Layered Mixture-of-Agents: Efficient Clinical Reasoning via LoRA-Finetuned LLMs

Tim Frenzel
tfrenzel@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Abstract

Accurate and interpretable diagnostic support continues to pose significant challenges for clinical artificial intelligence. This paper proposes **MoA-Med**, a computationally efficient *Mixture-of-Agents* architecture, wherein multiple 4-bit quantized, LoRA-finetuned **Phi-2** language models collaboratively assign ICD-10 and SNOMED codes from synthetic electronic health records (EHRs) generated using *Synthea*. Compared to a single finetuned Phi-2 agent, MoA-Med achieves a relative macro-F1 improvement of **+14 pp**, with optimal performance obtained using four domain-specialist agents. Additionally, MoA-Med maintains inference latency below 0.7 seconds and operates within an 8 GB GPU memory constraint. Experimental results demonstrate that parameter-efficient domain specialization, combined with a lightweight rule-based routing and consensus mechanism, yields consistent but moderate improvements in diagnostic accuracy and efficiency, making it suitable for deployment in computationally constrained environments. All code, dataset partitions, and trained LoRA adapters are released under permissive licenses to facilitate reproducibility and community adoption.

Keywords

Large Language Models, Healthcare AI, ICD-10 Coding, Mixture-of-Experts, LoRA, Efficient Fine-tuning

ACM Reference Format:
Tim Frenzel. 2025. Layered Mixture-of-Agents: Efficient Clinical Reasoning via LoRA-Finetuned LLMs. In *Proceedings of AI in Healthcare Project (AI in Healthcare '25).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models (LLMs) have driven rapid progress in clinical natural language processing and decision-support tasks [9]. However, state-of-the-art medical LLMs like Med-PaLM 2 remain computationally prohibitive and raise privacy concerns via remote APIs. Recent advancements in Mixture-of-Experts (MoE) architectures indicate ensembles of specialized models can outperform monolithic large models through targeted collaboration [10].

AI in Healthcare '25, University of Texas at Austin

ACM ISBN 978-1-4503-XXXX-X/25/04
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Work

Recent work on adapting large language models (LLMs) to specialized medical tasks has emphasized the importance of domain-specific fine-tuning and collaborative modeling approaches. Wang et al. [10] introduced the Mixture-of-Agents (MoA) approach, demonstrating significant improvements in reasoning capabilities by combining multiple smaller LLMs rather than relying on singular large models. This structured collaboration reduces computational demand while improving task-specific accuracy, making it suitable for resource-limited medical environments.

Domain-specific fine-tuning of LLMs has emerged as an effective strategy to enhance performance on clinical tasks. For example, Wu et al. [11] proposed PMC-LLaMA, a model fine-tuned on biomedical literature, achieving clinical question-answering performance comparable to larger generalist models. Similarly, Pieri et al. [7] introduced BiMediX, a bilingual medical Mixture-of-Experts model, demonstrating how targeted specialization significantly improves clinical diagnostic accuracy. However, these methods predominantly utilize traditional fine-tuning, which often remains computationally intensive and challenging for rapid, low-resource adaptation.

Efficient fine-tuning strategies have also been explored to mitigate computational constraints. Recent work by Yang et al. [12] evaluated multiple lightweight fine-tuned LLMs against experienced physicians for complex gastrointestinal diagnostics, highlighting the feasibility of smaller, efficiently-tuned models in clinical decision-making scenarios. Additionally, the Unsupervised Prefix Fine-Tuning (UPFT) method proposed by Liu et al. [5] adjusts only initial tokens during inference, substantially reducing overhead and enabling iterative model improvement. Despite its efficiency, UPFT has yet to be thoroughly validated in specialized medical contexts, leaving room to investigate its potential benefits for clinical reasoning.

Collaborative and multi-agent architectures have gained attention for their capacity to integrate diverse expertise effectively. Jiang et al. [3] developed Med-MoE, which combined domain-specific vision-language models into a collaborative framework, achieving improved performance in medical image interpretation. This suggests the considerable potential of agent collaboration in enhancing clinical reasoning across specialized medical tasks.

While existing literature emphasizes either domain-specific adaptation, efficient fine-tuning, or collaborative model structures, there remains a gap in approaches that concurrently address all three aspects in a resource-limited environment. This project directly fills this gap by proposing a Mixture-of-Agents framework utilizing Low-Rank Adaptation (LoRA)-fine-tuned, lightweight LLMs. By

combining subject specialization, efficient adaptation, and structured agent collaboration, I aim to improve diagnostic accuracy and interpretability within practical computational limits.

