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 MEDAGENTS, a novel multi-disciplinary collaboration framework for the medical domain. MedAgents leverages LLM-based agents in a role-playing setting that participate in a collaborative multi-round discussion, thereby enhancing LLM proficiency and reasoning capabilities. This training-free 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 MEDAGENTS framework excels at mining and harnessing the medical expertise within LLMs, as well as extending its reasoning abilities. 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 tend to encounter domain-specific tasks that necessitate a combination of domain expertise and complex reasoning abilities (Moor et al., 2023; Wu et al., 2023b; 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., 2023b; Bao et al., 2023; Singhal et al., 2023a).

Two major challenges prevent LLMs from effectively handling tasks in the medical sphere: (i) Limited volume and specificity of medical training data compared to the vast general text data, due to cost and privacy concerns (Thirunavukarasu et al., 2023). (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, particularly within math and coding, have been made to enhance prompting methods (Besta et al., 2023; Lála et al., 2023; Wang et al., 2023b), strategies used in the medical field have been shown to induce hallucination (Umapathi et al., 2023; Harris, 2023; Ji et al., 2023), indicating the need for more robust approaches.

Meanwhile, recent research has surprisingly witnessed the success of multi-agent collaboration (Xi et al., 2023; Wang et al., 2023d) which highlights the simulation of human activities (Du et al., 2023; Liang et al., 2023; Park et al., 2023) and optimizes the collective power of multiple agents (Chen et al., 2023; Li et al., 2023d; Hong et al., 2023). Through such design, 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.

---

1 Although Med-PaLM 2 (Singhal et al., 2023b) serves as a specialized medical LLM fine-tuned on the basis of PaLM 2, it is closed-sourced and not publicly accessible yet.
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.

This process subsequently enhances the model’s reasoning capabilities throughout multiple rounds of interaction (Wang et al., 2023c; Fu et al., 2023).

Motivated by these notions, we pioneer a multi-disciplinary collaboration framework (MedAgents) specifically tailored to the clinical domain. Our objective centers on unveiling the intrinsic medical knowledge embedded in LLMs and reinforcing reasoning proficiency in a training-free manner. As shown in Figure 1, the MedAgents framework gathers experts from diverse disciplines and reaches consistent conclusions through collaborative discussions.

Based on our MedAgents framework, we conduct experiments on nine datasets, 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). To better align with real-world application scenarios, our study focuses on the zero-shot setting, which can serve as a plug-and-play method to supplement existing medical LLMs such as Med-PaLM 2 (Singhal et al., 2023b). Encouragingly, our proposed approach outperforms settings for both chain-of-thought (CoT) and self-consistency (SC) prompting methods. Most notably, our approach achieves better performance under the zero-shot setting compared with the 5-shot 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. Targeted refinements focused on mitigating these particular shortcomings would enhance the model’s proficiency and reliability.

Our contributions are summarized as follows:

(i) To the best of our knowledge, we are the first to propose a multi-agent framework within the medical domain and explore how multi-agent communication within the medical setting can lead to a consensus decision, adding a novel dimension to the current literature on medical question answering.

(ii) Our proposed MedAgents framework enjoys enhanced faithfulness and interpretability by harnessing role-playing and collaborative agent discussion. And we demonstrate that role-playing allows LLM to explicitly reason with accurate knowledge, without the need for retrieval augmented generation (RAG). Examples illustrating interpretability are shown in Appendix B.

(iii) Experimental results on nine datasets

---

2We follow the evaluation setting from Faln-PaLM (Singhal et al., 2023a).
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 MedAgents, a multi-disciplinary collaboration framework. 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 to concisely represent 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.

2 Method

This section presents the details of our proposed multi-disciplinary collaboration MedAGENTS framework. Figures 1 and 2 give an overview and an illustrative example of its pipeline, respectively. Our proposed MedAGENTS 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

3Details about all guideline prompts and roles are shown in Section D for clarification.

2.1 Expert Gathering

Given a clinical question \( q \) and a set of options \( op = \{o_1, o_2, \ldots, o_k\} \) where \( k \) is the number of options, the goal of the Expert Gathering stage is to recruit a group of question domain experts \( QD = \{qd_1, qd_2, \ldots, qd_m\} \) and option domain experts \( OD = \{od_1, od_2, \ldots, od_n\} \) where \( m \) and \( n \) represent the number of question domain experts and option domain experts, respectively. 4

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:

\[
QD = LLM\left(q, r_{qd}, prompt_{qd}\right),
\]

\[
OD = LLM\left(q, op, r_{od}, prompt_{od}\right),
\]

where \( (r_{qd}, prompt_{qd}) \) and \( (r_{od}, prompt_{od}) \) stand for the system role and guideline prompt to gather domain experts for the question \( q \) and options \( op \).

