MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning

Anonymous authors
Paper under double-blind review

Abstract

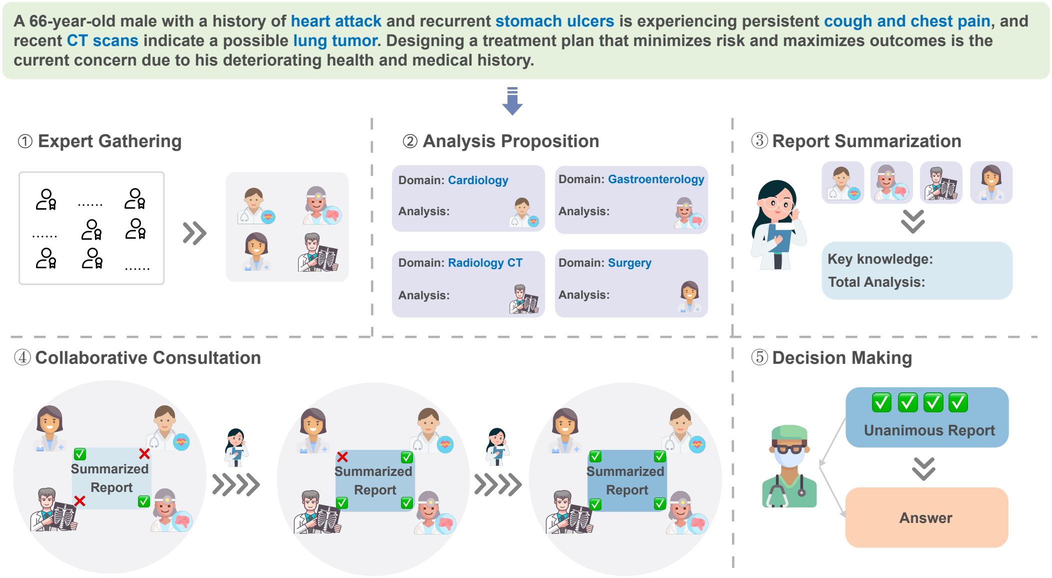
Large Language Models (LLMs), despite their remarkable progress across various general domains, encounter significant barriers in medicine and healthcare. This field faces unique challenges such as domain-specific terminologies and reasoning over specialized knowledge. To address these issues, we propose a novel **M**ulti-disciplinary **C**ollaboration (**MC**) framework for the medical domain that leverages role-playing LLM-based agents who participate in a collaborative multi-round discussion, thereby enhancing LLM proficiency and reasoning capabilities. This training-free and interpretable framework encompasses five critical steps: gathering domain experts, proposing individual analyses, summarising these analyses into a report, iterating over discussions until a consensus is reached, and ultimately making a decision. Our work focuses on the zero-shot setting, which is applicable in real-world scenarios. Experimental results on nine datasets (MedQA, MedMCQA, PubMedQA, and six subtasks from MMLU) establish that our proposed MC framework excels at mining and harnessing the medical expertise within LLMs, as well as extending its reasoning abilities.

1 Introduction

Large language models (LLMs) (Brown et al., 2020; Scao et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; OpenAI, 2023) have exhibited notable generalization abilities across a wide range of tasks and applications (Lu et al., 2023; Zhou et al., 2023; Park et al., 2023), with these capabilities stemming from their extensive training on vast comprehensive corpora covering diverse topics. However, in real-world scenarios, LLMs are inclined to encounter domain-specific tasks that necessitate a combination of domain expertise and complex reasoning abilities (Moor et al., 2023; Wu et al., 2023a; Singhal et al., 2023a; Yang et al., 2023). Amidst this backdrop, a noteworthy research topic lies in the adoption of LLMs in the medical field, which has gained increasing prominence recently (Zhang et al., 2023a; Bao et al., 2023; Singhal et al., 2023a).

Two principal challenges prevent LLMs from effectively handling tasks in the medical sphere: (i) Limited *volume and specificity* of training data in medicine, compared to the vast general text data, owing to cost and privacy considerations (U.S. Department of Health and Human Services, 1996). While Google's Med-PaLM 2, a specialized medical LLM finetuned from PaLM 2, exists, it is not publicly accessible. (ii) The demand for *extensive domain knowledge* (Schmidt and Rikers, 2007) and *advanced reasoning skills* (Liévin et al., 2022) makes eliciting medical expertise via simple prompting challenging (Kung et al., 2023; Singhal et al., 2023a). Although numerous attempts have been made to enhance prompting methods, like GoT and RAG, particularly within math and coding, strategies used in the medical field have been shown to induce 'hallucinations', indicating the need for more robust approaches.

At the same time, as opposed to the conventional single *input-output* paradigms, recent research has surprisingly observed LLM-based succeeding in a broad array of tasks (Xi et al., 2023; Wang et al., 2023a). Among such work, the design of multi-agent collaboration favorably stands out by highlighting the simulation of human activities (Du et al., 2023; Liang et al., 2023; Park et al., 2023) and optimizing the collective power of multiple agents (Chen et al., 2023; Li et al., 2023a; Hong et al., 2023).



