SAMAC: Collaborating with Multi-Disciplinary Team for Adaptive Medical Decision-Making

Maosheng Li[1], Xiao Wu[1], Hu Wang[2], Imran Razzak[1], Yutong Xie[1]*

[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Khalifa University
{maosheng.li,xiao.wu,imran.razzak,yutong.xie}@mbzuai.ac.ae
hu.wang@ku.ac.ae

Abstract

Multi-agent large language model systems offer a promising paradigm for multidisciplinary medical decision support by integrating heterogeneous expert perspectives. However, existing frameworks often rely on loosely structured interactions, where disagreement is insufficiently diagnosed and discussion may collapse into superficial consensus or parallel opinion aggregation. We propose Structured Adaptive Multidisciplinary Collaboration (SAMAC), a structured and strategy-aware multi-agent framework for adaptive MDT-style medical reasoning. SAMAC introduces Structured Deliberation to detect opinion disagreement and majority not confident, along with a Consensus Purification mechanism that transforms disagreement into constructive and collaborative refinement. When disagreement arises, SAMAC adaptively selects between Self-refinement and Counter-questioning to enable evidence-grounded negotiation. Experiments on three medical datasets—RADCURE, HANCOCK, and OSCCPI—show that SAMAC consistently improves overall performance and collaborative gains over existing multi-agent baselines. In particular, SAMAC improves the Average Score by 4.29 points on the RADCURE dataset and achieves higher Relative Self-Correction Scores (RSS), indicating more effective resolution of disagreement and stable collective decisions. These results highlight the importance of controlled and diverse interaction dynamics in multi-agent medical reasoning. Code will be released upon publication.Our code is publicly available at: https://github.com/MaoshengLi-fox/SAMAC.

1 Introduction

Healthcare is increasingly improved by the quality, timeliness, and precision of its medical decisions. Clinical decision-making encompasses diagnostic evaluation and treatment strategy, which ultimately affect patients' survival and quality of life (Djulbegovic and Elqayam, 2017; Elwyn et al., 2012). At the same time, patients may have heterogeneous comorbidities, individualized risk profiles, and contextual factors that often fall outside the strict boundaries of individual clinicians (Boast et al., 2018; McDuff et al., 2025). As a result, clinical decision-making is no longer a straightforward application of evidence, but a process of fully interpreting, weighing, and reconciling competing considerations based on multidisciplinary teams (MDTs) (Bates et al., 2025; Hendriks et al., 2024; Taberna et al., 2020). In this work, we study this problem from the perspective of LLM-based MDT-style reasoning, focusing on how multiple expert agents should deliberate when their opinions or confidence levels diverge.

Recent large language model (LLM)-based MDT approaches aim to emulate this collaborative process by leveraging LLMs to integrate structured and unstructured clinical data and generate multi-disciplinary decision support in natural language (Clusmann et al., 2023; Kim et al., 2024a,b; Li et al., 2024; Tang et al., 2024; Wu et al., 2025; Yan et al., 2024). For example, MedAgents (Tang et al., 2024) proposes a role-playing-based (Li et al., 2023) MDT collaboration framework in which LLM agents representing different medical specialties iteratively discuss and refine their analyses until they reach a consensus report. In addition, MDAgents (Kim et al., 2024a) dynamically assigns individual or team-based configurations based on the estimated medical complexity of the query, emulating real-world medical decision-making flows. More recently, KAMAC (Wu et al., 2025) has proposed a knowledge-driven adaptive collaboration strategy, in which LLM agents dynamically detect knowledge gaps during discussion and recruit additional specialists to expand the expert team, allowing flexible cross-specialty reasoning in complex clinical scenarios.

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 18474–18485
October 24-29, 2026 ©2026 Association for Computational Linguistics

Existing adaptive MDT-style systems mainly focus on adapting who participates in the collaboration, such as assigning individual or team-based configurations or recruiting additional specialists when knowledge gaps are detected. However, after the expert team is formed, the interaction process itself remains under-controlled: agents are often asked to discuss, debate, or vote under a fixed protocol, regardless of the current disagreement state (Du et al., 2024; Liang et al., 2024).

To address these limitations, we propose Structured Adaptive Multi-disciplinary Collaboration (SAMAC), a structured framework for adaptive MDT-style medical reasoning with multi-disciplinary expert agents. SAMAC addresses this complementary problem by adapting when agents should deliberate and how disagreement should be resolved, thereby providing an explicit interaction-control layer for MDT-style reasoning. As shown in Fig. 1(b), disagreement within an MDT can arise from two complementary sources: opinion disagreement among experts and confidence disagreement defined as Majority Not Confident (MNC), where more than half of the agents report neutral or not confident. Based on this, we introduce a Structured Deliberation that systematically organizes expert opinions and confidence levels (Chen et al., 2024; Yang et al., 2024), providing explicit, interpretable guidance for disagreement diagnosis. By jointly examining opinion disagreement and MNC, the framework determines whether further discussion is necessary, guiding agents toward focused and meaningful deliberation rather than superficial exchanges. To further enhance collaboration quality, we propose a Consensus Purification mechanism that treats disagreement as a diagnostic trigger. When disagreements arise, SAMAC initiates differentiated behaviors that self-refine viewpoints, counter-question opinions, and update evidence, transforming superficial opinion exchange into structured multi-agent negotiation. Our contributions are threefold:

- We propose SAMAC, a structured interaction-control framework for LLM-based MDT-style medical reasoning. Rather than focusing only on expert recruitment or final aggregation, SAMAC explicitly models the collaboration state and regulates when and how agents deliberate.
- We design Structured Deliberation, a confidence-aware triggering mechanism that jointly captures answer-level disagreement and epistemic uncertainty, thereby preventing multi-agent collaboration from degenerating into implicit majority voting. We further introduce Consensus Purification, which routes disagreement resolution into two complementary behaviors—Self-refinement and Counter-questioning—to support evidence-grounded negotiation.
- Extensive experiments on three medical benchmarks, RADCURE, HANCOCK, and OSCCPI, demonstrate that our SAMAC consistently enhances multi-agent collaborative performance.