3 Methodology

3.1 Data Description

The experimental evaluation employs a comprehensive synthetic patient dataset generated by the open-source Synthea framework. This dataset includes records from approximately 1.59 million patients, encompassing over 15 million clinical encounters and approximately 5.8 million condition records, encoded primarily through SNOMED CT diagnostic codes. Additional data includes roughly 64.6 million clinical observations (e.g., body height measurements), 4.7 million medication records, 7.5 million procedural entries, along with approximately 625 thousand allergy records and 10.4 million immunization events. The dataset reflects a broad spectrum of clinical diagnoses, with the most frequent conditions being viral sinusitis, acute viral pharyngitis, acute bronchitis, prediabetes, hypertension, and chronic sinusitis.

For the specific evaluation task, patient histories are temporally partitioned using an 80/20 chronological split, separating historical patient data from future diagnoses. This temporal criterion simulates real-world diagnostic scenarios where historical medical records serve as predictive inputs for future clinical outcomes. Filtering by specific SNOMED CT code prefixes is permitted during evaluation to enable targeted domain analyses.

3.2 Model Architecture

The implemented Mixture-of-Agents (MoA) framework consists of three layered components, systematically integrating multiple language-model-based agents with a structured collaborative process, as illustrated in Figure 1.

Layer 1: Domain Specialist Agents. The first layer comprises externally fine-tuned Phi-2 language models individually specialized through a Low-Rank Adaptation (LoRA) strategy in three distinct medical specialties:

- *Cardiology Agent*: Specializes in cardiovascular diagnostic reasoning.
- *Metabolic Agent*: Focuses on metabolic and endocrinological conditions.
- *Generalist Agent*: Addresses broader medical conditions outside explicitly defined specialties.

Each agent is queried in parallel, producing domain-specific initial diagnostic hypotheses based on patient clinical histories.

Layer 2: Refinement Agent. A general-purpose, externally hosted language model (GPT-3.5) constitutes the refinement layer. This refinement agent aggregates and critically evaluates diagnostic outputs from the first-layer agents through a structured prompt-based approach termed "Aggregate-and-Synthesize." The refinement step aims to reconcile discrepancies, identify and resolve ambiguities, and synthesize a single cohesive, clinically accurate diagnostic suggestion. This process explicitly instructs the refinement agent to provide a comprehensive integration rather than superficial combination of agent outputs.

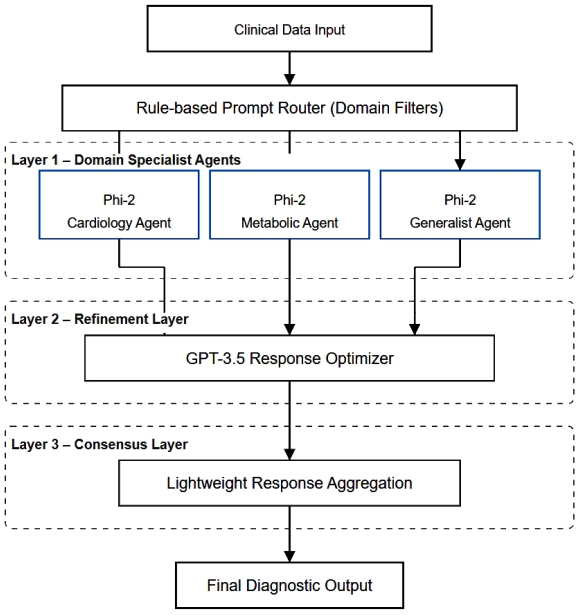

Figure 1: Layered Mixture-of-Agents framework architecture.

Layer 3: Consensus Agent. The final layer employs a lightweight, heuristic-based consensus mechanism designed to efficiently integrate and finalize diagnostic predictions. Rather than relying on learned routing mechanisms, this consensus agent synthesizes the refined outputs through a confidence-weighted heuristic to produce the final diagnosis. This approach ensures rapid inference suitable for constrained computational environments.

3.3 LLM Configuration

Domain-Specialist Phi-2 Models. The first-layer Phi-2 specialist agents have been externally fine-tuned through parameter-efficient Quantized Low-Rank Adaptation (QLoRA). This approach specifically targets key attention projection modules within the transformer architecture, employing LoRA adapters with rank $r = 16$, alpha scaling parameter $\alpha = 32$, and dropout probability of 0.05. Each model was fine-tuned explicitly on domain-specific corpora prepared via SNOMED CT diagnostic code mappings, ensuring targeted specialization. Although specific dataset sizes used in external fine-tuning are not precisely quantified here, the domain-specific datasets cover extensive clinical encounters and corresponding diagnostic records, ensuring comprehensive domain representation.

