A Knowledge-driven Adaptive Collaboration of LLMs for Enhancing Medical Decision-making

Xiao Wu[1,2] **Ting-Zhu Huang**[2*] **Liang-Jian Deng**[2] **Yanyuan Qiao**[3]
Imran Razzak[1] **Yutong Xie**[1*]

[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
[2]University of Electronic Science and Technology of China
[3]Swiss Federal Institute of Technology Lausanne (EPFL)

{xiao.wu, imran.razzak, yutong.xie}@mbzuai.ac.ae tingzhuhuang@126.com
liangjian.deng@uestc.edu.cn yanyuan.qiao@epfl.ch

Abstract

Medical decision-making often involves integrating knowledge from multiple clinical specialties, typically achieved through multidisciplinary teams. Inspired by this collaborative process, recent work has leveraged large language models (LLMs) in multi-agent collaboration frameworks to emulate expert teamwork. While these approaches improve reasoning through agent interaction, they are limited by static, pre-assigned roles, which hinder adaptability and dynamic knowledge integration. To address these limitations, we propose KAMAC, a **Knowledge-driven Adaptive Multi-Agent Collaboration** framework that enables LLM agents to dynamically form and expand expert teams based on the evolving diagnostic context. KAMAC begins with one or more expert agents and then conducts a knowledge-driven discussion to identify and fill knowledge gaps by recruiting additional specialists as needed. This supports flexible, scalable collaboration in complex clinical scenarios, with decisions finalized through reviewing updated agent comments. Experiments on two real-world medical benchmarks demonstrate that KAMAC significantly outperforms both single-agent and advanced multi-agent methods, particularly in complex clinical scenarios (*i.e.*, cancer prognosis) requiring dynamic, cross-specialty expertise. Our code is publicly available at: https://github.com/XiaoXiao-Woo/KAMAC.

1 Introduction

In healthcare, diagnosis, prognosis, and a variety of clinical treatments are guided by medical decision-making processes that require the application of complex medical knowledge (Sutton et al., 2020). An individual professional medical perspective is not enough to meet the needs of patients. Multidisciplinary teams (MDTs) or integrated care teams may participate in disease treatment in practical clinical processes (Kodner and Spreeuwenberg, 2002).

Recently, large language models (LLMs), owing to their powerful reasoning and knowledge synthesis capabilities, have demonstrated promising potential in emulating the roles of clinicians and supporting medical decision-making (Tang et al., 2023; Kim et al., 2024; Chen et al., 2025). Multi-agent collaboration (MAC) based on LLMs has emerged as a key paradigm, enhancing the reasoning performance of individual agents through collective deliberation. For instance, (Tang et al., 2023) verified that a training-free collaboration framework in which multiple LLM-based agents simulate a multidisciplinary medical team through role-playing and multi-round discussions, and achieved strong performance across medical question answering (QA) datasets. In addition, (Chen et al., 2025) further leveraged medical multi-agents and implemented cumulative consultation strategies using retrieval augmentation generation (RAG), which enhances model outputs by retrieving external medical knowledge to support clinical reasoning and improve diagnostic accuracy.

Some multi-LLM debate frameworks are also closely related to collaboration (Kaesberg et al., 2025; Chen et al., 2023b; Abdelnabi et al., 2024; Liang et al., 2023). Among them, a framework for iterative collaboration between agents to make decisions is proposed, which stimulates higher quality answers (compared to a single model) by involving multiple models in the discussion. These works explore the potential application of LLMs and the possibility of their use in medical MDT decisions.

Although these MAC methods enable agents to tackle problems that are difficult or unsolvable by a single agent by learning new contexts and actions through interactions with peers or known information, the challenge remains unresolved. It mainly stems from the use of static, pre-assigned roles based on inherent domain knowledge, which limits

[*]Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 33495–33512
November 4-9, 2025 ©2025 Association for Computational Linguistics

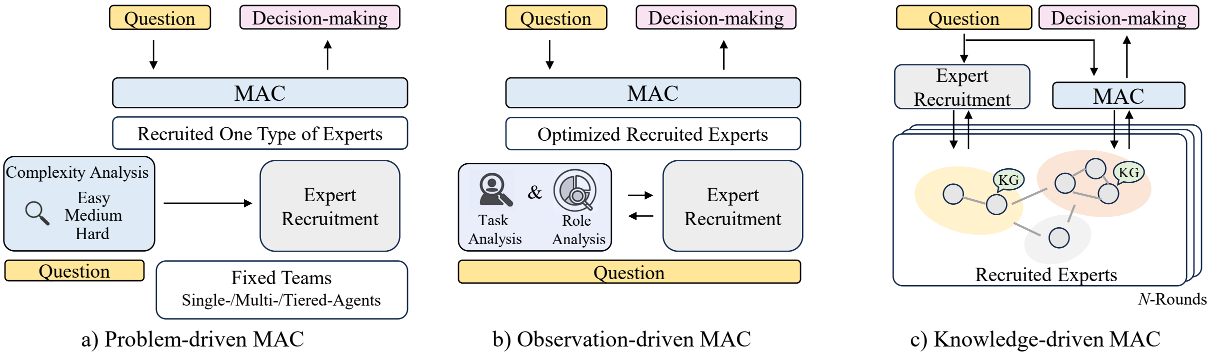


Figure 1: **Comparison of multi-agent collaboration (MAC) strategies in medical decision-making.** (a) *Problem-driven MAC* (Kim et al., 2024; Yang et al., 2024b) uses predefined question-complexity tiers (easy, medium, hard) to assemble static single- or multi-tier expert teams (b) *Observation-driven MAC* (Chen et al., 2023c,a) dynamically analyzes task and role characteristics from initial observations to optimize expert recruitment for each question. (c) Our proposed *Knowledge-driven MAC* adaptively expands the expert team during discussion by detecting knowledge gaps (KG), enabling scalable and flexible collaboration for complex, cross-domain clinical scenarios.

the system's adaptability during collaboration. As discussions progress, each agent tends to produce increasingly fine-grained analyses within its fixed specialty. For example, in the evaluation of a patient presenting with chest pain, a radiology agent may focus solely on imaging findings suggestive of pulmonary embolism, while a cardiology agent may emphasize electrocardiogram (ECG) changes indicative of myocardial infarction. Without a mechanism to reconcile or adapt these perspectives, the collaboration degenerates into a juxtaposition of isolated preferences rather than a convergent diagnostic consensus. This fragmentation undermines the effectiveness of consensus strategies and restricts the system's ability to dynamically incorporate broader context or cross-domain reasoning.

Recent studies have attempted to improve MAC flexibility by incorporating novel expert recruitment strategies. For instance, problem-driven MAC (Kim et al., 2024) (Figure 1a) assigns expert teams based on question complexity, while observation-driven MAC (Chen et al., 2023a,c) (Figure 1b) selects experts according to task and role analysis. However, these methods still rely on static or pre-optimized expert pools and cannot adapt during multi-round interactions. As a result, even when new, fine-grained insights emerge over multiple discussion rounds, no new experts are brought in. The limitations in these works still hinder truly scenario-specific collaboration, especially in dynamic and diverse clinical environments.

