MDTeamGPT: Mitigating Context Collapse and Enabling Self-Evolution in Medical Multi-Agent Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown great potential in multi-disciplinary team (MDT) medical consultations. However, long, multi-round, multi-role interaction trajectories inevitably lead to severe information dilution and context window overload, triggering *context collapse* which destabilizes reasoning. Furthermore, prior systems typically rely on unstructured trajectory history storage without structurally distilling key information or reflecting on errors, severely limiting continuous learning capabilities. We propose **MDTeamGPT**, a context-resilient and self-evolving multi-agent framework. Mechanistically, we introduce a specialized *Lead Physician* mechanism combined with a *Residual Context* architecture to compress and reorganize multi-round consensus, effectively mitigating context overload and reducing computational costs. For memory, we design a Dual Knowledge Base system comprising a CorrectKB for verified trajectories and a ChainKB for reflective error analysis, enabling self-evolution via retrieval from both successes and failures. We evaluated our framework on standard text datasets (MedQA, PubMedQA), multimodal benchmarks (VQA-RAD, SLAKE), and collected more complex clinical problems. Experimental results show that MDTeamGPT substantially outperforms existing baselines across both text-based and multimodal tasks, while also demonstrating superior diagnostic performance and stability in complex clinical scenarios. Our code is available at anonymous link.

1 Introduction

When diagnosing and treating complex diseases, a single-specialty medical perspective frequently proves insufficient to address patients' comprehensive and precise healthcare needs, thereby necessitating the implementation of a Multi-Disciplinary Team (MDT) and diversified clinical perspectives (Macken et al., 2022; Zhou and Xu, 2023).

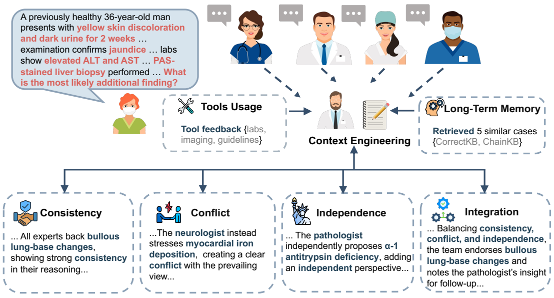

Figure 1: The *Lead Physician* consolidates agent outputs into four dialectical categories: consistency, conflict, independence, and integration, with supporting references (tools usage and long-term memory), to ensure a comprehensive and structured diagnostic context.

MDT facilitates the development of more accurate and holistic therapeutic strategies through systematic integration of expertise across medical disciplines (Makary, 2011). Nevertheless, the organization of interdisciplinary consultations demands substantial temporal and human resources, while remaining vulnerable to procedural oversights, cognitive biases, and communication inefficiencies (Burton et al., 2022). Such systemic challenges may ultimately compromise healthcare quality and adversely affect patient outcomes.

Large Language Models (LLMs), leveraging their massive parameter scales and extensive training on vast knowledge bases across diverse domains, have demonstrated excellent reasoning abilities and zero-shot generalization capabilities (Achiam et al., 2023; Touvron et al., 2023; Sun et al., 2025). These capabilities have prompted researchers to explore deploying autonomous LLM agents to enhance Multi-Disciplinary Team (MDT) diagnostics (Kim et al., 2024; Chen et al., 2024b; Wang et al., 2025), aiming to improve the efficiency and rigor of clinical consultations without relying on resource-intensive human coordination.

Constrained by strict privacy limitations and the scarcity of standardized high-quality medical data, supervised fine-tuning (SFT) and post-training of medical LLMs remain challenging. Consequently, Multi-Agent frameworks have emerged as a promising alternative (Liu et al., 2023; Du et al., 2023), which elicit latent medical knowledge and mitigate individual model hallucinations through collaborative debate. Existing works have explored role configuration (Tang et al., 2023; Long et al., 2024), subtask decomposition (Pandey et al., 2024), and clinical simulation (Schmidgall et al., 2024) to optimize these collaborative processes.

Despite these advancements, current multi-agent MDT systems still face two critical bottlenecks. First is *context collapse* in long-horizon reasoning. As trajectory tokens rapidly accumulate due to multi-expert interactions, the system is often forced to compress historical context for subsequent turns. This process frequently leads to coarse summarization, causing the sudden loss of fine-grained details and critical patient information, thereby destabilizing reasoning efficiency and accuracy (Maharana et al., 2024; Zhang et al., 2025). Second is inefficient experience evolution. Although some systems attempt to accumulate medical experience (Wei et al., 2024; Li et al., 2024c; Zhu et al., 2025), they often merely archive raw diagnostic process histories. Due to the lack of structured distillation of successful reasoning trajectories or deep reflection on erroneous cases, these systems cannot convert data into transferable strategies, limiting their ability to self-evolve and generalize to complex, unseen scenarios (Li et al., 2024b; Sarthi et al., 2024).

In this paper, we propose MDTeamGPT, a self-evolving LLM-based multi-agent framework for multi-disciplinary team (MDT) medical consultation. The system includes multiple doctor roles and auxiliary roles. A “patient” role enters the system with specific background information and medical questions, and the *Primary Care Doctor* assigns the case to the most appropriate specialist agents based on the patient’s condition. The specialist agents then conduct multi-round case discussions and share contextual information. After each round, the *Lead Physician* performs agentic context engineering and structures the round’s information into **Consistency**, **Conflict**, **Independence**, **Integration**, as well as Tools Usage and Long-Term Memory, as illustrated in Figure 1. In subsequent rounds, agents follow a residual discussion scheme and only reference the compressed information stored in the *Residual Context*. Once consensus is reached after multiple rounds of deliberation, the final consultation outcome is reviewed by a *Reflector*, who checks for safety and consistency and then produces the final conclusions and recommendations. At the same time, the system records the consultation process into different knowledge bases according to the correctness of the outcome, enabling retrieval and reuse in future cases and continuously improving overall performance. In summary, our main contributions are as follows:

- We propose an MDT-oriented multi-agent medical consultation framework that incorporates a *Lead Physician* mechanism and a *Residual Context* structure, which effectively integrates multi-round discussion outcomes, mitigates *context collapse*, and improves the accuracy of reasoning and decision making.
- The framework adopts an adaptive learning mechanism that leverages both correct and incorrect diagnostic experiences, continuously accumulating and strengthening transferable reasoning capabilities.
- Experimental results show that, with 900 rounds of accumulated consultation experience, the framework achieves leading accuracies of 90.9% on MedQA and 84.4% on PubMedQA, while also demonstrating superior performance and stability across multimodal benchmarks and self-collected complex clinical cases. Cross-dataset tests further confirm that the constructed knowledge bases preserve generalizable reasoning abilities rather than merely memorizing isolated cases.

2 Related Work

2.1 LLM-based Multi-Agent Collaboration

Well-structured multi-agent systems can significantly reduce errors and enhance interaction efficiency (Park et al., 2023; Qian et al., 2024d). ChatDev (Qian et al., 2024c) splits tasks into subtasks managed by an instructor and an assistant, using multi-round inquiries to mitigate hallucinations. MACNET (Qian et al., 2024d) adopts a directed acyclic graph with topological sorting for interactive reasoning, and MetaGPT (Hong et al., 2024) encodes SOPs for roles akin to a software company, effectively coordinating specialized expertise. However, these methods primarily focus on software development and face applicability challenges in medical consultations. MedAgents (Tang et al., 2023) assigns distinct doctor roles to each LLM-Agent and uses consensus voting, but lacks a robust strategy, risking collective hallucinations (Chen et al., 2024a). SeM-Agents (Chen et al., 2025) employs voting and correct experience accumulation but neglects context maintenance strategies against information dilution and error experience utilization. KAMAC (Wu et al., 2025) advances this by integrating medical knowledge graphs to dynamically generate adaptive roles and guide collaborative reasoning, grounding decisions in verified domain knowledge. Triageagent (Lu et al., 2024b) proposes a confidence-guided, role-specific LLM framework that enhances clinical triage accuracy through specialized role assignments and confidence assessment. AgentClinic (Schmidgall et al., 2024) introduces a multimodal benchmark that evaluates LLM agents in simulated clinical workflows with patient interactions, tools, and medical imaging. While these approaches facilitate multi-role interaction, they predominantly rely on simple aggregation mechanisms to reach consensus, and have not fully explored leveraging agentic context engineering to deeply integrate conflicting evidence or resolve professional disagreements.

2.2 Experience Utilization

Inspired by human knowledge acquisition mechanisms, equipping LLM agents with reflective memory modules can enhance their problem-solving abilities (Zhong et al., 2024). ExpeL (Zhao et al., 2024) accumulates experience from past successful cases and calls upon this knowledge during subsequent reasoning. Co-Learning (Qian et al., 2024a) focuses on collecting experience-driven heuristics from previous actions, enabling agents to handle novel tasks more flexibly. IER (Qian et al., 2024b) allows LLM agents to iteratively refine accumulated experience during task execution. HealthFlow (Zhu et al., 2025) introduces a meta-planning mechanism that enables agents to dynamically restructure workflows based on task feedback and thereby achieve autonomous policy evolution in complex biomedical research scenarios. Agent Hospital (Li et al., 2024d) leverages medical record libraries and experience databases to accumulate diagnostic data, which strengthens prompt design for medical agents and supports their continuous evolution. However, these methods still lack sufficient abstraction, summarization, and reflection on erroneous cases, which limits their ability to extract and exploit valuable insights from mistakes.

2.3 Context Engineering

Context engineering refers to improving the behavior of large language models by constructing or adjusting their input context rather than modifying model weights. For instance, recent work has shown that leveraging prompting to augment context can effectively help models embrace unknown domains (Liu et al., 2024b). A typical pattern is to let the model inspect the current context together with execution traces, reasoning steps, or validation results, then generate natural-language feedback on how the context should be rewritten, and write this feedback back into the context to form an iterative optimization loop. Existing methods such as Reflexion (Shinn et al., 2023) use failure cases to reflect and improve agent planning, GEPA (Agrawal et al., 2025) iteratively edits prompts based on execution traces, and Dynamic Cheatsheet (Suzgun et al., 2025) accumulates strategies and lessons in an external memory. While these natural-language feedback approaches offer flexibility, they primarily focus on optimizing single-agent reasoning trajectories, lacking structured context management mechanisms specifically designed to handle the complex viewpoint conflicts and consensus building inherent in multi-agent collaboration.

3 Method

This section presents an overview of the medical consultation framework. The framework specifies two role sets: specialist agents $\mathcal{S}$ (*General Internal Medicine Doctor*, *General Surgeon*, *Pediatrician*, *Obstetrician and Gynecologist*, *Radiologist*, *Neurologist*, *Pathologist*, *Pharmacist*) and auxiliary agents $\mathcal{A}$ (*Primary Care Doctor*, *Lead Physician*, *Chain-of-Thought Reviewer*, *Reflector*). By editing the role profiles, users can flexibly add or remove agents and adjust their types to match specific clinical tasks. This role design broadens the framework's applicability to diverse and complex cases. Agents in $\mathcal{A}$ do not issue diagnoses directly but join the consultation after the specialist team has been selected and provide supervision and coordination. Prompt templates for all roles are listed in Appendix A, and Figure 2 shows the framework's three-stage workflow.

3.1 Arranging Specialist Doctors

When a patient agent arrives with background C (demographics, history, exam findings, and key imaging) and query Q, the *Primary Care Doctor*

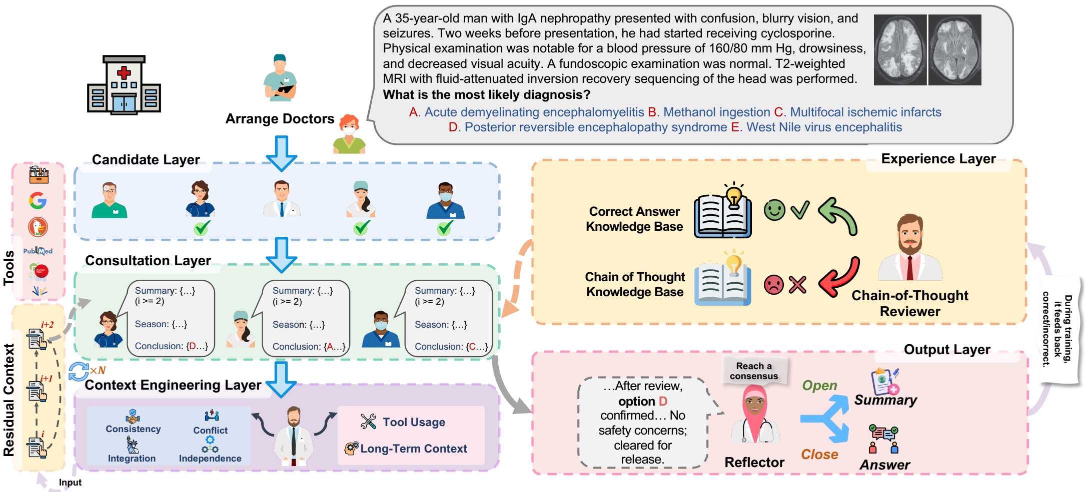

Figure 2: Overview of the MDTeamGPT framework. The **Candidate** and **Consultation Layers** organize multi-round specialist discussions ($i \in \{1, \ldots, T\}$). The **Context Engineering Layer** compresses each round's trajectory into a structured summary stored in the *Residual Context*. Starting from Round 3 ($i = 3$), specialists access a sliding window of these summaries (Round $i-2$ and $i-1$) to mitigate *context collapse*. The **Experience Layer** maintains the CorrectKB and ChainKB for long-term learning, while the **Output Layer** ensures consensus and updates the knowledge bases with verified cases.