A 66-year-old male with a history of heart attack and recurrent stomach ulcers is experiencing persistent cough and chest pain, and recent CT scans indicate a possible lung tumor. Designing a treatment plan that minimizes risk and maximizes outcomes is the current concern due to his deteriorating health and medical history.

Figure 1: Diagram of our proposed MC framework. Given a medical question as input, the framework performs reasoning in five stages: (i) expert gathering; (ii) analysis proposition; (iii) report summarization; (iv) collaborative consultation; and (v) decision making.

Through the design of multi-agent collaboration, the expertise implicitly embedded within LLMs, or that the model has encountered during its training, which may not be readily accessible via traditional prompting, is effectively brought to the fore. This process subsequently enhances the model's reasoning capabilities throughout multiple rounds of interaction (Wang et al., 2023a;b; Du et al., 2023; Fu et al., 2023).

Motivated by these notions, we pioneer a **Multi-disciplinary Collaboration (MC)** framework specifically tailored to the clinical domain. Our objective centers on unveiling the intrinsic medical knowledge embedded in LLMs and reinforcing reasoning proficiency in an interpretable, training-free manner. As is shown in Figure 1, the MC framework is based on five pivotal steps (i) Expert gathering: gather experts from distinct disciplines according to the clinical question. (ii) Analysis proposition: domain experts put forward their analyses with their expertise. (iii) Report summarization: compose a summarized report based on a previous series of analyses. (iv) Collaborative consultation: engage the experts in discussions over the summarized report. The report will be revised iteratively until an agreement from all experts is reached. (v) Decision making: derive a final decision from the unanimous report.

Having established the theoretical foundation of our approach, we conduct experiments on nine datasets Singhal et al. (2023a), including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019) and six medical subtasks from MMLU (Hendrycks et al., 2020), similar to Flan-PaLM (Singhal et al., 2023a). To better align with real-world application scenarios, our study focuses on the zero-shot setting. Encouragingly, our proposed approach outperforms settings for both chain-of-thought (CoT) and self-consistency prompting methods. Most notably, our approach demonstrates better performances under the zero-shot setting compared with the few-shot (5-shot) strong baselines.

Based on our results, we further investigate the influence of agent numbers and conduct human evaluations to pinpoint the limitations and issues prevalent in our approach. We find four common categories of errors: (i) lack of domain knowledge; (ii) mis-retrieval of domain knowledge; (iii) consistency errors; and (iv) CoT errors. Further refinements focused on mitigating these particular shortcomings would enhance the model's proficiency and reliability.

Our contributions are summarized as follows: (a) To the best of our knowledge, we are the first to propose a multi-agent framework within the medical domain. Our method harnesses role-playing and collaborative agent discussion for avoiding hallucinations and ensuring faithful CoT reasoning. This strategic approach notably enhances the interpretability of models. (b) Experimental results on nine datasets demonstrate the general effectiveness of our proposed MC framework. (c) We identify and categorize common error types in our approach through rigorous human evaluation to shed light on future studies.

2 Related Work

2.1 LLMs in Medical Domains

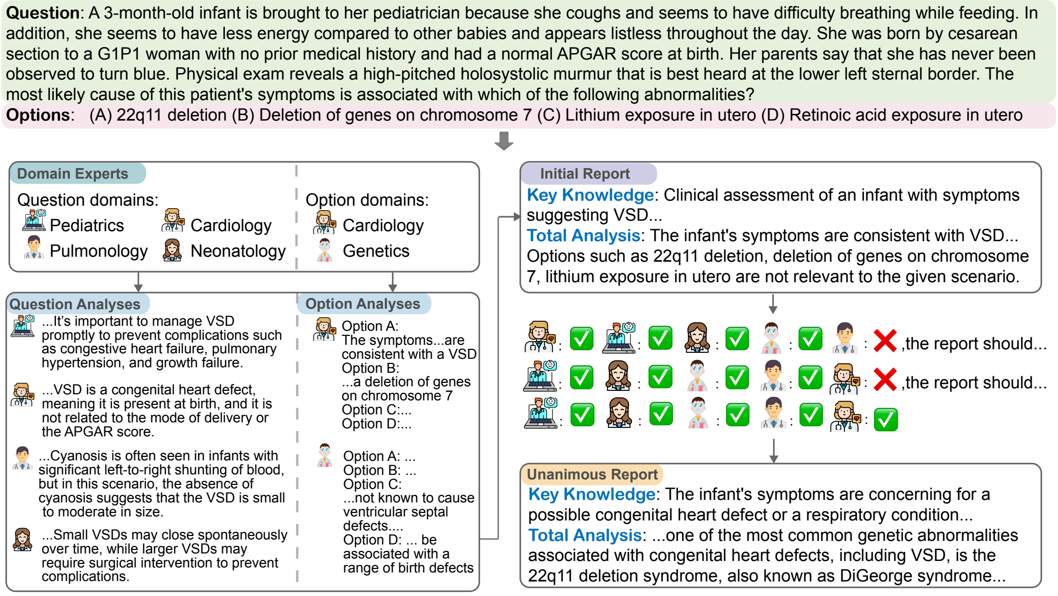
Recent years have seen remarkable progress in the application of LLMs (Wu et al., 2023a; Singhal et al., 2023a; Yang et al., 2023), with a particularly notable impact on the medical field (Bao et al., 2023; Nori et al., 2023; Rosoł et al., 2023). Although LLMs have demonstrated their potential in distinct medical applications encompassing diagnostics (Singhal et al., 2023a; Han et al., 2023), genetics (Duong and Solomon, 2023; Jin et al., 2023), pharmacist (Liu et al., 2023), and medical evidence summarization (Tang et al., 2023a;b; Shaib et al., 2023), concerns persist when LLMs encounter clinical inquiries that demand intricate medical expertise and decent reasoning abilities (Umapathi et al., 2023; Singhal et al., 2023a). Thus, it is of crucial importance to further arm LLMs with enhanced clinical reasoning capabilities. Currently, there are two major lines of research on LLMs in medical domains, tool-augmented methods and instruction-tuning methods.

