



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

Xiangru Tang^{1,*}, Anni Zou^{2,*}, Zhuosheng Zhang², Yilun Zhao¹, Xingyao Zhang,
Arman Cohan¹, Mark Gerstein¹

¹Yale University, ²Shanghai Jiao Tong University
{xiangru.tang, yilun.zhao, arman.cohan, mark.gerstein}@yale.edu

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 the reasoning over specialized knowledge. To address these obstinate issues, we propose a novel Multi-disciplinary Collaboration (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 particularly focuses on the zero-shot scenario, our results on nine data sets (MedQA, MedMCQA, PubMedQA, and six subtasks from MMLU) establish that our proposed MC framework excels at mining and harnessing the medical expertise in LLMs, as well as extending its reasoning abilities. Based on these outcomes, we further conduct a human evaluation to pinpoint and categorize common errors within our method, as well as ablation studies aimed at understanding the impact of various factors on overall performance. Our code can be found at <https://github.com/gersteinlab/MedAgents>.

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., 2023c; 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. With remarkable progress in the general domain, adaptation of LLMs to the medical field has gained increasing prominence (Zhang et al., 2023b; Bao et al., 2023; Singhal et al., 2023a).

Currently, there are two dominant challenges that hinder current LLMs from achieving medical-related tasks: (i) The volume and specificity of training data in the medical field are limited compared with general web data used to train LLMs. (ii) High performance in this field requires extensive domain knowledge and sophisticated reasoning abilities. On one hand, although the general-domain

*Contributed equally.

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.

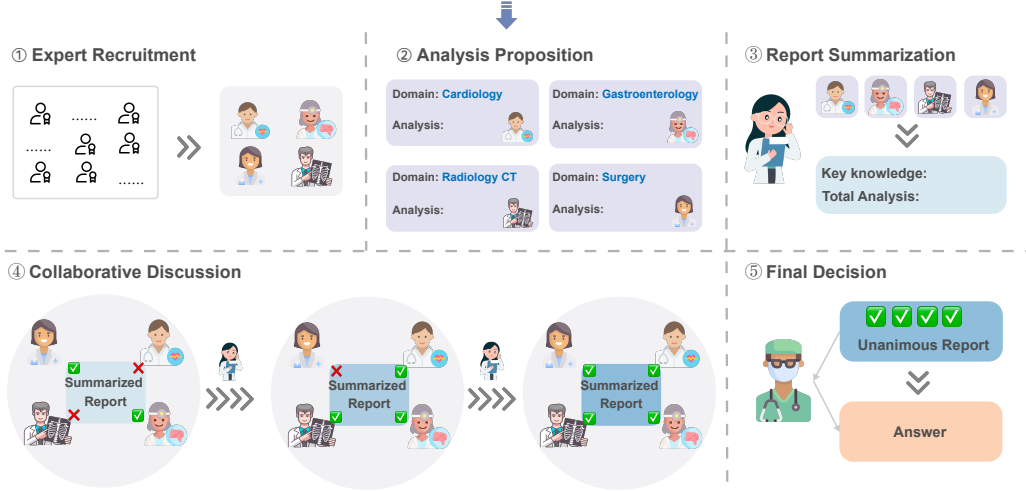


Figure 1: A Diagram of our proposed multi-disciplinary collaboration framework. Given a medical question as input, the framework performs reasoning in five stages: (i) expert gathering: distinct domain experts are assembled based on the nature of the clinical question; (ii) analysis proposition: each expert, using their specific disciplinary knowledge, puts forward an analysis; (iii) report summarization: a consolidated report is generated incorporating all experts’ analyses; (iv) collaborative discussion: the experts review, discuss, and refine the report iteratively until a consensus is reached; (v) decision making: a final decision, reflecting unanimous agreement among all experts, is derived from the report.

knowledge within LLMs has more or less permeated into distinct specific realms (Yu et al., 2022; Chen et al., 2023), it remains insufficient for models to understand or recall the required medical expertise via simple and direct prompting (Kung et al., 2023; Singhal et al., 2023a). On the other hand, given the abundance of sophisticated terminology in medical knowledge (Schmidt and Rikers, 2007), LLMs endure heightened demands when attempting to navigate this knowledge and reason upon it, potentially resulting in errors within their reasoning processes (Liévin et al., 2022).