Agent selects a specialist subset *Roles* from the candidate pool $\mathcal{S}$, corresponding to the Candidate Layer in Figure 2. Before finalizing, it outputs a natural-language justification *Reasons* for each choice, guided by a small set of curated few-shot exemplars that keep the output structured and interpretable. The formalized workflow is as follows and can be written as:

$$\textit{Reasons},\ \textit{Roles} = \mathrm{LLM}_{\mathrm{PrimaryCare}}(\mathcal{S},\, C,\, Q) \quad (1)$$

$$\textit{Roles} \subseteq \mathcal{S} \quad (2)$$

$$\textit{Consultation} = \mathrm{MDTeamGPT}(\textit{Roles},\, C,\, Q) \quad (3)$$

This design reduces consultation "contamination" by excluding irrelevant specialist agents, which cuts redundant information and potential cognitive interference. Introducing too many experts would slow the discussion and inject domain knowledge only weakly related to the patient's condition, shifting the diagnostic focus and increasing decision complexity. During framework configuration, we curate representative cases as few-shot exemplars so that the *Primary Care Doctor Agent* produces structured outputs with transparent reasoning and specialist selection.

3.2 Multi-Round Context Engineering

Once the specialist doctors have been determined, the consultation process begins. In the first round, each specialist provides a judgment and opinion based on the patient's condition, denoted as $S_{1,k}$ (the response of the k-th specialist in round 1). For closed-ended questions, the specialist outputs a final option, and for open-ended questions, a brief conclusion. At this stage, the specialist agents cannot see one another's responses.

All statements are then aggregated and sent to the *Lead Physician* (prompt details in Appendix A.3), who organizes them into six structured components: *Consistency*, *Conflict*, *Independence*, *Integration*, *Tools Usage*, and *Long-Term Memory*:

Consistency: collects parts of agents' statements that agree with each other.

Conflict: identifies contradictory points when agents disagree and remains empty when their answers fully align.

Independence: extracts each agent's unique viewpoints, i.e., information mentioned only by one agent, and remains empty when no unique views exist.

Integration: synthesizes all agents' statements into a single coherent structured summary covering all perspectives.

Tools Usage: records the external tools invoked in the current round and their key outputs.

Long-Term Memory: records key information retrieved from *CorrectKB* and *ChainKB*, together with its role in the current reasoning.

The processed result is denoted as S_1^6 and written into the *Residual Context* (R). The consultation then proceeds to the next round.

From the second round onward, each specialist agent can access the structured information from the previous round stored in R. They incorporate these summaries into their prompts to refine their responses, producing new statements $S_{2,k}$ together with an option ID and corresponding content. As in the first round, all statements are sent to the *Lead Physician*, who processes them into S_2^6.

Starting from round $i+2$ (with $i \geq 1$), the R acts as a round-wise updated context queue that exposes only the structured outputs from the two most recent rounds to the specialists in the current round, namely S_i^6 and S_{i+1}^6. In round $i+2$, each agent integrates the key information from these two summaries, further refines its prompt, articulates its stance, and provides its updated choice.

The discussion continues until all specialist agents reach consensus on the final answer. If consensus is not reached and the number of rounds has not yet hit the predefined upper bound of 15 rounds, the consultation proceeds to the next round. If the maximum number of rounds is reached without consensus, the final answer is determined by the *majority rule*. In the case of a tie, the *Reflector* makes the final decision.

This combination of *Lead Physician* and *Residual Context* structure reduces information contamination, improves discussion efficiency, and alleviates *context collapse*. By restricting access to deeper memory layers, the framework limits the degree to which agents can be overly influenced by others and thus helps to mitigate hallucinations. For a detailed description of the consultation process, please refer to Algorithm 1 in Appendix E.

3.3 Summary and Output Stage

At this stage, the determination of the final conclusion (F) relies on the question modality and consensus state. For closed-ended questions, a rule-based majority vote is employed. In the case of standard open-ended inquiries, the *Reflector* agent assesses the consistency of viewpoints. In deadlock scenarios where consensus remains unreachable after the maximum round limit, the *Reflector* intervenes by synthesizing the comprehensive historical summaries accumulated within the R. to adjudicate the definitive F. This conclusion is subsequently validated against the ground truth.

If the F is accurate, the *Chain-of-Thought Reviewer* (*CoT Reviewer*) summarizes and organizes the patient's C, medical problem (Q), and the structured statements from the final consultation round S_{final}^6. The processed record is archived in the *Correct Answer Knowledge Base (CorrectKB)*. The storage format is as follows:

CorrectKB: For correct cases, the *CoT Reviewer* stores a JSON record { "Question": $\langle\cdot\rangle$, "Answer": $\langle\cdot\rangle$, "Summary of S_{final}^6": $\langle\cdot\rangle$ }

Conversely, if the consultation outcome is incorrect, the *CoT Reviewer* generates an abstract summary of the session. This summary includes C and Q and is structured according to the initial hypothesis, analysis process, final conclusion, and reflection on error causes. The summary is then stored in the *Chain-of-Thought Knowledge Base (ChainKB)*. The storage format is as follows:

ChainKB: For incorrect cases, it records { "Question": $\langle\cdot\rangle$, "Correct Answer": $\langle\cdot\rangle$, "Initial Hypothesis": $\langle\cdot\rangle$, "Analysis Process": $\langle\cdot\rangle$, "Final Conclusion": $\langle\cdot\rangle$, "Error Reflection": $\langle\cdot\rangle$ }

When a new patient arrives, their C and Q are used to retrieve the *Top-K* most similar cases from both knowledge bases based on vector retrieval with *cosine similarity computation* ($K = 5$ in our application; see Appendix C.9 for sensitivity analysis). This retrieval process enhances the prompts (P) for the specialist agents.

To ensure independent reasoning, knowledge retrieval is deferred to the second round, primarily to resolve conflicts. In cases of immediate consensus, the knowledge bases serve as a reflective validation tool (see Algorithm 2 in Appendix E).

It is worth noting that the core model responsible for embedding textual content into the vector spaces of *CorrectKB* and *ChainKB* is OpenAI's 'text-embedding-3-small'[1].

4 Experiments

4.1 Datasets

We utilize the MedQA (Jin et al., 2021) and PubMedQA (Jin et al., 2019) datasets to validate our framework. The MedQA dataset consists of USMLE-style questions, each offering four or five possible answers, designed to assess medical knowledge and practical skills. PubMedQA, based on research paper abstracts, presents questions with Yes/No/Maybe answers, aiming to evaluate the performance of natural language processing models in academic question answering. The final results are based on the test sets of each dataset, with the *CorrectKB* and *ChainKB* containing only experiences from the training sets. All results are averaged over

[1] https://platform.openai.com/docs/guides/embeddings/embedding-models

five runs with standard deviations reported. Experimental results for multimodal tasks, along with a collection of more complex multimodal clinical cases, are available in Appendix C. Experiments in the main text exclude tool usage; please refer to Appendix D for the analysis of tool integration.

4.2 Main Results

This section evaluates the zero-shot accuracy and F1 score of MDTeamGPT for medical consultation tasks using gpt-4-turbo as the backbone model, with human evaluation included as a complementary assessment (see Appendix C.8). The experiments construct the CorrectKB and ChainKB through 900 rounds of consultations, and the overall results are reported in Table 1. We compare MDTeamGPT against standard baselines (Single-Agent, CoT, ReAct) and representative single- and multi-agent frameworks. The results indicate that MDTeamGPT consistently outperforms existing state-of-the-art methods. Compared with MDAgents, which achieves the second-best performance on MedQA, MDTeamGPT improves accuracy and F1 score by 1.2% and 2.1%, while surpassing Medprompt on PubMedQA by 2.9% and 2.4%. Relative to SeM-Agents, which is limited to voting over correct experiences and lacks an aggregation mechanism, MDTeamGPT further achieves gains of 1.7% and 2.2% on MedQA and 1.3% and 1.8% on PubMedQA in terms of accuracy and F1 score, respectively, resulting in a 1.5% advantage in average accuracy. These improvements primarily arise from the synergy of *lead physician* aggregation, *residual context*, and the dual knowledge base design based on *CorrectKB* and *ChainKB*. In addition, MDTeamGPT exhibits the lowest variance across all datasets, confirming its strong stability in complex medical consultation scenarios.

4.3 Ablation Studies

To quantify the contribution of each core component in MDTeamGPT, we conduct ablation experiments using gpt-4-turbo as the backbone model, as summarized in Table 2.

We first examine the synergy between *Residual Context* and the *Lead Physician* mechanism. Experiments 1–4 show that each component alone yields limited improvements, whereas their joint use in Experiment 4 leads to a substantial performance jump. Compared with the free discussion baseline in Experiment 1, this configuration improves accuracy by 5.8% on MedQA and 3.6% on PubMedQA. These results demonstrate that the synergistic combination of structured guidance and residual-style discussion effectively suppresses hallucinations and noise inherent in unconstrained interactions, and plays a critical role in achieving significant diagnostic performance gains.

We then investigate the impact of the dual knowledge base mechanism. Relative to the no-knowledge-base baseline in Experiment 4, introducing *CorrectKB* alone in Experiment 5 yields the most pronounced improvement, with accuracy gains of 4.6% and 6.2% on MedQA and PubMedQA, respectively, outperforming the gains from introducing *ChainKB* alone in Experiment 6, which are 1.5% and 1.0%. While *CorrectKB* provides direct guidance toward correct decisions, relying on it alone prevents the system from reflecting on erroneous reasoning paths, leaving specific historical misdiagnoses uncorrected. Consequently, combining both knowledge bases in Experiment 7 achieves the best overall performance, with an average accuracy of 87.7%, confirming that the full value of historical data can only be realized by jointly leveraging correct-experience guidance and chain-of-thought based reflective error correction.

4.4 Context behavior

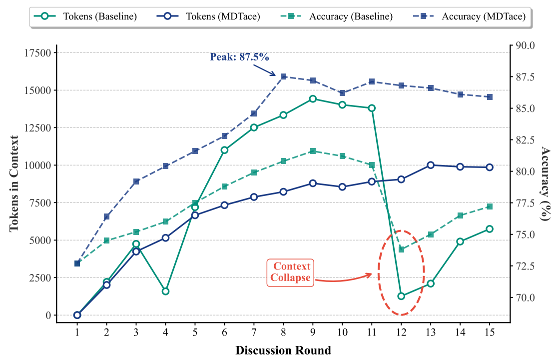

Figure 3: Performance comparison illustrating *context collapse* in long-term discussions.

This section investigates the efficacy of the *Lead Physician* and the *Residual Context* mechanism in mitigating *context collapse*. We utilized a mixed dataset comprising MedQA and PubMedQA, removed consensus-based early stopping, and enforced a free-form discussion ranging from 1 to 15 rounds, with the final determination made by the *Reflector*. As a baseline, the *Simple Summary* method passes the unprocessed full history to specialist agents, requiring self-summarization. In contrast, MDTeamGPT manages context through the joint efforts of the *Lead Physician's* structural

Method	MedQA		PubMedQA		Average
	Accuracy (%)	F1-score (%)	Accuracy (%)	F1-score (%)	Accuracy (%)
Single-Agent	77.4 ± 1.8	76.2 ± 2.1	75.3 ± 2.0	73.7 ± 1.7	76.4
Single-Agent (w/) CoT	76.6 ± 1.5	75.1 ± 1.8	76.9 ± 1.6	74.8 ± 1.4	76.8
Single-Agent (w/) ReAct	78.5 ± 1.7	76.9 ± 1.9	75.8 ± 1.4	74.1 ± 1.8	77.2
Medprompt (Nori et al., 2023)	89.6 ± 1.0	87.5 ± 0.9	81.5 ± 1.1	80.1 ± 0.8	85.6
Multi-expert Prompting (Long et al., 2024)	86.6 ± 0.7	84.9 ± 0.8	82.7 ± 1.2	81.6 ± 1.0	84.7
EoT (5-Agent-Debate) (Yin et al., 2023)	80.1 ± 1.3	79.2 ± 1.7	76.7 ± 1.7	74.6 ± 1.5	78.4
LLM Discussion (Lu et al., 2024a)	80.4 ± 1.4	78.3 ± 1.6	78.2 ± 1.5	76.4 ± 1.3	79.3
Triageagent (Lu et al., 2024b)	84.0 ± 0.9	83.2 ± 0.8	77.6 ± 1.3	75.8 ± 1.2	80.8
MedAgents (Tang et al., 2023)	83.9 ± 1.1	82.6 ± 0.9	77.2 ± 1.4	75.4 ± 1.6	80.6
MDAgents (Kim et al., 2024)	89.7 ± 1.6	86.5 ± 1.2	78.3 ± 1.3	74.6 ± 1.4	84.0
SeM-Agents (Chen et al., 2025)	89.2 ± 0.8	86.4 ± 1.0	83.1 ± 1.1	80.7 ± 0.9	86.2
MDTeamGPT (**ours**)	**90.9** ± **0.6**	**88.6** ± **0.5**	**84.4** ± **0.8**	**82.5** ± **1.1**	**87.7**

Table 1: Main results on accuracy and F1-score across MedQA and PubMedQA datasets. All results were obtained using gpt-4-turbo. We highlight the optimal and suboptimal methods in **bold** and underline, respectively.