To alleviate this, as illustrated in Figure 1c, we propose a **Knowledge-driven Adaptive Multi-Agent Collaboration (KAMAC)** framework for enhancing medical decision-making. Specifically, KAMAC dynamically increases the number of medical expert team members required for patients by exploring additional expert knowledge during the discussion process. KAMAC begins with an initial consultation involving one or more experts. It then engages in a knowledge-driven collaborative discussion, which assesses whether additional expertise is needed by detecting knowledge gaps (KG) and dynamically recruits appropriate experts to fill the knowledge gaps, enabling scalable and flexible collaboration for complex, cross-domain clinical scenarios. Finally, a moderator is responsible for reviewing updated agent comments to complete the decision-making process. Such progressive collaboration and flexible team expansion allow the model to adaptively allocate resources and produce more accurate, context-aware decisions. In contrast to prior methods, the proposed method enables the system to adapt to the evolving clinical treatment in the real world and provide more nuanced and comprehensive support to patients.

Our contributions include three folds:

1. We propose the KAMAC framework that dynamically extends a single expert agent into multiple expert agents to form a multi-disciplinary team for medical decision-making.

2. We design a knowledge-driven collaborative discussion mechanism that enables agents to dynamically expand the team to fill knowledge gaps, aiming to improve adaptability and decision accuracy in complex clinical scenar-

ios.

3. Extensive experiments on two medical benchmarks, MedQA and Progn-VQA, demonstrating that our KAMAC improves single-agent and advanced multi-agent collaboration frameworks.

2 Related Work

Advanced LLMs such as GPT-4 (Achiam et al., 2023), DeepSeek (Liu et al., 2024; Guo et al., 2025), and Gemini (Team et al., 2024) have demonstrated strong reasoning capabilities and have been used as agents with considerable computational investment in various medical tasks such as question answering (Kim et al., 2024; Tang et al., 2023), diagnosis (Zhang et al., 2024), and report generation (Thawakar et al., 2024; Hyland et al., 2023). We list two main related areas of work:

LLM-Based Agentic Medical Decision-Making Medical decision-making systems leverage multiple LLM experts, each assigned a predefined clinical specialty to mimic real-world multidisciplinary teams. Early work demonstrated that consensus among expert agents yields higher diagnostic accuracy than any single model or simple voting schemes (Tang et al., 2023; Liang et al., 2023; Chen et al., 2023b). Some recent works mainly focused on diagnostic findings (Kim et al., 2024; Li et al., 2024b) and knowledge integration (Xiong et al., 2024; Nori et al., 2023; Kim et al., 2024). MediQ (Li et al., 2024b) designs a system to seek methods to guide the deepening of interactions between patients and experts. For instance, (Kim et al., 2024) verified that expert collaboration has better accuracy for medical decision-making than a single expert, showcasing that consensus is superior to a voting strategy in various clinical applications. More recently, MDteamGPT (Chen et al., 2025) adds a leader agent, historical dialogues, and RAG to integrate information and supplementation strategies to assist in decision-making.

Multi-Agent Collaboration in Medical Decision-Making Researchers have demonstrated that multi-agent collaborative research can enhance the reasoning capabilities of these LLMs (Yue et al., 2024; Wang et al., 2024; Li et al., 2024a). Well-designed strategies can enhance autonomous multi-agent systems for task-solving capabilities, such as debate (Chan et al., 2023; Abdelnabi et al., 2024), consensus (Kaesberg et al., 2025; Chen et al., 2023b), conflict-solving, generation/evolution (Yuan et al., 2024; Chen et al., 2023c,a), and encouragement (Liang et al., 2023; Tran et al., 2025). Some multi-LLM debate frameworks are also closely related to collaboration (Kaesberg et al., 2025; Chen et al., 2023b; Abdelnabi et al., 2024; Liang et al., 2023). Among them, a framework for iterative collaboration between agents to make decisions is proposed, which stimulates higher quality answers (compared to a single model) by involving multiple models in the discussion. These works explore the potential application of LLMs and the possibility of their use in medical MDT diagnostics. Although assigning experts can effectively improve the performance of specific tasks, the rationality of expert assignment in multi-agent collaboration is still insufficient. (Chen et al., 2023a,c) introduces optimal expert generation strategies in the initial expert recruitment stage, but it does not consider the relationship between expert knowledge and cooperation between experts. (Yuan et al., 2024) introduces a dynamic evolution strategy for the existing experts but relies on a large initial population and requires additional investment. These limitations make it unsuitable for medical decision-making.

3 Method

3.1 Overview

Figure 2 presents the KAMAC framework, which comprises three main stages: (1) Initial Consultation: KAMAC begins with a single/multiple expert agents, which evaluate the case and provide initial feedback for ongoing discussion; (2) Knowledge-driven Collaborative Discussion: Agents engage in a structured, knowledge-guided dialogue to determine whether further expertise is required and then adaptively expands the team and promotes structured discussions among agents, guided by domain knowledge and the evolving diagnostic context, and (3) Decision Making: A designated moderator coordinates the final decision process by initiating a voting mechanism among agents. The pseudocode of KAMAC is shown in Algorithm 1. More details in all prompts refer to Appendix D.

3.2 Initial Consultation

Given a clinical problem $\mathcal{Q}$, KAMAC first performs an initial consultation by recruiting one or more expert agents from a predefined expert pool.