To close this gap, there is a surging trend of methods striving to endow LLMs with enhanced proficiency in medical knowledge by instruction tuning (Han et al., 2023; Li et al., 2023d; Wu et al., 2023a; Zhang et al., 2023b). These approaches resort to either external knowledge bases (Li et al., 2023d; Wu et al., 2023a) or self-prompted data (Zhang et al., 2023b) to create instruction datasets, which are subsequently leveraged to tune LLMs, ensuring better alignment with human intent in the medical field. Nevertheless, obtaining high-quality instruction tuning data in the medical domain is expensive, prone to privacy issues (U.S. Department of Health and Human Services, 1996), and thus not scalable. On the other hand, self-generated instruction data often lack sufficient quality and may need further human verification (Xu et al., 2023). Furthermore, such instruction tuning methods inflict additional training costs and are not applicable to black-box LLMs. In addition, such methods do not necessarily emphasize improving the reasoning capabilities of the models (Liang et al., 2023).

At the same time, as opposed to the conventional single *input-output* pattern, recent research has surprisingly witnessed the success of LLM-based agents across a spectrum of tasks (Xi et al., 2023; Wang et al., 2023b). 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 coordinating the potential of multiple agents (Chen et al., 2023; Li et al., 2023c; Hong et al., 2023). 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. In turn, this process enhances the model’s reasoning capabilities over the course of multi-round interaction (Wang et al., 2023b,a; Du et al., 2023; Fu et al., 2023).

Inspired by the above ideas, we propose a **Multi-disciplinary Collaboration (MC)** framework in the clinical domain, aiming to unveil the intrinsic clinical knowledge from LLMs as well as bolster the reasoning competence in a training-free and interpretable manner. Specifically, 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 own analysis with their expertise; (iii) report summarization: compose a summarized report on the basis of 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 the 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 MultiMedQA multiple-choice dataset [Singhal et al. \(2023a\)](#), including MedQA ([Jin et al., 2021](#)), MedMCQA ([Pal et al., 2022](#)), PubMedQA ([Jin et al., 2019](#)) and MMLU medical topics ([Hendrycks et al., 2020](#)), similar to Flan-PaLM ([Singhal et al., 2023a](#)). To better align with real-world application scenarios, our study focuses on a zero-shot setting. Encouragingly, our results reveal that across all tasks, our proposed approach outperforms settings for both chain-of-thought (CoT) and self-consistency prompting methods. Most notably, our approach demonstrates better performance under the zero-shot setting compared with the few-shot (5-shot) capabilities of 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.

To sum up, our work has three major contributions as follows:

- (i) We propose a multi-disciplinary collaboration framework for question-answering tasks in the medical domain. This novel approach endeavors to unveil the inherent clinical expertise present in LLMs and enhance their reasoning competence.
- (ii) We present our experimental results on nine datasets. The results demonstrate the general effectiveness of the MC framework and show that our proposed MC framework excels at mining and harnessing the medical expertise in LLMs.
- (iii) 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 witnessed the impressive advancements brought about by LLMs across various domains ([Ling et al., 2023](#); [Wu et al., 2023c](#); [Singhal et al., 2023a](#); [Yang et al., 2023](#)), among which a promising and noteworthy application lies in the medical domain ([Bao et al., 2023](#); [Nori et al., 2023](#); [Rosol 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., 2023b,a](#); [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](#)). Consequently, it is of crucial importance to further tap into the medical expertise to arm LLMs with enhanced clinical reasoning capabilities. Currently, there are two major lines of research on LLMs in medical domains, namely 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.