For tool-augmented approaches, recent studies rely on external tools to acquire additional information for clinical reasoning. For instance, GeneGPT (Jin et al., 2023) guided LLMs to leverage the Web APIs of the National Center for Biotechnology Information (NCBI) to meet various biomedical information needs. Zakka et al. (2023) proposed Almanac, a framework that is augmented with retrieval capabilities for medical guidelines and treatment recommendations. Kang et al. (2023) introduced a method named KARD to improve small LMs on specific domain knowledge by fine-tuning small LMs on the rationales generated from LLMs and augmenting small LMs with external knowledge from a non-parametric memory.

Current instruction tuning research predominantly leverages external clinical knowledge bases and self-prompted data to obtain instruction datasets (Tu et al., 2023; Zhang et al., 2023b; Singhal et al., 2023b; Tang et al., 2023c). These datasets are then employed to fine-tune LLMs within the medical field (Singhal et al., 2023b). Some of these models utilize a wide array of datasets collected from medical and biomedical literature, fine-tuned with specialized or open-ended instruction data (Li et al., 2023b; Singhal et al., 2023b). Others focus on specific areas such as traditional Chinese medicine or large-scale, diverse medical instruction data to enhance their medical proficiency (Tan et al., 2023; Zhang et al., 2023a). Unlike these methods, our work emphasizes harnessing latent medical knowledge intrinsic to LLMs and improving reasoning in a training-free setting.

2.2 LLM-based Multi-agent Collaboration

The development of LLM-based agents has made significant progress in the community by endowing LLMs with the ability to perceive surroundings and make decisions individually (Wang et al., 2023c; Yao et al., 2022; Nakajima, 2023; Xie et al., 2023; Zhou et al., 2023). Beyond the initial single-agent mode, the multi-agent pattern has garnered increasing attention recently (Xi et al., 2023; Li et al., 2023a; Hong et al., 2023) which further explores the potential of LLM-based agents by learning from multi-turn feedback and cooperation. In essence, the key to LLM-based multi-agent collaboration is the simulation of human activities such as role-playing (Wang et al., 2023a; Hong et al., 2023) and communication (Wu et al., 2023b; Qian et al., 2023; Li et al., 2023c;d). For instance, Solo Performance Prompting (SPP) (Wang et al., 2023a) managed to combine the strengths of multiple minds to improve performance by dynamically identifying and engaging multiple personas over the course of task-solving. Camel (Li et al., 2023c) leveraged role-playing to enable chat agents to communicate with each other for task completion. Several recent works attempt to incorporate adversarial collaboration including debates (Du et al., 2023; Xiong et al., 2023) and negotiation (Fu et al., 2023) among multiple agents to further boost performance. Liang et al. (2023) proposed a multi-agent debate framework in which various agents put forward their statements in a *tit for tat* pattern. Inspired by the multi-disciplinary consultation mechanism which is common and effective in hospitals, we are thus inspired to apply this mechanism to medical reasoning tasks through LLM-based multi-agent collaboration.



Question: A 3-month-old infant is brought to her pediatrician because she coughs and seems to have difficulty breathing while feeding. In addition, she seems to have less energy compared to other babies and appears listless throughout the day. She was born by cesarean section to a G1P1 woman with no prior medical history and had a normal APGAR score at birth. Her parents say that she has never been observed to turn blue. Physical exam reveals a high-pitched holosystolic murmur that is best heard at the lower left sternal border. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?