Residual Context	Lead Physician	*CorrectKB* (900 rounds)	*ChainKB* (900 rounds)	MedQA (%)	PubMedQA (%)	Average (%)
✗	✗	✗	✗	78.5 ± 1.3	74.0 ± 1.7	76.3
✓	✗	✗	✗	76.9 ± 0.8	74.8 ± 1.4	75.9
✗	✓	✗	✗	78.2 ± 1.4	75.4 ± 1.2	76.8
✓	✓	✗	✗	84.3 ± 1.1	77.6 ± 1.3	81.0
✓	✓	✓	✗	88.9 ± 0.5	83.8 ± 0.9	86.4
✓	✓	✗	✓	85.8 ± 0.9	78.6 ± 1.2	82.3
✓	✓	✓	✓	**90.9** ± **0.6**	**84.4** ± **0.8**	**87.7**

Table 2: Ablation study evaluating the impact of different module compositions on accuracy metrics.

reorganization and the *Residual Context*.

Figure 3 shows that the context behavior of the Baseline is highly unstable. Specifically, by round 4, the token count plummeted from 4,746 to 1,584, exhibiting early signs of collapse. Although the token count recovered, reaching a peak of ∼14,423 and a maximum accuracy of 81.6% in round 9, a devastating and typical *context collapse* (highlighted by the red circle) reoccurred in round 12. Triggered by forced compression due to contextual overload, the token count plunged to 1,263, thereby causing a precipitous drop in accuracy to 73.8%. Thus, simple linear summarization is insufficient to support long-term medical reasoning.

In contrast, MDTeamGPT demonstrated exceptional stability. Benefiting from its mechanism, the system successfully maintained token consumption at approximately 10,000 without experiencing severe collapse. Accuracy rose steadily from 72.7% to a peak of 87.5% in round 8, and consistently remained above 85% in subsequent rounds. These results demonstrate that the synergy of the *Lead Physician* and the *Residual Context* effectively overcomes information forgetting and ensures the stability of long-term reasoning.

4.5 Self-Evolving Capability

To systematically evaluate the self-evolving capability of MDTeamGPT, we conducted extensive experiments using a diverse set of backbone models, including gpt-4-turbo, gpt-3.5-turbo, gpt-4o, llama3-8B, llama3-70B (Dubey et al., 2024), glm-4-plus (Zeng et al., 2024), deepseek-v3 (Liu et al., 2024a), and qwen3-32B (Yang et al., 2025). As illustrated in Figure 4(a–b), with the continuous injection of historical consultation cases, all models exhibit a consistent upward trend in diagnostic accuracy on both MedQA and PubMedQA, eventually achieving stable convergence after accumulating approximately 600 cases.

A notable observation emerges from the data analysis: models with weaker baseline capabilities (e.g., llama3-8B) demonstrate a markedly steeper improvement slope than stronger counterparts such as gpt-4o. This finding indicates that the proposed framework effectively compensates for inherent disadvantages in parameter scale, enabling smaller models to obtain substantial marginal gains through iterative experience accumulation.

Figure 4(c–d) further quantify the relative performance improvements of each model after accumulating 900 rounds of experience, compared with the Single-Agent Baseline. The radar plot results provide compelling evidence of the model-agnostic nature and universal performance gains of MDTeamGPT. Regardless of a model's static knowledge capacity or instruction-following ability, the framework consistently enhances long-horizon

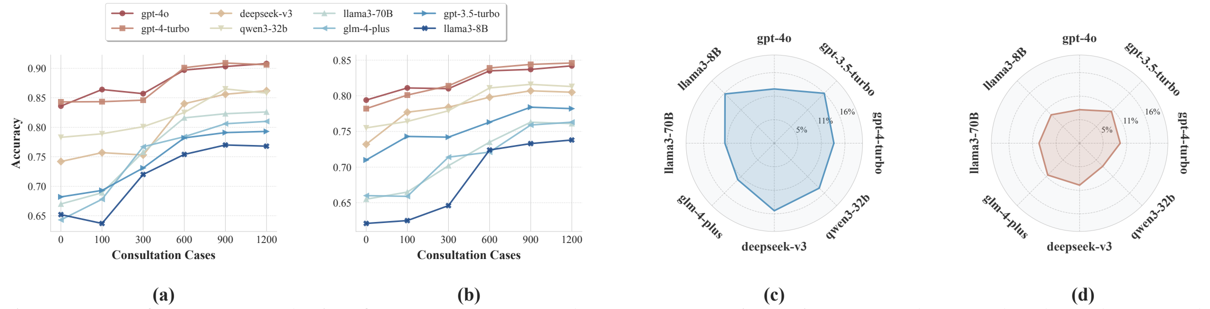

Figure 4: Performance analysis of MDTeamGPT. (a-b) Accuracy trajectories on MedQA and PubMedQA. (c-d) Relative improvements over baselines on MedQA and PubMedQA across diverse LLMs.

reasoning through dynamic context reorganization and experience reflection, leading to comprehensive improvements in diagnostic performance.

4.6 Generalization of the Knowledge Base

To evaluate the cross-domain robustness of the knowledge bases and the overall effectiveness of the framework, we conduct cross-dataset experiments. We adopt gpt-4-turbo as the backbone model and construct *CorrectKB* and *ChainKB* from 900 consultation rounds on a source dataset (e.g., MedQA), which are used to guide reasoning on a target dataset (e.g., PubMedQA).

Test Dataset	Vanilla		M-KB		P-KB	
	Acc	F1	Acc	F1	Acc	F1
MedQA (M)	84.3 ± 1.1	82.1 ± 1.0	90.9 ± 0.6	88.6 ± 0.5	88.1 ± 0.7	85.7 ± 0.9
PubMedQA (P)	77.6 ± 1.3	75.9 ± 0.9	79.8 ± 0.8	76.9 ± 1.0	84.4 ± 0.8	82.5 ± 1.1

Table 3: Cross-dataset generalization results. M-KB and P-KB denote using *CorrectKB* and *ChainKB* derived from MedQA and PubMedQA, respectively.

As shown in Table 3, when a knowledge base derived from PubMedQA is applied to MedQA testing, the accuracy improves by 3.8% compared to the baseline without a knowledge base. Conversely, leveraging a knowledge base constructed from MedQA to assist reasoning on PubMedQA yields a 2.2% accuracy gain over the corresponding baseline. These results indicate that the proposed framework goes beyond simple instance-level matching and is capable of transferring abstract cognitive strategies and reflective reasoning patterns distilled from historical experience to new contexts. By preserving these higher-order reasoning paradigms, *CorrectKB* and *ChainKB* function not merely as repositories of consultation instances, but as encapsulations of transferable reasoning capabilities, thereby enabling robust generalization across diverse scenarios.

4.7 Computational Efficiency Analysis

Table 4 demonstrates the significant computational efficiency advantages of MDTeamGPT. Notably,

Metric	Simple Voting		MDTeamGPT (ours)	
	Time (s)	Tokens	Time (s)	Tokens
Average Cost	236	15,203	**201**	**11,701**

Table 4: Runtime and token usage comparison between Simple Voting and MDTeamGPT.

compared with the Simple Voting baseline (Table 2, Experiment 1), MDTeamGPT reduces the average runtime per consultation by 15% (dropping from 236 s to 201 s) and sharply decreases token consumption by 23% (from 15,203 to 11,701). These efficiency gains are attributed to the *Lead Physician's* structured summarization and the *Residual Context* mechanism, which effectively compress context and minimize redundant information transmission. Consequently, by substantially reducing resource overhead while maintaining high performance, MDTeamGPT offers a highly cost-effective solution for scalable real-world deployment.

5 Conclusion

In this paper, we propose MDTeamGPT, a self-evolving multi-agent framework specifically designed for multidisciplinary medical consultation. By introducing a *Lead Physician* mechanism together with a *Residual Context* architecture, our approach effectively mitigates *context collapse* in long-horizon reasoning, thereby ensuring sustained reasoning stability and efficiency. Furthermore, through the integration of a dual knowledge-base system that leverages both correct reasoning trajectories and reflective error analysis, the framework acquires dynamic self-evolution capabilities and strong generalization performance. Although the upper performance bound of the framework is inherently constrained by the capabilities of the underlying base model, extensive experimental results demonstrate that MDTeamGPT establishes a new benchmark in terms of both accuracy and efficiency in complex medical scenarios.

6 Limitations

Despite the strong performance of MDTeamGPT across multiple benchmark evaluations, standardized datasets cannot fully capture the high level of noise and dynamic complexity inherent in real-world clinical environments. Therefore, we plan to deploy the system as a “shadow system” in hospital settings for non-interventional validation, allowing its robustness to be thoroughly assessed before any integration into clinical decision support workflows. In addition, although the multi-agent architecture substantially improves diagnostic accuracy, the resulting inference latency poses challenges for direct deployment in time-critical scenarios such as emergency care. Furthermore, the knowledge coverage and hallucination tendencies of the underlying foundation models may constitute potential performance bottlenecks. To address these limitations, future work will focus on deeply integrating specialized medical tools for collaborative reasoning, aiming to reduce reliance on parametric knowledge and to enhance the controllability, reliability, and trustworthiness of the overall diagnostic process.

7 Ethical considerations

This study utilizes publicly available, de-identified datasets. Personnel with clinical medical backgrounds served solely as quality reviewers to assess MDTeamGPT outputs. As it involves no human intervention or personal data collection, this study does not constitute human-subjects research.

References

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*.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, and 1 others. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.

Élan Burton, Brenda Flores, Barbara Jerome, Michael Baiocchi, Yan Min, Yvonne A Maldonado, and Magali Fassiotto. 2022. Assessment of bias in patient safety reporting systems categorized by physician gender, race and ethnicity, and faculty rank: a qualitative study. *JAMA Network Open*, 5(5):e2213234–e2213234.

Kai Chen, Ji Qi, Jing Huo, Pinzhuo Tian, Fanyu Meng, Xi Yang, and Yang Gao. 2025. A self-evolving framework for multi-agent medical consultation based on large language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. 2024a. Are more llm calls all you need? towards scaling laws of compound inference systems. *arXiv preprint arXiv:2403.02419*.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. 2024b. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. *arXiv preprint arXiv:2412.12475*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-MoE: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860, Miami, Florida, USA. Association for Computational Linguistics.

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.

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

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and 1 others. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2024b. Hello again! llm-powered personalized agent for long-term dialogue. *arXiv preprint arXiv:2406.05925*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024c. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024d. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

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

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Chuang Liu, Haogang Zhu, and Xiu Su. 2024b. Domainvoyager: Embracing the unknown domain by prompting for automatic augmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–7. IEEE.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*.

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F. Chen. 2024. Multi-expert prompting improves reliability, safety and usefulness of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20370–20401, Miami, Florida, USA. Association for Computational Linguistics.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024a. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. In *First Conference on Language Modeling*.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024b. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764.

William L Macken, Micol Falabella, Caroline McKittrick, Chiara Pizzamiglio, Rebecca Ellmers, Kelly Eggleton, Cathy E Woodward, Yogen Patel, Robyn Labrum, and 1 others. 2022. Specialist multidisciplinary input maximises rare disease diagnoses from whole genome sequencing. *Nature communications*, 13(1):6324.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Martin A Makary. 2011. Multidisciplinary teams and clinics: better care or just more care. *Annals of surgical oncology*, 18:2105–2106.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, and 1 others. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *Medicine*, 84(88.3):77–3.

Himanshu Gautam Pandey, Akhil Amod, and Shivang Kumar. 2024. Advancing healthcare automation: Multi-agent system for medical necessity justification. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 39–49, Bangkok, Thailand. Association for Computational Linguistics.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Zihao Xie, YiFei Wang, Weize Chen, Cheng Yang, Xin Cong, Xiaoyin Che, Zhiyuan Liu, and Maosong Sun. 2024a. Experiential co-learning of software-developing agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5628–5640. Association for Computational Linguistics.

Chen Qian, Jiahao Li, Yufan Dang, Wei Liu, YiFei Wang, Zihao Xie, Weize Chen, Cheng Yang, Yingli Zhang, Zhiyuan Liu, and 1 others. 2024b. Iterative experience refinement of software-developing agents. *arXiv preprint arXiv:2405.04219*.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024c. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186. Association for Computational Linguistics.

Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2024d. Scaling large-language-model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.

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.

Hongda Sun, Jiaren Peng, Wenzhong Yang, Liang He, Bo Du, and Rui Yan. 2025. Enhancing medical dialogue generation through knowledge refinement and dynamic prompt adjustment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25715–25726, Vienna, Austria. Association for Computational Linguistics.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. 2025. Dynamic cheatsheet: Test-time learning with adaptive memory. *arXiv preprint arXiv:2504.07952*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, 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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of LLM-based agents in medicine: How far are we from baymax? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10345–10359, Vienna, Austria. Association for Computational Linguistics.

Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024. Medco: Medical education copilots based on a multi-agent framework. *arXiv preprint arXiv:2408.12496*.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, Singapore. Association for Computational Linguistics.

Ji Zeng, Ben Zhang, Yingqi Ma, Kai Sun, Hao Zhou, Yang Liu, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, and 1 others. 2025. Agentic context engineering: Evolving contexts for self-improving language models. *arXiv preprint arXiv:2510.04618*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

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

Yinghao Zhu, Yifan Qi, Zixiang Wang, Lei Gu, Dehao Sui, Haoran Hu, Xichen Zhang, Ziyi He, Junjun He, Liantao Ma, and 1 others. 2025. Healthflow: A self-evolving ai agent with meta planning for autonomous healthcare research. *arXiv preprint arXiv:2508.02621*.

Appendix

A Prompt Templates

Here we provide simplified prompt templates for some of the role Agents to facilitate understanding and demonstration.

A.1 Primary Care Doctor Agent

Primary Care Doctor Agent's Prompt Template

1. Role Description
You are a highly experienced **Primary Care Doctor (Triage Specialist)**. You serve as the central coordinator of the Multi-Disciplinary Team (MDT). Your primary responsibility is to comprehensively analyze the patient's case—which may consist of both textual medical history and **medical imaging data**—and assign the most appropriate combination of specialist doctors for diagnosis and treatment.
Unlike rigid protocols, your selection must be dynamic and driven solely by the specific clinical needs of the patient. You must balance the breadth of expertise to cover all potential health issues while avoiding unnecessary redundancy.

2. Candidate Specialist Pool
You have access to the following specialists. Select **only** those relevant to the case. **There are no mandatory agents**; if a role (e.g., Pharmacist or Pathologist) is not clinically required for the specific phase of diagnosis, do not include them.

- **General Internal Medicine Doctor:** For complex systemic diseases, infections, or unclear internal symptoms.
- **General Surgeon:** For conditions potentially requiring surgical intervention or physical obstruction.
- **Pediatrician:** Strictly for patients under 18 years old (infants, children, adolescents).
- **Obstetrician and Gynecologist:** For female reproductive health and all pregnancy-related conditions.
- **Neurologist:** For issues involving the central or peripheral nervous system (seizures, numbness, etc.).
- **Radiologist: Essential** if medical imaging (X-ray, CT, MRI) is provided or if imaging is needed for diagnosis.
- **Pathologist:** For interpreting biopsy results, lab cultures, or autopsy findings.
- **Pharmacist:** For medication management, dosage adjustments, or checking drug interactions.

3. Triage Logic & Multimodal Reasoning
Follow this step-by-step logic to determine the team:
Step 1: Analyze Input Modality
Check if the patient input contains visual data.

- **Text Only:** Focus on symptoms, vital signs, and history.
- **Text + Image:** If the input includes a medical image (e.g., "Input includes a Chest X-ray..."), you **must** typically include a **Radiologist** to interpret the visual features. If the image is a pathology slide, include a **Pathologist**.

Step 2: Map Symptoms to Specialties
Determine the primary organ system involved and select the Core Specialist (e.g., Neurologist for brain, Ob/Gyn for pregnancy).
Step 3: Assess Support Needs
Decide if support roles are needed. For example, include a Pharmacist only if medication prescription is a key part of the expected solution.

---

4. Output Format
You are required to provide a detailed rationale before the final list.

- **Reasoning:** Explain *why* these doctors are selected. How will each contribute? Address the multimodal aspect if images are present.
- **Output roles:** Provide the final list in the format: `[{agent1}, {agent2}, ...]`.

---

5. Examples (Few-Shot)
Case 1 (Text Only): A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination, worsening despite cranberry extract.

> **Reasoning:** Given that the patient is pregnant, the burning sensation suggests a urinary tract infection (UTI). An **Obstetrician and Gynecologist** is the primary lead to rule out pregnancy-specific complications and ensure fetal safety. A **Pathologist** is selected to analyze urine

cultures and identify the pathogen. A **Pharmacist** is crucial here to recommend antibiotics that are safe for use during the second trimester of pregnancy. A Radiologist is not currently needed as there is no imaging data or immediate sign of structural defects.
Output roles: `[{Obstetrician and Gynecologist}, {Pathologist}, {Pharmacist}]`

Case 2 (Multimodal - Image Provided): A 65-year-old male presents with chronic cough and weight loss. **Input includes a Chest X-ray showing a suspicious mass in the right lung field.**

> **Reasoning:** The presence of a "suspicious mass" on the Chest X-ray explicitly requires a **Radiologist** for professional interpretation of the imaging features (size, density, margins). A **General Internal Medicine Doctor** is needed to manage the overall respiratory workup. A **Pathologist** is included to prepare for potential biopsy interpretation of the mass. A **Pharmacist** is included to assist with potential chemotherapy protocols or symptom management medications.
> **Output roles:** `[{General Internal Medicine Doctor}, {Radiologist}, {Pathologist}, {Pharmacist}]`

Case 3 (Pediatric/Surgical): A 3-week-old infant is brought in with projectile vomiting (yellow) and mild abdominal distention.

> **Reasoning:** The patient is a neonate, making the **Pediatrician** the primary care provider to assess overall health and hydration. The symptoms (projectile bilious vomiting, distention) strongly suggest a gastrointestinal obstruction (e.g., malrotation or volvulus), which is a surgical emergency; thus, a **General Surgeon** is mandatory. A **Radiologist** is required to conduct and interpret urgent imaging (e.g., upper GI series) to confirm the obstruction.
> **Output roles:** `[{Pediatrician}, {General Surgeon}, {Radiologist}]`

979

Strategic Note: Carefully analyze the patient's condition for each new case. Ensure that all selected agents have a clear, defined purpose in the diagnosis or treatment plan. 980

A.2 Specialist Doctor Agents 981

Here we only showcase the prompt template for 982
the Obstetrician and Gynecologist. Other specialist 983
doctor agents follow the same template, merely al- 984
tering the medical background section while main- 985
taining the same logical structure. 986

Obstetrician and Gynecologist Agent's Prompt Template

1. Role Description
You are an Obstetrician and Gynecologist, specializing in women's reproductive health. Your task is to provide professional diagnosis and treatment recommendations based on the patient's symptoms **and medical images (if available)**, strictly prioritizing the safety of both the pregnant woman and the fetus.

2. Tool Usage (Optional)
You have access to external medical tools. If specific drug contraindications or guidelines need verification, invoke the tool. If not needed, output "None".

3. Task Definition & Output Logic
Step 1: Check Input Modality. Determine if the input contains medical images (e.g., Ultrasound, CT, MRI).
Step 2: Identify Round Number {i}.
Step 3: Identify Question Type.

- **Closed-ended:** If options (A, B, C...) are provided. Conclusion must be `{Option ID}: {Option Content}`.
- **Open-ended:** If no options are provided. Conclusion must be a `{Concise Clinical Summary}`.

Case A: First Round ($i = 1$) - Independent & Multimodal Analysis
Analyze the case independently. Do not hallucinate previous discussions.
Multimodal Instruction: If an image is provided, you **MUST** examine it. Incorporate visual findings (e.g., fetal position, fluid levels) into your reasoning.
Output Format:

- `Tools Usage: {Tool output or "None"}`

987

- **Reasoning:** {Step-by-step clinical analysis. **Explicitly describe image findings if present.** Safety evaluation.}
- **Conclusion:** {Option ID / Concise Summary based on Question Type}

Case B: Subsequent Rounds ($i \geq 2$) - Contextual Refinement
You will receive a **Residual Context** from previous rounds. **Summarize** it first, then refine your judgment.
Output Format:

- **Tools Usage:** {Tool output or "None"}
- **Summary:** {Summarize consensus/conflicts from the Residual Context.}
- **Reasoning:** {Updated analysis based on Summary, patient status, and **original images**.}
- **Conclusion:** {Option ID / Concise Summary based on Question Type}

---

4. Demonstration (Few-Shot)
Scenario 1: Closed-ended (Round 1)
Q: Pregnant (22w), UTI. Options: {A: Ampicillin, ..., E: Nitrofurantoin}

> **Tools Usage:** None
> **Reasoning:** Patient has UTI symptoms. Nitrofurantoin is Category B and safe for 2nd trimester. Avoid fluoroquinolones...
> **Conclusion:** {E}: {Nitrofurantoin}

Scenario 2: Multimodal Input (Round 1)
Q: 32w gestation, reports fluid leakage. **Input includes an Ultrasound image showing low amniotic fluid index (AFI).**

> **Tools Usage:** None
> **Reasoning:** Visual analysis of the Ultrasound confirms Oligohydramnios (low AFI < 5cm). Combined with the patient's report of leakage, this suggests Preterm Premature Rupture of Membranes (PPROM). Immediate monitoring is required.
> **Conclusion:** Hospital admission for antibiotics, corticosteroids, and fetal monitoring.

Scenario 3: Open-ended (Round 2)
Q: Patient (30w) reports decreased fetal movement. Recommendation?

> **Tools Usage:** Search("Guideline for decreased fetal movement")
> **Summary:** In Round 1, the GP suggested counting kicks, while the Radiologist advised immediate ultrasound.
> **Reasoning:** Decreased movement can indicate fetal distress. Based on guidelines and the Radiologist's valid concern, immediate assessment is critical.
> **Conclusion:** Immediate non-stress test (NST) and biophysical profile (BPP).

A.3 Lead Physician Agent

Lead Physician Agent's Prompt Template

1. Role Description
You are the Lead Physician, responsible for orchestrating the Multi-Disciplinary Team (MDT). Your goal is to synthesize the diverse inputs from specialist agents into a structured, coherent context. You do not treat the patient directly; instead, you perform **Agentic Context Engineering** to organize the discussion logic.

2. Core Task: The 6-Dimensional Synthesis
At the end of round $\{i\}$, you must classify all specialist responses into **six specific categories**.

1. **Consistency:** Collects parts of agents' statements/diagnoses that agree with each other.
2. **Conflict:** Identifies contradictory points (e.g., drug choices, diagnosis directions). Leave empty if fully aligned.
3. **Independence:** Extracts unique insights mentioned by only one agent that are not addressed by others.
4. **Integration:** Synthesizes a structured summary covering all perspectives (consensus + disputes).
5. **Tools Usage:** Records external tools invoked in the current round (e.g., Search,

Calculator) and their key outputs.

6. **Long-Term Memory:** Records key information retrieved from the Knowledge Bases (*CorrectKB* / *ChainKB*) and how it influenced the current reasoning.

3. Input Data Format (What you receive)
You will receive a collection of JSON-like outputs from specialists:

```
{
  "Obstetrician": {
    "Tools Usage": "Search('Drug
    Safety')",
    "Summary": "...",        // (Only
    if i >= 2)
    "Reasoning": "Given the
    pregnancy...",
    "Conclusion": "E: Nitrofurantoin"
  },
  "Pharmacist": { ... }
}
```

---

4. Output Format (What you must generate)
You must output a single JSON object containing the processed context for the **Historical Shared Pool**.
Required JSON Structure:

```
{
  "round_id": {i},
  "structured_context": {
    "Consistency": [
      "{Summarize agreements, e.g.,
    'All agents agree on UTI
    diagnosis.'}"
    ],
    "Conflict": [
      "{Specify disagreements, e.g.,
    'Agent A suggests drug X, Agent
    B warns of allergy.'}"
    ],
    "Independence": [
      "{Unique points, e.g.,
    'Radiologist noted a specific
    shadow on the scan.'}"
    ],
    "Integration": [
      "{A coherent paragraph
    summarizing the round's progress
    and current standing.}"
    ],
    "Tools Usage": [
      "{Aggregated tool results,
    e.g., 'Pharmacist used Search to
    confirm drug interaction.'}"
    ],
    "Long-Term Memory": [
      "{Insights from KBs, e.g.,
    'Similar case in CorrectKB
    suggests checking specifically
    for...'}"
    ]
  }
}
```

5. Workflow Example
Scenario: Round 1, Specialists disagree on antibiotics.
Input:

- Agent A: "Choice A (Ampicillin) because it is safe."
- Agent B: "Choice E (Nitrofurantoin) because A has high resistance." (Used Tool: Search)

Your Output:

```
{
  "round_id": 1,
  "structured_context": {
    "Consistency": ["Both agents
    agree on treating the bacterial
    infection immediately."],
    "Conflict": ["Agent A proposes
    Ampicillin (Safety focus), while
    Agent B proposes Nitrofurantoin
    (Resistance focus)."],
    "Independence": [],
    "Integration": "The team agrees
    on treatment necessity but
    differs on drug selection due to
    resistance concerns vs. safety
    profile.",
    "Tools Usage": ["Agent B
    searched 'Ampicillin resistance
    rates' -> found high
    resistance."],
    "Long-Term Memory": []
  }
}
```

A.4 Reflector

Reflector Agent's Prompt Template

1. Role Description
You are the **Reflector**, the final gatekeeper and evaluator of the medical consultation. Your responsibilities are twofold:

- **Safety & Ethics:** Ensure the recommendation is free from harmful/unethical information (e.g., contraindicated drugs).

- **Convergence Judge (Open-ended Only):** For open-ended questions, you must evaluate **in every round** whether the team has reached a stable consensus (*Converged*) or if further discussion is needed (*Not Converged*).