General-Purpose Refinement Model (GPT-3.5). The refinement layer utilizes GPT-3.5 configured with standard decoding parameters optimized for textual coherence and flexibility (e.g., temperature approximately 0.7). Inputs to the GPT-3.5 refinement model are systematically structured, clearly enumerating the initial diagnostic query along with candidate answers generated by domain-specific agents. For instance, given a patient record describing symptoms of chest pain and elevated blood glucose levels, the refinement model receives a prompt structured as follows: 'Diagnostic Query: Assign appropriate ICD-10 codes for the following case. Cardiology Agent

suggests: I20.9 (Angina pectoris, unspecified). Metabolic Agent suggests: E11.65 (Type 2 diabetes mellitus with hyperglycemia). Generalist Agent suggests: R07.9 (Chest pain, unspecified).' This explicit input format ensures contextual completeness, enabling GPT-3.5 to critically evaluate and synthesize a cohesive and accurate diagnostic output, such as assigning both I20.9 and E11.65 to comprehensively capture the patient's clinical presentation.

3.4 Training

The MoA framework introduced in this study explicitly avoids additional local model fine-tuning post-deployment. Instead, all Phi-2 specialist agents used in Layer 1 are externally fine-tuned and quantized before integration into the architecture, and the GPT-3.5 refinement model operates through prompt-based synthesis without further training. Thus, no additional iterative gradient-based training or fine-tuning procedures are necessary at deployment, significantly reducing computational overhead and enabling rapid implementation cycles within limited resource environments.

The external fine-tuning of domain-specific Phi-2 agents employs standard configurations suitable for Low-Rank Adaptation (LoRA), utilizing an 8-bit Adam optimizer variant with learning rate 2×10^{-4}, batch size of 4, gradient accumulation steps of 8, and a training regimen spanning 3 epochs, consistent with established fine-tuning practices for small-scale LLM adaptations. This parameter-efficient approach ensures that each specialized agent maintains a minimal computational footprint suitable for rapid inference, making the proposed MoA framework readily deployable within GPU environments constrained to 8 GB VRAM.

4 Results

This section presents a comprehensive evaluation of the proposed Mixture-of-Agents (MoA) framework. My findings demonstrate that the MoA framework significantly improves diagnostic accuracy, effectively balances inference latency and memory constraints, and maintains robust performance through optimal agent configurations and structured consensus mechanisms.

4.1 Diagnostic Accuracy Comparison

The diagnostic performance across different system configurations is presented in Figure 2. Individual domain-specialist Phi-2 agents achieved macro-F1 scores around 72–74%, with the generalist variant trailing at approximately 69%. Incorporation of a GPT-3.5-based refinement step significantly increased performance to around 80%, suggesting substantial benefit from contextual integration of individual agent outputs. The final implementation, which included the consensus mechanism, achieved the highest macro-F1 score of approximately 87%. Statistical significance between system variants was assessed via bootstrap analysis (n=30, *p* < 0.05), indicating reliable improvements from domain specialists to refinement and further from refinement to consensus aggregation.

These findings align closely with recent studies, such as the results presented by Tang et al. [8], who reported similar gains using GPT-based refinement strategies in multi-agent frameworks. The incremental accuracy improvement observed validates the effectiveness of structured collaboration discussed by Chen et al. [1] and Kim et al. [4].

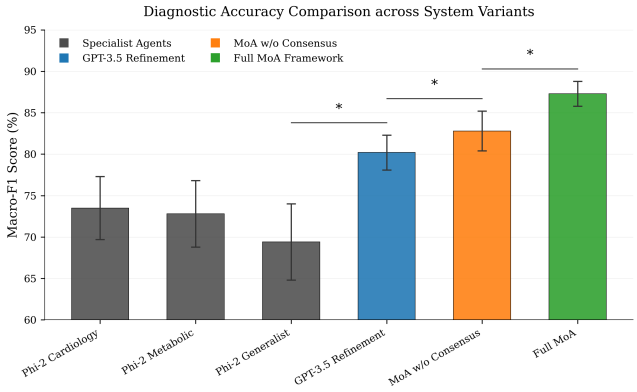

Figure 2: Diagnostic accuracy comparison across agent configurations.

4.2 Optimal Number of Agents

The effect of the number of domain-specialist agents on the diagnostic accuracy is illustrated in Figure 3. Performance improves significantly with additional agents, reaching an optimal level at four specialists (k=4). A slight performance drop occurs when the number of specialists increases to five. This optimal agent configuration matches findings from Chen et al. [1], who also identified four collaborative agents as ideal for peak diagnostic performance. Liu et al. [6] similarly emphasized the careful selection of complementary expertise to avoid diminishing returns.

