Adaptive Collaboration Strategy for LLMs in Medical Decision Making

Yubin Kim[1] **Chanwoo Park**[1] **Hyewon Jeong**[1♮] **Yik Siu Chan**[1]
Xuhai Xu[1] **Daniel McDuff**[2] **Cynthia Breazeal**[1] **Hae Won Park**[1]
[1]Massachusetts Institute of Technology [2]Google Research
{ybkim95, cpark97, hyewonj, yiksiuc, xoxu, breazeal, haewon}@mit.edu
dmcduff@google.com

Abstract

Foundation models have become invaluable in advancing the medical field. Despite their promise, the strategic deployment of LLMs for effective utility in complex medical tasks remains an open question. Our novel framework, Medical Decision-making Agents (**MDAgents**) aims to address this gap by automatically assigning the effective collaboration structure for LLMs. Assigned solo or group collaboration structure is tailored to the complexity of the medical task at hand, emulating real-world medical decision making processes. We evaluate our framework and baseline methods with state-of-the-art LLMs across a suite of challenging medical benchmarks: MedQA, MedMCQA, PubMedQA, DDXPlus, PMC-VQA, Path-VQA, and MedVidQA, achieving the best performance in 5 out of 7 benchmarks that require an understanding of multi-modal medical reasoning. Ablation studies reveal that MDAgents excels in adapting the number of collaborating agents to optimize efficiency and accuracy, showcasing its robustness in diverse scenarios. We also explore the dynamics of group consensus, offering insights into how collaborative agents could behave in complex clinical team dynamics. Our code can be found at `https://github.com/mitmedialab/MDAgents`

1 Introduction

Medical Decision-Making (MDM) is a multifaceted and intricate process, where clinicians navigate vast and diverse sources of information to arrive at precise conclusions under complexity [88]. Recently, Large Language Models (LLMs) have shown potential in transforming MDM [16, 57, 64] by digesting vast amounts of medical literature [67] and clinical information [1], thereby supporting probabilistic [85] and causal [32] reasoning processes crucial to medical practice. The severe implications of inaccuracies in healthcare are misdiagnoses [45] and inappropriate treatments [9, 79], which demand a uniquely careful and precise approach. Additionally, MDM involves the interpretation of complex and multi-modal data, such as imaging [5], electronic health records (EHR) [56], signals [20, 39], genetic information [55], and the rapid integration of new medical research into clinical practice [61, 70], highlighting the unique challenges in this field.

While decision-making tools including multi-agent systems [7, 80] have shown promise in various general domains [26, 27, 35, 37, 54, 58], their performance in healthcare has been limited. This limitation arises due to their generalist design, which lacks the specialized, in-depth medical knowledge and processes essential for accurate MDM [44]. In contrast, human clinicians apply an adaptive, collaborative, and tiered approach to MDM – considering the current and past history of the patient, available evidence from medical literature, and the clinicians' clinical expertise and experience [14].

[♮]Hyewon Jeong received her MD degree from Yonsei University College of Medicine, South Korea.

Preprint.

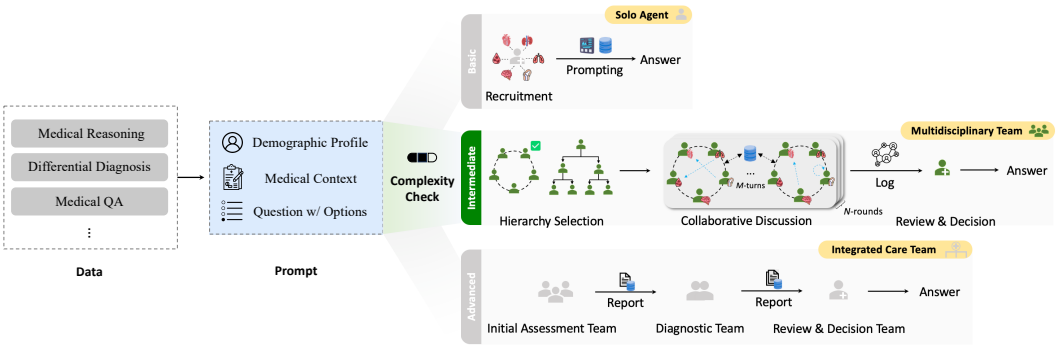


Figure 1: **MDAgents: Our medical agents decision making framework**. Given a medical query consisting of patient information and medical contexts, the framework performs 1) medical complexity check, 2) interaction type selection, 3) prompting/collaboration, and 4) decision making. Across multiple benchmarks that include multi-modal medical query, our framework performed the best performance in **5 out of 7** benchmarks.

For instance, patients are triaged in urgent care units by the severity and complexity of their medical conditions [8, 19, 81]. Examples of low-complexity cases include patients with single uncomplicated acute cases or stable chronic conditions that could be sent to their primary care physicians (PCP) [78]. In contrast, examples of high-complexity cases can be patients with complicated injuries or chronic conditions, including side effects of treatment or with superimposed diseases. These cases often require consultation with specialty physicians [21, 53]. Inspired by this real-world MDM process, we propose **M**edical **D**ecision-making **Agents** (**MDAgents**), an adaptive medical decision-making framework that leverages LLMs to emulate the hierarchical diagnosis procedures ranging from individual clinicians to collaborative clinician teams (Figure 1)

MDAgents unfolds in four stages: **1) Medical Complexity Check** - The system evaluates the intricacies of the medical query, stratifying the problem into *low*, *moderate*, or *high* categories based on the query's medical context. **2) Expert Recruitment** - Depending on the identified complexity level, the framework activates the appropriate diagnostic method. A simple solo approach is employed for issues with *low* complexity, while a Multidisciplinary Team (MDT) or Integrated Care Team (ICT) is convened for more scenarios with *moderate* or *high* complexities. **3) Inference Process** - For solo queries, the framework uses a set of prompting techniques (e.g. Chain-of-Thought (CoT), Self-Consistency (SC) and Medprompt) to provide answers. In cases of an MDT, multiple LLM agents with specialized medical expertise are brought together to form a consensus through collaborative discussion. For the most complex cases, an ICT synthesizes information from diverse domains to produce a comprehensive report, culminating in the final decision.

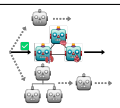
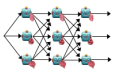
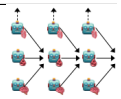
MDAgents is the first adaptive decision-making framework for LLMs that mirrors real-world MDM processes, allowing for dynamic collaboration among AI agents based on task complexity. We conduct experiments on seven medical question-answering datasets, including 1) *text-only*: MedQA [29], MedMCQA [52], PubMedQA [30], DDxPlus [65] 2) *image* + *text*: PMC-VQA [86], Path-VQA [25], and 3) *video* + *text*: MedVidQA [23]. Our framework demonstrates superior performance in accuracy over previous solo and group methods on **5 out of 7** medical benchmarks. Further analysis reveals that our framework can provide effective trade-off between the performance and the number of API calls by varying number of agents. Moreover, rigorous testing under various temperatures has demonstrated better robustness of our system than solo and group methods. Finally, we provide ablation studies evidencing that our framework finds the appropriate complexity level for each MDM instance.

2 Related Works

2.1 LLMs in Medical Decision Making

Large language models (LLMs) have been used in the medical field for a range of medical tasks and clinical applications [68, 87]. They can answer questions from medical exams [34, 42], biomedical

Table 1: Comparison between our MDAgents framework and other multi-agent collaboration methods. Among these works, MDAgents is the only one to perform all key dimensions of LLM decision making.

Method	**MDAgents (Ours)**	**Single**	**Voting** [75]	**Delphi** [64]	**Debate** [12]	**ReConcile** [6]
Interaction Type						
Multiple Roles	✓	✗	✓	✓	✓	✓
Early Stopping	✓	✗	✓	✓	✓	✗
Adaptive Structure	✓	✗	✗	✗	✗	✗
Refinement	✓	✗	✗	✓	✓	✗
Conversation Pattern	flexible	static	static	static	static	static

research [30], and clinical diagnosis [43, 59]. Some medical LLMs are also evaluated on generative tasks such as generating medical reports [71], describing medical images [73], and performing diagnostic dialogue with patients [69]. To advance the capabilities of medical LLMs, two main approaches have been explored: (1) training with domain-specific data [22], and (2) applying inference-time techniques such as prompt engineering [60] and Retrieval Augmented Generation (RAG) [83]. While initial research concentrated on pre-training and fine-tuning with medical knowledge, the rise of large general-purpose LLMs has enabled training-free methods where models leverage their latent medical knowledge. For example, GPT-4 [50], without any specialized prompt crafting, surpasses the passing score on USMLE by over 20 points and outperforms fine-tuned models including Med-PaLM [47, 48]. The promise of general-purpose models has thus inspired various techniques such as Medprompt and ensemble refinement to improve LLM reasoning [60], as well as RAG tools that use external resources to improve the factuality and completeness of LLM responses [31, 83]. Our approach leverages these techniques and the capabilities of general-purpose models, while acknowledging that a solitary LLM may not fully encapsulate the collaborative and multidisciplinary nature of real-world MDM. We thus emphasize joining multiple expert LLMs for effective collaboration in order to solve medical tasks with greater accuracy.