Options: (A) 22q11 deletion (B) Deletion of genes on chromosome 7 (C) Lithium exposure in utero (D) Retinoic acid exposure in utero

...It's important to manage VSD promptly to prevent complications such as congestive heart failure, pulmonary hypertension, and growth failure.

...VSD is a congenital heart defect, meaning it is present at birth, and it is not related to the mode of delivery or the APGAR score.

...Cyanosis is often seen in infants with significant left-to-right shunting of blood, but in this scenario, the absence of cyanosis suggests that the VSD is small to moderate in size.

...Small VSDs may close spontaneously over time, while larger VSDs may require surgical intervention to prevent complications.

Option A:
The symptoms...are consistent with a VSD
Option B:
...a deletion of genes on chromosome 7
Option C:...
Option D:...

Option A: ...
Option B: ...
Option C:
...not known to cause ventricular septal defects....
Option D: ... be associated with a range of birth defects

Key Knowledge: Clinical assessment of an infant with symptoms suggesting VSD...
Total Analysis: The infant's symptoms are consistent with VSD... Options such as 22q11 deletion, deletion of genes on chromosome 7, lithium exposure in utero are not relevant to the given scenario.

: ✓ : ✓ : ✓ : ✓ : ✗ ,the report should...
: ✓ : ✓ : ✓ : ✓ : ✗ ,the report should...
: ✓ : ✓ : ✓ : ✓ : ✓
Unanimous Report

Key Knowledge: The infant's symptoms are concerning for a possible congenital heart defect or a respiratory condition...
Total Analysis: ...one of the most common genetic abnormalities associated with congenital heart defects, including VSD, is the 22q11 deletion syndrome, also known as DiGeorge syndrome...

Figure 2: Illustrative example of our proposed Multi-disciplinary Collaboration (MC) framework.

3 Method

This section presents the details of our proposed Multi-disciplinary Collaboration (MC) framework. Figure 1 and 2 give an overview and an illustrative example of its pipeline. Our proposed MC framework works in five stages: (i) expert gathering: assemble experts from various disciplines based on the clinical question; (ii) analysis proposition: domain experts present their own analyses with their expertise; (iii) report summarization: develop a report summary on the basis of previous analyses; (iv) collaborative consultation: hold a consultation over the summarized report with the experts. The report will be revised repeatedly until every expert has given their approval. (v) decision making: derive a final decision from the unanimous report.

3.1 Expert Gathering

Given a clinical question q and a set of options $op = \{o_1, o_2, \ldots, o_k\}$, the goal of the Expert Gathering stage is to recruit a group of question domain experts $\mathcal{QD} = \{qd_1, qd_2, \ldots, qd_m\}$ and option domain experts $\mathcal{OD} = \{od_1, od_2, \ldots, od_n\}$. Specifically, we assign a role to the model and provide instructions to guide the model output to the corresponding domains based on the input question and options, respectively:

$$\begin{aligned} \mathcal{QD} &= \text{LLM}\left(q, \mathsf{r}_{\mathsf{qd}}, \mathtt{prompt}_{\mathsf{qd}}\right), \\ \mathcal{OD} &= \text{LLM}\left(q, op, \mathsf{r}_{\mathsf{od}}, \mathtt{prompt}_{\mathsf{od}}\right), \end{aligned} \tag{1}$$

where $\left(\mathsf{r}_{\mathsf{qd}}, \mathtt{prompt}_{\mathsf{qd}}\right)$ and $\left(\mathsf{r}_{\mathsf{od}}, \mathtt{prompt}_{\mathsf{od}}\right)$ stand for the system role and guideline prompt to gather domain experts for the question q and options op.

Table 1: Summary of the Datasets.

Dataset	Format	Choice	Testing Size	Domain
MedQA	Question + Answer	A/B/C/D	1273	US Medical Licensing Examination
MedMCQA	Question + Answer	A/B/C/D and Explanations	6.1K	AIIMS and NEET PG entrance exams
PubMedQA	Question + Context + Answer	Yes/No/Maybe	500	PubMed paper abstracts
MMLU	Question + Answer	A/B/C/D	1089	Graduate Record Examination & US Medical Licensing Examination

3.2 Analysis Proposition

After gathering domain experts for the question q and options op, we aim to inquire experts to generate corresponding analyses prepared for later reasoning: $\mathcal{QA} = \{qa_1, qa_2, \ldots, qa_m\}$ and $\mathcal{OA} = \{oa_1, oa_2, \ldots, oa_n\}$.