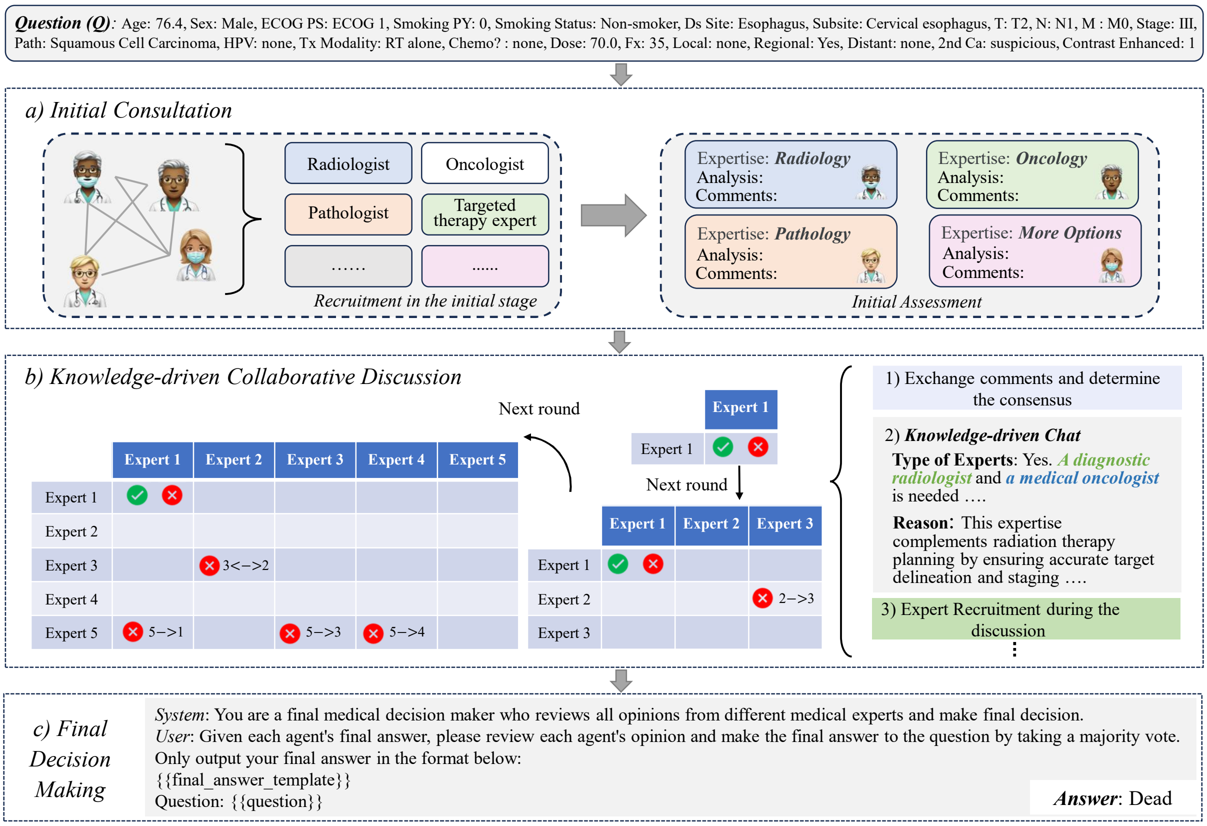


Figure 2: **Schematic diagram of Knowledge-driven Adaptive Multi-Agent Collaboration (KAMAC) framework for medical decision-making.** The KAMAC includes three parts: (a) Initial Consultation: One or more expert agents (*e.g.*, radiologist, pathologist) are selected based on the clinical question to provide initial assessments; (b) Knowledge-driven Collaborative Discussion: Agents iteratively exchange views to refine reasoning. If a knowledge gap is detected, KAMAC dynamically recruits additional specialists, and the expanded team continues the dialogue until consensus is reached or the round limit is met; and (c) Final Decision Making: A moderator reviews all agent responses and produces the final answer. The symbols ✓ and ✗ indicate agreement/disagreement with the current expert's comment, respectively. Only when a disagreement occurs, (✗ i -> j) or (✗ i <-> j) is used to denote a one-way or two-way discussion between expert i and expert j, respectively.

These agents represent diverse clinical roles (*e.g.*, radiologist, cardiologist) and are selected based on their relevance to the query using an expert recruitment prompt P_1. Each recruited agent independently analyzes the problem using an initial assessment prompt P_2, producing diagnostic opinions or treatment suggestions. The individual responses are aggregated into a consolidated feedback signal, which serves as the basis for initiating collaborative discussion in the next stage. This step simulates a typical initial clinical encounter, where specialists offer their perspectives before deliberation begins.

3.3 Knowledge-driven Collaborative Discussion

In this stage, KAMAC facilitates multi-round, knowledge-driven discussions among the recruited expert agents. Each round begins with agents exchanging their views based on the evolving shared context. Using the agent interaction prompt P_3, they critique each other's responses, resolve inconsistencies, and collaboratively refine their reasoning and comments.

At the end of each round, the currently assigned experts are prompted to assess whether a knowledge gap (KG) remains—that is, whether their collective expertise is sufficient to fully address the problem. This self-assessment is facilitated by the KG detection prompt P_4, which takes as input the current discussion and feedback. If a gap is detected, KAMAC triggers expert recruitment by issuing a targeted recruitment prompt P_5, allowing the system to enlist additional domain-specific agents to address the identified deficiency.

The newly recruited agents receive contextual examples (*i.e.*, the current discussion history) as few-shot input and respond to the original question

Algorithm 1: Knowledge-driven Adaptive Multi-Agent Collaboration (KAMAC) Decision-making

Input: Problem Q
Result: Answer ans.
Initialize: KAMAC $\leftarrow$ [].
$\triangleright$ Define prompts. More details in all prompts refer to Appendix D.
$r \leftarrow 1$, $Consensus$ $\leftarrow$False, KG $\leftarrow$False.
P_1: Expert Recruitment Prompt; P_2: Initial Assessment Prompt.
P_3: Agent Interaction Prompt; P_4: KG Prompt for Recruited Experts.
P_5: KG Prompt for Expert Recruitment; P_6: Agent Update Prompt.
P_7: Final Decision Prompt.
$\triangleright$ **Initial consultation.**
$\triangleright$ Recruit N clinician agents.
$(Agent_1, Agent_2, \cdots, Agent_N) \leftarrow$ Recruit(Q, KAMAC, P_1)
$\triangleright$ Clinician agents consist of a multi-disciplinary team.
KAMAC $\leftarrow (Agent_1, Agent_2, \cdots, Agent_N)$
$\triangleright$ Initial assessment.
$(Option_1, Option_2, \cdots, Option_N) \leftarrow$ Chat(Q, KAMAC, P_2)
$\triangleright$ Concat all options as feedback.
$Feedback \leftarrow$ Concat($Option_1, Option_2, \cdots, Option_N$)
$\triangleright$ **Knowledge-driven collaborative discussion.**
while $r \leq R$, and *not* $Consensus$, and *not* KG **do**
  $\triangleright$ Exchange agent's comments and determine the consensus.
  $Consensus, Feedback \leftarrow$ Chat(Q, KAMAC, $Feedback$, P_3)
  $\triangleright$ Assess whether any additional specialist is needed to fill a knowledge or diagnostic gap.
  $KG \leftarrow$ Chat(Q, KAMAC, $Feedback$, P_4)
  if KG **then**
    $\triangleright$ Expert recruitment for recruiting additional experts during the discussion.
    $(Agent_{N+1}, Agent_{N+2}, \cdots, Agent_M) \leftarrow$ Recruit(Q, KAMAC, P_5)
    $\triangleright$ Review all options and provide comments as feedback.
    $(Option_{N+1}, Option_{N+2}, \cdots, Option_M) \leftarrow$ Chat($(Agent_{N+1}, Agent_{N+2}, \cdots, Agent_M)$, $Feedback$, P_2)
    $Feedback \leftarrow$ Concat($Feedback, Option_{N+1}, Option_{N+2}, \cdots, Option_M$)
    $\triangleright$ Exchange agent's comments and determine the consensus.
    $Consensus, Feedback \leftarrow$ Chat(Q, $(Agent_{N+1}, Agent_{N+2}, \cdots, Agent_M)$,
    $Feedback$, P_3)
    $\triangleright$ Update KAMAC.
    KAMAC $\leftarrow$ (KAMAC, $\cdots, Agent_{N+1}, \cdots, Agent_M$)
    KG $\leftarrow$False
  $\triangleright$ Update agent's comments.
  $Feedback \leftarrow$ Chat(Q, KAMAC, $Feedback$, P_6)
  $r \leftarrow r + 1$
$\triangleright$ **Make the final decision by LLMs.**
$ans \leftarrow$ Moderator(Q, $Feedback$, P_7)
return ans

using the assessment prompt P_2, conditioned on the ongoing feedback. Their outputs are appended to the current feedback buffer and integrated into the group discussion in the subsequent round. This recursive process allows progressive team expansion, enabling KAMAC to dynamically adapt to the evolving complexity of the diagnostic scenario.