2.2 Multi-Agent Collaboration

An array of studies have explored effective collaboration frameworks between multiple LLM agents [38, 80] to enhance the capabilities of individual LLMs [72]. A common framework is role-playing, where each agent adopts a specific role (e.g. an Assistant Agent or a User Agent) and break down a task into sub-steps to solve it collaboratively [38, 80]. While role-playing focuses on collaboration and multi-step problem-solving [82], another framework called “multi-agent debate" prompts each agent to solve the task independently [12]. Then, they reason through other agents’ answers to converge on a shared response in order to improve factuality and performance in mathematical and reasoning tasks [12, 40]. Similar frameworks include voting [75], Delphi consensus [64], ReConcile group discussions, and negotiating [17]. Table 1 compares existing setups across key dimensions in multi-agent interaction. Although these frameworks have shown improvement in the respective tasks, they rely on a pre-determined number of agents and interaction settings. When applied to settings with a wider variety of tasks, such as medical question-answering [87], this static architecture may lead to suboptimal multi-agent configuration for certain tasks and result in limited performance [41]. Additionally, it could be computationally inefficient and expensive to employ multiple agents for simpler tasks without noticeable performance gains [12]. Given that different models and frameworks could generalize better to different tasks [84], we propose a framework that dynamically assigns the optimal collaboration strategy at inference time based on the complexity of the query. We target our strategy at the relatively untapped field of MDM, which requires a team effort and is believed to benefit from multi-agent collaboration [64].

Algorithm 1 Adaptive Medical Decision Making Framework

Require: Problem Q

$Complexity \leftarrow \text{COMPLEXITYCHECK}(Q)$ ▷ Determine the complexity of the medical input
if $Complexity = low$ **then**
  $Agent \leftarrow \text{RECRUIT}(Q, Complexity)$ ▷ Recruit a single agent
  $ans \leftarrow Agent(Q)$

else if $Complexity = moderate$ **then**
  $MDT \leftarrow \text{RECRUIT}(Q, Complexity)$ ▷ Recruit a Multi-disciplinary Team
  $Agent \leftarrow \text{RECRUIT}(Q, Complexity, MDT)$
  $r \leftarrow 0$
  $Consensus \leftarrow$ False
  $Interaction \leftarrow []$
  while $r \leq R$ and not $Consensus$ **do**
    $Consensus, Log \leftarrow \text{COLLABORATIVEDISCUSSION}(Q, MDT)$ ▷ Iterative discussions
    if not $Consensus$ **then**
      for all $Agent \in MDT$ **do**
        $Feedback \leftarrow Moderator(Interaction, Agent)$ ▷ Moderator reviews and provides feedback
        $Agent.\text{UPDATE}(Feedback)$ ▷ Update agent with the feedback
      end for
      $Interaction \leftarrow Interaction + [Log] + [Feedback]$
    end if
    $r \leftarrow r + 1$
  end while
  $ans \leftarrow Agent(Q, Interaction)$ ▷ Moderator agent makes the final decision

else
  $ICT \leftarrow \text{RECRUIT}(Q, Complexity)$ ▷ Recruit Integrated Care Team
  $Reports \leftarrow []$
  for $Team \in ICT$ **do**
    $Report \leftarrow \text{GENERATEREPORT}(Q, Team)$ ▷ Each Team curates a report
    $Reports \leftarrow Reports + [Report]$
  end for
  $ans \leftarrow Agent(Q, Reports)$ ▷ Final decision made
end if

return ans

3 Methods

3.1 Query Complexity Assessment (Line 1 of Algorithm 1)

To operationalize our adaptive medical decision-making framework, we first determine the complexity of the medical query using *moderator* LLM which functions as a *generalist practitioner*. The moderator LLM aims to play the role of a classifier to return the complexity level of the given medical query. The moderator LLM gets the information on how medical complexity should be defined and is instructed to classify one of three different complexity levels:

1. ***low*** - This level is characterized by straightforward, well-defined medical issues that can be resolved by a single primary care physician (PCP) LLM. These typically include common, acute illnesses or stable chronic conditions where the medical needs are predictable and require minimal interdisciplinary coordination.
2. ***moderate*** - At this level, the medical issues involve multiple interacting factors, necessitating a collaborative approach among a multidisciplinary team (MDT). These scenarios typically require the integration of diverse medical knowledge areas and coordination between specialists to develop effective care strategies.
3. ***high*** - This category includes highly complex medical scenarios that demand extensive coordination and combined expertise from an Integrated Cure Team (ICT). These cases often involve multiple chronic conditions, significant dependencies on medical technology, and

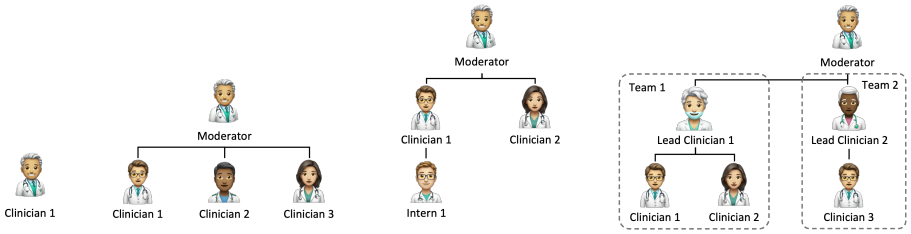


Figure 2: Simplified agent structure examples assigned during the expert recruitment process ranging from (a) A Primary Care Clinician (PCC), (b) Multi-disciplinary Team (MDT), (C) MDT w/ hierarchy to (d) Integrated Care Team (ICT).

complex decision-making that integrates various healthcare services, potentially including patient support considerations.

The method for determining complexity levels by LLM is analogous to the techniques used in real-world clinical decision-making [2, 3, 15, 63, 77]. Classification to these levels reflects the multifaceted nature of medical care, accommodating the spectrum of scenarios that healthcare professionals encounter, from routine visits to complex diagnoses that require in-depth analysis and risk assessment.

3.2 Expert Recruitment (Line 3, 7, 17 of Algorithm 1)

Given a medical query, the goal of the *recruiter* LLM is to recruit domain experts as individuals, in groups, or as multiple teams, based on the complexity levels determined by the *moderator* LLM. Specifically, we assign medical expertise and roles to multiple LLMs, instructing them to either act independently as solo medical agents or collaborate with other medical experts in a team. In Figure 8 in Appendix, we also provide frequently recruited agents for each benchmark as a reference.

3.3 Medical Collaboration and Refinement

The initial assessment protocol of our decision-making framework categorizes query complexity into *low*, *moderate*, and *high*. This categorization is grounded in established medical constructs such as acuity [18] for straightforward cases, comorbidity [62] and case management complexity [10] for intermediate and multi-disciplinary care requirements, and severity of illness [11] for high complexity cases requiring comprehensive management. The following outlines the specific refinement approach for each category:

***low* - straightforward cases (Line 2-4 of Algorithm 1)** For queries classified under Low complexity, characterized by straightforward clinical decision pathways, a single primary care physician (PCP) agent (Figure 2 (a)) is deployed by the definition in Section 3.1. The domain expert who is recruited by *recuriter* LLM, applies few-shot prompting to the problem. The output answer, denoted as ans, is directly obtained from the agent's response to Q without the need for iterative refinement, formalized as $ans = Agent(Q)$, with $Agent$ representing the engaged PCP agent.

***moderate* - intermediate complexity cases (Line 6-14 of Algorithm 1)** In addressing more complex queries, the utilization of a Multi-Disciplinary Team (MDT) (Figure 2 (b) and (c)) approach has been increasingly recognized for its effectiveness in producing comprehensive and nuanced solutions [36]. The MDT framework leverages the collective expertise of professionals from diverse disciplines, facilitating a holistic examination of the query at hand. This collaborative method is particularly advantageous in scenarios where the complexity of a problem transcends the scope of a single domain, necessitating a fusion of insights from various specialties [3, 63]. The MDT approach not only enhances decision-making quality through the integration of multidimensional perspectives but also significantly improves the adaptability and efficiency of the problem-solving process [15].

Building upon this foundation, our framework specifically addresses queries of moderate complexity through a structured, multi-tiered collaborative approach. A multidisciplinary team (MDT) recruited

Table 2: Summary of the Datasets. *t*: text, *i*: image, *v*: video. In Appendix A, we provide detailed sample information for each benchmark.

