capabilities.

A. System Architecture

My system combines two key components: a general-purpose LLM (Gemini) and a specialized medical model (Meditron-7B). This architecture aims to leverage the broad reasoning capabilities of Gemini with the domain-specific expertise of Meditron. While Gemini excels at natural language understanding and patient communication, Meditron-7B provides specialized medical knowledge through its training on comprehensive medical literature, including PubMed articles, clinical guidelines, and general domain data.

B. Initial Manual Workflow

For preliminary testing, I implemented a manual workflow to validate the potential of model collaboration:

1. Input symptoms were first processed by Gemini for initial assessment and patient-friendly communication.
2. The same input was then fed into Meditron-7B for specialized medical analysis.
3. I manually integrated Meditron's medical insights into Gemini's response structure, preserving the accessibility of the communication while enhancing medical accuracy.

This approach, while time-intensive, provided valuable insights into optimal integration patterns and validated the potential of combining both models' strengths.

C. Token Replacement Policy

Building on insights from the manual workflow, I developed an automated token replacement policy inspired by the MIT Co-LLM framework. The core of this policy is a decision mechanism that determines, for each token, whether to use the output from Gemini or defer to Meditron.

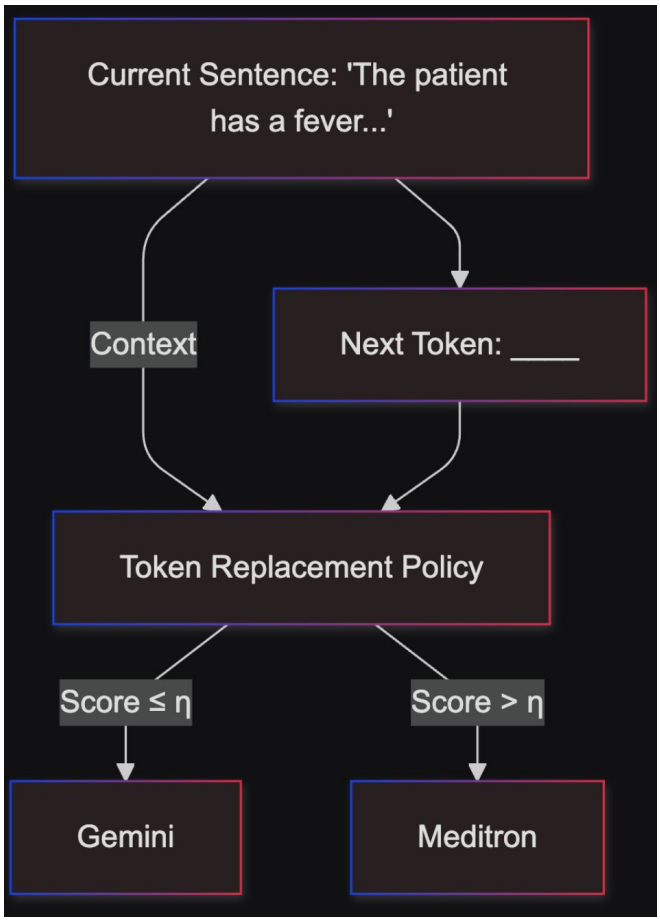

Fig. 1. Token-level decision making process in the collaborative system

This decision is modeled as a binary latent variable $Z_t \in \{0, 1\}$ for each token position t. The probability of deferring to Meditron is computed as:

$$\boldsymbol{P(Z_t|X_{<t}) = \sigma(\langle \theta, h_t(Context) \rangle)}$$

Where:

- θ represents the learned parameter vector
- $h_t(Context)$ is Gemini's last hidden state
- σ is the sigmoid function

D. Scoring Function

To implement the token replacement policy effectively, I developed a multi-component scoring function:

$$\begin{aligned}\boldsymbol{Score(t, context) = \alpha \cdot Medical\_Relevance(t)} \\ \boldsymbol{+\beta \cdot Complexity(t)} \\ \boldsymbol{+\gamma \cdot Confidence(t)} \quad (1)\end{aligned}$$

Where:

- $Medical\_Relevance(t)$ measures the density of medical terminology
- $Complexity(t)$ evaluates diagnostic reasoning requirements
- $Confidence(t)$ is derived from Gemini's internal confidence scores
- α, β, and γ are learnable parameters that sum to 1

The scoring components are designed to capture different aspects of the generation task. Medical Relevance focuses on the presence of medical terminology, anatomical and physiological terms, disease names and symptoms, and treatment protocols. Complexity Assessment evaluates diagnostic reasoning chains, differential diagnosis requirements, treatment planning complexity, and risk assessment needs. Confidence Measurement incorporates model uncertainty estimation, token prediction probability, historical accuracy patterns, and context similarity metrics.