For instruction tuning methods, current research makes use of external clinical knowledge bases and self-prompted data to obtain instruction datasets (Tu et al., 2023; Zhang et al., 2023a; Singhal et al., 2023b; Tang et al., 2023c), which are then employed to tune LLMs on medical domains. For example, LLaVA-Med (Li et al., 2023a) leveraged a broad-coverage biomedical figure-caption dataset collected from PubMed Central and took advantage of GPT-4 to self-instruct open-ended instruction-following data from the captions in order to fine-tune a large general-domain vision-language model. MedChatZH (Tan et al., 2023) served as a dialogue model for traditional Chinese medical QA, which was pre-trained on Chinese traditional medical books and finetuned with an elaborated medical instruction dataset. AlpaCare (Zhang et al., 2023b) benefited from its large-scale and diverse medical instruction-following data MedInstruct-52k, resulting in remarkable generality and medical proficiency.

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 (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; Wang et al., 2023b; Li et al., 2023c; Hong et al., 2023) which further explores the potential of LLM-based agents by learning from multi-turn feedback and mutual cooperation. In essence, the key to LLM-based multi-agent collaboration is the simulation of human activities such as role-playing (Wang et al., 2023b; Hong et al., 2023) and communication (Wu et al., 2023b; Qian et al., 2023; Li et al., 2023b). For instance, Solo Performance Prompting (SPP) (Wang et al., 2023b) 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., 2023b) 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; Liang 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. Fu et al. (2023) let two LLMs play the roles of a buyer and a seller and negotiate with each other.

3 Method

This section presents the details of our proposed Multi-disciplinary Collaboration (MC) framework. As is shown in Figure 1, the MC framework works in five stages: (i) expert gathering: gather experts from distinct disciplines according to the clinical question; (ii) analysis proposition: domain experts put forward their own analysis with their expertise; (iii) report summarization: compose a summarized report on the basis of 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 the experts is reached; (v) decision making: derive a final decision from the unanimous report.²

3.1 Expert Gathering

Given a clinical question q and a set of options $op = \{o_1, o_2, \dots, o_k\}$, the goal of the Expert Gathering stage is to recruit a group of question domain experts $\mathcal{QD} = \{qd_1, qd_2, \dots, qd_m\}$ and option domain experts $\mathcal{OD} = \{od_1, od_2, \dots, od_n\}$. Specifically, we assign a role to the model and provide instructions to guide the model output the corresponding domains based on the input question and options: $\mathcal{QD} = \text{LLM}(q, r_{qd}, \text{prompt}_{qd})$, and $\mathcal{OD} = \text{LLM}(q, op, r_{od}, \text{prompt}_{od})$, where $(r_{qd}, \text{prompt}_{qd})$ and $(r_{od}, \text{prompt}_{od})$ stand for the system role and guideline prompt to gather domain experts for the question q and options op respectively. An example of r_{qd} and r_{od} is shown

²Details about all guideline prompts are shown in Section A for clarification.