Dataset	Modality	Format	Choice	Testing Size	Domain
MedQA	*t*-only	Question + Answer	A/B/C/D	1273	US Medical Licensing Examination
MedMCQA	*t*-only	Question + Answer	A/B/C/D and Explanations	6.1K	AIIMS and NEET PG entrance exams
PubMedQA	*t*-only	Question + Context + Answer	Yes/No/Maybe	500	PubMed paper abstracts
DDxPlus	*t*-only	Question + Answer	A/B/C/D/ $\cdots$	134 K	Pathologies, Symptoms and Antecedents from Patients
PMC-VQA	*i* + *t*	Question + Answer	A/B/C/D	50 K	VQA pairs across Images, spanning diverse Modalities and Diseases
Path-VQA	*i* + *t*	Question + Answer	Yes/No	3391	Open-ended Questions from Pathology Images
MedVidQA	*v* + *t*	Question + Answer	A/B/C/D	155	First Aids, Medical Emergency, and Medical Education Questions

by *recruiter* LLM starts an iterative discussion process aimed at reaching a consensus with at most R rounds (Line 10-12). For every round, consensus within the MDT is manually verified. In the event of a disagreement, the moderator agent, consistent with the one described in Section 3.1 reviews the MDT's discourse and formulates feedback for each agent.

***high* - complex care cases (Line 17-24 of Algorithm 1)** In contrast to the MDT approach, the Integrated Care Team (ICT) (Figure 2 (d)) paradigm is essential for addressing the highest tier of query complexity in healthcare. This structured progression through the ICT ensures a depth of analysis that is specialized and focused at each stage of the decision-making process. Beginning with the Initial Assessment Team, moving through various diagnostic teams, and culminating with the Final Review & Decision Team, our ICT model aligns specialist insights into a cohesive narrative that informs the ultimate decision. This phased approach is implemented in Line 19-21. This phased approach, supported by evidence from recent healthcare studies, has been shown to enhance the precision of clinical decision-making, as each team builds upon the foundation laid by the previous, ensuring a meticulous and refined examination of complex medical cases [28]. The resultant reports are not only reflective of comprehensive medical evaluations but also of a systematic and layered analysis that is critical in the management of intricate health scenarios [13].

3.4 Decision Making

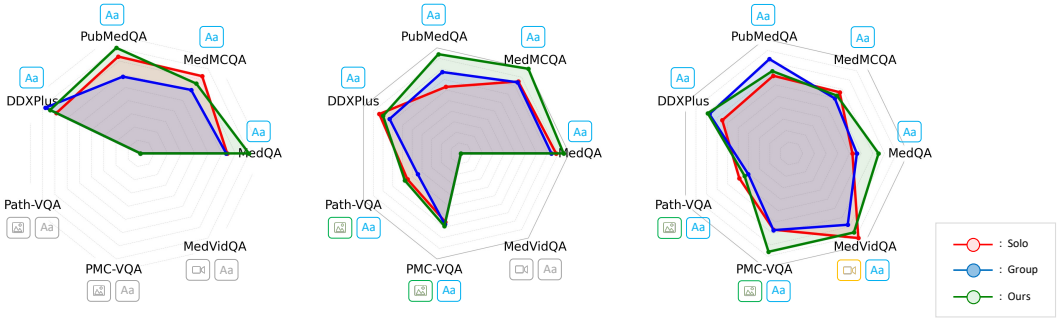
In the final stage of our framework, the role of the decision-maker LLM is crucial. This agent synthesizes the diverse inputs generated throughout the decision-making process to arrive at a well-informed final answer to the medical query q. This synthesis involves several components depending on the complexity level of the query:

1. *low*: Directly utilizes the initial response from the primary decision-making agent.
2. *moderate*: Incorporates the conversation history (*Interaction*) between the recruited agents to understand the nuances and disagreements in their responses.
3. *high*: Considers detailed reports (*Reports*) generated by the agents, which include comprehensive analyses and justifications for their diagnostic suggestions.

The decision-making process is formulated as:

$$ans = Agent(\cdot) \tag{3.1}$$

Where the final answer, *ans* is determined by integrating the outputs across different complexities. This integration employs sophisticated ensemble techniques such as temperature ensembles, and decision strategies like majority and weighted voting to ensure the decision is robust and reflects a consensus among the models when applicable.



(a) GPT-3.5 (b) GPT-4 (V) (c) Gemini-Pro (Vision)

Figure 3: Our adaptive setting outperforms Solo and Group settings on multiple medical benchmarks across LLMs. Note that with GPT-4 (V), Our adaptive setting performs the best in all medical benchmarks except for the DDXPlus.

4 Experiments and Results

In this section, we evaluate our framework and baseline models using state-of-the-art LLMs across multiple medical benchmarks in Solo, Group, and Adaptive settings. Our experiments and ablation studies highlight the framework's dynamic performance, demonstrating robustness and efficiency by modulating agent numbers and temperature. Results also show a beneficial convergence of agent opinions in collaborative settings.

Table 3: The best result of each LLM with different agent settings evaluated on medical datasets. *t*: text, *i*: image, *v*: video.

		GPT-3.5			**GPT-4(V)**			**Gemini-Pro(Vision)**		
Dataset	**Modality**	Solo	Group	Ours	Solo	Group	Ours	Solo	Group	Ours
MedQA	*t*-only	64.0	63.3	**80.0**	88.0	83.7	**95.0**	52.5	56.3	**74.8**
MedMCQA	*t*-only	**73.0**	60.0	66.0	85.0	84.0	**100.0**	**66.7**	60.0	63.0
PubMedQA	*t*-only	73.4	58.2	**80.0**	63.0	77.1	**94.0**	67.9	**82.5**	72.0
DDxPlus	*t*-only	68.8	**77.8**	74.0	**83.8**	73.3	80.0	65.1	77.3	**78.9**
PMC-VQA	*i* + *t*	-	-	-	55.0	44.4	**57.6**	**48.9**	40.5	44.0
Path-VQA	*i* + *t*	-	-	-	64.9	67.1	**68.8**	66.7	67.1	**86.0**
MedVidQA	*v* + *t*	-	-	-	-	-	-	**92.3**	77.8	86.2

4.1 Setup

To verify the effectiveness of our framework, we conduct comprehensive experiments with baseline methods on seven datasets including MedQA, MedMCQA, PubMedQA, DDXPlus, Path-VQA, PMC-VQA and MedVidQA. Table 2 summarizes the data statistics. More information about the datasets and tasks is presented in Appendix A.

In all the experiments, we evaluate the performance of agents driven by GPT-3.5 [4, 51], GPT-4 (v) [50], and Gemini-pro (vision) [66] across various tasks. For all the quantitative experiments in this section, we compare three settings: (1) **Solo**: Using a single LLM agent in the decision-making state. (2) **Group**: Implementing a group of agents to collaborate during the decision-making process. (3) **Adaptive**: Our proposed method MDAgents, adaptively constructs the inference structure from PCP to MDT and ICT.

Medical Question Answering With MedQA [29], MedMCQA [52] and PubMedQA [30], we focus on question answering through text, involving both literature-based and conceptual medical knowledge questions. Specifically, PubMedQA tasks models to answer questions using abstracts from PubMed, requiring synthesis of biomedical information. MedQA and MedMCQA test model's ability to understand and respond to multiple-choice questions derived from medical educational

materials and examinations. These tasks evaluate the AI's medical reasoning and comprehension capabilities.

Diagnostic Reasoning DDxPlus [65] involves clinical vignettes that require differential diagnosis, closely mimicking the diagnostic process of physicians. The task tests the model's ability to reason through symptoms and clinical data to suggest possible medical conditions, thus evaluating the AI's diagnostic reasoning abilities similar to a clinical setting.

Medical Visual Interpretation Path-VQA [25], PMC-VQA [86] and MedVidQA [23] datasets challenge models to interpret medical images and videos, requiring integration of visual data with clinical knowledge. In detail, PathVQA focuses on answering questions based on pathology images, testing AI's capability to interpret complex visual information from medical images. PMC-VQA evaluates AI's proficiency in deriving answers from both text and images found in scientific publications, requiring a multifaceted understanding of visual and textual content. Finally, MedVidQA extends to video-based content, where AI models need to process information from medical procedure videos, combining visual cues, and narrative understanding to answer related questions.