2 Related Work

Multidisciplinary decision-making with LLMs. LLM-based MDT-style medical reasoning has increasingly shifted from single-agent reasoning toward multidisciplinary collaboration. Recent MDT-style systems use LLM agents to represent different clinical specialties, enabling the integration of structured clinical variables, unstructured reports, and other patient evidence into natural-language decision support (Clusmann et al., 2023; Kim et al., 2024a,b; Li et al., 2024; Tang et al., 2024; Wu et al., 2025; Yan et al., 2024). This paradigm is particularly suitable for complex clinical tasks, where no single specialty may fully cover all relevant diagnostic, prognostic, and treatment-related factors. Representative work includes MedAgents (Tang et al., 2024), which builds a role-playing-based (Li et al., 2023) MDT framework for iterative expert discussion and consensus formation.

Adaptive collaboration in multi-agent medical reasoning. Recent multi-agent medical systems increasingly seek to move beyond static expert panels toward adaptive collaboration. One direction focuses on the dynamic composition of multidisciplinary teams (MDTs). For example, MDAgents (Kim et al., 2024a) dynamically assigns individual or team-based configurations according to the estimated complexity of the medical query, while KAMAC (Wu et al., 2025) further introduces a knowledge-driven adaptive collaboration framework that detects knowledge gaps during discussion and recruits additional specialists when necessary. Such methods make the expert team more flexible than fixed-role settings and allow the MDT structure to evolve with the diagnostic context.

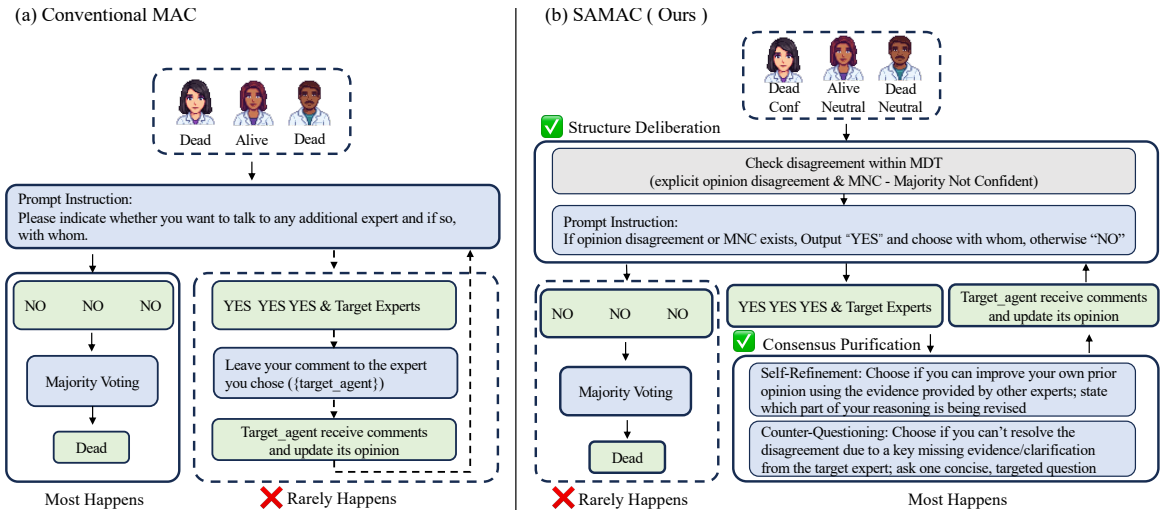


Figure 1: (a) Conventional multi-agent collaboration (MAC) relies on prompt-triggered discussion and majority voting. (b) Our SAMAC features Structured Deliberation and Consensus Purification for adaptive MDT-style reasoning. Blue boxes denote system-level instructions and green boxes mean agent actions during interaction.

Another important aspect of adaptivity lies in the interaction process among expert agents. Existing multi-agent reasoning frameworks usually rely on a single predefined interaction behavior throughout collaboration. For example, some methods aggregate agent outputs through majority voting or consensus-based decision-making (Kaesberg et al., 2025), while debate-based frameworks construct opposing roles or iterative argumentation to resolve conflicts (Du et al., 2024; Liang et al., 2024). Other medical multi-agent systems mainly adopt prompt-triggered discussion, where agents are asked whether further interaction or additional expertise is needed (Kim et al., 2024a; Wu et al., 2025). While these studies have substantially advanced LLM-based MDT collaboration, most of them focus either on expert composition or on a fixed interaction protocol. In contrast, SAMAC focuses on an interaction-control layer that diagnoses the current collaboration state and routes agents to different disagreement-resolution behaviors after expert opinions are available.

3 Method

3.1 Overview

As shown in Fig. 2, our SAMAC mainly includes five stages: (1) Recruit Stage: SAMAC is built on the dynamic recruitment framework of KAMAC (Wu et al., 2025) and inherits its recruitment backbone. We further extend this stage by requiring each recruited expert agent to produce not only an opinion but also an explicit confidence level and supporting evidence; (2) Structured Deliberation: SAMAC explicitly diagnoses disagreement by jointly evaluating opinion disagreement among experts and Majority Not Confident (MNC); (3) Consensus Purification and Opinion Update: When disagreement is detected, SAMAC initiates structured collaboration and adaptively selects between Self-refinement and Counter-questioning strategies to regulate how conflicts are resolved. The updated opinions and evidence are iteratively refined through this process; (4) Knowledge-Gap (KG) Identification: After discussion, SAMAC follows KAMAC (Wu et al., 2025) to examine whether unresolved knowledge gaps remain. If gaps are detected, SAMAC dynamically recruits additional specialists to address them and provide targeted clarification before the final decision. (5) Final Results: SAMAC aggregates refined opinions and evidence into a unified clinical recommendation.

3.2 Structured Deliberation through Explicit Guidance

A core component of SAMAC is explicit guidance that structures multi-agent deliberation and prevents collaboration from degenerating into independent reasoning followed by majority voting. Instead of allowing agents to observe peer responses only and implicitly conform to dominant opinions, SAMAC introduces Structured Deliberation that explicitly diagnoses disagreement and regulates when and how collaboration should proceed.