Question Analyses Given a question q and a question domain $qd_i \in \mathcal{QD}$, we ask LLM to serve as an expert specialized in domain qd_i and derive the analyses for the question q following the guideline prompt $\mathtt{prompt}_{\mathsf{qa}}$:

$$qa_i = \text{LLM}\left(q, qd_i, \mathsf{r}_{\mathsf{qa}}, \mathtt{prompt}_{\mathsf{qa}}\right). \tag{2}$$

Option Analyses Now that we have an option domain od_i and question analyses $\mathcal{QA}$, we can further analyze the options by taking into account both the relationship between the options and the relationship between the options and question. Concretely, we deliver the question q, the options op, a specific option domain $od_i \in \mathcal{OD}$, and the question analyses $\mathcal{QA}$ to the LLM:

$$oa_i = \text{LLM}\left(q, op, od_i, \mathcal{QA}, \mathsf{r}_{\mathsf{oa}}, \mathtt{prompt}_{\mathsf{oa}}\right). \tag{3}$$

3.3 Report Summarization

In the Report Summarization stage, we attempt to summarize and synthesize previous analyses from various domain experts $\mathcal{QA} \cup \mathcal{OA}$. Given question analyses $\mathcal{QA}$ and option analyses $\mathcal{OA}$, we ask LLMs to play the role of a medical report assistant, allowing it to generate a synthesized report by extracting key knowledge and total analysis based on previous analyses:

$$Repo = \text{LLM}\left(\mathcal{QA}, \mathcal{OA}, \mathsf{r}_{\mathsf{rs}}, \mathtt{prompt}_{\mathsf{rs}}\right). \tag{4}$$

3.4 Collaborative Consultation

Since we have a preliminary summary report $Repo$, the objective of the Collaborative Consultation stage is to engage distinct domain experts in multiple rounds of discussions and ultimately render a summary report that is recognized by all experts. During each round of discussions, the experts give their votes (*yes*/*no*) as well as modification opinions if they vote *no* for the current report. Afterward, the report will be revised based on the modification opinions. Specifically, during the i-th round of discussion, we note the modification comments from the experts as Mod_i, then we can acquire the updated report as $Repo_i = \text{LLM}\left(Repo_{i-1}, Mod_i, \mathtt{prompt}_{\mathsf{mod}}\right)$. In this way, the discussions are held iteratively until all experts vote *yes* for the final report $Repo_f$.

3.5 Decision Making

In the end, we demand LLM act as a medical decision maker to derive the final answer to the clinical question q referring to the unanimous report $Repo_f$:

$$ans = \text{LLM}\left(q, op, Repo_f, \mathsf{r}_{\mathsf{dm}}, \mathtt{prompt}_{\mathsf{dm}}\right). \tag{5}$$

4 Experiments

4.1 Setup

Tasks and Datasets. We evaluate our MC framework on three benchmark datasets MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019), as well as six subtasks most relevant to the medical domain from MMLU datasets (Hendrycks et al., 2020) including anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and college biology. Table 1 summarizes the data statistics. MedQA consists of USMLE-style questions with four or five possible answers. MedMCQA encompasses four-option multiple-choice questions from Indian medical entrance examinations (AIIMS/NEET). MMLU (Massive Multitask Language Understanding) covers 57 subjects across various disciplines, including STEM, humanities, social sciences, and many others. The scope of its assessment stretches from elementary to advanced

Table 2: Main results on MedQA, MedMCQA, PubMedQA, and six subtasks from MMLU including anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and college biology (Acc). SC denotes the self-consistency prompting method. Results in **bold** are the best performances.

Method	MedQA	MedMCQA	PubMedQA	Anatomy	Clinical knowledge	College medicine	Medical genetics	Professional medicine	College biology	Avg.
				`Flan-Palm`						
Few-shot CoT	60.3	53.6	77.2	66.7	77.0	83.3	75.0	76.5	71.1	71.2
Few-shot CoT + SC	67.6	57.6	75.2	71.9	80.4	88.9	74.0	83.5	76.3	75.0
				`GPT-3.5`						
**few-shot setting*										
Few-shot	54.7	56.7	67.6	65.9	71.3	59.0	72.0	75.7	73.6	66.3
Few-shot CoT	55.3	54.7	71.4	48.1	65.7	55.5	57.0	69.5	61.1	59.8
Few-shot CoT + SC	62.1	58.3	73.4	70.4	76.2	69.8	78.0	79.0	77.2	71.6
**zero-shot setting*										
Zero-shot	54.3	56.3	73.7	61.5	76.2	63.6	74.0	75.4	75.0	67.8
Zero-shot CoT	44.3	47.3	61.3	63.7	61.9	53.2	66.0	62.1	65.3	58.3
Zero-shot CoT + SC	61.3	52.5	**75.7**	**71.1**	75.1	68.8	76.0	**82.3**	75.7	70.9
myblue`MC framework` (**Ours**)	**64.1**	**59.3**	72.9	65.2	**77.7**	**69.8**	**79.0**	82.1	**78.5**	**72.1**
gray!25	`GPT-4`									
**few-shot setting*										
Few-shot	76.6	70.1	73.4	79.3	89.5	75.6	**93.0**	91.5	91.7	82.3
Few-shot CoT	73.3	63.2	74.9	75.6	89.9	61.0	79.0	79.8	63.2	73.3
Few-shot CoT + SC	82.9	73.1	75.6	80.7	90.0	**88.2**	90.0	95.2	93.0	85.4
**zero-shot setting*										
Zero-shot	73.0	69.0	76.2	78.5	83.3	75.6	90.0	90.0	90.0	80.6
Zero-shot CoT	61.8	69.0	71.0	82.1	85.2	80.8	92.0	93.5	91.7	80.8
Zero-shot CoT + SC	74.5	70.1	75.3	80.0	86.3	81.2	**93.0**	94.8	91.7	83.0
myblue`MC framework` (**Ours**)	**83.7**	**74.8**	**76.8**	**83.5**	**91.0**	87.6	**93.0**	**96.0**	**94.3**	**86.7**