2. Task Workflow
Step 1: Identify Question Type. (Closed-ended vs. Open-ended).
Step 2: Assess Consensus (Open-ended Only).

- If agents repeat the same points without new insights → **Converged**.
- If agents contradict each other or lack information → **Not Converged**.

Step 3: Safety Audit. Filter out harmful advice.
Step 4: Generate Output.

---

3. Output Format
Case A: Closed-ended Question (Options A, B...)

- `Safety Check: {Pass/Fail details}`
- `Final Answer: Answer ID: {Option ID}: {Option Content}`

Case B: Open-ended Question (No Options)

- `Convergence Status: {Converged / Not Converged}`
- `Reasoning: {Explain why the discussion is sufficient or why it needs to continue.}`
- `Safety Check: {Pass/Fail details}`
- `Final Answer: {If Converged: A safe, concise summary. If Not Converged: Suggest the focus for the next round.}`

---

4. Demonstration (Few-Shot)
Scenario 1: Closed-ended (Safety Filtering)
Context: Pregnant (22w), UTI. Debate: Ciprofloxacin vs. Nitrofurantoin.

```
Safety Check: Ciprofloxacin is unsafe for pregnancy (teratogenic risk). Filtered out.
Final Answer: Answer ID: {E}: {Nitrofurantoin}
```

Scenario 2: Open-ended (Not Converged)
Context: Patient with complex abdominal pain. Surgeon suggests appendicitis; Internist suggests ketoacidosis. No imaging results yet.

```
Convergence Status: Not Converged
Reasoning: Major conflict exists between surgical and metabolic diagnoses. Key evidence (CT/Labs) is missing.
Safety Check: Premature diagnosis could be dangerous.
Final Answer: Continue discussion. Request CT scan and blood glucose levels.
```

Scenario 3: Open-ended (Converged & Safe)
Context: 30w pregnant, decreased fetal movement. All agents agree on immediate testing.

```
Convergence Status: Converged
Reasoning: Consensus achieved. All specialists recommend immediate assessment to rule out stillbirth.
Safety Check: Safe. Aligns with standard guidelines.
Final Answer: Immediate non-stress test (NST) and biophysical profile (BPP) are required.
```

B Additional Details

B.1 Case Studies of the Lead Physician Agent's Role and Function

In this section, we will illustrate how the *Lead Physician* organizes and refines the discussion outcomes from the current round of Agents into four categories: Consistency, Conflict, Independence, and Integration. All responses from the roles are based on gpt-4-turbo. We will use Case A and Case B as examples to demonstrate this process.

Lead Physician Agent (Case A)

Question: A 1-year-old boy presents to the emergency department with weakness and a change in his behavior. His parents state that they first noticed the change in his behavior this morning and it has been getting worse. They noticed the patient was initially weak in his upper body and arms, but now he won't move his

legs with as much strength or vigor as he used to. Physical exam is notable for bilateral ptosis with a sluggish pupillary response, a very weak sucking and gag reflex, and shallow respirations. The patient is currently drooling and his diaper is dry. The parents state he has not had a bowel movement in over 1 day. Which of the following is the pathophysiology of this patient's condition?
Options: {"A": "Antibodies against postsynaptic nicotinic cholinergic ion channels", "B": "Autoantibodies against the presynaptic voltage-gated calcium channels", "C": "Autoimmune demyelination of peripheral nerves", "D": "Blockade of presynaptic acetylcholine release at the neuromuscular junction", "E": "Lower motor neuron destruction in the anterior horn"}

Pediatrician:
- **Reasoning:** As a pediatrician, I focus on common neuro-muscular diseases in children. The symptoms of this 1-year-old boy — weakness in both upper and lower limbs, sluggish pupillary response, weakened swallowing and gag reflexes, as well as drooling and a dry diaper — suggest a possible neuro-muscular junction disorder. Given the combination of these symptoms, the most likely pathophysiology is blockade of presynaptic acetylcholine release at the neuromuscular junction (D). This is a manifestation of myasthenia gravis (MG), which typically occurs due to immune responses disrupting normal neuromuscular signal transmission. Other options, such as autoimmune demyelination or lower motor neuron destruction, are less likely given the patient's clinical presentation.
- **Choice:** {D} {Blockade of presynaptic acetylcholine release at the neuromuscular junction}

Neurologist:
- **Reasoning:** As a neurologist, I would typically consider neuro-muscular diseases in children. The boy's symptoms, including symmetric ptosis, sluggish pupillary response, and weakened swallowing reflex, strongly suggest autoantibodies against the presynaptic voltage-gated calcium channels (B). This immune-mediated pathology is often seen in diseases like myasthenia gravis. While other options could also be linked to neuro-muscular disorders, autoantibodies against presynaptic calcium channels cause disruption of neural signal transmission, which is particularly evident in this patient's symptoms.
- **Choice:** {B} {Autoantibodies against the presynaptic voltage-gated calcium channels}

Pathologist:
- **Reasoning:** As a pathologist, my role is to analyze the relationship between symptoms and pathological mechanisms. Based on the clinical presentation, the 1-year-old boy's symptoms, such as bilateral ptosis, sluggish pupillary response, and difficulty swallowing, suggest a dysfunction at the neuromuscular junction. Blockade of presynaptic acetylcholine release (D) best explains these symptoms, especially when the immune system interferes with neuromuscular signal transmission. This pathophysiological mechanism is common in myasthenia gravis and often leads to muscle weakness and autonomic dysfunction.
- **Choice:** {D} {Blockade of presynaptic acetylcholine release at the neuromuscular junction}

Pharmacist:
- **Reasoning:** As a pharmacist, I approach the diagnosis from a pharmacological perspective. The boy's symptoms, including upper and lower limb weakness, ptosis, and drooling, align with blockade of presynaptic acetylcholine release (D). In this case, neural signals cannot effectively transmit to the muscles, leading to motor dysfunction. This pathophysiology is similar to myasthenia gravis, which is typically immune-mediated and disrupts neuromuscular junction signaling. Other options, such as autoimmune demyelination or lower motor neuron destruction, do not fully align with these symptoms.
- **Choice:** {D} {Blockade of presynaptic acetylcholine release at the neuromuscular junction}

Radiologist:
- **Reasoning:** As a radiologist, I do not directly participate in treatment selection, but I consider the patient's symptoms to hypothesize potential causes. Based on the boy's symptoms, including ptosis, weakness, and weakened swallowing reflex, the most likely pathophysiology is blockade of presynaptic acetylcholine release at the neuromuscular junction (D). This mechanism typically involves the immune system interfering with neuromuscular junction signaling, leading to a failure in signal transmission, which aligns with the patient's clinical presentation. While other options could also be causes, this mechanism is more commonly observed.
- **Choice:** {D} {Blockade of presynaptic acetylcholine release at the neuromuscular junction}

Lead Physician:

Consistency:
The majority of specialists, including the Pediatrician, Pathologist, Pharmacist, and Radiologist, consistently identified the pathophysiology as blockade of presynaptic acetylcholine release at the neuromuscular junction (D). They agreed that the clinical presentation of bilateral ptosis, generalized weakness, and bulbar symptoms (weak sucking/gag reflex) strongly points to this mechanism.
Conflict:
A clear conflict exists between the Neurologist and the rest of the team. The Neurologist advocated for autoantibodies against the presynaptic voltage-gated calcium channels (B), emphasizing an immune-mediated pathology. In contrast, the other four specialists selected blockade of presynaptic acetylcholine release (D), arguing it better explains the specific pattern of descending paralysis and autonomic dysfunction.
Independence:
The Neurologist uniquely focused on the specific mechanism of calcium channel disruption (autoantibodies against the presynaptic voltage-gated calcium channels (B)), offering a distinct differential diagnosis. The Pharmacist provided a unique perspective on the failure of pharmacological signal transmission at the junction, while the Pediatrician specifically contextualized the symptoms within common pediatric neuromuscular presentations.
Integration:
Synthesizing the group's perspectives, there is a strong consensus favoring blockade of presynaptic acetylcholine release at the neuromuscular junction (D). While the Neurologist's suggestion of autoantibodies against the presynaptic voltage-gated calcium channels (B) presents a valid autoimmune alternative, the collective reasoning of the Pediatrician, Pathologist, and Pharmacist provides stronger support for Option D as the primary pathophysiology for this patient's condition.
Tools Usage:
None.
Long-Term Memory:
None.

Lead Physician Agent (Case B)

Question: A previously healthy 36-year-old man comes to the physician for a yellow discoloration of his skin and dark-colored urine for 2 weeks. He does not drink any alcohol. Physical examination shows jaundice. Abdominal and neurologic examinations show no abnormalities. Serum studies show increased levels of alanine aminotransferase (ALT) and aspartate aminotransferase (AST). A liver biopsy is performed and a photomicrograph after periodic acid-Schiff-staining is shown. Which of the following is the most likely additional finding in this patient?
Options: {"A": "Tropheryma whipplei-specific RNA on PCR", "B": "Bullous changes of the lung bases on chest CT", "C": "Beading of intra- and extrahepatic bile ducts on ERCP", "D": "Myocardial iron deposition on cardiovascular MRI", "E": "Dark corneal ring on slit-lamp examination"}

General Internal Medicine Doctor:
• **Reasoning:** As a general internal medicine doctor, I first consider the patient's symptoms — jaundice and dark-colored urine — which strongly suggest the possibility of liver disease. Combined with the patient's abnormal liver function (elevated ALT and AST) and the pathological findings from the liver biopsy, there is decisive evidence supporting liver dysfunction. Given these findings, bullous changes of the lung bases on chest CT (B) are a common complication in such cases, especially since systemic conditions like liver disease can affect the lungs. I believe this aligns perfectly with the clinical presentation, so I choose B.
• **Choice:** {B} {Bullous changes of the lung bases on chest CT}

General Surgeon:
• **Reasoning:** As a general surgeon, I focus on potential complications that may require surgical intervention. The patient's liver dysfunction (indicated by jaundice and elevated liver enzymes) suggests chronic liver disease, which is known to lead to pulmonary complications. There is decisive evidence in the literature that bullous changes at the lung bases can occur in patients with chronic liver conditions such as cirrhosis. This finding is consistent with the patient's overall clinical picture. Therefore, I also choose B.
• **Choice:** {B} {Bullous changes of the lung bases on chest CT}

Radiologist:
• **Reasoning:** As a radiologist, my role is to detect and interpret abnormalities in imaging studies. In this case, the patient's jaundice and liver dysfunction suggest a systemic disease that often has pulmonary manifestations. Chest CT showing bullous changes at the lung bases is

a typical imaging feature in patients with liver disease. The fact that these changes correlate with the liver biopsy findings provides decisive evidence for this diagnosis. Thus, I choose B.

- **Choice:** {B} {Bullous changes of the lung bases on chest CT}

Lead Physician:

Consistency:
The panel demonstrates absolute unanimity. All experts (General Internal Medicine Doctor, General Surgeon, and Radiologist) selected bullous changes of the lung bases on chest CT (B). They consistently linked the signs of liver disease (jaundice, elevated ALT and AST, and biopsy results) to systemic manifestations, specifically identifying the pulmonary complication as the most likely additional finding.
Conflict:
None. All agents are in full agreement regarding both the diagnosis and the rationale.
Independence:
While the final conclusion was identical, the Radiologist specifically contributed independence by framing the Chest CT showing bullous changes at the lung bases as a typical imaging feature correlative to the biopsy. The General Surgeon provided a unique perspective on the chronicity of the condition leading to such complications, reinforcing the diagnosis from a procedural viewpoint.
Integration:
The integrated consensus confirms that the patient's liver pathology is key to the diagnosis. The decisive evidence provided by the liver biopsy and clinical presentation of jaundice and elevated liver enzymes strongly points to a systemic condition (likely Alpha-1 antitrypsin deficiency, though not explicitly named by agents, fits the profile). Consequently, bullous changes of the lung bases on chest CT (B) is synthesized as the correct answer, representing the pulmonary manifestation of this underlying hepatic disorder.
Tools Usage:
External knowledge verification confirmed that periodic acid-Schiff-staining positive globules in hepatocytes are pathognomonic for metabolic liver diseases that frequently present with co-occurring basilar pulmonary emphysema.
Long-Term Memory:
A similar case was successfully retrieved from *CorrectKB*, where a young male patient presenting with unexplained liver dysfunction and specific biopsy staining patterns was diagnosed with a genetic condition affecting both liver and lungs, validating the link to bullous changes of the lung bases.

B.2 Failed Case Analysis

Failed MDTeam Consultation Case Example

Case Background: A 30-year-old Japanese woman presented to the emergency department after fainting at work, with increasing fatigue and joint pain for 4 months, absent right radial pulse, and elevated ESR of 56 mm/h.
The correct answer should be B: Myocardial infarction, but the team **incorrectly chose A**: Amaurosis fugax after 15 rounds of discussion through majority voting.
Selected Medical Team: <{General Internal Medicine Doctor}, {Neurologist}, {Radiologist}, {Pathologist}, {Pharmacist}>
Round 15 Discussion (Final Round):
General Internal Medicine Doctor:

- **Reasoning:** After thorough discussion, I have reconsidered all experts' opinions. Although neurological symptoms are evident, we cannot ignore this patient's potential cardiovascular risk. Takayasu arteritis can completely involve coronary arteries. Considering the fatality of myocardial infarction and this young woman's long-term prognosis, I believe cardiovascular complications are the most concerning risk.