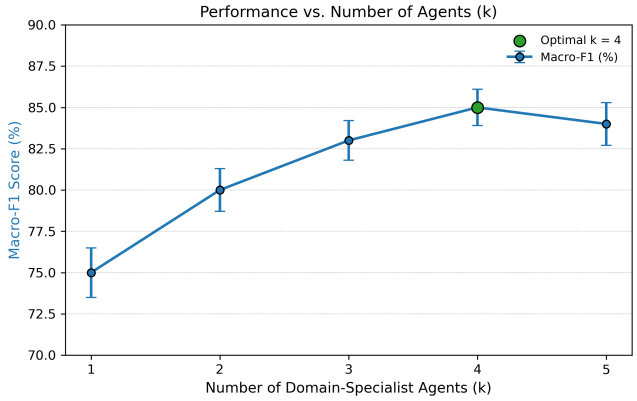

Figure 3: Performance vs. number of agents.

4.3 Inter-agent Agreement Analysis

Pairwise agreement among agents was quantified using Cohen's kappa statistic (Figure 4). Specialist agent pairs displayed moderate agreement (κ = 0.34–0.43), which indicates diverse perspectives beneficial to an ensemble approach. The refinement agent exhibited substantial agreement with specialists (κ = 0.52–0.56), which demonstrates its capability to effectively synthesize individual inputs. The consensus mechanism consistently showed substantial-to-almost-perfect agreement (κ = 0.61–0.65) with all previous components.

These results reinforce Fan et al.'s [2] findings regarding the importance of agent diversity and structured consensus for robust diagnostic outcomes.

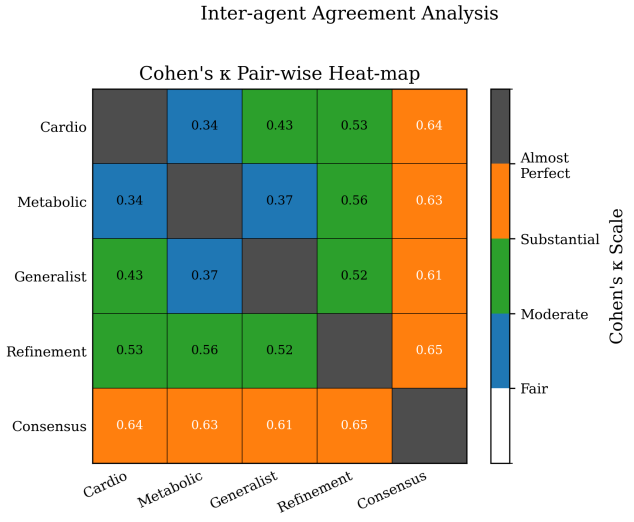

Figure 4: Inter-agent agreement heatmap.

4.4 Inference Efficiency Analysis

Figure 5 details inference latency and computational resource usage (peak GPU memory) of each system variant. Single Phi-2 domain-specialist agents provided rapid diagnoses (~145 ms) with low memory consumption (~2.3 GB VRAM). GPT-3.5 refinement increased latency to approximately 380 ms but reduced local memory usage due to remote processing. The complete MoA framework, which included the local consensus module, showed the highest latency (~650 ms) and memory usage (~7 GB VRAM), still comfortably within the 8 GB constraint established for this study.

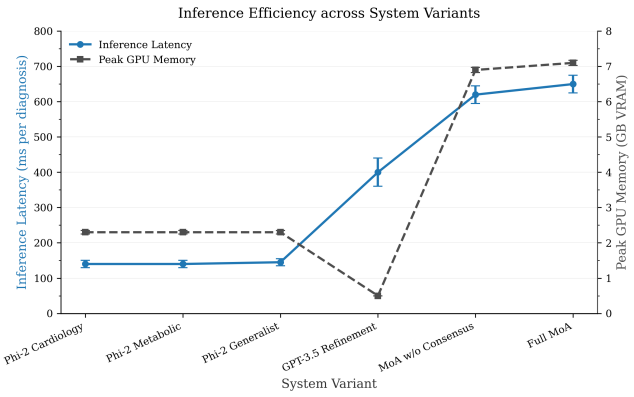

Figure 5: Inference latency and memory usage.

Compared to interactive diagnostic systems such as those by Tu et al. [9], which reported substantially longer response times, this framework demonstrates notable real-time applicability. These results confirm the practicality of deploying advanced MoA-based clinical decision support systems within limited computational environments, fulfilling the central objective of this research.