Baseline Methods

- **Solo:** The baseline methods considered for the Solo setting include the followings: Zero-shot [33] directly incorporates a prompt to facilitate inference, while Few-shot [4] involves a small number of examples. Few-shot CoT [76] integrates rationales before deducing the answer. Few-shot CoT-SC [74] builds upon Few-shot CoT by sampling multiple chains to yield the majority answer. Ensemble Refinement (ER) [60] is a prompting strategy that conditions model responses on multiple reasoning paths to bolster the reasoning capabilities of LLMs. Medprompt [49] is a composition of several prompting strategies that enhances the performance of LLMs and achieves state-of-the-art results on multiple benchmark datasets, including medical and non-medical domains.
- **Group:** We tested five group-based decision-making methods: Voting [75], Multidisciplinary Collaboration (MC) [64], Reconcile [6], AutoGen [80], and DyLAN [41]. Autogen was based on four agents, with one User Agent, one Clinician, one Medical Expert, and one Moderator. Each agent was given one round of respond [2]. DyLAN setup followed the base implementations of four agents with no specific roles and four maximum rounds of interaction. [3]

4.2 Main Results

In Table 3, we report the classification accuracy on MedQA, MedMCQA, PubMedQA, DDXPlus, Path-VQA, PMC-VQA and MedVidQA dataset. We compare our method (Adaptive) with several baselines in both Solo and Group settings.

Adaptive method outperforms Solo and Group settings. As depicted in Figure 3 and Table 3, MDAgents outperforms baseline methods within both Solo and Group settings, showing best performance in **5 out of 7** medical benchmarks tested. This reveals the effectiveness of adaptive strategies integrated within our system, particularly when navigating through the text-only (e.g., MedQA where it outperformed Solo by up to 22.3% and Group by up to 18.5%) and text-image datasets (e.g., Path-VQA and PMC-VQA). Our approach not only comprehends textual information with high precision but also adeptly synthesizes visual data, a pivotal capability in medical diagnostic evaluations.

Why Adaptive Decision Making Framework Works Well We hypothesize that an LLM, functioning as a classifier or a generalist doctor, will select the optimal complexity level for each MDM problem. This hypothesis is supported by Figure 4, which illustrates that the model appropriately matches the complexity levels—low, moderate, and high—of the given problem.

Formally, for any given problem P, we denote the likelihood that the correct answer can be solved by a specific complexity level as $p_{\text{complexity-level}}(P)$, where complexity-level $\in \{\text{low}, \text{moderate}, \text{high}\}$. The

[2] https://microsoft.github.io/autogen/docs/notebooks/agentchat_groupchat

[3] https://github.com/SALT-NLP/DyLAN/tree/main

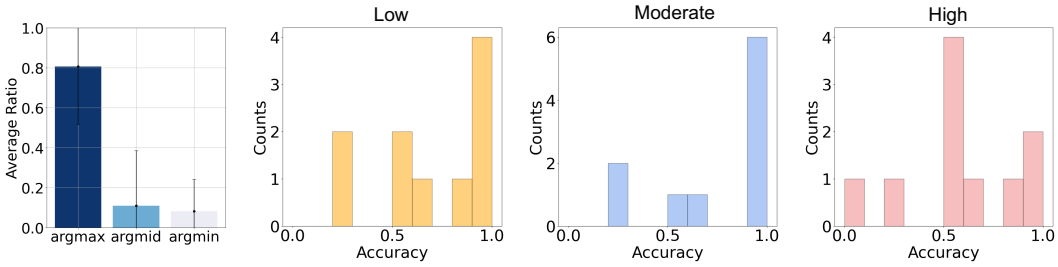


Figure 4: Experiment with MedQA dataset with N=10 randomly sampled questions. (a) The average plot for the probability of moderator provides the complexity level with the best / middle / worst choice for each problem. (b-d) The distribution of accuracy level of low / moderate / high solver. We solved each question with a specific complexity level 10 times.

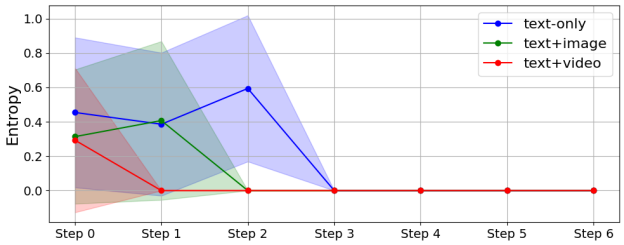


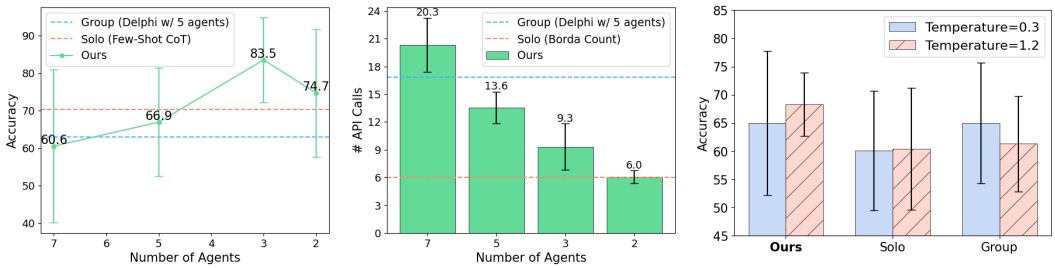
Figure 5: An illustration of consensus entropy in group collaboration process of MDAgents (w/ Gemini-Pro (Vision), N=30 for each dataset) on medical benchmarks with different modality inputs.

classifier LLM selects the complexity level according to the highest, middle, and lowest probabilities, referred to as $\arg\max$, $\arg$ mid, and $\arg\min$ respectively, with probabilities a, b, and c. Therefore, the accuracy of our system for problem P can be described by the equation $a \cdot p_{\text{argmax}}(P) + b \cdot p_{\text{argmid}}(P) + c \cdot p_{\text{argmin}}(P)$, and also the overall accuracy would be $\mathbb{E}_P[a \cdot p_{\text{argmax}}(P) + b \cdot p_{\text{argmid}}(P) + c \cdot p_{\text{argmin}}(P)]$. The calculated values of a, b, c are $a = 0.81 \pm 0.29$, $b = 0.11 \pm 0.28$, and $c = 0.08 \pm 0.16$, which indicates that LLM can provide an optimal complexity level with probability at least 80%.

These findings suggest that a classifier LLM can implicitly simulate various complexity levels and optimally adapt to the complexity required for each medical problem, as shown in Figure 4. This ability to adjust complexity dynamically proves to be crucial for applying LLMs effectively in MDM contexts as shown by the competitiveness of our Adaptive approach.

Solo vs. Group setting. In Table 4 and 5 in Appendix, we show performance comparison within Solo and Group settings. When individual agents demonstrate competent performance, such as in MedQA, MedVidQA, and DDxPlus, the precision of solo decision-making often outperforms group methods, attributing to a reduced cognitive load and avoidance of *groupthink*, a state where the desire for harmony in a decision-making group results in irrational or dysfunctional outcomes [46]. On the other hand, as the complexity of tasks increases, group settings can suffer from *polythink*, where excessive perspectives lead to confusion and collective judgment errors, reflecting the challenges seen in self-debugging processes under complex conditions [24]. The group strategies, while robust in information-rich environments, appear vulnerable to compounded inaccuracies during collaborative reasoning, especially in complex problem domains [24]. Our adaptive approach avoids these shortcomings by flexibly choosing the reasoning strategies, effectively merging the benefits of both individual and collective decision-making. It constantly improves its performance by discussion and review, which ensures consistently high accuracy across various medical datasets.

Convergence Trends in Consensus Dynamics. As depicted in Figure 5, the trend towards consensus among MDAgents is evident across different data modalities. The *text+video* modality demonstrates a rapid convergence, potentially reflecting the agents' efficient processing of combined textual and visual cues. On the other hand, the *text+image* and *text*-only modalities display a more



(a) (b) (c)

Figure 6: 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.

gradual decline in entropy, indicating a progressive narrowing of interpretative diversity among the agents. Despite the differing rates and initial conditions, all modalities exhibit a convergence of agent opinions over time. This uniformity in reaching consensus highlights the MDAgents' capability to integrate and reconcile information, irrespective of the data format. The entropy H here serves as an indicator of consensus progression, quantified as:

$$H = -\sum_{i=1}^{M} p(x_i) \log_2 p(x_i) \tag{4.1}$$

where M is the total number of unique answers, x_i represents a unique answer, and $p(x_i)$ is the probability of occurring among all answers.

4.3 Ablation Studies

Impact of Number of Agents in Group Setting. In Figure 6, we examine the effects of varying the number of agents in a collaborative Group setting. The result shows that a higher number of agents does not leads to better performance; rather, our Adaptive method, which intelligently calibrates the number of collaborating agents, achieves optimal performance with fewer agents (N=3), as shown by its peak accuracy of 83.5%. This not only indicates efficiency in decision-making but also computational and economic benefits, considering the reduced number of API calls needed, especially when contrasted with the Solo and Group settings.
With regards to computational efficiency and cost, the Solo setting (5-shot CoT-SC) resulted in a 6.0 and Group setting (Delphi with N=5) resulted in a 20.3 API calls, suggesting a high computational cost without a corresponding increase in accuracy. On the other hand, our Adaptive method exhibits a more economical use of resources, demonstrated by fewer API calls (13.6 with five agents) while maintaining high accuracy, a critical factor in deploying scalable and cost-effective medical AI solutions.