Throughout the discussion, all agents, including the initial and newly recruited ones, update their reasoning using the agent update prompt P_6, which ensures alignment with the current collective context. This process continues until either (1) a consensus is reached via iterative agreement checks using P_3, or (2) a maximum number of discussion rounds R is reached.

3.4 Decision Making

In the final stage, KAMAC invokes a moderator agent, typically a general-purpose LLM, to generate the final decision. The moderator receives the latest set of agent comments and the full discussion history and synthesizes a response via a decision prompt (P_7).

4 Experiments

4.1 Datasets

To evaluate the proposed KAMAC framework, we conduct experiments on the testing sets of two publicly available medical question answering (QA) datasets: MedQA (Jin et al., 2021) and Progn-VQA (Welch et al., 2023).

MedQA We use all 1273 samples in the testing set. This dataset describes the United States Medical Licensing Examination and includes questions, multiple-choice questions, and answers.

Progn-VQA We use all 750 Visual Question Answering (VQA) pairs in the testing set. The dataset includes head and neck cancer Computed Tomography (CT) image volumes collected from 2005-2017 treated with definitive radiotherapy at the University Health Network in Toronto, Canada. It also contains the corresponding regions of interest (ROIs) and structured patient information in RT-STRUCT format with standardized descriptions, including demographic, clinical, and treatment information based on the $7th$ edition TNM staging system and AJCC (American Joint Committee on Cancer). The dataset contains patient information, CT image volumes and ROIs, and the patient's survival status at the last follow-up. Please see Table 7 for understanding the clinical and imaging information used in the dataset. For CT input, we selected the axial slice with the largest cross-sectional area of the ROI. More details on the input clinical and imaging variables are provided in Appendix C.

4.2 Implementation Details

We use GPT-4.1-mini [1] as the primary model for all experiments, with the temperature set to 0 to ensure deterministic outputs. In addition, we compare our proposed method with DeepSeek-R1 (Guo et al., 2025), as shown in Table 2. For each medical question, we store the corresponding chat history in a local file. When revisiting the same question, the system loads the saved file to regenerate consistent initial medical comments from each role before resuming the collaborative discussion. The final decision is made solely based on the proposed collaboration method. The maximum number of discussion rounds R is set to 3. The initial number of experts is set to 1. We select GPT-4.1-mini due to its strong medical reasoning capabilities, low latency, predictable computational cost, and fully deterministic behavior. These advantages make it preferable for our controlled evaluation setting, in contrast to larger models such as GPT-4 or retrieval-enhanced models like DeepSeek-R1, which often entail higher overhead and less consistent outputs.

4.2.1 Comparison Methods

The compared methods include: (1) **Single-agent**, which uses an LLM for decision-making, where the question and the answer template are input to output an answer, (2) **Chain of Thought** (CoT) (Wei et al., 2022), which combines the single-agent backbone with a step-by-step prompt to conduct analysis and decision-making, (3) **Majority Voting**, which is used in multi-agent decision-making methods (Chen et al., 2023b; Yang et al., 2024a; Kaesberg et al., 2025) for making final decision with more that 50% votes. (4) **Consensus**, which is also adopted in (Kaesberg et al., 2025; Kim et al., 2024). (5) **MDAgents** (Kim et al., 2024) is an advanced multi-agent framework that performs problem-driven expert recruitment, MAC, and consensus decision to output the final results.

4.3 Evaluation Metrics

We evaluate the proposed method using four standard metrics: accuracy (Acc), precision (Prec), specificity (Spec), and recall score (Recall).

5 Results and Analysis

5.1 Comparisons with State-of-the-Arts

In Table 1, the proposed method achieves improved results on four metrics compared to multiple meth-

[1] https://openai.com/index/gpt-4-1/

Table 1: Main results on four common metrics across MedQA and Progn-VQA datasets, evaluated using GPT-4.1-mini. **Bold** values indicate the best performance. Here, 'SA' means the single-agent methods, and 'MA' means the multi-agent methods. Gray-highlighted cells indicate the average score.

Methods	**Types**	**MedQA** Acc	Prec	Spec	Recall	**Avg**	**Progn-VQA** Acc	Prec	Spec	Recall	**Avg**
Single-agent	SA	79.50	79.65	94.86	79.36	83.34	86.00	86.28	14.79	97.21	71.07
+CoT	SA	84.21	84.82	96.03	84.02	87.27	84.67	86.29	15.52	97.32	70.95
Majority Voting	MA	86.49	86.93	96.60	86.38	89.10	86.27	86.12	12.17	**99.84**	71.10
Consensus	MA	80.68	80.70	95.15	80.59	84.28	86.86	86.81	31.85	98.86	76.09
MDAgents	MA	87.74	87.92	96.92	87.55	90.03	87.01	88.83	33.70	96.21	76.44
KAMAC	MA	**88.14**	**88.30**	**97.02**	**88.11**	**90.39**	**87.20**	**89.79**	**40.52**	95.74	**78.31**

Table 2: Performance comparison of Baseline and KAMAC on MedQA and Progn-VQA using DeepSeek-R1 and GPT-4.1-mini across four metrics and their average. Gray-highlighted cells indicate the average score, with relative improvements shown in small colored text. Where 'Baseline' means single-agent+CoT.

Method	**MedQA** Acc	Prec	Spec	Recall	**Avg**	**Progn-VQA** Acc	Prec	Spec	Recall	**Avg**
Baseline (DeepSeek-R1)	88.14	88.12	97.03	88.00	90.32	77.87	88.11	37.07	85.33	72.10
KAMAC (DeepSeek-R1)	**89.63**	**89.53**	**97.41**	**89.50**	**91.52** (+1.20)	86.13	88.41	31.03	96.21	75.45 (+3.35)
Baseline (GPT-4.1-mini)	84.21	84.82	96.03	84.02	87.27	84.67	86.29	15.52	**97.32**	70.95
KAMAC (GPT-4.1-mini)	88.14	88.30	97.02	88.11	90.39 (+3.12)	**87.20**	**89.79**	**40.52**	95.74	**78.31** (+7.36)

Table 3: Discussion on the number of initial agents on the MedQA and Progn-VQA datasets. Gray-highlighted cells indicate the average score.