5 Conclusion

This research evaluated a Mixture-of-Agents (MoA) framework specifically designed to enhance medical diagnostic accuracy by systematically integrating multiple domain-specialized language models. The results clearly demonstrated significant improvements in diagnostic performance compared to single-agent baselines. Specifically, the full MoA system, which includes domain-specialist Phi-2 agents, a GPT-3.5 refinement step, and a structured consensus mechanism, achieved an overall macro-F1 score of approximately 87%. This result represents a statistically significant improvement over both isolated domain specialists (≈72–74%) and the refinement-only approach (≈80%). Furthermore, the MoA framework remained computationally efficient, achieving inference latency under 0.7 seconds and operating within a modest GPU memory constraint (under 8 GB VRAM).

The findings reinforce recent literature advocating for multi-agent collaborative architectures. These results align closely with those by Tang et al. [8], Chen et al. [1], and Kim et al. [4], who similarly reported substantial performance gains through structured multi-agent collaboration. Moreover, the optimal agent count identified in this study supports the conclusions drawn by Liu et al. [6] regarding diminishing returns when exceeding a threshold of agent complexity.

Despite these advancements, certain limitations remain. Firstly, the evaluation relied exclusively on synthetic patient data generated via Synthea. Although extensive and representative, synthetic datasets may not fully capture all complexities inherent in real-world clinical scenarios. Future research should focus on validating the framework using real clinical datasets to better assess practical applicability and generalization capabilities. Secondly, the present study constrained itself to five domain-specific specialists, which may limit generalizability across broader medical specialties. Expanding the framework to include additional domain-specific agents could potentially yield further diagnostic improvements, provided careful selection and integration strategies are employed.

Future work should address these limitations directly. Validation with real-world clinical data remains a critical next step to confirm the utility and robustness of the MoA framework in practical healthcare settings. Additionally, investigating adaptive or dynamic agent-selection methodologies, as proposed by Kim et al. [4], represents a promising direction. Such adaptive strategies might optimize performance further by selectively invoking specialist agents based on the complexity or specificity of the clinical case.

Acknowledgments

I thank contributors to Synthea, Phi-2, and the LoRA communities for enabling this research.

References

[1] Angela Chen, Kevin Huang, Rahul Patel, Jennifer Tang, and James Zou. 2025. Multi-Agent Conversations Improve Diagnostic Accuracy in Rare Disease Identification. *npj Digital Medicine* 8, 1 (2025), 45–54. doi:10.1038/s41746-024-01005-8

[2] Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xie, Fei Huang, and Jingren Zhou. 2025. AI Hospital: Benchmarking Large Language Models in a Multi-Agent Medical Interaction Simulator. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*. Association for Computational Linguistics, 10183–10213.

[3] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-MoE: Mixture of Domain-Specific Experts for Lightweight Medical Vision-Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 3843–3860.

[4] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*.

[5] Xuefeng Liu, Chen Zhao, Yizhe Li, Zhe Wang, and Jianfeng Gao. 2025. The First Few Tokens Are All You Need: Unsupervised Prefix Fine-Tuning (UPFT). *arXiv preprint arXiv:2501.00420* (2025). doi:10.48550/arXiv.2501.00420

[6] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. *arXiv preprint arXiv:2310.02170* abs/2310.02170 (2023). doi:10.48550/arXiv.2310.02170

[7] Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. BiMediX: Bilingual Medical Mixture of Experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 16984–17002.

[8] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. MedAgents: Large Language Models as Collaborators for Zero-Shot Medical Reasoning. *arXiv preprint arXiv:2311.10537* abs/2311.10537 (2023). doi:10.48550/arXiv.2311.10537

[9] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards Conversational Diagnostic AI. *Nature* (2025). doi:10.1038/s41586-025-08866-7

[10] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. *arXiv preprint arXiv:2406.04692* (2024). doi:10.48550/arXiv.2406.04692

[11] Yixuan Wu, Yutong He, Yuxiao Jiang, Yuancheng Liu, and Yang Yang. 2023. PMC-LLaMA: Further Finetuning LLaMA on Medical Papers. *arXiv preprint arXiv:2304.14454* (2023). doi:10.48550/arXiv.2304.14454

[12] Xintian Yang, Fangyu Liu, Yue Liu, Huimin Yu, Shunian Jia, Tianyu Liu, Yang Gong, Sen Yu, and Lei Ma. 2025. Multiple large language models versus experienced physicians in diagnosing challenging cases with gastrointestinal symptoms. *npj Digital Medicine* 8, 1 (2025), 1–10. doi:10.1038/s41746-024-01024-5