Specifically, Structured Deliberation aggregates all the latest experts' opinions and computes an objective "system disagreement signal" vector from the current opinions, including:

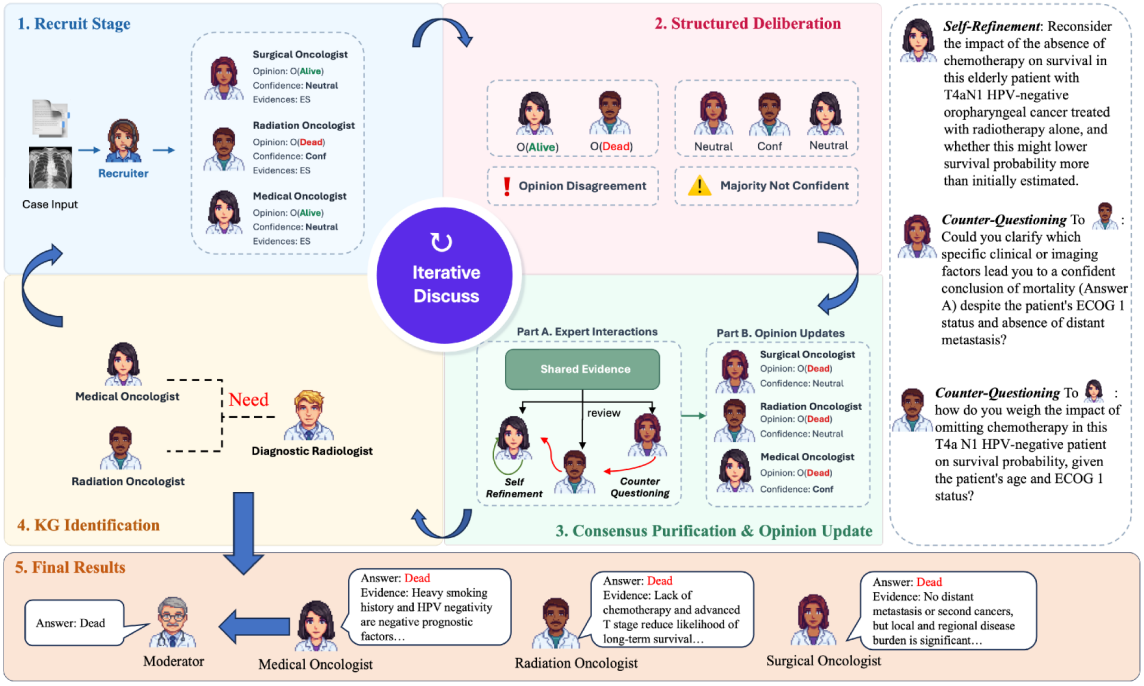


Figure 2: Overview of the proposed SAMAC framework for adaptive collaboration of multi-agent MDT-style medical reasoning. Given a clinical question and patient evidence, SAMAC first recruits a configurable number of domain-specific agents to provide independent predictions with confidence estimates. It then performs structured deliberation to identify opinion disagreement or insufficient consensus among agents, where insufficient consensus is determined when more than half of the agents provide Not-confident predictions. Once triggered, the agents enter an iterative collaboration process involving self-refinement, counter- questioning, and opinion updates, allowing them to reassess uncertain evidence and resolve conflicting judgments. When necessary, SAMAC further identifies missing knowledge and recruits an additional specialist before producing the final summarized decision.

- **Opinion disagreement**: Whether multiple distinct answers exist among experts.

- **Majority Not Confident (MNC)**: Whether more than half of the experts are not confident, determined through majority voting.

If neither substantive opinion disagreement nor insufficient confidence is identified, the system proceeds directly to generating the final answer. Otherwise, the disagreement is routed to the Consensus Purification module. These disagreement signals serve as explicit guidance for deliberation rather than passive observation. Agents are instructed to examine not only the final answers but also the confidence levels and underlying reasoning chains, thereby fostering proactive collaboration. By structurally guiding attention toward opinion divergence and epistemic uncertainty, Structured Deliberation prevents superficial consensus formation and mitigates premature convergence to majority voting.

3.3 Consensus Purification through Disagreement Diagnosis

Medical diagnosis requires structured reasoning grounded in established clinical evidence rather than purely empirical inference. Evidence-based agentic frameworks, such as MedAgent-Pro (Wang et al., 2025), emphasize aligning decision-making with retrieved clinical criteria to ensure reliability and interpretability. Inspired by this paradigm, we posit that disagreement in multi-agent collaboration may arise either from evidential insufficiency or from reasoning inconsistency. Meanwhile, recent studies have shown that agent generalization is closely related to the model's ability to perform self-refinement based on feedback (Fu et al., 2025; Madaan et al., 2023; Shinn et al., 2023). Motivated by these insights, we introduce a strategy-aware interaction mechanism that explicitly distinguishes between **Self-refinement** and **Counter-questioning** during multi-agent collaboration. Specifically, when disagreement signals are detected, SAMAC adaptively selects an appropriate interaction strategy from Self-refinement or

Counter-questioning. The strategy is chosen from a predefined set based on the nature of the disagreement and the availability of evidence. The supported strategies include:

- **Self-refinement**: Applied if the Agent can improve its own prior opinion using the evidence provided by other experts; Then state which part of its reasoning is being revised. After the current round of interaction concludes, the agent revises its assumptions or inference chain before updating its conclusion.

- **Counter-questioning**: Applied if the Agent can't resolve the disagreement due to a key missing evidence/clarification from the target expert; Then ask one concise, targeted question. After the current round of interaction concludes, the questioned expert provides an answer to the targeted question based on their own domain-specific expertise.

After each interaction round, expert opinions and confidence levels are updated, and disagreement signals are recomputed. The process iterates until consensus is reached or a maximum of three refinement rounds is reached.