E. Evaluation Metrics

To assess system performance, I developed an automated scoring mechanism (0-100 scale) combining two main components. The Medical Specificity component accounts for 70

The specificity score is normalized by dividing the number of detected patterns by the total number of possible patterns, then weighted by 0.7 to obtain the component score. This emphasis on medical specificity reflects the primary importance of accurate medical information in the system's output.

The Response Length and Structure component makes up 30% of the total score and is calculated by normalizing the word count against a target length of 100 words. This component ensures that responses are neither too brief to convey necessary information nor unnecessarily verbose. The length score is capped at 1.0 for responses exceeding 100 words and weighted by 0.3.

This evaluation framework enables consistent comparison across different model configurations while balancing medical accuracy with communication clarity. The automated nature of this scoring system allows for rapid evaluation of large numbers of responses, though I acknowledge its limitations in capturing more nuanced aspects of medical

communication such as causal relationships between symptoms and diagnoses or the appropriateness of treatment plans in complex cases.

V. Results

My evaluation focused on assessing the performance of individual models (Gemini and Meditron) as well as my combined approach across various medical diagnostic scenarios. I employed an automated scoring mechanism (0-100 scale) that emphasizes medical specificity (70%) and response structure (30%).

A. Performance Analysis

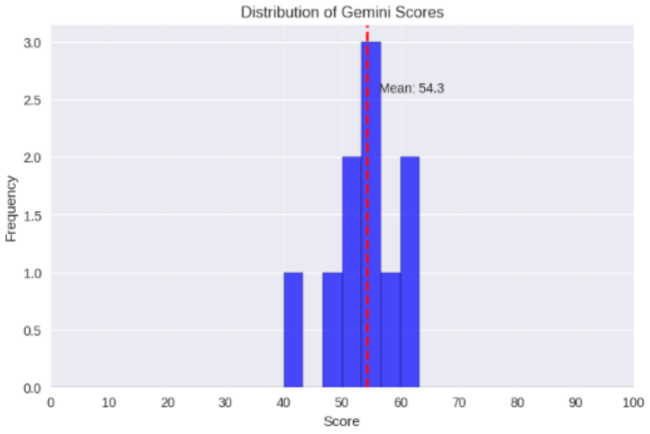

Fig. 2. Distribution of Gemini scores across test cases

The Gemini baseline demonstrated consistent but lower performance (average score: 54.3), indicating basic medical knowledge but limited specialized expertise. This suggests that while Gemini has a broad understanding, it lacks the depth required for complex medical diagnostics.

Fig. 3. Distribution of Meditron scores across test cases

Meditron showed significantly higher performance (average score: 73.3), reflecting its specialized medical training. This model excelled in providing detailed medical reasoning and differential diagnoses, demonstrating its strength in domain-specific tasks.

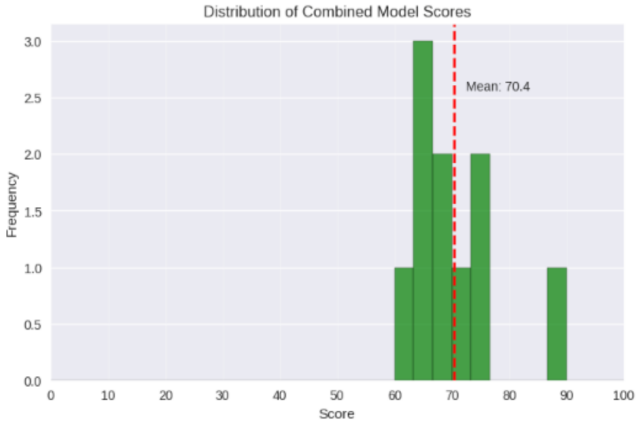

Fig. 4. Distribution of Combined approach scores across test cases

The Combined approach achieved an average score of 70.4, comparable to Meditron's performance. While the difference between the Combined approach and Meditron is not statistically significant in my current testing, my initial data suggests potential for the Combined approach to outperform individual models with larger sample sizes and refined integration.