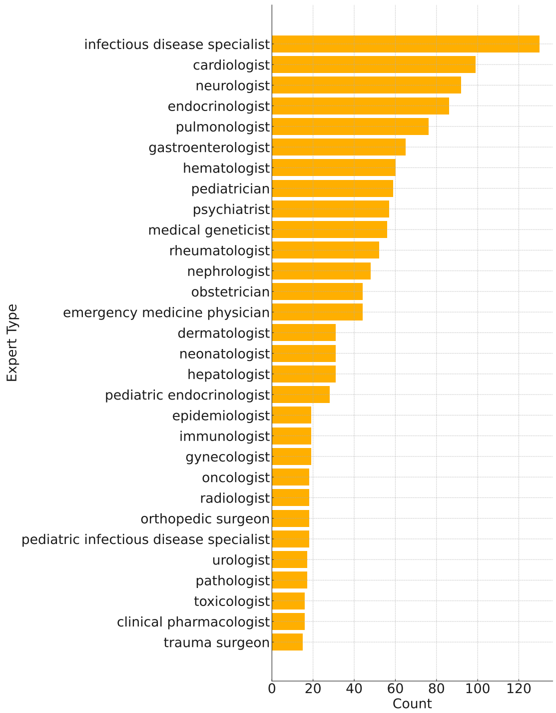
Initial Agents Number	**MedQA** Acc	Prec	Spec	Recall	**Avg**	**Progn-VQA** Acc	Prec	Spec	Recall	**Avg**
1	88.14	88.30	97.02	88.11	90.39	87.20	89.79	40.52	95.74	78.31
3	87.98	88.16	96.98	87.82	90.24	86.40	90.43	45.69	93.85	79.09
5	80.28	80.31	95.06	80.13	83.95	89.10	89.54	35.43	96.69	77.69

ods on the MedQA dataset. For the Progn-VQA dataset, the proposed method achieves better results on the Acc, Prec, and Spec metrics. In addition, KAMAC leverages knowledge-driven prompts to detect KG and expand experts to form multi-agent collaborative discussions. Focusing on multi-agent-based methods, both the majority voting and consensus are set to five experts, while MDAgents adopts a single agent, a multi-disciplinary team with five experts, and an integrated care team with nine experts. Compared with them, the proposed method can achieve better results, which demonstrates that our method overcomes the limitation of knowledge in the single-agent model and has a more suitable multi-disciplinary team to enhance multi-agent reasoning and collaboration.

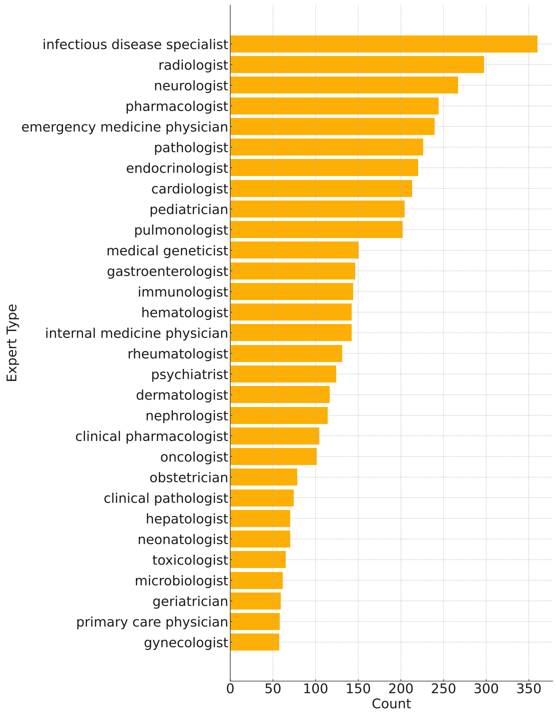
In Table 2, we further evaluate our method on another model, DeepSeek-R1. In our method, the initial number of experts is set to 1, which is consistent with the baseline method (single-agent + CoT), but the experimental results are better than the baseline method. This improvement shows that our method can be generalized to more LLM models. In addition, this improvement aligns with the actual clinical treatment scenario, where clinical treatment allows the dynamic addition of experts according to the patient's clinical treatment situation, thereby carrying out more effective treatment. It contributes to optimizing the best treatment options and limited clinical resources in medical applications.

5.2 Discussion of KAMAC

Results of Different Multi-agent Methods In Table 1, we present different multi-agent methods. Our approach outperforms the static multi-agent methods (majority voting and Consensus) and the MDAgent method on both datasets. In the static multi-agent methods, five experts were set to conduct a multi-agent collaboration. In MDAgents,



(a) Number of Initial Experts: 1



(b) Number of Initial Experts: 5

Figure 3: Histogram illustrating the impact of initial expert settings on the final top-30 expert distribution in our method on the MedQA dataset. "Count" denotes the total frequency of each expert type. An 80% overlap in expert types is observed between the 1- and 5-expert settings.

there were 2.41 and 4.34 experts per case. Compared to the above methods, KAMAC recruited 1.28 and 2.41 experts per case, respectively, which is still significantly more efficient than the full multi-agent approach and is 53% and 56% lower than MDAgents, respectively. In addition, we also perform statistical analysis and cost-efficiency in Appendix B.

Comparison of Initial Number of Experts As shown in Table 3, using one initial expert outperforms other configurations. This indicates that starting with a single agent promotes more targeted recruitment. With only one initial agent, KAC-MAF more accurately identifies knowledge gaps and recruits the most relevant experts. In contrast, beginning with five agents may introduce overlapping or irrelevant perspectives early, increasing redundancy and noise in subsequent decisions. Thus, fewer initial experts allow for more precise adaptation and reduce early overfitting. In Figure 3, we show the expert distribution for KAMAC with 1 versus 5 initial experts. Among the top 30 experts, 24 are common to both settings (80% overlap). However, the total number of experts remains lower when starting with one expert compared to five.

6 Conclusion

This work presents KAMAC, a knowledge-driven adaptive multi-agent collaboration framework that brings structured, dynamic reasoning into medical decision-making with LLMs. By allowing agents to actively assess their own limitations and request additional expertise when needed, KAMAC overcomes the rigidity of traditional multi-agent setups and more faithfully mirrors real-world clinical workflows. Our experiments on two real-world medical QA benchmarks demonstrate that KAMAC consistently outperforms both single-agent and existing multi-agent baselines.

Beyond accuracy improvements, KAMAC offers deeper insights into AI collaboration: decision quality improves not merely through more parameters or agents, but through adaptive, feedback-driven interaction grounded in knowledge awareness. This framework brings multi-agent LLM systems closer to real-world clinical workflows, where expert composition evolves with case complexity. Future directions include modeling agent uncertainty and integrating clinician-in-the-loop feedback to further support real-time deployment in medical environments.