Through this iterative consensus purification process, weakly supported arguments are progressively revised or discarded, hidden reasoning inconsistencies are exposed, and experts converge toward evidence-aligned conclusions. In the context of MDT-style medical reasoning, this mechanism reduces superficial disagreement, mitigates premature consensus, and promotes principled multi-agent negotiation.

4 Experiments and Results

4.1 Datasets

To comprehensively evaluate our framework under multi-modal clinical scenarios, we conducted experiments on three publicly available datasets: RADCURE (Welch et al., 2024), HANCOCK (Dörrich et al., 2025), and OSCCPI (Guan et al., 2025). For RADCURE, we use 750 cases from the official test split. Each case includes contrast-enhanced CT images along with demographic, clinical, and treatment information. We formulate the task as survival status prediction in a radiotherapy setting, requiring multi-modal reasoning over imaging and structured clinical variables. From the HANCOCK cohort, we select all 298 cases of oropharyngeal cancer. Each case provides hematoxylin–eosin (HE) stained whole-slide images, immunohistochemistry (IHC) images, structured clinical data, pathological reports, and tissue microarray measurements. Similarly to RADCURE, we evaluate survival status prediction in a multimodal oncology scenario. For OSCCPI, we used 200 cases from the OSCCPI test split. Each case contains high-resolution histopathology images covering both tumor core and edge regions. We formulate the task as recurrence prediction, enabling pathological reasoning evaluation.

4.2 Metrics

We evaluate the performance of our framework using a combination of standard classification metrics and collaboration-aware improvement metrics.

Classification Metrics. We report the following standard metrics: Accuracy (Acc), Specificity (Spec), Precision (Prec), and Recall (Rec). Also, we calculate the Average Score (Avg) for a clear presentation:

$$\text{Avg} = \frac{\text{Acc} + \text{Prec} + \text{Spec} + \text{Recall}}{4}.$$

Collaboration-aware Correction Metrics. To quantify the effect of multi-agent discussion, we adopt and extend the framework of Yang et al. (2025) and report Confidence Level (CL), Critique Score (CS), and Relative Self-Correction Score (RSS). For each case i, let $a_i, b_i \in \{0, 1\}$ indicate whether the prediction is correct before and after collaboration. We define

$$\text{Acc}_{\text{pre}} = \frac{1}{N}\sum_i a_i, \quad \text{Acc}_{\text{post}} = \frac{1}{N}\sum_i b_i.$$

CL measures the proportion of initially correct predictions preserved after collaboration, while CS measures the proportion of initially incorrect predictions corrected after collaboration:

$$\text{CL} = P(b = 1 \mid a = 1),$$
$$\text{CS} = P(b = 1 \mid a = 0).$$

Accordingly, the post-collaboration accuracy can be decomposed as

$$\text{Acc}_{\text{post}} = \text{CL} \cdot \text{Acc}_{\text{pre}} + \text{CS} \cdot (1 - \text{Acc}_{\text{pre}}).$$

Following Yang et al. (2025), RSS further normalizes the collaborative gain by the available correction space:

$$\text{RSS} = \frac{\text{Acc}_{\text{post}} - \text{Acc}_{\text{pre}}^2}{2\text{Acc}_{\text{pre}}(1 - \text{Acc}_{\text{pre}})}.$$

An RSS value of 0.5 corresponds to no net accuracy change after collaboration; $\mathrm{RSS} > 0.5$ indicates positive collaborative gain, while $\mathrm{RSS} < 0.5$ indicates performance degradation.

4.3 Image Analysis before Collaboration

Since the focus of this work is the collaboration protocol rather than medical image representation learning, we use a fixed image-to-text preprocessing pipeline to convert visual evidence into textual clinical findings. We first converted image information into textual evidence using GPT-4.1-mini. For each case, GPT-4.1-mini analyzed the associated medical images together with the corresponding bounding-box annotations, focusing on the regions of interest indicated by the bbox. The model extracted clinically relevant visual findings from both the global image context and the localized bbox regions. These image-derived findings were concatenated with the patient's structured clinical indicators to construct the final input question, which was then provided to the recruited expert agents for collaborative reasoning, confidence estimation, evidence extraction, and final decision-making. This design keeps the visual evidence extraction stage identical across all compared methods, allowing the evaluation to isolate the effect of the multi-agent collaboration mechanism.

4.4 Results

Comparison with Different LLM-based Methods. Table 1 presents the comparative results on three datasets. For fair comparison, the initial number of experts in the Majority Voting baseline is set to 3, consistent with KAMAC-3 and SAMAC-3. All experiments are conducted using GPT-4.1-mini (OpenAI, 2023) as the LLM backbone to ensure consistency across methods.

Overall, SAMAC-3 achieves the best Average Score on all three datasets, reaching 82.11 on RADCURE, 81.54 on HANCOCK, and 72.78 on OSCCPI. Compared with KAMAC-3, SAMAC-3 improves the Average Score by 4.29, 2.40, and 4.17 points, respectively. These consistent gains demonstrate the effectiveness and robustness of the proposed strategy-aware interaction mechanism across different clinical datasets. The improvements are particularly notable on RADCURE and OSCCPI, suggesting that SAMAC-3 is especially beneficial in clinically complex and heterogeneous scenarios where agents may need more targeted discussion to resolve disagreements.

In terms of individual metrics, although some baselines such as MDAgent and KAMAC-3 obtain slightly higher Recall on certain datasets, their relatively low Specificity leads to inferior overall performance. In contrast, SAMAC-3 substantially improves Specificity, especially on RADCURE, HANCOCK, and OSCCPI, indicating that it better balances sensitivity and specificity rather than simply favoring positive predictions.

Ablation Study. We conduct ablation experiments on RADCURE to evaluate the contribution of each component in SAMAC, as shown in Table 2. Starting from the KAMAC-3 baseline, enabling either opinion disagreement detection (OD) or majority-not-confident detection (MNC) improves both Average Score and RSS, indicating that answer-level disagreement and insufficient collective confidence are complementary signals for deciding when further deliberation is needed. Combining OD and MNC as Structured Deliberation (SD) further increases the discussion rate and improves the Average Score to 79.80, but the RSS does not consistently increase, suggesting that triggering more discussions alone is not sufficient.