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

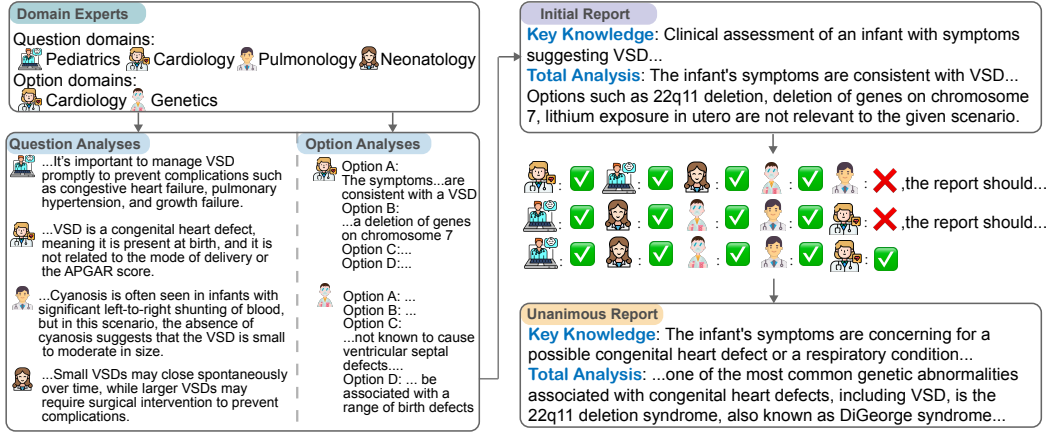


Figure 2: Illustrative example of our proposed Multi-disciplinary Collaboration (MC) framework applied to a pediatric medical problem. The questions and options are first presented, with domain experts subsequently gathered. The recruited experts conduct thorough Question and Option analyses based on their respective fields. An initial report synthesizing these analyses is then prepared as a concise representation of the performed evaluations. The assembled LLM experts, possessing respective disciplinary backgrounds, engage in a discussion over the initial report, voicing agreements and disagreements. Ultimately, after iterative refinement and consultation, a unanimous report is generated that best represents the collective expert knowledge and reasoning on the given medical problem.

below:

r_{qd} : You are a medical expert who specializes in categorizing a specific medical scenario into specific areas of medicine.

r_{od} : As a medical expert, you possess the ability to discern the two most relevant fields of expertise needed to address a multiple-choice question encapsulating a specific medical context.

3.2 Analysis Proposition

After gathering domain experts for the question q and options op , we aim to inquire experts to generate corresponding analyses, which are prepared for later reasoning.

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 prompt_{qa} : $qa_i = \text{LLM}(q, qd_i, r_{qa}, \text{prompt}_{qa})$. As such, we manage to attain a set of question analyses $\mathcal{QA} = \{qa_1, qa_2, \dots, qa_m\}$. An example of r_{qa} can be shown as:

r_{qa} : You are a medical expert in the domain of qd_i . From your domain, your goal is to scrutinize and diagnose the symptoms presented by patients in specific medical scenarios.

Option Analyses Now that we have an option domain od_i and question analyses \mathcal{QA} , we are able to 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: $oa_i = \text{LLM}(q, od_i, \mathcal{QA}, r_{oa}, \text{prompt}_{oa})$. In this way, we acquire a series of option analyses $\mathcal{OA} =$

$\{oa_1, oa_2, \dots, oa_n\}$. An example of r_{oa} appears below:

r_{oa} : You are a medical expert specialized in the domain od_i . You are adept at comprehending the nexus between questions and choices in multiple-choice exams and determining their validity. Your task is to analyze individual options with your expert medical knowledge and evaluate their relevancy and correctness.

3.3 Report Summarization

In Report Summarization stage, we attempt to summarize and synthesize previous analyses from various domain experts $QA \cup OA$. Given question analyses QA and option analyses 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: $Repo = \text{LLM}(QA, OA, r_{rs}, \text{prompt}_{rs})$, where $Repo$ can be formulated as: [Key Knowledge : extracted knowledge; Total Analysis : synthesized analysis]. An example of r_{rs} is illustrated below:

r_{rs} : You are a medical report assistant who excels at summarizing and synthesizing.

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 personal 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 $Repo_i = \text{LLM}(Repo_{i-1}, Mod_i, \text{prompt}_{mod})$. 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 $Repo_f$: $ans = \text{LLM}(q, op, Repo_f, r_{dm}, \text{prompt}_{dm})$. An example of r_{dm} is demonstrated below:

r_{dm} : You are a medical decision maker skilled in making decisions based on summarized reports.

4 Experiments

4.1 Setup

Tasks and Datasets. We evaluate our MC framework on two 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

Table 1: Summary of the Datasets. Part of the values are from the appendix of Singhal et al. (2023a).

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

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.