professional levels, evaluating both world knowledge and problem-solving capabilities. While the subject areas tested are diverse, encompassing traditional fields like mathematics and history, as well as more specialized areas like law and ethics, we deliberately limit our selection to the sub-subjects within the medical domain for this exercise, following (Singhal et al., 2023a).

Implementation. We utilize the popular and publicly available GPT-3.5-Turbo and GPT-4 (OpenAI, 2023) from Azure OpenAI Service.[1] All experiments are conducted in the **zero-shot** setting. The temperature is set to 1.0 and *top_p* to 1.0 for all generations. The number of SC iterations is 5. The number k of options is 4 except for PubMedQA (3). The numbers of domain experts for the question and options are set as: $m = 5, n = 2$ except for PubMedQA ($m = 4, n = 2$). We randomly sample 300 examples for each dataset and conduct experiments on them. Statistically, the cost of our method is \$1.41 for 100 QA examples (about ¢1.4 per question) and the inference time per example is about $40s$.

4.2 Main Results

Table 2 presents the main results on the nine datasets, including MedQA, MedMCQA, PubMedQA, and six subtasks from MMLU. We compare our method with several baselines in both zero-shot and few-shot settings. Notably, our proposed MC framework outperforms the zero-shot baseline methods by a large margin, indicating the effectiveness of our MC framework in real-world application scenarios. Furthermore, our approach achieves comparable performance under the zero-shot setting compared with the strong baseline *Few-shot CoT+SC*. Interestingly, we notice that adding CoT results in performance degradation in some cases. We discover that using CoT alone can easily lead to *hallucinations* in specific domains, while our multi-agent role-playing method is able to circumvent this weakness.

[1] `https://learn.microsoft.com/en-us/azure/ai-services/openai/`

Table 3: Ablation study for different processes in our MC framework. Anal: Analysis proposition, Summm: Report summarization, Cons: Collaborative consultation.

Method	Accuracy(%)
Direct Prompting	49.0
CoT Prompting	55.0
w/ MedAgents	
+ Anal	62.0(↑ 7.0)
+ Anal & Summ	65.0(↑ 10.0)
+ Anal & Summ & Cons	67.0(↑ 12.0)

Table 4: Optimal number of agents on MedQA, MedMCQA, PubMedQA, and MMLU.

Dataset	MedQA	MedMCQA	PubMedQA	MMLU
#Question agents	5	5	4	5
#Option agents	2	2	2	2

5 Analysis

5.1 Ablation Study

Since our MC framework simulates a multi-disciplinary collaboration process that contains multiple intermediate steps, a natural question is whether each intermediate step contributes to the ultimate result. To investigate this, we ablate three major processes, namely *analysis proposition*, *report summarization* and *collaborative consultation*. Results in Table 3 show that all of these processes are non-trivial. Notably, the proposition of MedAgents substantially boosts the performance (i.e., 55.0%→62.0%), whereas the subsequent processes achieve relatively slight improvements over the previous one (i.e., 62.0%→65.0/67.0%). This suggests that the initial role-playing agents are responsible for exploring medical knowledge of various levels and aspects within LLMs, while the following processes play a role in further verification and revision.

5.2 Number of agents

As our proposed MC framework involves multiple agents that play certain roles to acquire the ultimate answer, we explore how the number of collaborating agents influences the overall performance. We vary the number of question agents and option agents while fixing the other variable to observe the performance trends on the MedQA dataset. Figure 3 and Table 4 illustrate the optimal number of different agents. Besides, our key observation lies in that the performance improves significantly with the introduction of any number of expert agents compared to our baseline, thus verifying the consistent contribution of multiple expert agents. [2]

5.3 Error Analysis

Based on our results, we conduct a human evaluation to pinpoint the limitations and issues prevalent in our model. We distill these errors into four major categories: (i) **Lack of Domain Knowledge**: these errors occur when the model demonstrates an inadequate understanding of the specific medical knowledge necessary to provide an accurate response; (ii) **Mis-retrieval of Domain Knowledge**: the model has the necessary domain knowledge but fails to retrieve or apply it correctly in the given context; (iii) **Consistency Errors**: such errors arise when the model provides differing responses to the same statement. The inconsistency suggests confusion in the model's understanding or application of the underlying knowledge; (iv) **CoT Errors**: errors under this category pertain to flawed reasoning sequences or lapses in logical cohesion. The model may form and follow inaccurate rationales, leading to incorrect conclusions.