Limitations

While KAMAC demonstrates promising results, it has several limitations. The current framework focuses on textual and imaging inputs; future work could incorporate additional modalities such as genomic or longitudinal clinical data to support a wider range of medical tasks. Although KAMAC achieves strong performance without fine-tuning the underlying LLMs, domain-specific fine-tuning may further improve accuracy and agent-role fidelity. However, this would introduce significant computational overhead and is challenged by the scarcity of high-quality, labeled medical data. Balancing accuracy gains with efficiency and data availability remains an important direction for future fine-tuning efforts.

Acknowledgments

This work was supported in part by Natural Science Foundation of Sichuan Province (2024NSFSC0038) and the NSFC (12171072).

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, competition, and maliciousness: Llm-stakeholders interactive negotiation. *Advances in Neural Information Processing Systems*, 37:83548–83599.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. 2023a. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023b. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Kai Chen, Xinfeng Li, Tianpei Yang, Hewei Wang, Wei Dong, and Yang Gao. 2025. Mdteamgpt: A self-evolving llm-based multi-agent framework for multi-disciplinary team medical consultation. *arXiv preprint arXiv:2503.13856*.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, and 1 others. 2023c. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, and 1 others. 2023. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. *arXiv preprint arXiv:2502.19130*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Dennis L Kodner and Cor Spreeuwenberg. 2002. Integrated care: meaning, logic, applications, and implications–a discussion paper. *International journal of integrated care*, 2:e12.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024a. More agents is all you need. *Transactions on Machine Learning Research*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448.

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

ML Welch, S Kim, A Hope, SH Huang, Z Lu, J Marsilla, M Kazmierski, K Rey-McIntyre, T Patel, B O'Sullivan, and 1 others. 2023. Computed tomography images from large head and neck cohort (radcure). *The Cancer Imaging Archive*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Joshua C Yang, Damian Dalisan, Marcin Korecki, Carina I Hausladen, and Dirk Helbing. 2024a. Llm voting: Human choices and ai collective decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1696–1708.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2024b. Confidence vs critique: A decomposition of self-correction capability for llms. *arXiv preprint arXiv:2412.19513*.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. 2024. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms. In *NeurIPS Workshop on Open-World Agents*.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. 2024. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

A Clinical Impact and Challenges

In Section 6, we discussed the limitations of KAMAC. Now, we will explore the clinical implications and challenges of multi-agent collaboration.

Adaptive Multi-agent Collaboration Clinical decision-making often involves the emergence of new information and continuous refinement of diagnoses. The dynamic multi-agent framework is well-suited to such scenarios, as it enables progressive deepening of the diagnostic process and allows selective inclusion of additional experts when necessary. This flexibility helps ensure that complex or ambiguous cases receive input from the most relevant specialists, thereby improving diagnostic accuracy and comprehensiveness. However, given the inherent uncertainties associated with large language models, the dynamics and stability of the system in real-world clinical applications warrant further investigation in future studies.

Consensus and Deliberation In clinical practice, the effective integration of diverse expert opinions is essential. Our framework facilitates progressive discussions by starting with a single agent, allowing the reasoning process to unfold incrementally and making disagreements or uncertainties more explicit. This structure encourages more thorough deliberation. It should be noted that we did not attempt to force consensus among experts; rather, additional experts were introduced selectively to ensure that the collective recommendations are as comprehensive and well-supported as possible.

B More Discussion of KAMAC

Component Analysis Table 1 also presents a component-wise comparison to evaluate the contribution of each module. Specifically, KAMAC extends the baseline (Single-agent + CoT) by integrating a knowledge gap detection mechanism that dynamically recruits multiple experts for collaborative reasoning. This knowledge-aware adaptive collaboration replaces both the predefined recruitment strategy used in majority voting and the problem-driven recruitment in MDAgent. KAMAC consistently outperforms the alternatives across nearly all evaluation metrics, demonstrating the efficacy of its design.

Consensus Strategies In this part, we compare two consensus strategies: Ensemble Refinement (Singhal et al., 2022) and Majority Voting. In the Ensemble Refinement strategy, the moderator agent aggregates all expert responses and assigns weights to each in an effort to produce a more nuanced consensus. However, as shown in Table 4, this approach yields lower performance than Majority Voting. Specifically, it tends to amplify the influence of individual expert comments and is more susceptible to noise. In contrast, Majority Voting adopts a simpler and more robust mechanism: it treats each expert's opinion equally and determines the final decision purely based on vote counts. This strategy effectively mitigates the impact of individual biases or outlier predictions, resulting in more stable performance across diverse cases. Accordingly, Majority Voting is adopted as the default consensus strategy in our framework, as it consistently outperforms Ensemble Refinement within the KAMAC architecture.

B.1 Computational Cost Analysis

In Section 4, we introduce static multi-agent methods, MDAgent and KAMAC, and report their average number of experts on two test sets: 5 and 5 for static multi-agent methods, 2.41 and 3.34 for MDAgent, and 1.28 and 2.14 for KAMAC. We also compare inference time, average LLM API calls, and total cost across single-agent and multi-agent methods (Table 5).

The single-agent CoT baseline uses one expert per case, while static multi-agent methods rely on a fixed pool of five experts, resulting in higher latency, API usage, and cost. These static methods lack scalability and fail to balance accuracy and efficiency. In contrast, KAMAC dynamically recruits fewer experts (1.28 on average for MedQA), reducing expert usage, API calls, and reasoning time by 73–79% compared to static approaches. Compared to MDAgent, KAMAC improves accuracy and inference speed while reducing expert usage by 47%, API calls by 24%, and total cost by 21%, demonstrating better scalability and cost-effectiveness.

B.2 Statistical Analysis

Since the output of the large language model varies greatly across multiple runs, this high variability affects the direct pairwise comparison between KAMAC and MDAgent, affecting the judgment of statistical significance. To avoid this, we adopt an alternative statistical significance test strategy. Specifically, we compared the two methods independently against a shared baseline configuration (single agent + CoT). Then, we repeat the two meth-

Table 4: Discussion on the consensus strategy of the KAMAC method on the MedQA and Progn-VQA datasets. Gray-highlighted cells indicate the average score.

Consensus Strategy	**MedQA**				**Avg**	**Progn-VQA**				**Avg**
	Acc	Prec	Spec	Recall		Acc	Prec	Spec	Recall	
Ensemble Refinement	87.35	87.40	96.83	87.23	89.70	86.27	90.91	49.14	93.06	79.85
Majority Voting	88.14	88.30	97.02	88.11	90.39	87.20	89.79	40.52	95.74	78.31

Table 5: Comparison of Baseline, Major Voting, MDAgent, and KAMAC on metrics, expert usage (Number), inference time per case (Time (s)), API calls, and cost on the MedQA dataset.

