A Demonstration of Adaptive Collaboration of Large Language Models for Medical Decision-Making

Yubin Kim¹ Chanwoo Park¹ Hyewon Jeong¹ Cristina Grau-Vilchez¹ Yik Siu Chan¹ Xuhai Xu² Daniel McDuff³ Hyeonhoon Lee⁴ Cynthia Breazeal¹ Hae Won Park¹

¹ Massachusetts Institute of Technology

- ² Columbia University
- ³ Google Research
- ⁴ Seoul National University Hospital

1. Introduction

Medical Decision-Making (MDM) is a multi-faceted process that requires clinicians to assess complex multi-modal patient data patient, often collaboratively (Zhou and Xu, 2023). Large Language Models (LLMs) (OpenAI, 2024; Anthropic, 2024; Reid et al., 2024) promise to streamline this process by synthesizing vast medical knowledge and multi-modal health data (Singhal et al., 2022; Kim et al., 2024b). However, single-agent are often ill-suited for nuanced medical contexts requiring adaptable, collaborative Our $MDAgents^1$ (Kim et al., problem-solving. 2024a) addresses this need by dynamically assigning collaboration structures to LLMs based on task complexity, mimicking real-world clinical collaboration and decision-making Parekh et al. (2011); Grembowski et al. (2014). This framework improves diagnostic accuracy and supports adaptive responses in complex, real-world medical scenarios, making it a valuable tool for clinicians in various healthcare settings, and at the same time, being more efficient in terms of computing cost than static multi-agent decision making methods.

2. Method

As shown in Figure 1-(a), MDAgents operates in four steps, adapting to each medical query:

- 1. Complexity Check: A moderator agent evaluates the query and assigns a complexity level (low, moderate, or high) based on established clinical guidelines.
- 2. **Recruitment**: A recruiter agent then forms an appropriate team. Low-complexity cases go to a single Primary Care Physician (PCP), while more complex cases have Multi-disciplinary Teams (MDTs) or Integrated Care Teams (ICTs).
- 3. Analysis and Synthesis: Solo agents use techniques like Chain-of-Thought (CoT) (Wei et al., 2022) prompting for analysis, while MDTs refine answers through rounds of discussion. ICTs use a tiered decision process for the high complexity cases, engaging specialists in sequential stages.
- 4. Final Decision: The decision-maker agent aggregates inputs and synthesizes a final answer, leveraging diverse agent insights, conversational history, and moderator feedback.

For up-to-date decision-making, MDAgents incorporates MedRAG (Xiong et al., 2024), which accesses

YBKIM95@MIT.EDU CPARK97@MIT.EDU HYEWONJ@MIT.EDU CRISGRAU@MEDIA.MIT.EDU YIKSIUC@MIT.EDU XX2489@COLUMBIA.EDU DMCDUFF@GOOGLE.COM HHOON@SNU.AC.KR CYNTHIAB@MEDIA.MIT.EDU HAEWON@MEDIA.MIT.EDU

^{1.} Our code and demo can be found at https:// mdagents2024.github.io/



Figure 1: Overview. (a) The system assesses complexity, recruits agents, and iteratively refines answers through solo, MDT, and ICT configurations, emulating real-world clinical collaborations. (b) Our method outperforms Solo and Group settings across different medical benchmarks.

recent biomedical data to enhance accuracy. The model leverages GPT-4 for reasoning and was tested across medical benchmarks, including text-only (e.g., MedQA (Jin et al., 2020)), image-based (e.g. Path-VQA (He et al., 2020)), and multi-modal (e.g. Med-VidQA (Gupta et al., 2022)) datasets.

To evaluate the effectiveness of our framework, we conducted extensive experiments using baseline methods across ten datasets, including MedQA, Pub-MedQA, and others. Each dataset was tested with 50 samples, and we measured the average inference times for varying complexities: low (14.7 seconds), moderate (95.5 seconds), and high (226 seconds). Our experiments compared three configurations: (1) Solo, utilizing a single LLM agent for decision-making; (2)Group, where multiple agents collaborate; and (3)Adaptive, our proposed MDAgents method, which dynamically adjusts the inference structure. For low-complexity cases, we employed 3-shot prompting, while moderate and high-complexity cases used zero-shot prompting across all settings.

3. Results

As shown in Figure 1-(b), MDAgents demonstrated state-of-the-art performance, achieving the highest accuracy in **7 out of 10** benchmarks and outperforming both solo LLMs and static multi-agent methods. Across different medical benchmarks, MDAgents achieved up to 4.2% improvement over solo and group setting. From our ablations in Figure 2-3 in Appendix, we present MDAgent requires fewer API calls than larger static multi-agent models, achieving the same accuracy with fewer resources. Specifically, using 3 agents yielded optimal performance with a significantly lower number of API calls than singleagent CoT or larger multi-agent setups (e.g., MedAgents with 5 agents), highlighting our framework's computational efficiency. We also present that our framework shows robust performance in extreme temperature values (T=0.3, T=1.2) (Figure 2) and the agent's opinion convergence in multi-agent collaboration settings (Figure 3).