Method	MedQA	MedMCQA	PubMedQA	Anatomy	Clinical knowledge	College medicine	Medical genetics	Professional medicine	College biology
Flan-Palm									
- Few-shot CoT	60.3	53.6	77.2	66.7	77.0	83.3	75.0	76.5	71.1
- Few-shot CoT + SC	67.6	57.6	75.2	71.9	80.4	88.9	74.0	83.5	76.3
GPT-3.5									
- Zero-shot	50.4	53.6	71.8	63.7	75.5	79.2	74.0	77.6	77.6
- Zero-shot CoT	52.1	50.2	72.3	64.9	76.6	80.1	71.6	77.1	75.5
- Few-shot	55.9	56.7	67.6	65.2	78.1	78.5	83.0	76.8	67.0
- Few-shot CoT	60.7	54.7	71.4	60.7	74.7	80.6	70.0	75.4	67.0
- Few-shot CoT + SC	64.0	59.7	73.4	64.4	78.5	84.7	76.0	82.0	74.0
MC framework (Ours)									
- GPT-3.5	63.7	58.3	72.9	65.3	77.8	81.3	79.2	82.7	79.4
- GPT-4	83.0	72.0	75.0	81.0	89.0	95.0	94.0	96.0	81.0

anatomy, clinical knowledge, college medicine, medical genetics, professional medicine, and college biology. 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 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). Table 1 summarizes the data statistics.

Implementation. We utilize the popular and publicly available GPT-3.5-Turbo and GPT-4 (OpenAI, 2023) from Azure OpenAI Service.³ 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 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 MedMCQA ($m = 4, n = 2$). Considering the costly API expenses, we randomly sample 100 examples for each dataset and conduct experiments with GPT-4 on them. Details about the prompt templates involved in this study are listed in Appendix A.

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 including CoT and self-consistency prompting in both zero-shot and few-shot settings. We select GPT-3.5-Turbo and an instruction-tuning variant Flan-Palm (Chung et al., 2022) as the backbones of baselines following Singhal et al. (2023b). Notably, our proposed MC framework outperforms all the zero-shot baseline methods by a large margin, indicating the effectiveness of our MC framework in real-world application scenarios. Furthermore, our approach surprisingly demonstrates comparable performance under the zero-shot setting compared with the strong baseline *Few-shot CoT+SC*.

5 Analysis

5.1 Number of agents

As our work proposes a Multi-disciplinary Collaboration (MC) framework in which multiple agents play certain roles to derive the ultimate answer, we explore how the number of collaborating agents in our MC framework influences the overall performance. We fix the number of option agents as 2 and vary the number of question agents from 1 to 8. We run the experiments on 50 samples from MedQA, MedMCQA and PubMedQA datasets. Table 3 shows that the optimal number of question

³<https://learn.microsoft.com/en-us/azure/ai-services/openai/>

agents is 5 for MedQA, MedMCQA and 4 for PubMedQA, beyond which there may be diminishing returns or even potential confusion caused by information overload.

Table 3: Optimal number of agents for the question on MedQA, MedMCQA, and PubMedQA.

Dataset	MedQA	MedMCQA	PubMedQA
Number of agents	5	5	4

5.2 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.

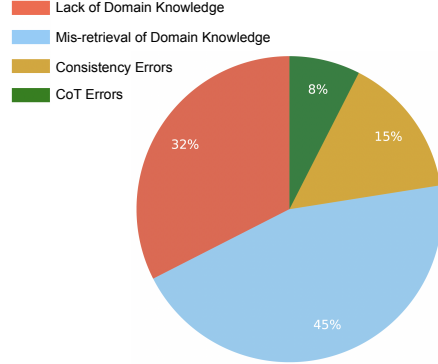


Figure 3: Ratio of different categories in error cases.