Robustness of MDAgents with different parameters. In Figure 6, we highlight the robustness of different LLM settings under low (T=0.3) and high (T=1.2) temperature conditions. Our Adaptive setting shows remarkable resilience to changes in temperature, with performance improving under higher temperatures. This suggests that our model can utilize the creative and diverse outputs generated at higher temperatures to enhance decision-making, a property that is not as pronounced in the Solo and Group settings. Such robustness is particularly valuable in real-world medical applications, where varying degrees of uncertainty and ambiguity in data are the norm.

Impact of Complexity Selection. Figure 7 shows the impact of query complexity on model accuracy. Our adaptive method significantly outperforms the static complexity assignments (Low, Moderate, High) across benchmarks with different modality inputs, highlighting the necessity for dynamic complexity assessment in MDM. In text-only queries, the Low setting presents a close second to the Adaptive method, underscoring that for less complex queries, simpler models can

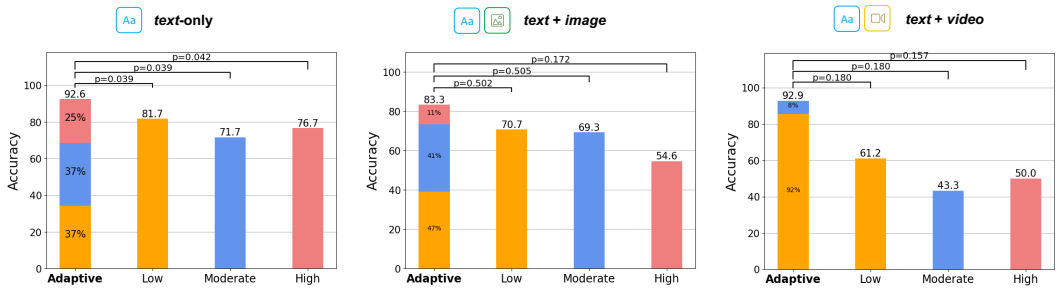


Figure 7: Impact of complexity selection of the query. Accuracy of each ablation on *text*-only (left), *text+image* (center) and *text+video* (right) benchmarks are reported.

perform adequately. However, the real testament to the Adaptive model's superiority is within the text+image and text+video modalities, where it achieves 83.3% and 92.9% accuracy, respectively, far exceeding the Low and other static models. This emphasizes the model's agility in handling the compounded difficulty of multi-modal medical data.

5 Conclusion

This paper introduces **MDAgents**, a framework designed to enhance the utility of LLMs in complex medical decision-making by dynamically structuring effective collaboration models. To reflect the nuanced consultation aspects in clinical settings, MDAgents adaptively assigns LLMs either to roles independently or within groups, depending on the task's complexity. This emulation of real-world medical decision processes has been comprehensively evaluated, with MDAgents outperforming previous solo and group methods in **5 out of 7** medical benchmarks. The case study illustrates the practical efficacy and collaborative dynamics of our proposed framework, providing insights into how differing expert opinions are synthesized to reach a more accurate diagnosis. This is evidenced by our agents' ability to converge on the correct diagnosis despite initially divergent perspectives. Ablation studies further elucidate the individual contributions of agents and strategies within the system, revealing the critical components and interactions that drive the framework's success. By harnessing the strength of multi-modal reasoning and fostering a collaborative process among LLM agents, our framework opens up new possibilities for enhancing LLM-assisted medical diagnosis systems, pushing the boundaries of automated clinical reasoning.

Limitations and Future Works

Despite the successes of our framework in showing promising performance in medical decision-making tasks, we recognize several limitations that open pathways for future research.

Medical Focused Foundation Models. An essential enhancement would be to incorporate the foundation models specifically trained on medical data, like Med-PaLM 2 [60]. These models will serve as a more accurate and up-to-date knowledge base, tailored to understand and generate medically accurate content with reduced error margins. By grounding the LLMs in the medical domain, we aim to diminish hallucinations and enrich the model's expertise.

Patient-Centered Diagnosis. A primary limitation lies in the fact that our current framework operates within the confines of multi-choice question answering and does not account for patient-physician interactions. Real-world medical diagnostics are deeply rooted in patient-centered approaches, where differential diagnosis is informed by a continuous exchange between the patient's narrative and the physician's expertise. To bridge this gap, future iterations of our framework will aim to incorporate a more interactive system that not only assists physicians but also engages directly with patients. By embedding real physicians and patients within the feedback loop, the system can be trained to become more nuanced and patient-centric, thereby enhancing the quality and personalization of medical decision support.

Updating Medical Knowledge. Another limitation is the static nature of the LLMs' medical knowledge. As medicine is a rapidly evolving field, the information within the LLM must be continually updated to reflect the latest medical insights and findings. To address this, we propose the integration of a medical Retrieval-Augmented Generation (RAG) technique. This addition would allow the framework to be pulled from up-to-date medical literature, ensuring that the model's responses and decisions are informed by the most current data and evidence-based practices.

Acknowledgments

We thank Yoon Kim at MIT, Vivek Natarajan at Google, WonJin Yoon at Harvard Medical School, Seonghwan Bae at Sacheon Public Health Center, Hyeonhoon Lee and Hui Dong Lim at Seoul National University Hospital for their revisions, feedback, and support.

References

[1] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*, 2022.

[2] Ofir Ben-Assuli, Nanda Kumar, Ofer Arazy, and Itamar Shabtai. The use of analytic hierarchy process for measuring the complexity of medical diagnosis. *Health Informatics Journal*, 26(1):218–232, 2020.

[3] Justin Bitter, Elizabeth van Veen-Berkx, Hein G Gooszen, and Pierre van Amelsvoort. Multidisciplinary teamwork is an important issue to healthcare professionals. *Team Performance Management: An International Journal*, 19(5/6):263–278, 2013.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–14, 2019.

[6] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms, 2024.

[7] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023.

[8] Michael Christ, Florian Grossmann, Daniela Winter, Roland Bingisser, and Elke Platz. Modern triage in the emergency department. *Deutsches Ärzteblatt International*, 107(50):892, 2010.

[9] Darryl S Chutka, Paul Y Takahashi, and Robert W Hoel. Inappropriate medications for elderly patients. In *Mayo Clinic Proceedings*, volume 79, pages 122–139. Elsevier, 2004.

[10] Jane Cioffi and Roslyn Markham. Clinical decision-making by midwives: managing case complexity. *Journal of advanced nursing*, 25(2):265–272, 1997.

[11] Lesley F Degner and Jeffrey A Sloan. Decision making during serious illness: what role do patients really want to play? *Journal of clinical epidemiology*, 45(9):941–950, 1992.

[12] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[13] Carolyn Ee, James Lake, Joseph Firth, Fiona Hargraves, M De Manincor, Tanya Meade, Wolfgang Marx, and Jerome Sarris. An integrative collaborative care model for people with mental illness and physical comorbidities. *International Journal of Mental Health Systems*, 14:1–16, 2020.

[14] Glyn Elwyn, Dominick Frosch, Richard Thomson, Natalie Joseph-Williams, Amy Lloyd, Paul Kinnersley, Emma Cording, Dave Tomson, Carole Dodd, Stephen Rollnick, et al. Shared decision making: a model for clinical practice. *Journal of general internal medicine*, 27:1361–1367, 2012.

[15] Nancy E Epstein. Multidisciplinary in-hospital teams improve patient outcomes: A review. *Surgical neurology international*, 5(Suppl 7):S295, 2014.

[16] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis, 2024.

[17] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback, 2023.

[18] Amy L Garcia. Variability in acuity in acute care. *JONA: The Journal of Nursing Administration*, 47(10):476–483, 2017.

[19] Nicki Gilboy, Paula Tanabe, Debbie Travers, Alexander M Rosenau, et al. Emergency severity index (esi): A triage tool for emergency department care, version 4. implementation handbook 2012 edition. *AHRQ publication*, 12, 2011.

[20] Travis R Goodwin and Sanda M Harabagiu. Multi-modal patient cohort identification from eeg report and signal data. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1794. American Medical Informatics Association, 2016.

[21] 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.

[22] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021.

[23] Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering, 2022.

[24] Daisuke Hamada, Masataka Nakayama, and Jun Saiki. Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cognitive Research: Principles and Implications*, 5:1–15, 2020.

[25] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020.

[26] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[27] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework, 2023.

[28] M. Jimenez-Lara. Reaping the benefits of integrated health care. stanford social innovation review. *Stanford Social Innovation Review.*, 2016.

[29] 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.

[30] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[31] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2), February 2024.

[32] Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

[33] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[34] Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12, 02 2023.

[35] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[36] Danielle L LaFrance, Mary Jane Weiss, Ellie Kazemi, Joanne Gerenser, and Jacqueline Dobres. Multidisciplinary teaming: Enhancing collaboration through increased understanding. *Behavior analysis in practice*, 12(3):709–726, 2019.