4. Discussion

Developing MDAgents required balancing agent count and response accuracy while minimizing resource use. The framework's effectiveness lies in its adaptive decision-making, which **mirrors real clinical practice** by scaling collaboration with case complexity. Ablations showed that a 3-agent setup optimizes performance, minimizing unnecessary complexity for simpler cases while supporting complex ones. Key insights include verification steps to prevent hallucinations and the use of tools like MedRAG and moderator oversight to enhance diagnostic accuracy. MDAgents is currently being evaluated for integration in hospital settings to support clinical workflows.

Future Directions. We plan to integrate doctorin-the-loop feedback, keeping MDAgents aligned with clinical knowledge and enhancing reliability to reduce diagnostic errors and improve patient outcomes.

Acknowledgments

C.P. acknowledges support from the Takeda Fellowship, the Korea Foundation for Advanced Studies, and the Siebel Scholarship.

References

- Anthropic. Claude 3.5 sonnet. https://www. anthropic.com/news/, 2024. Accessed: 2024-10-30.
- David Grembowski, Judith Schaefer, Karin E Johnson, Henry Fischer, Susan L Moore, Ming Tai-Seale, Richard Ricciardi, James R Fraser, Donald Miller, Lisa LeRoy, et al. A conceptual model of the role of complexity in the care of patients with multiple chronic conditions. *Medical care*, 52:S7–S14, 2014.
- Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering, 2022.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making, 2024a. URL https://arxiv.org/abs/2404.15155.
- Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. Health-llm: Large language models for health prediction via wearable sensor data, 2024b.

OpenAI. Gpt-4 technical report, 2024.

Anand K Parekh, Richard A Goodman, Catherine Gordon, Howard K Koh, and HHS Interagency Workgroup on Multiple Chronic Conditions. Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. *Public health reports*, 126(4):460– 471, 2011.

- Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022. URL https://arxiv.org/abs/2212.13138.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrievalaugmented generation for medicine. arXiv preprint arXiv:2402.13178, 2024.
- Junbin Zhou and Xiao Xu. The difficulty of medical decision-making: should patients be involved? *Hepatobiliary Surgery and Nutrition*, 12(3):407, 2023.

Appendix A. Demonstration Details

The demonstration, accessible at https: //mdagents2024.github.io/demo.html, showcases the MDAgents framework's capability to adapt to cases of varying complexity. For each case:

- Low Complexity (Figure 4): A single general practitioner agent manages a straightforward query on GERD management.
- Moderate Complexity (Figure 5): The framework recruits multiple agents (MDT)—a neurologist, oncologist, and radiologist—for collaborative reasoning in a case involving headache and brain lesion.



Figure 2: Impact of the number of agents on (a) Accuracy, (b) Number of API Calls on medical benchmarks with GPT-4 (V) and (c) Performance of three different settings under low (T=0.3) and high (T=1.2) temperatures on medical benchmarks. Our Adaptive setting shows better robustness to different temperatures and even takes advantage of higher temperatures.



Figure 3: An illustration of consensus entropy in group collaboration process of MDAgents (w/ Gemini-Pro (Vision), N=30 for each dataset) on medical benchmarks with different modality inputs.

• High Complexity (Figure 6): MDAgents coordinates several specialized teams (ICT) across disciplines (e.g., neurology, pulmonology, psychology) to tackle a multisystem case of muscle fatigue and vision problems, emphasizing interdisciplinary collaboration.

Currently, this demo is powered by pre-collected LLM responses, illustrating how MDAgents works in different scenarios. For users interested in realtime usage, demos can be executed by following the instructions at https://github.com/mitmedialab/ MDAgents. A DEMONSTRATION OF ADAPTIVE COLLABORATION OF LARGE LANGUAGE MODELS FOR MEDICAL DECISION-MAKING

Case A 54-yr gastrod cimetic improv experit The pa his epi anywh examir He is s sympto medica (A) Irree (C) Irre (C) Irre (C) Re	e 1: GERD Man ear-old man comes sophageal reflux d dine for his contition ed. However, follow ending more freque distribution of the source gaster region follow agaster region follow agaster region follow most distribution auton demonstrates subsequently preser owns. What is the mo ation? eversible inhibition of versible inhibition of versible inhibition of	to the clinic for m isease (GERD). H and reports that ing his recent mor it episodes of che uning in quality, a uning in quality, a d whon he takes a a healthy male w tibed a new medic sist likely mechanis n of cyclooxygenar h Hr/K+-ATPase a h stamine H2 rec	Leve Complexity inagement of his elis currently on its GERD initially te to Michigan, he is st pain (4-5 times a week), nd is concentrated around it. does not radiati thu unrearkable findings, and of his en of action of his new se enzyme th parteria cells splors	 Details Compositive Composition of the complexity of	(Myscher Grubberger Gescher Gesch
t ep 1: N ne medica eason) T gastroer ultidiscipl	Aedical Compl al complexity of th This query is prima terology or pharn linary team. Thus,	exity Classifi iis case is LOW. arily pharmacolo nacology) can lik it's manageable	cation gical, asking about the mechanis ely answer this question. The ca by one agent without requiring o	sm of action of a medication used to treat GERD, which is a relatively comm see revolves around choosing between well-known drug mechanisms (e.g., j consultation with multiple experts or departments.	on condition. A single medical agent (or an experienced clinician with knowledg roton pump inhibitors or H2 receptor blockers), which doesn't necessitate a
tep 2: F	Recruitment			General Practitioner	
tep 2: F tep 3: C iteractio	Recruitment Collaboration on Log			General Practitioner	
tep 2: F ep 3: C teractiv	Recruitment Collaboration on Log (Speaker)	To (Listener)		General Practitioner	