Method	**Acc**	**Prec**	**Spec**	**Recall**	**Avg**	**Number**	**Time (s)**	**API Calls**	**Cost**
Baseline	84.21	84.82	96.03	84.02	87.27	1	9.56	2.20	4.01
Majority Voting	86.49	86.93	96.60	86.38	89.10	5	39.88	12.02	8.12
MDAgent	87.74	87.92	96.92	87.55	90.03	2.41	15.62	3.34	6.31
KAMAC	88.14	88.30	97.02	88.11	90.39	1.28	10.80	2.55	5.01

ods three times using the same baseline. Finally, each run of KAMAC and MDAgent is subjected to an independent sample t-test against the baseline, covering accuracy, precision, recall, and specificity.

In Table 6, results consistently demonstrate statistically significant improvements over the baseline (with most p-values < 0.01 and several p-values < 0.001). These findings indicate that both KAC-MAF and MDAgent achieved substantial performance gains, with KAC-MAF generally exhibiting larger standardized improvements and lower variance across repeated runs. For example, on the MedQA dataset, KAC-MAF yielded higher t-statistics for precision ($t = 33.05$ vs. 15.04) and recall ($t = 26.15$ vs. 13.78), corresponding to lower p-values ($p = 0.0009$ and $p = 0.0015$, respectively), which reflects more pronounced and consistent improvements relative to the baseline. On the Progn-VQA dataset, although both methods produced negative t-statistics for recall, due to the higher recall achieved by the single-agent CoT baseline, KAC-MAF still attained higher mean recall (0.6092 vs. 0.3164), with a statistically significant difference ($p = 0.0063$). Both methods also achieved significant gains in accuracy and specificity, with KAC-MAF maintaining more stable performance characterized by smaller standard deviations across runs. Overall, these results confirm that KAC-MAF not only achieved statistically significant improvements over baseline but also generally outperformed MDAgent in terms of both effect size and consistency across most evaluated metrics.

C More Details for Progn-VQA Dataset

According to the settings of (Welch et al., 2023), we provide the clinical and imaging information required for prognosis in Table 7. This information can fully describe the situation of patients with head and neck cancer.

D Prompt Template

We provide all prompts in our multi-agent medical decision-making framework, including expert recruitment, initial comments, collaborative discussion, and knowledge-driven prompts. For a single-agent setting, you can refer to (Kim et al., 2024). We also provide a full role-playing prompt (P2). The example below demonstrates this, with role-specific phrases highlighted in bold.

System: You are a **Radiation Oncologist** whose **expertise is strictly limited to radiation therapy planning and dosing for head and neck squamous cell carcinoma, especially HPV-positive cases.** Your job is to collaborate with other medical experts in a team.

User: Given the examplers, as a **Radiation Oncologist**, please return your answer to the medical query among the options provided. **You are not allowed to switch to any other medical specialty.**

Response:

Step 1: Initial Assessment of the CT Scan (without bounding box data)

The CT scan is a sagittal view of the head and neck region... *cervical vertebrae*, the airway (pharynx and larynx), and soft tissues of the neck... The airway appears patent but shows narrowing in the

Table 6: Statistical Significance Comparison of KAMAC and MDAgent vs Baseline: single agent + CoT

Dataset		KAMAC				MDAgent			
	Metric	Mean	Std	t-statistic	p-value	Mean	Std	t-statistic	p-value
MedQA	Acc	0.8761	0.0016	30.07**	0.0011**	0.8761	0.0011	**43.33*****	**0.0005*****
	Prec	**0.8765**	0.0015	**33.05*****	**0.0009*****	0.8756	0.0032	15.04**	0.0044**
	Recall	**0.8748**	0.0018	**26.15****	**0.0015****	0.8725	0.0031	13.78**	0.0052**
	Spec	0.9745	0.0076	**24.74****	**0.0016****	**0.9760**	0.0098	19.35**	0.0027**
Progn-VQA	Acc	**0.8700**	0.0024	**14.02****	**0.0051****	0.8656	0.0050	5.40*	0.0327*
	Prec	**0.8868**	0.0082	**6.92***	**0.0202***	0.8788	0.0077	5.88*	0.0277*
	Recall	**0.6092**	0.0267	–12.59**	0.0063**	0.3164	0.0180	**–41.62*****	**0.0006*****
	Spec	**0.9790**	0.0153	**12.26****	**0.0066****	0.9548	0.0052	29.47***	0.0011***

Note: * p<0.05; ** p<0.01; *** p<0.001. Bolded values denote the method with the higher absolute t-statistic (i.e., more significant difference).

Variable	**Description**
Age	Patient age
Sex	Patient sex
ECOG PS	ECOG Performance Status
Smoking PY	Cumulative smoking exposure (pack-years)
Smoking Status	Smoking status at initial consultation
Ds Site	Primary disease (cancer) site
Subsite	Subsite of the primary tumor
T	Tumor size and extent (AJCC 7th edition T category)
N	Regional lymph node involvement (AJCC 7th edition N category)
M	Distant metastasis (AJCC 7th edition M category)
Stage	Overall stage group (AJCC 7th edition)
Path	Pathological diagnosis or histological subtype
HPV	HPV status of the tumor, determined by p16 IHC with or without confirmation by HPV DNA PCR (blank if unavailable)
Tx Modality	Treatment modality
Chemo?	Whether concurrent chemoradiotherapy was administered
Dose	Total radiotherapy dose delivered (in Gy)
Fx	Number of radiotherapy fractions
Local	Indicator of local recurrence
Regional	Indicator of regional recurrence
Distant	Indicator of distant metastasis
2nd Ca	Indicator of second primary cancer
ContrastEnhanced	Indicator of whether contrast-enhanced imaging was used

Table 7: Descriptions of clinical and imaging variables included in the Progn-VQA dataset (Welch et al., 2023).

oropharyngeal region... The masked *region of interest (ROI)* is located anterior to the cervical spine...

Step 3: Clinical Reasoning

The patient likely has a head and neck *squamous cell carcinoma (HNSCC)...* a common site for *HPV-related carcinoma...* From a *radiation oncology* perspective, delineation of these volumes is critical... Airway narrowing may cause symptoms such as *dysphagia* or obstruction...

E Ethics Considerations

While our work strictly adheres to established benchmarks in the field of medical question answering, we acknowledge the potential risks associated with applying large language models (LLMs) to critical domains such as medicine. In particular, LLMs may exhibit inherent biases or generate inaccurate reasoning, which could lead to unintended consequences if applied without human oversight. Therefore, we emphasize that any deployment of such models in real-world medical settings should be conducted with caution and accompanied by expert validation.