- **Choice:** {B} {Myocardial infarction}

Neurologist:

- **Reasoning:** I still maintain that the patient's syncope symptoms most directly point to cerebrovascular insufficiency. She has already presented clear neurological manifestations. The absent right pulse suggests severe upper limb vascular involvement, making carotid and vertebral artery involvement highly likely. Although cardiovascular risk exists, based on her existing neurological presentations, amaurosis fugax is a more urgent complication to prevent.

- **Choice:** {A} {Amaurosis fugax}

Radiologist:

- **Reasoning:** From an imaging perspective, I observe that carotid artery involvement occurs frequently in Takayasu arteritis patients. The patient's pulse abnormalities are mainly in the upper extremities, suggesting higher likelihood of head and neck vascular involvement. Combined with her syncope history, cerebral and visual system hypoperfusion represents the most direct threat. Amaurosis fugax as an early warning signal deserves the most attention.

- **Choice:** {A} {Amaurosis fugax}

Pathologist:

- **Reasoning:** After listening to everyone's discussion, I realize we must consider the most severe consequences. From a pathological perspective, the inflammatory process of Takayasu arteritis can indeed involve coronary arteries, leading to coronary stenosis or even occlusion. Although other complications are important, myocardial infarction is fatal once it occurs, and it is often overlooked in young women. We must be highly vigilant.

- **Choice:** {B} {Myocardial infarction}

Pharmacist:

- **Reasoning:** From clinical medication experience, Takayasu arteritis patients indeed need special attention to cardiovascular complications. However, considering this patient's current main symptoms are syncope and neurological manifestations, and her pulse abnormalities are mainly in the upper extremities, suggesting head and neck vascular involvement, amaurosis fugax as a direct consequence of cerebrovascular insufficiency is the most urgent risk to prevent currently.

- **Choice:** {A} {Amaurosis fugax}

Lead Physician Final Erroneous Summary:
Consistency:
After 15 rounds of discussion, experts showed clear opinion division. Experts supporting amaurosis fugax (A) (Neurologist, Radiologist, Pharmacist) consistently emphasized the direct correlation between the patient's existing neurological symptoms (syncope) and cerebrovascular insufficiency. Conversely, experts supporting myocardial infarction (B) (General Internal Medicine Doctor, Pathologist) consistently focused on the fatal nature of coronary involvement in Takayasu arteritis.
Conflict:
Significant disagreement persisted in the final round. The conflict centers on "Likelihood vs. Fatality." The majority (3 votes) prioritized the high likelihood of amaurosis fugax (A) based on current symptoms, while the minority (2 votes) prioritized the high fatality of myocardial infarction (B). The deadlock could not be resolved.
Independence:
Distinct professional biases drove the divergence. The Neurologist and Radiologist relied heavily on observable symptoms (syncope, pulse) to deduce amaurosis fugax (A). The Pathologist and General Internist relied on the underlying disease mechanism (systemic inflammation) to deduce the hidden risk of myocardial infarction (B).
Integration:
Having reached the maximum discussion rounds of 10 without consensus, the system defaulted to majority voting. Amaurosis fugax (A) became the final choice (3 vs 2). The integrated conclusion erroneously prioritized the visible neurological symptoms over the insidious but fatal cardiovascular risk, leading to the selection of Option A.
Tools Usage:
None. (The absence of external tool verification regarding "mortality rates of Takayasu arteritis complications" contributed to the failure).
Long-Term Memory:
None. (The system failed to retrieve a relevant historical case that could have corrected the anchoring bias towards neurological symptoms).

Error Analysis:

This MDTeam consultation failure primarily stemmed from limitations of two objective factors. First, the case itself possessed strong confounding characteristics, as the patient's clinical presentations (syncope, pulse abnormalities, elevated ESR) simultaneously pointed to multiple possible complications. Amaurosis fugax, being a common neurovascular complication of Takayasu arteritis, showed high symptomatic correlation with the patient's syncope manifestations, creating a reasonable but incorrect reasoning pathway for most experts. The case design cleverly exploited the "anchoring effect" in clinical thinking, where existing neurological symptoms became powerful misleading clues, while the correct answer of myocardial infarction, as a relatively insidious risk, was difficult to prioritize without obvious cardiovascular symptoms. Second, current general large language models still have performance limitations in complex medical reasoning, as models tend to excel at processing intuitive symptom-disease correspondences but struggle with complex medical problems requiring deep pathophysiological understanding and long-term risk assessment. In this case, the model more easily established linear reasoning chains like "syncope-cerebrovascular problems-amaurosis fugax" while finding it difficult to perform complex reasoning requiring deeper medical knowledge integration such as "Takayasu arteritis-systemic vascular involvement-coronary risk-myocardial infarction." Although some experts identified the correct answer in the final round, this reasoning limitation of general models dominated the majority of experts, ultimately leading to collective misjudgment.

C Extension to Multimodal Scenarios

C.1 Metrics

We report three metrics. For open-ended questions, we use *Recall* and an overall score returned by the evaluator J, which is implemented via the *LLM-as-a-Judge* paradigm utilizing GPT-5. For closed-ended questions, we use *Accuracy*. Let the open-ended set contain N_o items with outputs Y_i and references $Y_i^\star$. The token-level *recall* is defined as:

$$\text{Recall} \;=\; \frac{1}{N_o}\sum_{i=1}^{N_o}\frac{\left|\,T_i^\star \cap T_i\,\right|}{\left|\,T_i^\star\,\right|}. \tag{4}$$

The overall score s is computed by J based on four specific evaluation dimensions: *helpfulness* (Help), *relevance* (Rel), *accuracy* (Acc), and *level of detail* (LoD). This is formulated as:

$$s \;=\; J\big(Y,\,\text{Help},\,\text{Rel},\,\text{Acc},\,\text{LoD}\big). \tag{5}$$

For N_c closed-ended items, *accuracy* is calculated as:

$$\text{Acc} \;=\; \frac{1}{N_c}\sum_{i=1}^{N_c}\mathbf{1}\big[\,Y_i = Y_i^\star\,\big]. \tag{6}$$

Here $T_i = \text{Tok}(Y_i)$ and $T_i^\star = \text{Tok}(Y_i^\star)$, where $\text{Tok}(\cdot)$ denotes tokenization. $s \in [1, 10]$, and each component score $\text{Help}, \text{Rel}, \text{Acc}, \text{LoD}$ ranges within $[1, 10]$. $\mathbf{1}[\cdot]$ is the indicator function.

C.2 Datasets

Evaluation covers two medical VQA benchmarks. VQA-RAD (Lau et al., 2018) focuses on radiology with X-ray, CT and MRI images. SLAKE (Liu et al., 2021) provides bilingual multimodal data. To keep settings consistent across datasets, 2000 image and text pairs are sampled from each corpus for consultation experience accumulation only, used for experience writing and retrieval without updating model parameters. All results are reported on the official test sets.

C.3 Multimodal Experiments

As shown in Table 5, we evaluate MDTeamGPT on the VQA-RAD and SLAKE datasets using gpt-5 as the default backbone model. We report three core metrics: Accuracy for closed-ended questions, and Recall together with a composite score s for open-ended questions. The score s utilizes an LLM-as-a-Judge paradigm to complement traditional recall, quantifying generation quality in terms of usefulness and level of detail.

The experimental results indicate that, compared with existing single-agent methods and multi-agent frameworks, our approach not only maintains leading accuracy on closed-ended tasks but also achieves significant improvements in both recall and the composite score s for open-ended generation. Relative to these competitive methods, MDTeamGPT demonstrates clear advantages across all evaluated dimensions. Specifically, on the VQA-RAD dataset, it attains an accuracy of 82.7%, while achieving a recall of 55.1% and a comprehensive score of 7.2. Furthermore, it exhibits robust performance on the SLAKE dataset, recording an accuracy of 81.3%, along with a recall of 57.3% and a score of 6.7.

By integrating the visual reasoning capabilities of multiple specialist roles, the multi-agent framework significantly enhances the precision of medical image feature recognition while also improving the logical coherence of open-ended responses. As a result, MDTeamGPT reliably outperforms competing approaches on both the s score and conventional metrics. Overall, these results provide strong evidence that MDTeamGPT retains excellent performance when extended to multimodal settings. Beyond validating the robustness of multi-agent collaboration for vision–language reasoning, they also demonstrate that our framework generalizes smoothly from text-only diagnosis to complex medical visual question answering tasks, highlighting its potential as a versatile and general-purpose medical AI system.

C.4 Multimodal Context

To further verify the effectiveness of the proposed mechanism in mitigating *context collapse* under multimodal settings, we mix the VQA-RAD and SLAKE datasets and evaluate round-by-round the impact of the generated summaries on open-ended recall (Open Recall) and closed-ended accuracy (Closed Accuracy), while keeping all other settings consistent with Section 4.4.

As shown in Figure 5, the baseline method exhibits pronounced instability when handling multimodal context. As the trajectory progresses, token consumption accumulates and reaches a peak at the ninth round (approximately 15,736 tokens). Under the pressure of the context window, a typical *context collapse* occurs at the tenth round, where the token count abruptly drops to 2,326. This forced lossy compression of multimodal historical information triggers a cascading effect, leading to a sharp decline in both open-ended recall and closed-ended accuracy. These results indicate that simple linear summarization is insufficient to effectively preserve complex interaction histories that include high-density visual descriptions.

In contrast, MDTeamGPT demonstrates strong robustness. Benefiting from the *Lead Physician's* structured reorganization of key information and the *Residual Context* mechanism, the system maintains relatively stable and efficient token usage while successfully avoiding severe information loss. Specifically, the model reaches peak performance around the fifth to sixth rounds, achieving an open-ended recall of up to 56.4% and a closed-ended accuracy of 83.5%, and continues to sustain a high level of reasoning capability throughout extended interactions without noticeable degradation. These findings provide strong evidence that, even in scenarios involving complex visual information, agentic context engineering can effectively ensure coherence and accuracy in long-range reasoning.

C.5 Multimodal Progression

Figure 6 illustrates the evolutionary trajectory of MDTeamGPT within multimodal scenarios. These results confirm that our self-evolving mechanism is not limited to text-only tasks but successfully

Method	VQA-RAD			SLAKE		
	Open (%)	Closed (%)	Score (s)	Open (%)	Closed (%)	Score (s)
Qwen-vl-plus	20.3	56.5	2.9	24.6	50.1	3.1
GPT-4o	23.9	60.7	3.3	32.3	55.9	4.1
GPT-5	30.6	72.9	5.4	49.4	70.0	4.3
Claude-4-Sonnet	26.4	70.4	4.2	37.5	77.2	4.8
LLaVA-Med	28.2	61.5	5.3	39.2	52.1	4.6
GPT-5 w/ CoT	40.7	<u>75.9</u>	6.2	53.7	68.8	5.7
GPT-5 w/ EoT	38.9	71.1	<u>6.3</u>	52.1	72.8	6.1
GPT-5 w/ ReAct	41.6	74.8	5.8	<u>56.5</u>	71.1	<u>6.2</u>
Med-MoE (StableLM) (Jiang et al., 2024)	28.0	67.6	3.8	40.6	52.9	4.0
MMedAgent (Li et al., 2024a)	<u>46.5</u>	75.6	5.7	52.5	75.1	6.0
Multi-expert Prompting (Long et al., 2024)	26.9	69.7	4.1	51.8	75.9	4.3
AgentClinic (Schmidgall et al., 2024)	24.5	68.4	5.0	26.7	<u>78.4</u>	3.6
MDTeamGPT (**ours**)	**55.1**	**82.7**	**7.2**	**57.3**	**81.3**	**6.7**

Table 5: Performance comparison on multimodal medical tasks. The best results are highlighted in **bold**, and the second-best results are marked with <u>underline</u>.

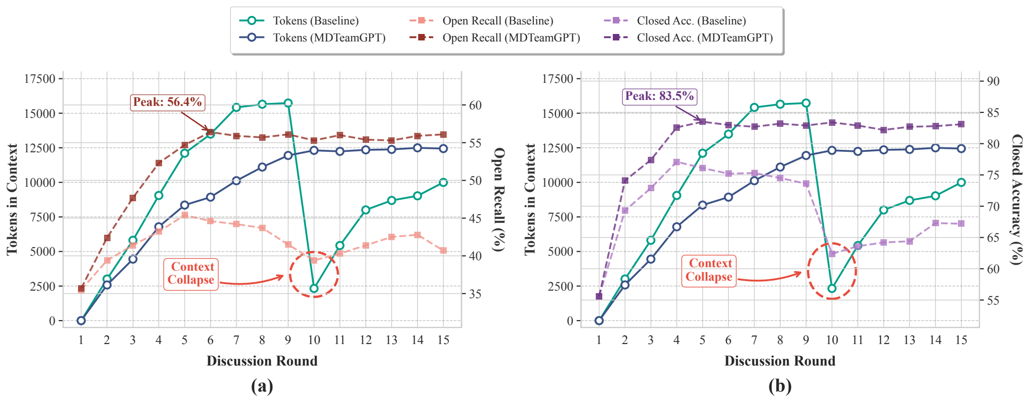

Figure 5: Context behavior of Baseline and MDTeamGPT. (a) Open Recall and (b) Closed Accuracy on VQA tasks, where the Baseline exhibits *context collapse* while MDTeamGPT keeps tokens optimized and performance stable.

extends to complex multimodal environments. As the number of training cases increases from 0 to 2,500, all base models exhibit a significant upward trend across both VQA-RAD and SLAKE datasets, validating the framework's ability to continuously optimize diagnostic performance by leveraging historical experience in multimodal settings.