[2] We find that the optimal number of agents is relatively consistent across different datasets, pointing to its potential applicability to other datasets beyond those we test on.

We randomly select 40 error cases in MedQA and MedMCQA datasets and analyze the percentage of different categories in these error cases. As is shown in Figure 4, the majority (77%) of the error examples are due to confusion about the domain knowledge (including the lack and mis-retrieval of domain knowledge), which illustrates that although our method further mines medical knowledge concealed within LLMs via multi-disciplinary consultation, there still exists a portion of domain knowledge that is explicitly beyond the intrinsic knowledge of LLMs, leading to a bottleneck of our proposed MC framework. As a result, our analysis sheds light on future directions to mitigate the aforementioned drawbacks and further strengthen the model's proficiency and reliability. One potential solution is incorporating credible medical knowledge sources to complement the existing shortcomings.

To illustrate the error examples more intuitively, we select four typical samples from the four error categories, which can be shown in Figure 4: (i) The first error is due to a lack of domain knowledge regarding *cutaneous larva migrans*, whose symptoms are not purely *hypopigmented rash*, as well as the fact that *skin biopsy* is not an appropriate test method, which results in the hallucination phenomenon. (ii) The second error is caused by mis-retrieval of domain knowledge, wherein the fact in green is not relevant to *Valsalva maneuver*. (iii) The third error is attributed to consistency errors, where the model incorrectly regards *20 mmHg within 6 minutes* and *20 mmHg within 3 minutes* as the same meaning. (iv) The fourth error is provoked by incorrect inference about the relevance of a fact and option A in CoT.

Our work cuts through the challenges observed in medicine domains, where the conventional Chain-of-thought (CoT) can induce 'hallucinations', or the generation of incorrect and irrelevant information due to limited understanding. Notably, in medical Query Answering (QA), CoTâĂŹs step-by-step logic often falters at generating accurate responses, a shortfall accentuated by a domain knowledge deficit. Our findings indicate a whopping 77% of errors could be traced back to domain knowledge shortfalls rather than CoT rationale.

However, our solution lies in the innovative use of role-playing within the MC framework, which equips the model to reason with precise knowledge. Surprisingly, this negates the need for Retrieval-Augmented Generation (RAG) in enhancing domain knowledge for medical QA, setting forth a novelty in our field.

6 Conclusion

This paper presents a novel multi-disciplinary collaboration framework for the medical domain that leverages role-playing LLM-based agents who participate in a collaborative multi-round discussion. The framework is training-free and interpretable, encompassing five critical steps: gathering domain experts, proposing individual analyses, summarising these analyses into a report, iterating over discussions until a consensus is reached, and ultimately making a decision. Experimental results on nine datasets show that our proposed framework outperforms all the zero-shot baselines by a large margin and demonstrates comparable performance with the strong few-shot baseline with self-consistency. According to our human evaluations on error cases, future studies may further improve the framework by mitigating the mistakes due to the lack of domain knowledge, mis-retrieval of domain knowledge, and addressing consistency errors and CoT errors.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100, 2022. URL `https://arxiv.org/abs/2211.05100`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311, 2022. URL `https://arxiv.org/abs/2204.02311`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971, 2023. URL `https://arxiv.org/abs/2302.13971`.

OpenAI. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. URL `https://arxiv.org/abs/2303.08774`.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. Webarena: A realistic web environment for building autonomous agents, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

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

Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. Precedent-enhanced legal judgment prediction with llm and domain-model collaboration, 2023a.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Mahdavi, Jason Wei, Hyung Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael SchÃd'rli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:1–9, 07 2023a. doi: 10.1038/s41586-023-06291-2.

Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning, 2023.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare:instruction-tuned large language models for medical application, 2023a.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. Disc-medllm: Bridging general large language models and real-world medical consultation, 2023.

U.S. Department of Health and Human Services. The hipaa privacy rule. `https://www.hhs.gov/hipaa/for-professionals/privacy/index.html`, 1996.

Henk G Schmidt and Remy MJP Rikers. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical education*, 41(12):1133–1139, 2007.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*, 2022.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, 2023a.

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

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

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

Yuan Li, Yixuan Zhang, and Lichao Sun. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents, 2023a.

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

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv: 2310.00746*, 2023b.

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

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR, 2022.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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

Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. Evaluation of the performance of gpt-3.5 and gpt-4 on the medical final examination. *medRxiv*, pages 2023–06, 2023.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander LÃűser, Daniel Truhn, and Keno K. Bressem. Medalpaca – an open-source collection of medical conversational ai models and training data, 2023.