[37] Beibin Li, Konstantina Mellou, Bo Zhang, Jeevan Pathuri, and Ishai Menache. Large language models for supply chain optimization. *arXiv preprint arXiv:2307.03875*, 2023.

[38] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society, 2023.

[39] Jingjing Li and Qiang Wang. Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene settings: Overview, challenges, and novel orientation. *Information Fusion*, 79:229–247, 2022.

[40] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate, 2023.

[41] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization, 2023.

[42] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions?, 2023.

[43] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[44] Mark A Musen, Blackford Middleton, and Robert A Greenes. Clinical decision-support systems. In *Biomedical informatics: computer applications in health care and biomedicine*, pages 795–840. Springer, 2021.

[45] David E Newman-Toker, Adam C Schaffer, C Winnie Yu-Moe, Najlla Nassery, Ali S Saber Tehrani, Gwendolyn D Clemens, Zheyu Wang, Yuxin Zhu, Mehdi Fanai, and Dana Siegal. Serious misdiagnosis-related harms in malpractice claims: the "big three"–vascular events, infections, and cancers. *Diagnosis*, 6(3):227–240, 2019.

[46] Bheema Shanker Neyigapula. Human-ai collaborative decision-making: A cognitive ergonomics approach. 2023.

[47] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023.

[48] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023.

[49] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. November 2023.

[50] OpenAI. Gpt-4 technical report, 2024.

[51] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[52] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022.

[53] 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.

[54] Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do llm agents have regret? a case study in online learning and games. *arXiv preprint arXiv:2403.16843*, 2024.

[55] Valerie F Reyna, Farrell J Lloyd, and Patrick Whalen. Genetic testing and medical decision making. *Archives of Internal Medicine*, 161(20):2406–2408, 2001.

[56] Max J Romano and Randall S Stafford. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. *Archives of internal medicine*, 171(10):897–903, 2011.

[57] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D. Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records, 2024.

[58] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

[59] 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, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera 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. *Nature*, 620(7972):172–180, 2023.

[60] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

[61] Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. *Medical decision making*. John Wiley & Sons, 2024.

[62] J Stairmand, Louise Signal, D Sarfati, C Jackson, L Batten, M Holdaway, and C Cunningham. Consideration of comorbidity in treatment decision making in multidisciplinary cancer team meetings: a systematic review. *Annals of Oncology*, 26(7):1325–1332, 2015.

[63] Miren Taberna, Francisco Gil Moncayo, Enric Jané-Salas, Maite Antonio, Lorena Arribas, Esther Vilajosana, Elisabet Peralvez Torres, and Ricard Mesía. The multidisciplinary team (mdt) approach and quality of care. *Frontiers in oncology*, 10:85, 2020.

[64] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning, 2024.

[65] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. Ddxplus: A new dataset for automatic medical diagnosis, 2022.

[66] Gemini Team. Gemini: A family of highly capable multimodal models, 2024.

[67] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[68] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940, 2023.

[69] 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. Towards conversational diagnostic ai, 2024.

[70] Sean R Tunis, Daniel B Stryer, and Carolyn M Clancy. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *Jama*, 290(12):1624–1632, 2003.

[71] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142, February 2024.

[72] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents, 2024.

[73] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models, 2023.

[74] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[75] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

[76] 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.

[77] Thomas Weida and Jane Weida. Outpatient e/m coding simplified. *Family Practice Management*, 29(1):26–31, 2022.

[78] Robin M Weinick, Rachel M Burns, and Ateev Mehrotra. Many emergency department visits could be managed at urgent care centers and retail clinics. *Health affairs*, 29(9):1630–1636, 2010.

[79] Dominic Wilkinson, Stavros Petrou, and Julian Savulescu. Expensive care? resource-based thresholds for potentially inappropriate treatment in intensive care. *Monash Bioethics Review*, 35(1):2–23, 2018.

[80] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

[81] Richard C Wuerz, Leslie W Milne, David R Eitel, Debbie Travers, and Nicki Gilboy. Reliability and validity of a new five-level triage instrument. *Academic emergency medicine*, 7(3):236–242, 2000.

[82] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023.

[83] Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, and William Hiesinger. Almanac: Retrieval-augmented language models for clinical medicine, 2023.

[84] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language, 2022.

[85] Haodi Zhang, Jiahong Li, Yichi Wang, and Yuanfeng Songi. Integrating automated knowledge extraction with large language models for explainable medical decision-making. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1710–1717. IEEE, 2023.

[86] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering, 2023.

[87] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. A survey of large language models in medicine: Progress, application, and challenge, 2024.

[88] Junbin Zhou and Xiao Xu. The difficulty of medical decision-making: should patients be involved? *Hepatobiliary Surgery and Nutrition*, 12(3):407, 2023.

A Dataset Information

We evaluate multi-agent collaboration frameworks across seven common medical question-answering datasets, which vary in question complexity. Generally, questions are deemed more complex if they involve multiple modalities or entail a lengthy, detailed diagnostic task. Below, we detail each dataset and provide a sample entry:

1. **MedQA.** The MedQA dataset consists of professional medical board exams from the US, Mainland China, and Taiwan [29]. Our study focuses on the English test set, comprising 1,273 questions sourced from the United States Medical Licensing Examination (USMLE). These questions are formatted as multiple-choice text queries with five options. Due to their textual nature and brevity, we categorize these questions as low.
   Sample Question: "A 47-year-old female undergoes a thyroidectomy for treatment of Graves' disease. Post-operatively, she reports a hoarse voice and difficulty speaking. You suspect that this is likely a complication of her recent surgery. What is the embryologic origin of the damaged nerve that is most likely causing this patient's hoarseness?"
   Options: "A: 1st pharyngeal arch, B: 2nd pharyngeal arch, C: 3rd pharyngeal arch, D: 4th pharyngeal arch, E: 6th pharyngeal arch"
2. **MedMCQA.** Similarly, the MedMCQA is derived from real-world medical entrance exams and includes approximately 194K multiple-choice questions across 21 medical subjects. We use 6.1K samples from the test set. Given the similarity to MedQA in terms of format and content, this dataset is also classified as having low complexity.
   Sample Question: "Retraction of mandible is achieved by: "
   Options: "Lateral pterygoid", "Temporalis", "Medial pterygoid", and "Masseter"
3. **PubMedQA.** PubMedQA is a QA dataset based on biomedical research [30]. It requires yes/no/maybe answers to questions grounded in PubMed abstract. The dataset comprises entries each containing a question, a relevant abstract minus the conclusion, and a ground truth label. We used 500 samples for testing. Given its binary choice format, we consider the complexity of this dataset to be low.
   Sample Question: "Can predilatation in transcatheter aortic valve implantation be omitted?"
   Context: "The use of a balloon expandable stent valve includes balloon predilatation of the aortic stenosis before valve deployment. The aim of the study was to see whether or not balloon predilatation is necessary in transcatheter aortic valve replacement (TAVI). Sixty consecutive TAVI patients were randomized to the standard procedure or to a protocol where balloon predilatation was omitted. There were no significant differences between the groups regarding early hemodynamic results or complication rates."
4. **DDxPlus.** DDxPlus is a medical diagnosis dataset using synthetic patient information and symptoms [65]. Each instance represents a patient, with attributes including age, sex, initial evidences, evidence, multiple options of possible pathologies, and a ground truth diagnosis. Due to its text-only and multiple-choice nature, we consider the complexity to be low.
   Sample Patient Information: Age: 96, Sex: F
   Evidences: ['e66', 'insp_siffla', 'j45', 'posttus_emesis', 'trav1_@_N', 'vaccination']"
   Initial evidence: 'posttus_emesis'
   Options: (A) Bronchite (B) Coqueluche
5. **PMC-VQA.** PMC-VQA is a large-scale medical visual question-answering dataset that contains 227K Visual Question Answering (VQA) pairs of 149K images [86]. It is structured as a multiple-choice QA task with one image input accompanying each question. Since the query requires a model to consider both text and image inputs, while maintaining medical expertise, we consider the complexity to be moderate.
   Sample Question: What is the appearance of the hyperintense foci in the basal ganglia on T1-weighted MRI image?
   Image: PMC8415802_FIG1.jpg
   Options: A: Hypodense, B: Hyperdense, C: Isointense, D: Hypointense
6. **Path-VQA.** PathVQA is a VQA dataset specifically on pathology images [25]. Different from PMC-VQA which consists of multiple choice questions, Path-VQA includes open-ended questions and binary "yes/no" questions. For the purpose of maintaining a

standardized accuracy evaluation, we use only the yes/no questions. Similar to PMC-VQA, we consider the complexity to be moderate.
Sample Question: Was a gravid uterus removed for postpartum bleeding?
Image: test_0273.jpg