Introducing resolution strategies further improves the effectiveness of collaboration. Self-refinement (SR) helps agents revise their own reasoning based on peer evidence, while Counter-questioning (CQ) encourages agents to request missing clarifications before updating their opinions. The full SAMAC integrating OD, MNC, SR, and CQ achieves the best overall performance, improving Avg from 77.82 to 82.11 and RSS from 0.480 to 0.577 over the baseline. These results indicate that SAMAC benefits from two complementary forms of interaction control: OD and MNC determine when discussion should be triggered, while SR and CQ determine how the triggered discussion should proceed. Thus the improvement of SAMAC is not merely due to more frequent discussion, but to the coordinated control of discussion triggering and disagreement resolution.

Effect of Initial Agent Number. Table 3 reports the impact of varying the initial number of participating agents ($N_{init} = 1, 3, 5$) on RADCURE dataset, examining its influence on expert recruitment (Rec.), discussion frequency (Dis.), and performance metrics (Avg and RSS). SAMAC consistently outperforms KAMAC on both Avg and RSS in all settings, demonstrating that our Structured Deliberation and Consensus Purification remain ef-

Datasets	Methods	Acc	Prec	Spec	Recall	Avg	RSS
RADCURE	SA_{CoT} (Wei et al., 2022)	86.80	88.82	33.62	96.53	76.44	–
	Majority voting (Wang et al., 2022)	88.40	89.93	40.52	97.16	79.00	–
	MDAgent (Kim et al., 2024a)	87.73	88.94	33.62	**97.63**	76.98	–
	KAMAC-3 (Wu et al., 2025)	88.13	89.32	36.21	**97.63**	77.82	0.480
	SAMAC-3 (Ours)	**89.87**	**91.39**	**50.00**	97.16	**82.11** (+4.29)	**0.577**
HANCOCK	SA_{CoT} (Wei et al., 2022)	82.71	86.87	35.85	92.98	74.60	–
	Majority voting (Wang et al., 2022)	82.83	87.26	37.74	92.62	75.11	–
	MDAgent (Kim et al., 2024a)	82.21	82.43	1.89	**99.59**	66.53	–
	KAMAC-3 (Wu et al., 2025)	88.93	88.41	39.62	**99.59**	79.14	0.482
	SAMAC-3 (Ours)	**89.26**	**89.89**	**49.06**	97.96	**81.54** (+2.4)	**0.549**
OSCCPI	SA_{CoT} (Wei et al., 2022)	78.28	78.42	12.77	**98.68**	67.04	–
	Majority voting (Wang et al., 2022)	78.00	79.14	17.02	96.73	67.72	–
	MDAgent (Kim et al., 2024a)	78.00	77.44	10.2	100	66.41	–
	KAMAC-3 (Wu et al., 2025)	77.50	79.12	22.45	95.36	68.61	0.439
	SAMAC-3 (Ours)	**80.00**	**81.71**	**34.69**	94.70	**72.78** (+4.17)	**0.530**

Table 1: Comparative experiments of different methods on three datasets. SA_{CoT} represents single-agent chain-of-thought reasoning; Majority voting aggregates multiple independent predictions by selecting the most frequent output; MDAgent uses multi-agent discussion to refine decisions; KAMAC-3 and SAMAC-3 both adopt an initial configuration of three agents, where KAMAC-3 follows a knowledge-aware collaboration framework and SAMAC-3 denotes our proposed self-adaptive multi-agent collaboration method. Results are reported on RADCURE, HANCOCK, and OSCCPI, with the best values highlighted in bold.

Methods	OD	MNC	SR	CQ	Acc	Prec	Spec	Recall	Avg	RSS
baseline	✗	✗	✗	✗	88.13	89.32	36.21	97.63	77.82	0.480
+OD	✓	✗	✗	✗	88.27	90.38	43.97	96.37	79.75	0.536
+MNC	✗	✓	✗	✗	88.27	90.27	43.10	96.53	79.54	0.536
+SD	✓	✓	✗	✗	88.00	90.48	44.83	95.90	79.80	0.524
SD+SR	✓	✓	✓	✗	88.40	90.76	46.55	96.06	80.44	0.543
SD+CQ	✓	✓	✗	✓	88.53	90.29	43.1	96.85	79.69	0.549
ours	✓	✓	✓	✓	**89.87**	**91.39**	**50.00**	**97.16**	**82.11**	**0.577**

Table 2: Ablation study of the key components in the proposed SAMAC framework on RADCURE. OD, MNC, SR, and CQ denote opinion disagreement detection, majority-not-confident detection, self-refinement, and counter-questioning, respectively. Starting from the baseline, different combinations of components are progressively enabled to evaluate their individual and joint contributions. The full model, which integrates all four components, achieves the best overall performance, obtaining the highest Average Score, and RSS.

fective across varying collaboration scales. When increasing N_{init} from 1 to 3, SAMAC achieves its best performance. However, increasing the number of agents does not yield further gains, as additional opinions tend to be redundant, adding interaction complexity without improving correction quality. Interestingly, while KAMAC shows performance fluctuations and decreasing RSS as N_{init} grows, SAMAC maintains stable RSS in all settings. This suggests that structured and strategy-aware interaction mitigates the instability typically introduced by larger multi-agent groups.

Multi-run comparison analysis. The multi-run results in Table 4 further confirm the robustness of SAMAC-3 over KAMAC-3. Across all three datasets, SAMAC-3 consistently achieves higher Avg and RSS scores. Specifically, SAMAC-3 improves Avg by +4.60, +1.97, and +4.22 on RADCURE, HANCOCK, and OSCCPI, respectively, while the corresponding RSS improvements are +0.103, +0.050, and +0.091. These consistent gains indicate that SAMAC-3 provides more stable performance improvements across repeated independent runs. The results suggest that our structured deliberation and consensus purification mechanisms can enhance collaborative correction capability across different datasets. Overall, SAMAC-3 improves both predictive performance and collaborative reliability compared with KAMAC-3.