Dat Duong and Benjamin D Solomon. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, pages 1–3, 2023.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*, 2023.

Zhengliang Liu, Zihao Wu, Mengxuan Hu, Bokai Zhao, Lin Zhao, Tianyi Zhang, Haixing Dai, Xianyan Chen, Ye Shen, Sheng Li, et al. Pharmacygpt: The ai pharmacist. *arXiv preprint arXiv:2307.10432*, 2023.

Xiangru Tang, Arman Cohan, and Mark Gerstein. Aligning factual consistency for clinical studies summarization through reinforcement learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 48–58, 2023a.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158, 2023b.

Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessy Li, and Byron C Wallace. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*, 2023.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R Dalal, Jennifer L Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, et al. Almanac: Retrieval-augmented language models for clinical medicine. *Research Square*, 2023.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *arXiv preprint arXiv:2305.18395*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, 2023.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023b.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. Gersteinlab at mediqa-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. *arXiv preprint arXiv:2305.05001*, 2023c.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023b.

Yang Tan, Mingchen Li, Zijie Huang, Huiqun Yu, and Guisheng Fan. Medchatzh: a better medical adviser learns from better instructions. *arXiv preprint arXiv:2309.01114*, 2023.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432*, 2023c.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Y Nakajima. Task-driven autonomous agent utilizing gpt-4, pinecone, and langchain for diverse applications. *See https://yoheinakajima. com/task-driven-autonomous-agent-utilizing-gpt-4-pinecone-and-langchain-for-diverse-applications (accessed 18 April 2023)*, 2023.

Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, et al. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634*, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023b.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023c.

Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023d.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining the inter-consistency of large language models: An in-depth analysis via debate. *arXiv e-prints*, pages arXiv–2305, 2023.

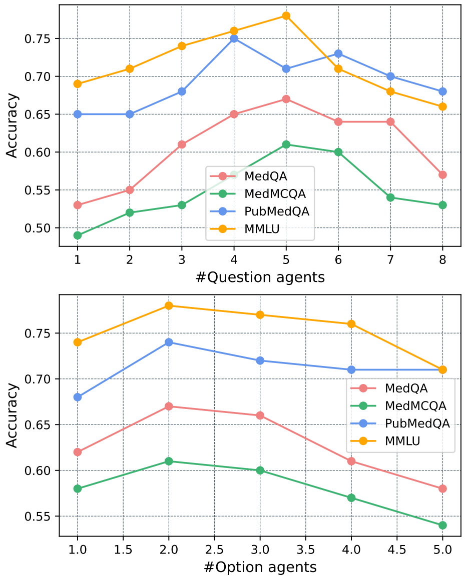


Figure 3: Influence of the number of question and option agents on various datasets.

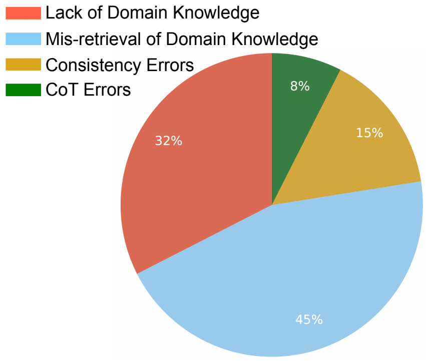


Figure 4: Ratio of different categories in error cases.

Category	Example	Interpretation
Lack of Domain Knowledge	...The hypopigmented rash ⊗ is a classic symptom of cutaneous larva migrans. To confirm the diagnosis, a skin biopsy ⊗ would be the most appropriate test.	About cutaneous larva migrans: 1. symptoms: ⊗ not simply hypopigmented rash 2. diagnostic method: ⊗ skin biopsy is not preferred
Mis-retrieval of Domain Knowledge	...The physician instructs the patient to stand from a supine position while still wearing the stethoscope. It is known as the "Valsalva maneuver" ⊗ During the Valsalva maneuver, ...	The patient is asked to merely stand from a supine position. It does not involve the Valsalva maneuver. ⊗
Consistency Errors	...Option A states that there is a decrease in systolic blood pressure of 20 mmHg within 6 minutes. This is a correct statement, as a drop in systolic blood pressure of at least 20 mmHg within 3 minutes of standing up is a diagnostic criterion for postural hypotension...	Correct statement: 20mmHg within 3 minutes Option A: 20mmHg within 6 minutes ⊗
CoT Errors	Q: Deciduous teeth do not show fluorosis because: ...(A) Placenta acts as a barrier: While it's true that placenta can act as a barrier for certain substances, this option is not relevant ⊗ to the question...	placenta can as a barrier for certain substances such as fluoride, which is part of the reason why deciduous teeth do not show fluorosis...

Figure 5: Examples of error cases from MedQA and MedMCQA datasets in four major categories including: lack of domain knowledge, mis-retrieval of domain knowledge, consistency errors, and CoT errors.