Expert Recruitment Prompt for MedQA (P_1)

System: You are an experienced medical expert who recruits a group of experts with diverse identities and asks them to discuss and solve the given medical query.
User:
Question: {{QUESTION}}
You can recruit {{NUM_AGENTS}} 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), or indicate if they are independent.
For example, if you want to recruit five experts, your answer can be like:
1. Pediatrician - Specializes in the medical care of infants, children, and adolescents. - Hierarchy: Independent
2. Cardiologist - Focuses on the diagnosis and treatment of heart and blood vessel-related conditions. - Hierarchy: Pediatrician > Cardiologist
3. Pulmonologist - Specializes in the diagnosis and treatment of respiratory system disorders. - Hierarchy: Independent
4. Neonatologist - Focuses on the care of newborn infants, especially those who are born prematurely or have medical issues at birth. - Hierarchy: Independent
5. Medical Geneticist - Specializes in the study of genes and heredity. - Hierarchy: Independent
Please answer in the above format, and do not include your reason.

Expert Recruitment Prompt for Progn-VQA (P_1)

System: 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.
User:
Question: {{QUESTION}}
Considering the medical question and the options for the answer, what kinds of experts will you recruit to better make an accurate decision? You also need to clearly specify the communication structure between experts or indicate if they are independent.
You must recruit exactly the following {{NUM_AGENTS}} experts, with no substitutions, no additional experts, and no omissions:
(e.g., Radiation Oncologist == Medical Oncologist == Pathologist == Surgical Oncologist (Recurrence/Secondary Cancers) == Targeted Therapy Expert),
Please strictly follow the format shown below, without adding any extra explanation or reasoning.
Format example if recruiting {{NUM_AGENTS}} experts:
1. Radiation Oncologist - Your expertise is strictly limited to radiation therapy planning and dosing for head and neck squamous cell carcinoma, especially HPV-positive cases.
- Hierarchy: Radiation Oncologist == Medical Oncologist
2. Medical Oncologist - Your expertise is strictly limited to systemic therapy decisions, including chemotherapy and immunotherapy in head and neck cancers.
- Hierarchy: Medical Oncologist == Radiation Oncologist
3. Surgical Oncologist (Recurrence/Secondary Cancers)—Your expertise is strictly limited to evaluating surgical options for recurrent or secondary malignancies in head and neck cancers.
- Hierarchy: Surgical Oncologist == Pathologist"
4. Pathologist - Your expertise is strictly limited to pathological diagnosis of head and neck squamous cell carcinoma, HPV status evaluation, and margin assessment post-surgery.
- Hierarchy: Pathologist == Surgical Oncologist
5. Targeted Therapy Expert - Your expertise is strictly limited to clinical application of EGFR inhibitors and novel agents targeting HPV-positive tumors.
- Hierarchy: Targeted Therapy Expert -> Medical Oncologist
Your answer must conform exactly to the format above.

Chain-of-thought Prompt for Initial Assessment (P_2)

System: You are a {{ROLE}} who {{DESCRIPTION}}. Your job is to collaborate with other medical experts in a team.
User: {{VISUAL COT INSTRUCTION}} (**Optional**)
Given the examplers, as a {{ROLE}}, please return your answer to the medical query among the options provided. You are not allowed to switch to any other medical specialty.
{{FEWSHOT_EXAMPLERS}}
Question: {{QUESTION}}
Your answer should be in the format below.
{{answer_template}}

Visual Chain-of-thought Prompt for Head and Neck CT Scan (Optional, only be used when input data include images.)

User: You will be provided with a head and neck CT scan that includes one or more masked regions of interest (ROIs). Alongside the scan, one or more 3D bounding box coordinates will be supplied, each defining specific volumetric regions within the scan. These coordinates identify either organs, disease regions, or cellular structures. Each bounding box is defined by its minimum and maximum values along the z, y, and x axes and is normalized relative to the original image size.
The given bounding box coordinates are: {{BBOX_COORDS}}.
Task Instructions:
1. **Initial Assessment**: Carefully analyze the CT scan image (without using the bounding box data). Describe any visible anatomical structures, patterns, abnormalities, and note the characteristics of the masked regions of interest (ROIs).
Do not use the bounding box data at this stage.
2. **Mapping Bounding Boxes**: Consider the bounding box coordinates and map them to the corresponding areas within the scan.
3. **Clinical Reasoning**: Summarize the patient's clinical context and findings in a clear, structured bullet-point format and reason through the patient's condition step by step.
4. **Integrated Conclusion**: Combine your findings from the image analysis, bounding box mapping, and masked ROI to concisely synthesize your final clinical impression.
Be thorough and precise in both your image-based observations and your clinical reasoning.

Agent Interaction Prompt (P_3)

User: Earlier in this conversation, a set of discussion opinions from other medical experts on your team was provided. Please do not forget those earlier opinions.
Now, additional new opinions have been provided. Considering both the earlier and the latest opinions together, please indicate whether you want to talk to any additional expert (yes/no).
Opinions: {{ASSESSMENT}}

Knowledge-driven Prompt for Recruited Experts (P_4)

User: You are part of the team: {{AGENTS}}. Earlier in this conversation, a set of discussion opinions from one or more medical experts on your team was provided. Please carefully review that information now. Based on your professional boundaries, determine whether there is a knowledge limitation or missing perspective that requires support from another specialist.
Please answer yes or no.
If yes, specify the type of expert needed and provide a short reason. Be specific and consider the multidisciplinary needs involved in managing complex patient information (e.g., diagnostic imaging, supportive care, pathology review, and other medical expertise).
It is acceptable to recognize areas of expertise already covered by current team members ({{AGENTS}}).
Do not recommend a specialist if their expertise is already represented in the team.

Knowledge-driven Prompt for Expert Recruitment (P_5)

User: Considering the medical question, discussion options, and the current expert team {{AGENTS}}, identify any that require recruiting new types of experts to ensure an accurate decision (exclude {{AGENTS}}).
You also need to clearly specify the communication structure between experts (e.g. Targeted Therapy Expert -> Medical Oncologist, Medical Oncologist == Radiation Oncologist)" or indicate if the new expert(s) will work independently.
Do not suggest removing, substituting, or duplicating existing experts. Only add new experts if necessary.
Format example if recruiting experts:
1. Medical Oncologist - Your expertise is strictly limited to systemic therapy decisions, including chemotherapy and immunotherapy in head and neck cancers. - Hierarchy: Independent
2. Other Medical Experts.
Your answer must conform exactly to the format above. If the existing expert team comprehensively have covered the necessary expertise for accurate decision, answer: <skip recruitment>

Agent Update Comments after Discussion Prompt (P_6)

User: Now that you've interacted with other medical experts, remind your expertise and the comments from other experts and make your final answer to the given question:{{QUESTION}}
Answer: {{ANSWER_TEMPLATE}}
Only output your final answer in the format below:
{{FINAL_ANSWER_TEMPLATE}}
Question: {{QUESTION}}

Final Decision Prompt (P_7)

System: You are a final medical decision maker who reviews all opinions from different medical experts and make final decision.
User: Given each agent's final answer, please review each agent's opinion and make the final answer to the question by taking a majority vote.
Only output your final answer in the format below:
{{FINAL_ANSWER_TEMPLATE}}
Question: {{QUESTION}}