We randomly select 40 error cases in MedQA and MedMCQA datasets and analyse the percentage of different categories in these error cases. As is shown in Figure 3, the majority (77%) of the error examples are due to confusion about the domain knowledge, which illustrates that although our method further mines medical knowledge concealed within LLMs by means of multi-disciplinary consultation, there still exists a portion of domain knowledge that is explicitly beyond the intrinsic knowledge of LLMs. As a result, our analysis sheds light on future directions to mitigate the aforementioned drawbacks and further strengthen the model’s proficiency and reliability.

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

- Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. [Disc-medllm: Bridging general large language models and real-world medical consultation](#). 1, 3
- 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. 2020. **Language models are few-shot learners**. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 1
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*. 2
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. **Palm: Scaling language modeling with pathways**. *ArXiv preprint*, abs/2204.02311. 1
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*. 7
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. **Improving factuality and reasoning in language models through multiagent debate**. 2, 4
- Dat Duong and Benjamin D Solomon. 2023. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, pages 1–3. 3
- Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. **Improving language model negotiation with self-play and in-context learning from ai feedback**. 2, 4
- Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander LÄuser, Daniel Truhn, and Keno K. Bresssem. 2023. **Medalpaca – an open-source collection of medical conversational ai models and training data**. 2, 3
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*. 3, 6
- Sirui Hong, Xiwu 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. 2023. **Metagpt: Meta programming for multi-agent collaborative framework**. 2, 4
- 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. 3, 6
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*. 3, 6
- Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*. 3
- Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *arXiv preprint arXiv:2305.18395*. 3
- 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. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198. 2
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*. 4
- Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*. 4
- Yuan Li, Yixuan Zhang, and Lichao Sun. 2023c. **Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents**. 2, 4

- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023d. **Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.** 2
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. **Encouraging divergent thinking in large language models through multi-agent debate.** 2, 4
- Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*. 2
- Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Tianjiao Zhao, et al. 2023. Beyond one-model-fits-all: A survey of domain specialization for large language models. *arXiv preprint arXiv:2305.18703*. 3
- Zhengliang Liu, Zihao Wu, Mengxuan Hu, Bokai Zhao, Lin Zhao, Tianyi Zhang, Haixing Dai, Xianyan Chen, Ye Shen, Sheng Li, et al. 2023. Pharmacygpt: The ai pharmacist. *arXiv preprint arXiv:2307.10432*. 3
- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*. 1
- Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265. 1
- Y Nakajima. 2023. 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). 4
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*. 3
- OpenAI. 2023. **Gpt-4 technical report.** *ArXiv preprint*, abs/2303.08774. 1, 7
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. 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. 3, 6
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. 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. Association for Computing Machinery. 1, 2
- Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*. 4
- Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. 2023. Evaluation of the performance of gpt-3.5 and gpt-4 on the medical final examination. *medRxiv*, pages 2023–06. 3
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. **Bloom: A 176b-parameter open-access multilingual language model.** *ArXiv preprint*, abs/2211.05100. 1
- Henk G Schmidt and Remy MJP Rikers. 2007. How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical education*, 41(12):1133–1139. 2
- Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessie Li, and Byron C Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*. 3
- 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 SchÄd’rli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and Vivek Natarajan. 2023a. **Large language models encode clinical knowledge.** *Nature*, 620:1–9. 1, 2, 3, 6, 7
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*. 4, 7