Method	Rec.	Dis.	Avg	RSS
KAMAC-1	67	0	76.39	0.512
SAMAC-1	141	104	78.18	0.570
KAMAC-3	36	27	77.82	0.480
SAMAC-3	3	153	**82.11**	**0.577**
KAMAC-5	24	27	78.39	0.448
SAMAC-5	36	261	80.03	0.500

Table 3: Effect of different initial agent numbers on RADCURE. KAMAC-n and SAMAC-n start with n initial agents. Rec. denotes the number of cases triggering additional expert recruitment, and Dis. denotes the number of cases entering the discussion process.

Dataset	Metric	KAMAC-3	SAMAC-3
RADCURE	Avg	$77.34_{\pm 0.49}$	$\mathbf{81.94}_{\pm 0.15}$
	RSS	$0.462_{\pm 0.016}$	$\mathbf{0.565}_{\pm 0.010}$
HANCOCK	Avg	$79.45_{\pm 0.45}$	$\mathbf{81.42}_{\pm 0.17}$
	RSS	$0.491_{\pm 0.012}$	$\mathbf{0.541}_{\pm 0.010}$
OSCCPI	Avg	$68.24_{\pm 0.52}$	$\mathbf{72.46}_{\pm 0.45}$
	RSS	$0.439_{\pm 0.001}$	$\mathbf{0.530}_{\pm 0.001}$

Table 4: Multi-run comparison of SAMAC-3 and KAMAC-3 across different datasets and metrics. Results are reported as mean$_{\pm}$standard deviation over repeated independent runs.

Backbone Robustness Analysis. To examine whether the effectiveness of SAMAC depends on a specific LLM backbone, we conduct additional experiments with DeepSeek and compare the results with GPT-4.1-mini. As shown in Table 5, SAMAC-3 consistently improves over KAMAC-3 on all three datasets with both GPT-4.1-mini and DeepSeek backbones. Although the absolute scores differ between the two LLMs, SAMAC-3 maintains higher average performance and generally better RSS scores. These results indicate that the effectiveness of SAMAC is not limited to a specific backbone and demonstrates good robustness across different LLMs.

Consensus Analysis under Initial Disagreement. We further analyze cases where the initially recruited agents produce conflicting opinions, and evaluate whether the subsequent multi-agent process can resolve these disagreements into a correct consensus. For a fair comparison, SAMAC and KAMAC are initialized with the same agent opinions and then execute their respective reasoning pipelines. As shown in Fig. 3, KAMAC leaves many cases unresolved, especially with five initial agents, where 79.5% of cases end with no consensus and only 15.4% reach a correct consensus. In contrast, SAMAC resolves disagreements more effectively, achieving 50.0% correct consensus and reducing no-consensus cases to 21.8% in the five-agent setting. Although SAMAC produces more incorrect consensus in this setting, KAMAC mainly avoids errors by failing to make a decision in most cases. Overall, these results show that structured and strategy-aware collaboration helps agents reach more decisive and accurate collective decisions under initial disagreement.

Dataset	Metric	KAMAC-3		SAMAC-3	
		GPT	DS	GPT	DS
RADCURE	Avg	77.82	77.84	82.11	81.39
	RSS	0.480	0.445	0.577	0.557
HANCOCK	Avg	79.14	76.11	81.54	78.64
	RSS	0.482	0.479	0.549	0.521
OSCCPI	Avg	68.61	68.44	72.78	70.10
	RSS	0.439	0.500	0.530	0.529

Table 5: Backbone robustness analysis comparing KAMAC-3 and SAMAC-3 with GPT-4.1-mini and DeepSeek (DS) across different datasets and metrics.

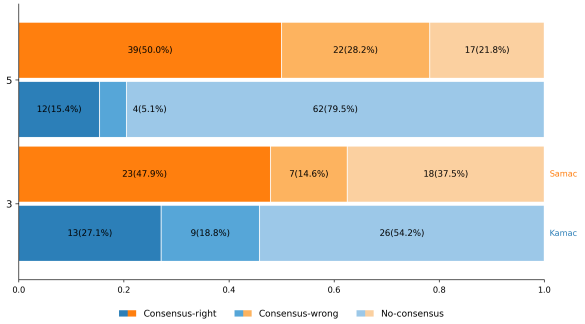


Figure 3: Consensus analysis under initial disagreement. The results are computed on cases where the initially recruited agents produce conflicting opinions. For a fair comparison, SAMAC and KAMAC start from the same set of initial agent opinions and then execute their complete reasoning pipelines. The bars show the final consensus status after the full process, including correct consensus, incorrect consensus, and no consensus, under different initial agent numbers.

5 Conclusion

We propose SAMAC, a structured framework for adaptive MDT-style medical reasoning. By explicitly detecting disagreement and adaptively applying Self-refinement and Counter-questioning, SAMAC transforms disagreement into constructive collaborative correction rather than passive aggregation. Experiments on three medical benchmarks show

that SAMAC improves average predictive performance and collaboration-aware correction metrics over competitive multi-agent baselines. The results suggest that controlling the deliberation process can benefit MDT-style reasoning workflows.

6 Limitations

Although SAMAC shows consistent improvements across three medical benchmark datasets, several aspects remain to be further explored. First, SAMAC uses agent-reported opinions and confidence signals to decide when structured discussion should be triggered. This provides an explicit and interpretable mechanism for identifying disagreement or insufficient confidence. However, confidence estimation in LLMs remains an active research problem. More calibrated confidence modeling or external verification modules may further improve the reliability of discussion triggering and strategy selection. Second, our experiments are conducted on three oncology-related datasets, including survival prediction and recurrence prediction tasks. These datasets provide representative multimodal clinical scenarios, but medical decision-making covers a broader range of tasks, such as differential diagnosis, treatment recommendation, and longitudinal patient management. Exploring its broader applicability to additional clinical workflows remains an interesting direction for future work. Finally, SAMAC is intended to support structured MDT-style reasoning rather than replace clinician judgment. In high-stakes clinical settings, more decisive consensus is not always preferable to abstention, especially when agents converge to an incorrect conclusion. Therefore, practical deployment would require clinician-in-the-loop review, external evidence verification, confidence calibration, and explicit abstention criteria for high-disagreement or low-confidence cases.