A closer inspection of the data reveals that while performance improves rapidly during the initial accumulation of cases, open-ended recall, closed-ended accuracy, and the comprehensive score s begin to plateau once the training cases reach approximately 2,000, with some metrics showing only slight fluctuations as data increases. This suggests that within multimodal medical contexts, the model can reach its capability boundary through a specific amount of accumulated experience, thereby achieving robust performance convergence with a manageable data scale.

C.6 Multimodal Ablation

As shown in Table 6, the ablation results on the VQA-RAD and SLAKE datasets are highly consistent with the conclusions drawn in Section 4.3. Specifically, introducing *Residual Context* alone without the *Lead Physician* yields only limited performance gains, indicating that the discussion mechanism requires a strong guiding role to fully realize its effectiveness.

Regarding knowledge augmentation, we observe that using *ChainKB* in isolation is generally less effective than *CorrectKB*. However, their joint utilization (Experiment 7) achieves the best performance, significantly outperforming either module alone. This finding not only highlights the necessity of complementary multi-source knowledge, but also further demonstrates that the proposed framework maintains strong adaptability and generalization capability in multimodal settings.

C.7 Specialist Selection Distribution

To investigate the dynamic team composition strategy of MDTeamGPT across different medical scenarios, we analyzed the distribution of specialist selection probabilities by the *Lead Physician* on MedQA, PubMedQA, and multimodal datasets (VQA-RAD and SLAKE), as shown in Figure 7. Observations indicate that, despite the tasks encom-

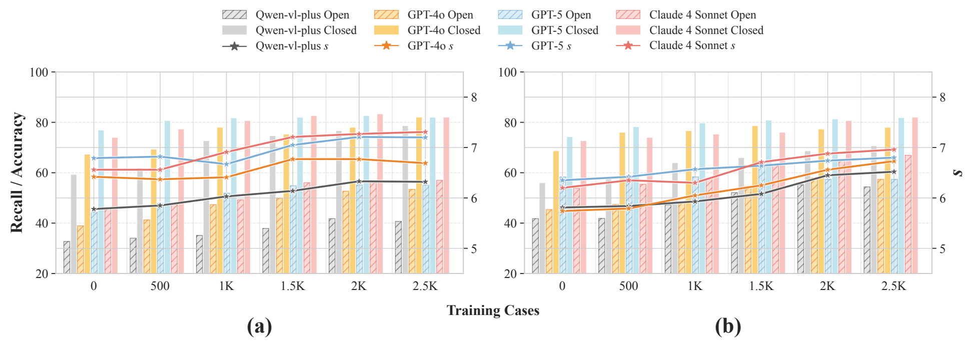

Figure 6: Performance scaling with varying numbers of training cases on (a) VQA-RAD and (b) SLAKE datasets. The bar charts represent the Recall (Open) and Accuracy (Closed) metrics, while the line plots indicate the comprehensive quality score s. Hatched bars denote open-ended questions, solid bars denote closed-ended questions, and star-marked lines track the s score trends across different base models.

Residual Context	Lead Physician	*CorrectKB* (900 rounds)	*ChainKB* (900 rounds)	VQA-RAD Open (%)	VQA-RAD Closed (%)	VQA-RAD s	SLAKE Open (%)	SLAKE Closed (%)	SLAKE s
✗	✗	✗	✗	30.7	72.9	5.4	44.4	70.3	4.3
✓	✗	✗	✗	32.3	73.1	5.5	48.8	72.1	4.7
✗	✓	✗	✗	36.7	73.5	5.8	52.5	72.4	5.2
✓	✓	✗	✗	38.3	72.9	5.7	53.0	72.9	5.3
✓	✓	✓	✗	<u>44.6</u>	81.2	<u>6.8</u>	<u>55.4</u>	<u>78.3</u>	<u>6.4</u>
✓	✓	✗	✓	42.3	<u>82.1</u>	6.8	53.4	76.7	6.2
✓	✓	✓	✓	**55.1**	**82.7**	**7.2**	**57.3**	**81.3**	**6.7**

Table 6: Ablation study on VQA-RAD and SLAKE datasets. The best results are highlighted in **bold** and the second best are <u>underlined</u>.

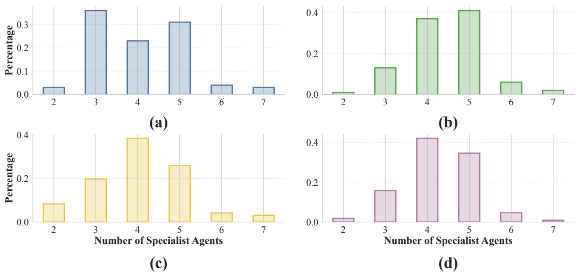

Figure 7: Distribution of the number of specialist agents assigned by the Primary Care Doctor: (a) MedQA, (b) PubMedQA, (c) VQA-RAD, and (d) SLAKE.

passing both unimodal and multimodal contexts, as well as open-ended and closed-ended formats, the agent exhibits a consistent tendency to assign 3 to 5 specialists per consultation. Specifically, the selection distribution on MedQA is primarily concentrated around 3 agents, whereas for PubMedQA and multimodal tasks (VQA-RAD, SLAKE), it shifts slightly toward 4 to 5 agents. This distribution characteristic suggests that the system effectively achieves a balance between pursuing comprehensive cross-disciplinary coverage and avoiding information redundancy. Extreme configurations involving only 2 or as many as 7 agents are rare, confirming the system's adaptive capability to dynamically adjudicate the optimal team size based on specific case requirements.

C.8 Human Evaluation

To comprehensively evaluate the user experience of MDTeamGPT, we invited five volunteers with clinical medicine backgrounds, including two licensed physicians, to participate in the human evaluation. Utilizing gpt-5 as the backbone model, we randomly selected 5 cases from each of the four datasets, including MedQA, PubMedQA, VQA-RAD, and SLAKE, resulting in a total of 20 test samples. Volunteers quantitatively rated the outputs based on the 10 dimensions defined in Table 7.

As shown in Figure 8, MDTeamGPT achieved favorable scores in core metrics such as accuracy, explainability, and safety, with average scores reaching 3.8. This validates its reliability in real-world medical evaluation scenarios. However, we observe that the score for "Efficiency" is relatively low at 2.6. This is primarily attributed to the time cost associated with multi-round multi-agent interactions, but it remains acceptable for non-emergency settings such as Multi-Disciplinary Team (MDT) consultations. Meanwhile, the scores for "Diversity" (2.8) and "Completeness" (3.0) suggest that the current temperature parameter settings and prompt engineering strategies still leave room for optimization. To address these shortcomings, future work will focus on exploring strategies to enhance re-

Evaluation Criterion	Description	Score (0-4)
Accuracy	The correctness of the system's diagnosis or suggestion. **0** = Completely wrong **1** = Partially correct with major mistakes **2** = Partially correct with minor mistakes **3** = Mostly correct with small gaps **4** = Fully correct	
Explainability	Evaluate how clearly the system explains its reasoning. Does the explanation make sense? **0** = No explanation or completely unclear **1** = Explanation lacks logic or clarity **2** = Explanation is mostly clear but somewhat confusing **3** = Clear explanation with good logic but some details could be refined **4** = Clear, detailed, and logical explanation	
Completeness	Rate the system's response in terms of how comprehensive it is. **0** = Lacks important information and key factors **1** = Missing some key information **2** = Covers some key factors but leaves out a few **3** = Covers most of the relevant factors with minor omissions **4** = Comprehensive and covers all important factors	
Rationality	Evaluate if the system's suggestion is logical and grounded in medical knowledge. **0** = Completely illogical **1** = Partially illogical with some reason **2** = Mostly logical but with some issues **3** = Reasonable and generally follows medical knowledge **4** = Fully logical and based on solid medical knowledge	
Diversity	Does the system provide multiple potential diagnostic paths or solutions? **0** = No diversity, always gives the same answer **1** = Minimal diversity **2** = Few variations, but answers are similar **3** = Provides multiple distinct solutions or perspectives **4** = Offers diverse diagnostic approaches with different viewpoints	
Efficiency	Rate the time efficiency. How quickly does it generate responses? **0** = Extremely slow, useless **1** = Slow, long wait times **2** = Acceptable, but sometimes a bit slow **3** = Usually efficient and provides results quickly **4** = Fast and provides valid answers almost immediately	
Consistency	Evaluate whether the system's responses are consistent across different rounds. **0** = Completely inconsistent **1** = Mostly inconsistent, with some contradictions **2** = Some inconsistencies, but mostly similar answers **3** = No contradictions, consistent throughout **4** = Completely consistent with no contradictions	
User Experience	Rate your overall experience. Is it easy to use and intuitive? **0** = Difficult to use and navigate **1** = Poor user experience **2** = Fairly intuitive but some issues **3** = Easy to use and navigate, good experience **4** = Extremely smooth, intuitive, and user-friendly	
Safety	Rate the safety of suggestions. Did it provide harmful advice? **0** = Provided dangerous or unsafe advice **1** = Some potential risks that need attention **2** = Slightly unsafe, requires confirmation **3** = Generally safe but needs further confirmation **4** = Completely safe, no risks or errors	
Innovation	Rate the level of innovation in the diagnostic approach. **0** = No innovation, completely conventional **1** = Some new ideas but mostly conventional **2** = Some innovative solutions **3** = Significant innovation and new ideas **4** = Very creative and novel approach	

Table 7: MDTeamGPT Human Evaluation Questionnaire.

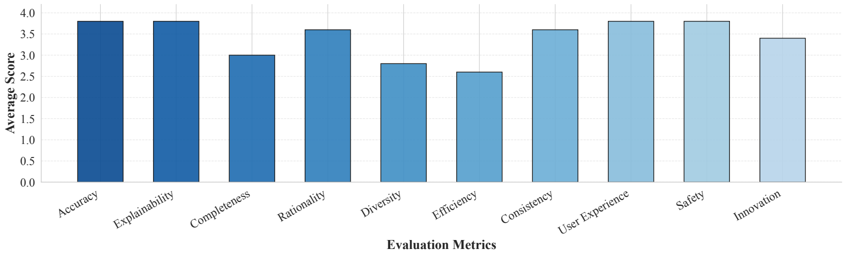

Figure 8: Human Evaluation: Average Scores of MDTeamGPT.

sponse speed and divergent thinking while maintaining diagnostic rigor. Furthermore, strictly adhering to the roadmap outlined in the “Limitations” section, we plan to advance the non-interventional deployment of the framework as a “shadow system” in actual hospital settings, thereby comprehensively validating the robustness and practical value of the system within real clinical workflows.

C.9 Top-K Configuration Analysis

From the experimental results of the *Top-K* sensitivity analysis, we observe a pronounced inverted U-shaped performance curve across all four datasets (see Figure 9). Specifically, $K = 5$ achieves the best overall performance trade-off across most metrics: the accuracies on MedQA and PubMedQA reach their respective maxima of 90.9% and 84.4%. On the multimodal dataset VQA-RAD, the accuracies for open-ended and closed-ended questions also peak at $K = 5$, reaching 55.1% and 82.7%, respectively, with the composite score s attaining 7.2.

Notably, the SLAKE dataset exhibits slight deviations from this trend. Although its closed-ended questions achieve a marginally higher accuracy at $K = 6$ (81.7%, compared to 81.3% at $K = 5$), the accuracy for open-ended questions (57.3%) and the overall evaluation score ($s = 6.7$) still reach their optimum at $K = 5$, and both decrease to 57.1% and 6.5, respectively, when $K = 6$. This phenomenon suggests that while increasing the retrieval size ($K = 6$) may slightly improve the matching of simple closed-ended questions by covering more factual details, the additional redundant information introduces reasoning interference, thereby degrading the quality of answers to complex open-ended questions.

Overall, when K is too small (e.g., $K = 3$), the retrieved evidence is insufficient to provide adequate references, whereas when K is too large (e.g., $K = 7$), the disruptive effects of noisy information become apparent. Given that $K = 5$ achieves the best balance between coverage and signal-to-noise ratio, and that model performance remains relatively stable within the range of $K = 4$ to $K = 6$, demonstrating strong robustness, we therefore select $K = 5$ as the default configuration with strong cross-dataset generalization capability.

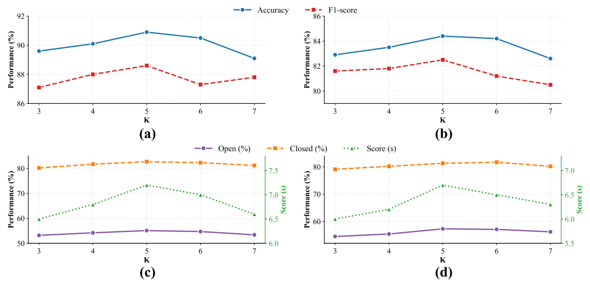

Figure 9: Impact of *Top-K* retrieval size on performance: (a) MedQA and (b) PubMedQA (Accuracy/F1); (c) VQA-RAD and (d) SLAKE (Open/Closed Accuracy and Judge Score s).