7. **MedVidQA.** MedVidQA dataset consists of 3,010 health-related questions with visual answers from validated video sources (e.g. medical school, health institutions, etc). We enhanced the dataset by using GPT-4 to generate multiple-choice answers, including one correct 'golden answer' and several false options, expanding its use for training and evaluating automated medical question-answering systems.
Sample Question: How to perform corner stretches to treat neck pain?
Video: h5MvX50zTLM.mp4
Options: A: By bending your knees and touching your toes, B: By performing jumping jacks, C: By leaning into a corner with your elbows up at shoulder level, D: By doing push-ups

B Prompt Templates

B.1 A single agent setting

Few-shot multiple choice questions

`{{instruction}}`
The following are multiple choice questions (with answers) about medical knowledge.
`{{few_shot_examples}}`
`{{context}}` **Question:** `{{question}}` `{{answer_choices}}` **Answer:**(

Chain-of-Thought multiple choice questions

`{{instruction}}`
The following are multiple choice questions (with answers) about medical knowledge.
`{{few_shot_examples w/ CoT Solutions}}`
`{{context}}` **Question:** `{{question}}` `{{answer_choices}}` **Answer:**(

Ensemble Refinement multiple choice questions

`{{instruction}}`
The following are multiple choice questions (with answers) about medical knowledge.
`{{few_shot_examples w/ CoT Solutions}}`
`{{context}}` **Question:** `{{question}}` `{{answer_choices}}`
`{{reasoning_paths}}` **Answer:**(

Medprompt multiple choice questions

`{{instruction}}`
The following are multiple choice questions (with answers) about medical knowledge.
`{{few_shot_examples w/ CoT Solutions from similarity calculation}}`
for *N* times do
`{{context}}` **Question:** `{{question}}` `{{shuffled_answer_choices}}`
Answer:(

Complexity check prompt

You are a medical expert who conducts initial assessment and your job is to decide the difficulty/complexity of the medical query.
Now, given the medical query as below, you need to decide the difficulty/complexity of it:
`{{question}}`
Please indicate the difficulty/complexity of the medical query among below options:
1) low: a single medical agent can output an answer.
2) moderate: number of medical experts with different expertise should dicuss and make final decision.
3) high: multiple teams of clinicians from different departments need to collaborate with each other to make final decision.
Answer:(

Recruiter prompt

You are an experienced medical expert who recruits a group of experts with diverse identity and ask them to discuss and solve the given medical query.
Now, given the medical query as below, you need to decide the difficulty/complexity of it:
`{{question}}`
You can recruit up to *N* experts in different medical expertise. Considering the medical question and the options for the answer, what kind of experts will you recruit to better make an accurate answer?
Also, you need to specify the communication structure between experts (e.g., Pulmonologist == Neonatologist == Medical Geneticist == Pediatrician > Cardiologist)

For example, if you want to recruit five experts, you answer can be like:
`{{examplers}}`
Please answer in above format, and do not include your reason.
Answer:(

B.2 Multi-agent setting

C Case Study

In the case study, we provide two examples of our framework with *moderate* (Figure 9) and *high* (Figure 10) complexity in PMC-VQA (*image+text*) and DDXPlus (*text*-only) respectively. These case studies reveals how our framework provides an environment for agents to collaborate, gather information, moderate and make final decisions in complex medical scenarios.

Agent initialization prompt

You are a `{{role}}` who `{{description}}`. Your job is to collaborate with other medical experts in a team.

Agent interaction prompt

Given the opinions from other medical agents in your team, please indicate whether you want to talk to any expert (yes/no). If not, provide your opinion. `{{opinions}}`

Next, indicate the agent you want to talk to: `{{agent_list}}`

Remind your medical expertise and leave your opinion to an expert you chose. Deliver your opinion once you are confident enough and in a way to convince other expert with a short reason.

Final decision prompt

You are a final medical decision maker who reviews all opinions from different medical experts and make final decision.

Given the `{{inputs}}`, please review the `{{inputs}}` and make the final answer to the question by `{{decision_methods}}`.
Answer:(

Table 4: Accuracy (%) on Medical benchmarks with **Solo** setting. CoT refers to Chain-of-Thought, SC refers to Self-Consistency, ER refers to Ensemble Refinement prompting. **Bold** number represents the best performance for each benchmark and model.

Model	Prompting	Dataset						
		MedQA	MedMCQA	PubMedQA	DDxPlus	Path-VQA	PMC-VQA	MedVidQA
GPT-3.5	Zero-shot	50.4	65.0	71.8	53.0	-	-	-
	Few-shot	55.9	57.0	67.6	46.0	-	-	-
	+ CoT	60.7	52.0	71.4	43.8	-	-	-
	+ CoT-SC	**64.0**	57.0	**73.4**	28.2	-	-	-
	ER	58.0	58.0	13.0	**68.8**	-	-	-
	Medprompt	56.0	**73.0**	70.0	55.0	-	-	-
GPT-4(V)	Zero-shot	74.0	78.0	60.0	69.0	56.0	52.0	-
	Few-shot	75.5	78.0	60.0	69.0	55.0	53.0	-
	+ CoT	83.0	81.0	**63.0**	**83.8**	62.2	49.0	-
	+ CoT-SC	**88.0**	**85.0**	50.0	55.6	**64.9**	**55.0**	-
	ER	**88.0**	82.0	39.0	77.6	**64.9**	52.0	-
	Medprompt	80.0	82.0	37.0	75.5	62.0	49.5	-
Gemini-Pro(Vision)	Zero-shot	48.0	57.1	51.0	48.8	63.8	41.2	40.0
	Few-shot	45.5	56.7	55.0	36.6	62.1	44.4	55.0
	+ CoT	48.5	**66.7**	66.7	**65.1**	50.0	41.3	30.4
	+ CoT-SC	50.0	61.3	60.0	60.4	63.3	**48.9**	88.9
	ER	**52.5**	63.5	**67.9**	34.8	**66.7**	45.7	**92.3**
	Medprompt	49.5	62.2	50.0	58.3	57.4	40.0	73.3

Table 5: Accuracy (%) on Medical benchmarks with **Group** setting. MC represents Multi-disciplinary Collaboration framework. **Bold** number represents the best performance for each benchmark and model type.

Model	Method	Dataset						
		MedQA	MedMCQA	PubMedQA	DDxPlus	Path-VQA	PMC-VQA	MedVidQA
GPT-3.5	Majority	**63.3**	52.7	**58.2**	51.9	-	-	-
	Weighted	57.8	52.7	56.0	56.8	-	-	-
	Borda Count	**63.3**	36.3	54.9	**77.8**	-	-	-
	MC [64]	63.2	**60.0**	35.4	36.0	-	-	-
GPT-4(V)	Majority	82.6	37.5	63.2	**73.3**	42.3	**44.4**	-
	Weighted	**83.7**	56.8	64.4	69.8	42.3	22.2	-
	Borda Count	81.5	37.5	63.2	**73.3**	42.3	33.3	-
	MC [64]	70.2	**84.0**	**77.1**	56.3	**54.6**	36.2	-
Gemini-Pro(Vision)	Majority	**56.3**	52.2	**82.5**	50.0	62.5	36.1	73.7
	Weighted	**56.3**	53.7	**82.5**	68.2	62.5	34.0	68.8
	Borda Count	**56.3**	44.8	**82.5**	**77.3**	62.5	36.1	**77.8**
	MC [64]	50.0	**60.0**	58.1	42.9	**67.1**	**40.5**	-
Multiple Models	Reconcile [6]	**70.9**	18.0	**77.4**	**73.6**	**59.3**	**31.9**	-
	AutoGen [80]	62.0	**64.0**	70.0	48.0	-	-	-
	DyLAN [41]	66.0	58.0	74.0	60.0	-	-	-

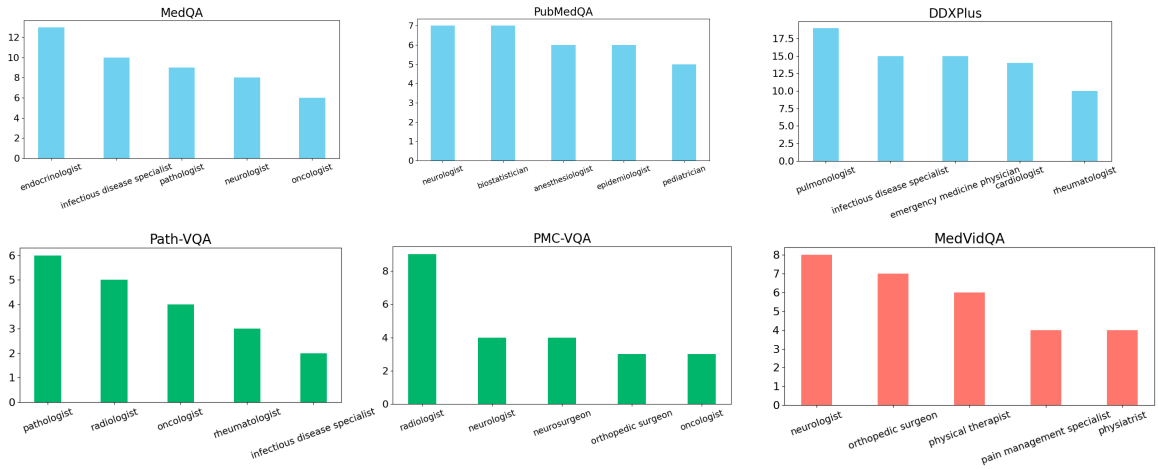


Figure 8: Top-5 most recruited medical experts in each benchmark.