7 Ethical Considerations

SAMAC is intended to support, not replace, qualified clinicians. Potential benefits include making disagreements explicit, eliciting missing evidence, and documenting why an agent revised its opinion. Risks include hallucinated or unsupported reasoning, false reassurance from confi- dent consensus, demographic or clinical biases in- herited from foundation models and datasets, pri- vacy leakage in real deployments, and increased computational and environmental cost. All ex- periments use existing de-identified datasets under their respective data-use terms; this work collected no new patient data. The reported benchmark gains do not es- tablish clinical validity, safety, fairness, or regula- tory compliance

References

Shona Marie Bates, Jialing Lin, Luke N Allen, Michael Wright, and Michael Kidd. 2025. Can multidisciplinary teams improve the quality of primary care? a scoping review. *EClinicalMedicine*, 88.

James Boast and 1 others. 2018. Making more of multimorbidity: an emerging priority. *Lancet*, 391(10131):1637.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Benjamin Djulbegovic and Shira Elqayam. 2017. Many faces of rationality: implications of the great rationality debate for clinical decision-making. *Journal of Evaluation in Clinical Practice*, 23(5):915–922.

Marion Dörrich, Matthias Balk, Tatjana Heusinger, Sandra Beyer, Hamed Mirbagheri, David J Fischer, Hassan Kanso, Christian Matek, Arndt Hartmann, Heinrich Iro, and 1 others. 2025. A multimodal dataset for precision oncology in head and neck cancer. *Nature Communications*, 16(1):7163.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Glyn Elwyn, Dominick Frosch, Richard Thomson, Natalie Joseph-Williams, Amy Lloyd, Paul Kinnersley, Emma Cording, Dave Tomson, Carole Dodd, Stephen Rollnick, and 1 others. 2012. Shared decision making: a model for clinical practice. *Journal of general internal medicine*, 27(10):1361–1367.

Dayuan Fu, Keqing He, Yejie Wang, Wentao Hong, Zhuoma Gongque, Weihao Zeng, Wei Wang, Jingang Wang, Xunliang Cai, and Weiran Xu. 2025. Agentrefine: Enhancing agent generalization through refinement tuning. *arXiv preprint arXiv:2501.01702*.

Jinquan Guan, Junhong Guo, Qi Chen, Jian Chen, Yongkang Cai, Yilin He, Zhiquan Huang, Yan Wang, and Yutong Xie. 2025. A high magnifications histopathology image dataset for oral squamous cell carcinoma diagnosis and prognosis. *Preprint*, arXiv:2507.16360.

Mathijs P Hendriks, Agnes Jager, Kees CWJ Ebben, Janine A van Til, and Sabine Siesling. 2024. Clinical decision support systems for multidisciplinary team decision-making in patients with solid cancer: composition of an implementation model based on a scoping review. *Critical Reviews in Oncology/Hematology*, 195:104267.

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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671.

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

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024b. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

M Taberna, F Gil Moncayo, E Jané-Salas, M Antonio, L Arribas, E Vilajosana, E Peralvez Torres, and R Mesía. 2020. The multidisciplinary team (mdt) approach and quality of care. front oncol. 2020; 10: 85.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning (2023). *arXiv preprint arXiv:2311.10537*.

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

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

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.

Mattea L Welch, Sejin Kim, Andrew J Hope, Shao Hui Huang, Zhibin Lu, Joseph Marsilla, Michal Kazmierski, Katrina Rey-McIntyre, Tirth Patel, Brian O'Sullivan, and 1 others. 2024. Radcure: An open-source head and neck cancer ct dataset for clinical radiation therapy insights. *Medical Physics*, 51(4):3101–3109.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. 2025. A knowledge-driven adaptive collaboration of llms for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33483–33500.

Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan

Zhao, Yixin Zhang, Renjun Zhang, and 1 others. 2024. Clinicallab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890*.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *arXiv preprint arXiv:2412.14737*.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2025. Confidence vs critique: A decomposition of self-correction capability for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3998–4014.

A Confidence Calibration And Analysis

Following recent work on verbalized confidence scores for LLMs (Yang et al., 2024), we evaluate whether the confidence signal used by SAMAC is empirically informative. We test three verbalized-confidence designs on RADCURE, group predictions by their reported confidence level, and compute the empirical accuracy of each group and the expected calibration error (ECE). For designs that output only a discrete confidence level, we map `Confident`, `Neutral`, and `Not_Confident` to representative scores of 0.85, 0.55, and 0.20, respectively, solely for computing ECE. For the level-score design, ECE is computed from the reported continuous scores. The available analysis provides stable group-accuracy estimates for Confident and Neutral predictions, which we report in Table 6; it does not provide a stable NotConfident subgroup estimate.

Confidence design	Confident	Neutral	ECE
Label-only	92.46%	70.92%	0.0918
Range-guided	89.91%	79.11%	0.1072
Level-score	89.73%	79.08%	0.0928

Table 6: Confidence calibration on RADCURE under three verbalized-confidence designs. Confident and Neutral columns report empirical group accuracy.

Label-only confidence asks the model to choose one of `Confident`, `Neutral`, and `Not_Confident`, accompanied by a short semantic description of each label. Range-guided confidence supplements the same labels with approximate ranges: 0.70–1.00 for `Confident`, 0.40–0.70 for `Neutral`, and 0.00–0.40 for `Not_Confident`. These ranges are prompting anchors rather than probabilities derived from model logits. Level-score confidence asks the model to return both a discrete confidence level and a continuous score between 0 and 1.

Across all three designs, predictions labeled Confident are more accurate than those labeled Neutral. Under label-only confidence, for example, the empirical accuracy is 92.46% for the Confident group and 70.92% for the Neutral group. The same ordering holds for the range-guided and level-score designs. Thus, although verbalized confidence is not a calibrated posterior probability, it contains meaningful information about answer reliability rather than behaving as an arbitrary label.