Figure 4: In a *low*-complexity case (GERD management), the MDAgents framework assigns a single general practitioner agent to handle a straightforward pharmacological query independently.

A DEMONSTRATION OF ADAPTIVE COLLABORATION OF LARGE LANGUAGE MODELS FOR MEDICAL DECISION-MAKING

management of his Le is summity on at his GERD initially move to Michigan, he is better that includes a work of the summit of the summit of a summit of a summit of both or control of his and summit of a summit of a summit of summit of a summit of a summit of a summit of a summit of a summit of a summit of a summit	with a complaint of headbach for the diffuse, duil in character, moderate in orning. It is not associated with fever so similar headbaches in the past. Her groups, offinin, and body mass index asamiation reveals normal liateral papildeemine is present. A (the brain reveals a solary testion in softmart hereinspiece is a solary testion in softmart hereinspiece is a solary testion in the high reveals normal liateral papildeemines. Which of the hereinspiece is a solary testion in softmart hereinspiece is a solary testion in softmart hereinspiece is a solary testion in the high reveals a solary testion in softmart hereinspiece is a solary testion in softmart hereinspiece is a solary testion in the high reveals a solary testion in softmart hereinspiece is a solary testion in the high reveals a solary testion in softmart hereinspiece is a solary testion in the high reveals a solary testion autouther (i) Anaphylistice (i) Chapter is the high reveals and the high reveals (ii) Sconstrolide	uade faigue and vision problems. Based on preses indicate most likely diagnosis. billé_mak dam_mg'.claud_mâchoire', d'alpopie', 'dyspr', 'fablesse_msm', faligabilié_msk', ome algue maging. Identifying the type of cancer mo an manifest in the brain. While a single ed for a thorough differential diagnosis, cr		
iffication IDERATE. Inical scenario that includes neurological symptoms (headache and p sperfise in both oncology and neurology, as the correct interpretation ore appropriate for a team of medical experts, such as oncologists ar	billedema) along with the discovery of a brain lesion with hemorrhage on in inges on an understanding of cancer metastasis patterns and how they ca neurologists, to discuss and reach a final decision due to the potential ner	maging, Identifying the type of cancer mo an manifest in the brain. While a single ed for a thorough differential diagnosis, co		
ist	Reologist	Radiologist		
To (Listener)	Message			
rologist A 52-year-old woman with morning headaches a	papilledema-signs of increased intracranial pressure. MRI shows a solitary hemor	rrhagic lesion.		
	Oncologist: Considering the hemorrhagic nature of the lesion, which cancers are most likely to cause this?			
ologist Oncologist: Considering the hemorrhagic nature				
sologist Oncologist: Considering the hemorrhagic nature rologist, Radiologist Melanoma is notorious for causing hemorrhagic	ain metastases. What are your thoughts on this, Neurologist?			
cologist Oncologist: Considering the hemorrhagic nature rologist, Radiologist Melanoma is notorious for causing hemorrhagic iologist Could imaging characteristics help differentiate t	ain metastases. What are your thoughts on this, Neurologist? ween metastases from different cancers?			
ologist Oncologist: Considering the hemorrhagic nature rologist, Radiologist Melanoma is notorious for causing hemorrhagic iologist Could imaging characteristics help differentiate t rologist Hemorrhagic lesions are common with melanom	ain metastases. What are your thoughts on this, Neurologist? ween metastases from different cancers? renal cell carcinoma, and choriocarcinoma. Among our options, melanoma fits best.			
Iroł	To (Listener) ogist A 52-year-old woman with morning headaches and	To (Listener) Message ogist A 52-year-old woman with morning headaches and papilledema—signs of increased intracranial pressure. MRI shows a solitary hemo		

Figure 5: In a *moderate*-complexity case (Headache & Brain Lesion), the framework recruits multiple agents (MDT)—a neurologist, oncologist, and radiologist—for collaborative reasoning to diagnose the likely cause of a brain lesion, requiring specialized input.

A DEMONSTRATION OF ADAPTIVE COLLABORATION OF LARGE LANGUAGE MODELS FOR MEDICAL DECISION-MAKING



Figure 6: In a *high*-complexity case (Muscle Fatigue & Vision Problems), MDAgents coordinates multiple MDTs (i.e. ICT) across disciplines (e.g., neurology, pulmonology, psychology) to address complex, multisystem symptoms, reflecting a comprehensive interdisciplinary approach.