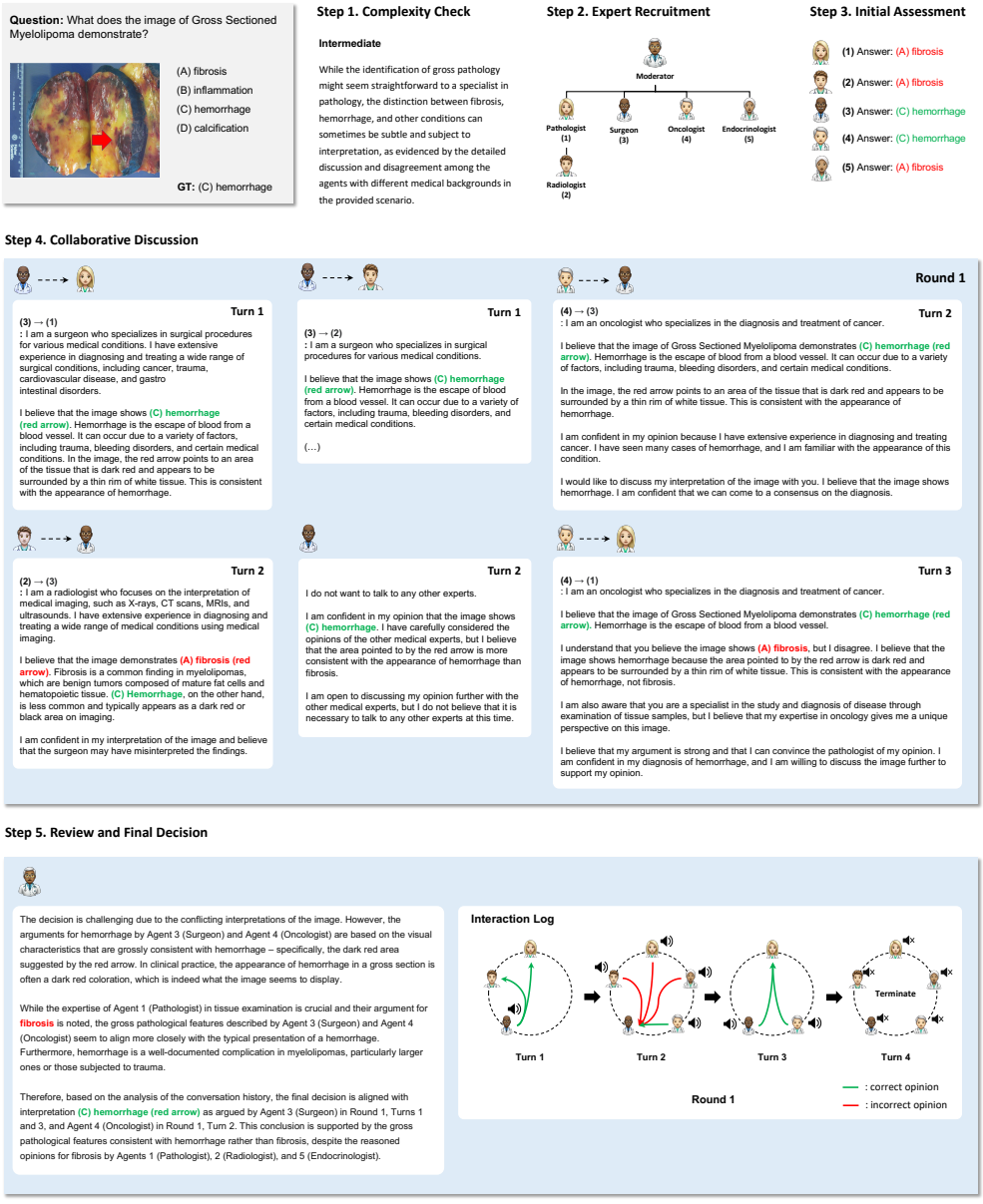


Figure 9: Illustration of our proposed framework in *moderate* complexity setting. Given a medical query (*image* + *text*) the framework performs reasoning in five steps: (i) complexity check, (ii) expert recruitment, (iii) initial assessment, (iv) collaborative discussion, and (v) review and final decision making. Green text represents the correct answer and the Red text represents the incorrect answer.

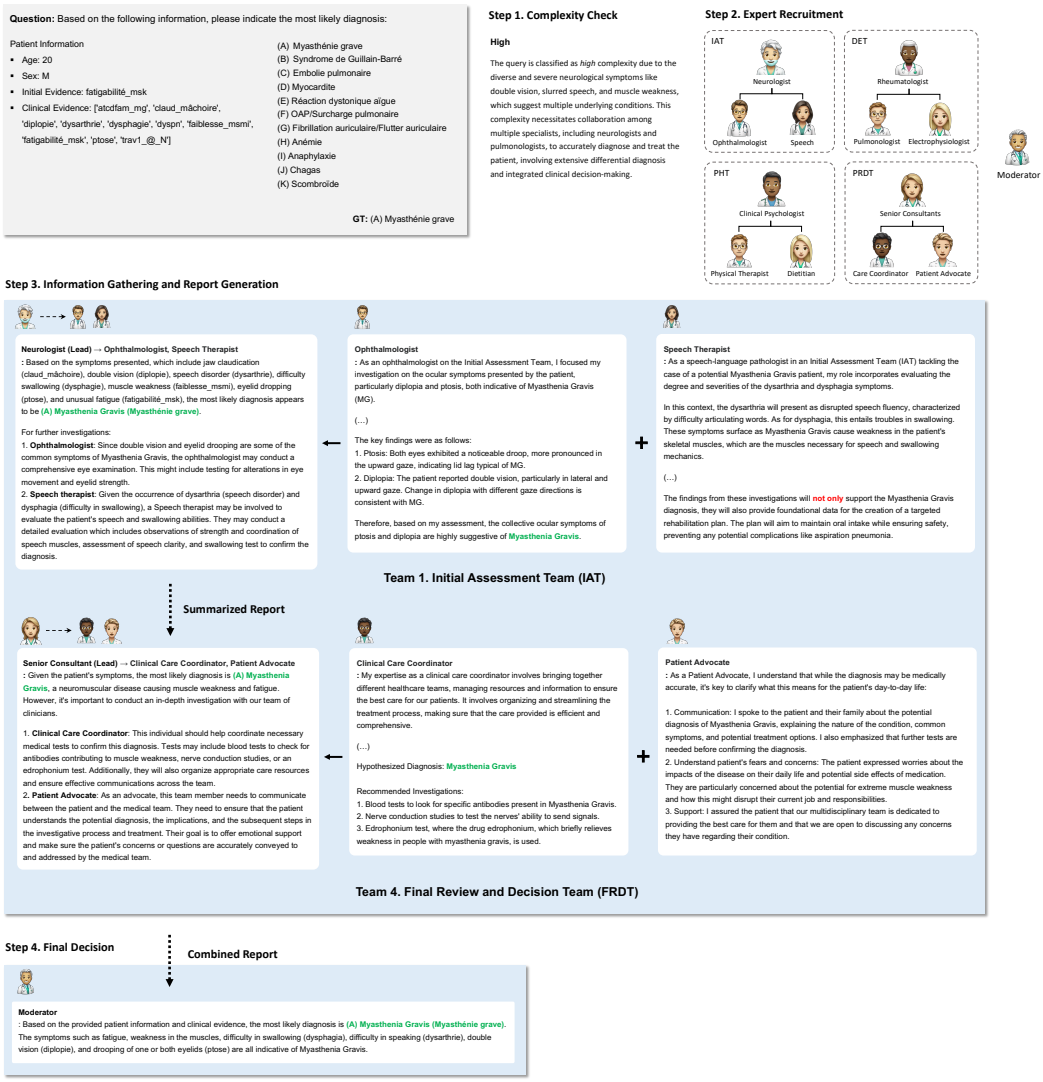


Figure 10: Illustration of our proposed framework in *high* complexity setting. Given a medical query (*text*-only) the framework performs reasoning in four steps: (i) complexity check, (ii) expert recruitment, (iii) information gather and report generation, (iv) final decision. Green text represents the correct answer.