- Yang Tan, Mingchen Li, Zijie Huang, Huiqun Yu, and Guisheng Fan. 2023. Medchatzh: a better medical adviser learns from better instructions. *arXiv preprint arXiv:2309.01114*. 4
- 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. 2023a. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158. 3
- Xiangru Tang, Arman Cohan, and Mark Gerstein. 2023b. Aligning factual consistency for clinical studies summarization through reinforcement learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 48–58. 3
- Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2023c. Gersteinlab at mediqua-chat 2023: Clinical note summarization from doctor-patient conversations through fine-tuning and in-context learning. *arXiv preprint arXiv:2305.05001*. 4
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. **Llama: Open and efficient foundation language models**. *ArXiv preprint*, abs/2302.13971. 1
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*. 4
- Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*. 3
- U.S. Department of Health and Human Services. 1996. The hipaa privacy rule. <https://www.hhs.gov/hipaa/for-professionals/privacy/index.html>. 2
- 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, Wenhui Chen, Jie Fu, and Junran Peng. 2023a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv: 2310.00746*. 2
- Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023b. **Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration**. 2, 4
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. **Pmc-llama: Towards building open-source language models for medicine**. 2
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023b. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*. 4
- Yiquan Wu, Siying Zhou, Yifei Liu, Weiming Lu, Xiaozhong Liu, Yating Zhang, Changlong Sun, Fei Wu, and Kun Kuang. 2023c. **Precedent-enhanced legal judgment prediction with llm and domain-model collaboration**. 1, 3
- Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*. 2, 4
- Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, et al. 2023. Openagents: An open platform for language agents in the wild. *arXiv preprint arXiv:2310.10634*. 4
- Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining the inter-consistency of large language models: An in-depth analysis via debate. *arXiv e-prints*, pages arXiv–2305. 4
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*. 2
- Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. **Investllm: A large language model for investment using financial domain instruction tuning**. 1, 3
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*. 4
- Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. **Legal prompting: Teaching a language model to think like a lawyer**. 2

- Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R Dalal, Jennifer L Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, et al. 2023. Almanac: Retrieval-augmented language models for clinical medicine. *Research Square*. 3
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*. 4
- Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023b. *Alpacare:instruction-tuned large language models for medical application*. 1, 2, 4
- Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, et al. 2023. *Webarena: A realistic web environment for building autonomous agents*. 1, 4

A Prompt Templates

Prompt templates involved in the experiments are as follows:

(1) $\text{prompt}_{\text{qd}}$ for gathering question domains: *You need to complete the following steps: 1. Carefully read the medical scenario presented in the question: question. 2. Based on the medical scenario in it, classify the question into five different subfields of medicine. 3. You should output in exactly the same format as: Medical Field: | .*

(2) $\text{prompt}_{\text{od}}$ for gathering option domains: *You need to complete the following steps: 1. 1. Carefully read the medical scenario presented in the question: question. 2. The available options are: options. Strive to understand the fundamental connections between the question and the options. 3. Your core aim should be to categorize the options into two distinct subfields of medicine. You should output in exactly the same format as: Medical Field: | .*

(3) $\text{prompt}_{\text{qa}}$ for deriving question analyses: *Please meticulously examine the medical scenario outlined in this question: question. Drawing upon your medical expertise, interpret the condition being depicted. Subsequently, identify and highlight the aspects of the issue that you find most alarming or noteworthy.*

(4) $\text{prompt}_{\text{oa}}$ for deriving option analyses: *Regarding the question: question, we procured the analysis of five experts from diverse domains. The evaluation from the question_domain expert suggests: question_analysis. The following are the options available: options. Reviewing the question's analysis from the expert team, you're required to fathom the connection between the options and the question from the perspective of your respective domain, and scrutinize each option individually to assess whether it is plausible or should be eliminated based on reason and logic. Pay close attention to discerning the disparities among the different options and rationalize their existence. A handful of these options might seem right on the first glance but could potentially be misleading in reality.*

(5) $\text{prompt}_{\text{rs}}$ for report summarization: *Here are some reports from different medical domain experts. You need to complete the following steps: 1. Take careful and comprehensive consideration of the following reports. 2. Extract key knowledge from the following reports. 3. Derive the comprehensive and summarized analysis based on the knowledge. 4. Your ultimate goal is to derive a refined and synthesized report based on the following reports. You should output in exactly the same format as: Key Knowledge;; Total Analysis:*

(6) $\text{prompt}_{\text{mod}}$ for modifying the report: *Here is advice from a medical expert specialized in domain: advice. Based on the above advice, output the revised analysis in exactly the same format as: Key Knowledge;; Total Analysis:*