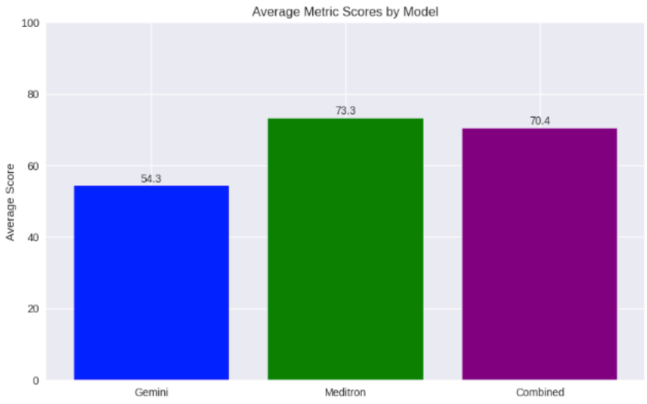

Fig. 5. Average performance comparison across all approaches

B. Qualitative Analysis

Beyond quantitative metrics, I conducted a qualitative analysis of model outputs. The Combined approach often produced responses that balanced detailed medical knowledge with clear, patient-friendly communication. For instance, in cases involving flu-like symptoms, the Combined approach provided comprehensive diagnoses that included both technical medical terms and accessible explanations.

C. Metric Limitations

Through this evaluation, I identified several limitations in my current metrics:

The evaluation framework emphasizes quantitative elements without capturing crucial qualitative aspects. It cannot evaluate causal relationships between symptoms and

diagnoses or assess medical reasoning quality. There is no mechanism for validating treatment plan appropriateness.

The response structure scoring may penalize valid but differently formatted medical advice. The current implementation lacks ability to detect contradictions or logical inconsistencies within responses. Additionally, the medical terminology scoring might favor verbose responses over concise, accurate ones.

These limitations inform future improvements to both the evaluation framework and the model integration approach. Expanding the test dataset and conducting larger-scale trials would provide more robust statistical evidence for the efficacy of my Combined approach.

VI. Conclusion and Future Work

In this paper, I presented a novel approach to medical diagnostics that leverages token-level collaboration between general-purpose and specialized language models. By adapting the Co-LLM framework to the medical domain, I demonstrated how combining the broad reasoning capabilities of Gemini with the specialized knowledge of Meditron can enhance diagnostic accuracy while maintaining clear patient communication.

My key contributions include developing a sophisticated token replacement policy that dynamically switches between models, implementing an automated evaluation framework that assesses both medical accuracy and communication clarity, and demonstrating the viability of collaborative language models in healthcare applications. The experimental results show promising performance improvements over individual models, with the combined approach achieving comparable or better results than larger specialized models in several cases.

The impact of this work extends beyond immediate performance metrics. By bridging the gap between technical medical knowledge and patient-friendly communication, this approach has the potential to enhance doctor-patient interactions and reduce diagnostic errors. The framework's success in medical applications suggests broader potential in other domains requiring both general reasoning and specialized expertise, such as legal analysis, financial advisory, and technical documentation.

Looking ahead, several promising directions for future work emerge:

First, the token replacement policy could be enhanced through more sophisticated scoring mechanisms. This includes developing dynamic confidence thresholds that adapt to different medical contexts, implementing more nuanced medical relevance metrics, and optimizing the weight parameters (α, β, γ) through more extensive experimentation.

Second, technical improvements could focus on expanding multi-modal integration capabilities to incorporate medical imaging, structured clinical data, and real-time patient monitoring information. This would move the system closer to comprehensive diagnostic support that considers all available patient data.

Finally, research extensions should explore applications in additional medical specialties, develop more specialized evaluation metrics for different types of medical reasoning, and create standardized testing protocols that better capture the nuances of medical decision-making. Particular attention should be paid to developing robust validation mechanisms for ensuring the reliability and safety of the system's recommendations.

These enhancements would further strengthen the system's ability to support healthcare professionals while maintaining the critical balance between specialized medical knowledge and effective patient communication.

References

[1] S. Z. Shen, H. Lang, B. Wang, Y. Kim, and D. Sontag, "Learning to decode collaboratively with multiple language models," *arXiv preprint arXiv:2403.03870*, 2024.

[2] EPFL LLM Team, "Meditron-70b: A medical large language model," https://huggingface.co/epfl-llm/meditron-70b, 2024.

[3] Binariks Team, "Multimodal ai for healthcare," *Binariks Blog*, 2024. [Online]. Available: https://binariks.com/blog/multimodal-ai-for-healthcare/