At the same time, ECE ranges from 0.0918 to 0.1072, indicating only moderate calibration. Therefore, SAMAC does not interpret confidence as an exact probability of correctness. Instead, it uses confidence as a relative uncertainty signal for collaboration routing and combines it with explicit answer disagreement. Confidence alone cannot determine the final answer, and the joint trigger reduces the risk that a noisy confidence report dominates the routing decision.

We further evaluate whether the three confidence designs change end-task performance. Table 7 shows that the variants remain broadly comparable: Accuracy spans 88.67–89.87, Average Score spans 82.11–84.43, and RSS spans 0.549–0.577.

Confidence design	Acc	Prec	Spec	Recall	Avg	RSS
Label-only	89.87	91.39	50.00	97.16	82.11	0.577
Range-guided	88.67	92.30	56.90	94.48	83.08	0.549
Level-score	89.20	93.00	61.21	94.32	84.43	0.557

Table 7: Final RADCURE performance under three verbalized-confidence designs.

B Efficiency and Cost

Method	Avg	Cost/case (USD)	Calls/case
SA_{CoT}	76.44	0.0018	2.00
Majority Voting	79.00	0.0079	6.001
MDAgent	76.98	0.0103	10.10
KAMAC-3	77.82	0.0142	11.69
SAMAC-3	82.11	0.0218	13.17

Table 8: RADCURE Average Score and provider-level per-case efficiency statistics.

Table 8 reports Average Score together with measured cost and API calls per case. Resource use generally rises with the amount of collaboration: SA_{CoT} is the least expensive method at $0.0018 and 2.00 calls per case, while SAMAC-3 has the highest cost and call count at $0.0218 and 13.17. The relationship between resource use and performance is not monotonic, however.

Relative to SA_{CoT}, KAMAC-3 uses 7.89× the cost and 5.85× the API calls, reflecting the substantial base cost of multi-agent collaboration. Relative to KAMAC-3, SAMAC-3 uses 1.54× the cost and 1.13× the calls while improving Avg by 4.29 points. Relative to Majority Voting, it uses 2.76× the cost and 2.19× the calls for a 3.11-point Avg gain. These comparisons expose a performance–efficiency trade-off rather than establishing unconditional dominance. The incremental SAMAC overhead is associated with strategy selection and the ensuing structured interaction; confidence and disagreement replace, rather than duplicate, the earlier trigger. Finally, the monetary values depend on the experimental provider, model pricing, prompts, and measurement period, so they should be treated as experiment-specific estimates rather than stable deployment costs.

C Complementary MedQA Evaluation

We additionally evaluate the English MedQA test setting (Jin et al., 2021). Unlike the primary oncology datasets, MedQA is text-only and covers broader medical knowledge. It is reported as complementary post-submission analysis, not as evidence for multimodal generality.

Method	Acc	Prec	Spec	Recall	Avg	RSS
SA_{CoT}	84.76	84.91	96.18	84.69	87.63	–
Majority Voting	84.52	85.53	96.32	85.10	87.87	–
MDAgent	84.59	84.48	96.15	84.49	87.43	–
KAMAC-3	85.31	85.42	96.32	85.07	88.03	0.530
SAMAC-3	**86.72**	**86.86**	**96.67**	**86.56**	**89.20**	**0.584**

Table 9: Complementary results on English MedQA.

Table 9 shows that SAMAC-3 attains the highest value on every reported predictive metric. Relative to KAMAC-3, it improves Accuracy by 1.41 points, Precision by 1.44 points, Specificity by 0.35 points, Recall by 1.49 points, and Avg by 1.17 points. It also raises the collaboration-aware RSS from 0.530 to 0.584, a gain of 0.054. Compared with Majority Voting, the strongest baseline for which RSS is not defined, SAMAC-3 improves Avg by 1.33 points. The consistent direction across Accuracy, Precision, Specificity, and Recall indicates that the aggregate improvement is not produced by a gain in only one component metric. These descriptive results suggest that the structured interaction mechanism can remain useful when the input is text-only and the questions cover broader medical knowledge.

D Heterogeneous-Model Analysis

To probe same-backbone role diversity, we compare GPT-only teams (GPT-4.1-mini) with heterogeneous teams combining GPT-4.1-mini, Qwen-3.7-Pro, and GLM-4.7.

Dataset	Composition	Acc	Prec	Spec	Recall	Avg	RSS
RADCURE	GPT-only	89.87	91.39	50.00	97.16	82.11	0.577
	Heterogeneous	85.45	94.11	69.83	88.31	84.42	0.521
HANCOCK	GPT-only	89.26	89.89	49.06	97.96	81.54	0.549
	Heterogeneous	68.58	93.14	76.92	66.80	76.36	0.530
OSCCPI	GPT-only	80.00	81.71	34.69	94.70	72.78	0.530
	Heterogeneous	73.85	85.40	57.45	79.05	73.94	0.513

Table 10: Verified heterogeneous-model results. OSCCPI results are subset-only and should not be compared with the full-set main table.

Table 10 shows that heterogeneous composition indeed changes multi-agent behavior, mainly through Precision and Specificity gains. Precision rises by 2.72–3.69 points and Specificity by 19.83–27.86 points, while Accuracy and Recall fall. Avg rises on RADCURE (+2.31) and the OSCCPI subset (+1.16), but drops on HANCOCK (−5.18) as Recall falls by 31.16 points. This suggests heterogeneous agents provide complementary tendencies and may mitigate shared same-model bias.

Lower Recall signals conservative judgments: fewer false positives but more missed positives. Thus heterogeneity is not uniformly better; clinical value depends on error costs.

E Prompt Design

Self-refine: Use this if your own reasoning may contain a mistake, missing factor, or weak interpretation, and you can revise it using the available evidence or another expert's evidence without needing further clarification. State what to reconsider.
Counter-questioning: Use this if one specific clarification from `{target_agent}` could help resolve the disagreement. Ask one concise question to that agent.