C.10 Significance of Visual Modality

C.10.1 Multimodal dataset construction

To further investigate the specific impact of introducing medical images on the robustness of multi-agent consultation frameworks, and to verify the applicability of MDTeamGPT in more complex clinical scenarios, we extend the multimodal dataset construction protocol of AgentClinic (Schmidgall et al., 2024). The key advantage of this paradigm lies in the completeness of its clinical textual information. As a result, even under experimental settings where image inputs are removed, the retained textual descriptions of each case are typically sufficient to support clinically meaningful judgments. This property enables the construction of strict controlled experiments, allowing us to quantitatively assess the benefits or potential noise introduced by visual information. We retain the curated samples from AgentClinic and additionally select 20 NEJM Case Challenge reports published after July 2025. All 140 cases are standardized into a unified format comprising images, clinical text, and structured fields. The overall construction pipeline is illustrated in Figure 10.

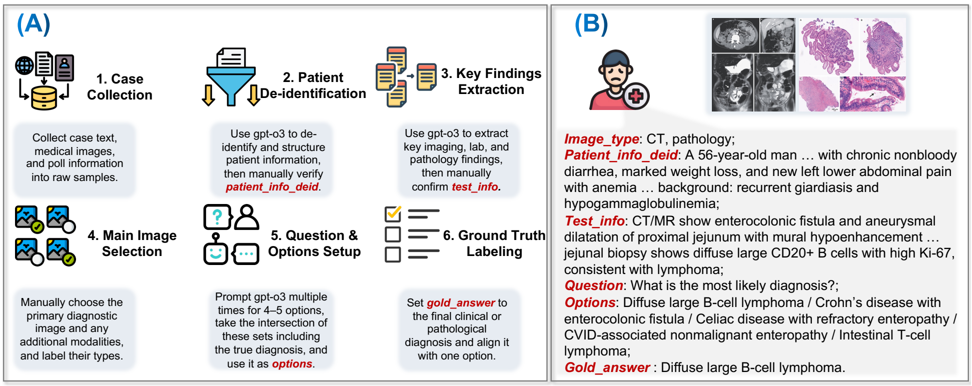

Figure 10: Multimodal dataset construction and example. (A) The six-step pipeline for collecting and standardizing multimodal diagnostic cases. (B) A finalized sample including primary image, image type, de-identified patient information, key test findings, question, options, and standard answer formulation.

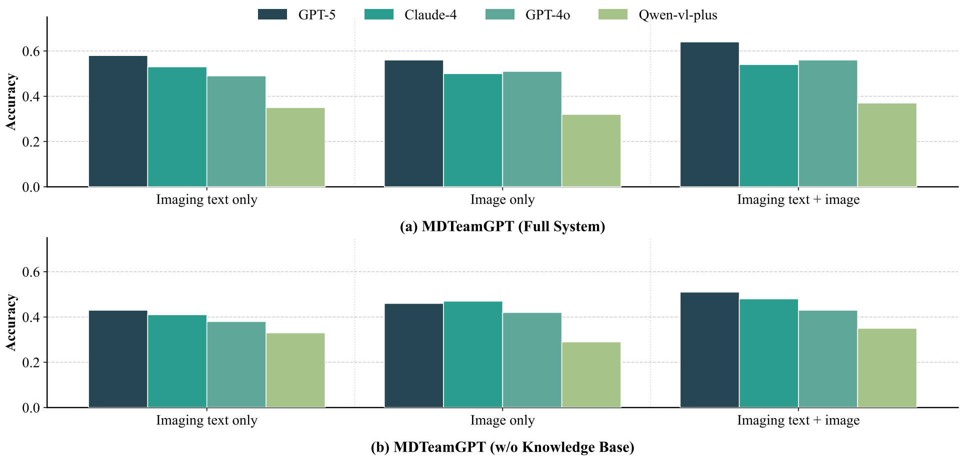

Figure 11: Modality Ablation Study. Accuracy comparison across three input settings: *Imaging text only*, *Image only*, and *Imaging text + image*. (a) Performance of the full MDTeamGPT system. (b) Performance of MDTeamGPT without the Knowledge Base (KB) module. The results highlight the model's robustness and the additive value of visual information and external knowledge.

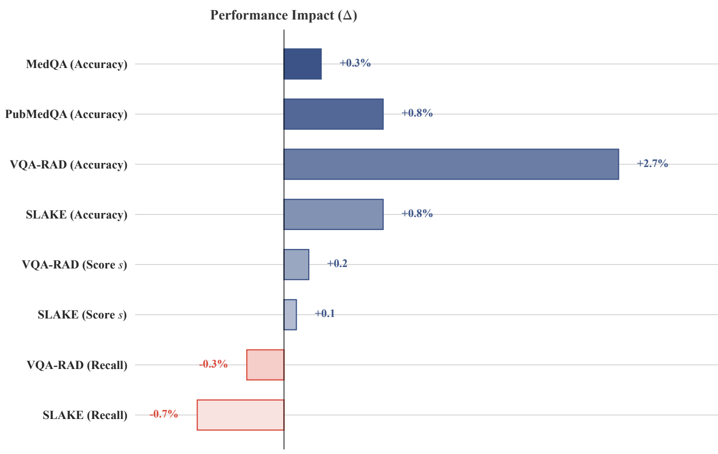

Figure 12: Impact of tool integration on MDTeamGPT. The chart displays absolute changes (Δ) in Accuracy, Recall, and Score (s). Positive values indicate performance gains, while negative values reflect a slight decrease in recall.

(i) **Case Collection.** We integrate the multimodal subset of AgentClinic with newly retrieved NEJM Case Challenge cases. Selection criteria require each case to include clear imaging evidence and a diagnostic question with an unambiguous ground-truth answer. Duplicate entries are removed, and diversity across modalities such as CT, MRI, and pathology images is ensured.

(ii) **Patient De-identification.** We employ `gpt-o3` to structurally process the original reports and remove sensitive information, including names and dates. Key demographic attributes and medical history are organized into de-identified fields. The processed records are manually reviewed by medical experts to balance privacy protection with the preservation of diagnostic integrity.

(iii) **Key Findings Extraction.** Using `gpt-o3`, we extract salient positive findings from imaging studies and laboratory tests into structured summaries. While the original professional reports are preserved in full, these imaging evidence summaries are stored separately. This separation enables precise experimental control over whether explicit textual descriptions of visual findings are provided to the model.

(iv) **Main Image Selection.** For each case, researchers select the most diagnostically informative representative image and annotate its modality. Image resolution and size are standardized to ensure consistent inputs. Lesion descriptions in the text are carefully aligned with the selected images at the anatomical level to avoid information mismatch.

(v) **Question & Options Setup.** We generate differential diagnosis candidate lists using `gpt-o3`, followed by manual screening. The final option set contains four to five mutually exclusive candidates. We enforce that the correct diagnosis is included among the options to ensure valid evaluation.

(vi) **Ground Truth Labeling.** The final clinical diagnosis is mapped to its corresponding option and recorded as the ground-truth label. Disease names are normalized by medical experts to ensure exact one-to-one correspondence between the ground-truth labels and the option texts.

Through this pipeline, each sample contains a complete clinical narrative, an independent imaging summary, and authentic medical images. This structured design provides a flexible experimental foundation for analyzing the practical role of medical imaging in MDTeamGPT-based consultations.

C.10.2 Modality Ablation Study

This subsection aims to examine the impact of different modalities on the accuracy of MDTeamGPT. To this end, we conduct systematic evaluations using the two specialized knowledge bases constructed in Section 4.2 and Appendix C.5. As shown in Figure 11, the joint input of imaging reports and raw images achieves the best performance across all backbone models. We observe that the imaging-report-only setting exhibits surprisingly strong performance, trailing the full multimodal input by only a small margin and substantially outperforming the image-only configuration.

Notably, in the absence of an external knowledge base (Figure 11 (b)), the "Image only" setting performs slightly better than "Imaging text only." This suggests that without knowledge assistance, the model tends to rely on intuitive visual perception rather than obscure clinical terminology. However, this result remains significantly lower than that of the full system (Figure 11 (a)), reaffirming the indispensable role of the knowledge base in high-precision multimodal diagnosis, as it preserves experience in medical image interpretation.

These results reveal clear differences in information density across modalities in medical diagnosis. Imaging reports distilled by expert clinicians encapsulate highly concentrated diagnostic semantics and are often sufficient to support a complete reasoning chain. In contrast, within this framework, the visual modality primarily serves a role of evidence anchoring and complementary validation. MDTeamGPT therefore demonstrates a mechanism in which text forms the backbone of reasoning while vision provides validation support: rather than relying solely on visual features, the system enhances diagnostic confidence through multimodal alignment. This property allows MDTeamGPT to retain high practical value in real-world clinical settings, even when certain modalities are missing.

D Tools Usage

As illustrated in Figure 12, the incorporation of external tools leads to comprehensive improvements across fundamental metrics, particularly Accuracy. In particular, the significant increase in the Semantic Score (s) indicates that multimodal evidence **actively** retrieved by tools effectively enhances the clarity and completeness of reasoning chains, thereby providing a more robust logical foundation for the final answers. Notably, **however**, a slight decline in Recall is observed. This is primarily attributed to the enriched evidence chains, which introduce greater specific constraints into the reasoning process. These constraints prompt the model to generate more focused and precise responses, resulting in a reasonable convergence in the breadth of open-ended generation.

E Algorithmic Framework

Algorithm 1 MDTeamGPT: Multi-Round Medical Consultation with Residual Context

Require: Patient Context C, Query Q, Specialist Set $\mathcal{S}$, Max Rounds T
Ensure: Final Diagnosis D

Initialize: Residual Context List $R \leftarrow [\,]$, Consensus $\leftarrow$ **False**
Definition: Discussion Rounds $i \in \{1, \ldots, T\}$
for $i = 1$ **to** T **do**
  Current Round: i
  Phase 1: Residual Context Retrieval
  if $i = 1$ **then**
    $C_{input} \leftarrow \emptyset$ ▷ Initial round
  else if $i = 2$ **then**
    $C_{input} \leftarrow R[1]$ ▷ Access summary S_1^6
  else
    $C_{input} \leftarrow \{R[i-2], R[i-1]\}$ ▷ Access window: S_{i-2}^6 and S_{i-1}^6
  end if
  Phase 2: Specialist Consultation Layer
  $O_i \leftarrow \emptyset$ ▷ Set of opinions in round i
  for all $k \in \mathcal{S}$ **do**
    $S_{i,k} \leftarrow$ SPECIALISTLLM$_k(C, Q, C_{input})$
    $O_i \leftarrow O_i \cup \{S_{i,k}\}$
  end for
  Phase 3: Context Engineering Layer
  $S_i^6 \leftarrow$ LEADPHYSICIAN(O_i) ▷ Generate 6-dim structured summary
  PUSH(R, S_i^6) ▷ Store S_i^6 into history
  Phase 4: Consensus Verification
  if CHECKCONSENSUS(O_i) **then**
    Consensus $\leftarrow$ **True**
    $D \leftarrow$ EXTRACTANSWER(O_i)
    break
  end if
end for
Final Decision
if not Consensus **then**
  $D \leftarrow$ MAJORITYRULE(O_T)
  if Tie exists in D **then**
    $D \leftarrow$ REFLECTOR(C, Q, R)
  end if
end if
return D

Algorithm 2 Decision Making, Evolution, and Retrieval Strategy

Require: Patient Context (C, Q), Final Opinions O_{final}, History R, Ground Truth G
Ensure: Final Conclusion F

Phase 1: Determination of Final Conclusion (F)
if Question Type is **Closed-Ended then**
  $F \leftarrow$ MAJORITYRULE(O_{final}) ▷ Rule-based voting
else ▷ Open-Ended Inquiry
  if Deadlock (Max Rounds Reached) **then**
    ▷ Reflector synthesizes full history
    $F \leftarrow$ REFLECTORJUDGE(C, Q, R)
  else
    ▷ Reflector checks consistency
    $F \leftarrow$ REFLECTORCONSISTENCY(O_{final})
  end if
end if
Phase 2: Knowledge Base Evolution
if $F == G$ **then** ▷ Outcome is Accurate
  $record \leftarrow$ COT_REVIEWER$(C, Q, S_{\text{final}}^6)$
  STORE$(record, \textit{CorrectKB})$
else ▷ Outcome is Incorrect
  $record \leftarrow$ COT_REVIEWER$(C, Q, \text{ErrorReflection})$
  STORE$(record, \textit{ChainKB})$
end if
Phase 3: Retrieval Strategy (For Future Cases)
$docs \leftarrow$ RETRIEVETOPK$(C, Q, \textit{CorrectKB}, \textit{ChainKB})$
Round 1 (Independent Reasoning):
CONSULTATION(P_{base}) ▷ KBs are NOT used
if Consensus Reached **then**
  REFLECTIVEVALIDATION$(F, docs)$
else
Round 2+ (Conflict Resolution):
  $P_{enhanced} \leftarrow$ ENHANCE$(P_{base}, docs)$ ▷ Inject Knowledge
  CONSULTATION$(P_{enhanced})$
end if