4We design domain experts for the question and options in order to leverage diverse agents to elicit multifaceted and comprehensive knowledge.
2.2 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: \( QA = \{ qa_1, qa_2, \ldots, qa_m \} \) and \( OA = \{ oo_1, oo_2, \ldots, oo_n \} \).

**Question Analyses**  Given a question \( q \) and a question domain \( qd_i \in QD \), 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 \( prompt_{qa} \):

\[
qa_i = \text{LLM}(q, qd_i, r_q, \text{prompt}_{qa}). \tag{2}
\]

**Option Analyses**  Now that we have an option domain \( od_i \) and question analyses \( QA \), 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 \( od_i \in OD \), and the question analyses \( QA \) to the LLM:

\[
oa_i = \text{LLM}(q, op, od_i, QA, r_o, \text{prompt}_{oa}). \tag{3}
\]

2.3 Report Summarization

In the 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:

\[
\text{Repo} = \text{LLM}(Q, OA, r_r, \text{prompt}_{r}). \tag{4}
\]

2.4 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. The overall procedure of this phase is presented in Algorithm 1. During each round of discussions, the experts give their votes (yes/no): \( vote = \text{LLM}(\text{Repo}, r_v, \text{prompt}_{vote}) \), 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 \( j \)-th round of discussion, we note the modification comments from the experts as \( Mod_j \), then we can acquire the updated report as \( \text{Repo}_j = \text{LLM}(\text{Repo}_{j-1}, Mod_j, \text{prompt}_{mod}) \). In this way, the discussions are held iteratively until all experts vote yes for the final report \( \text{Repo}_f \) or the discussion number attains the maximum attempts threshold.

2.5 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
3 Experiments

3.1 Setup

Tasks and Datasets. We evaluate our MEDAGENTS 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. More information about the evaluated datasets is presented in Appendix C.

Implementation. We utilize the popular and publicly available GPT-3.5-Turbo and GPT-4 (OpenAI, 2023) from Azure OpenAI Service. 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 iteration number and temperature of SC are 5 and 0.7, respectively. The number 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 ($m = 4, n = 2$). The number of maximum attempts $t$ is set as 5. 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 $\mathbf{1.4}$ per question) and the inference time per example is about 40s.6

Baselines. We have utilized models that are readily accessible through public APIs with the following baselines:

- **Setting w/o CoT**: Zero-shot (Kojima et al., 2022) directly incorporates the prompt *Let’s think step by step* after a question to facilitate inference. Few-shot CoT (Wei et al., 2022) adopts comparable methodologies to Few-shot but distinguishes itself by integrating rationales before deducing the answer.

- **Setting w/ SC**: SC (Wang et al., 2022) serves as an additional sampling method on Zero-shot CoT and Few-shot CoT, which yields the majority answer by sampling multiple chains.

3.2 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 MEDAGENTS framework outperforms the zero-shot baseline methods by a large margin, indicating the effectiveness of our MEDAGENTS 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, the introduction of CoT occasionally leads to a surprising degradation in performance.7 We have found that reliance on CoT in isolation can inadvertently result in hallucinations - spurious outputs typically associated with the misapplication of medical terminologies. In contrast, our multi-agent role-playing methodology effectively mitigates these issues, thus underscoring its potential as a more robust approach in medically oriented LLM applications.

4 Analysis

4.1 Ablation Study

Since our MEDAGENTS 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

---


7A more detailed analysis of CoT’s impact is provided in Appendix A.
Method & MedQA & MedMCQA & PubMedQA & Anatomy & Clinical knowledge & College medicine & Medical genetics & Professional medicine & College medicine & Avg. \\ \hline 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 \\ \hline GPT-3.5 \\ *few-shot setting* \\ Few-shot & 54.7 & 56.7 & 67.6 & 65.9 & 61.5 & 76.2 & 63.6 & 74.0 & 75.4 & 75.0 & 67.8 \\ 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 & 79.0 & 77.2 & 71.6 \\ \hline GPT-4 \\ *few-shot setting* \\ Zero-shot & 54.3 & 56.3 & 73.7 & 61.5 & 89.5 & 75.6 & 93.0 & 91.5 & 91.7 & 82.3 \\ 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 \\ MedA\textsc{gents (Ours)} & 64.1 & 59.3 & 72.9 & 65.2 & 77.7 & 69.8 & 79.0 & 82.1 & 78.5 & 72.1 \\ \hline Table 2: Main results (Acc). SC denotes the self-consistency prompting method. Results in bold are the best performances.

Method & Accuracy(%) \\ \hline Direct Prompting & 49.0 \\ CoT Prompting & 55.0 \\ + Med\textsc{agents} & \multicolumn{3}{c}{62.0(↑ 7.0)} \\ + Anal & \multicolumn{3}{c}{65.0(↑ 10.0)} \\ + Anal & Summ & Cons & \multicolumn{3}{c}{67.0(↑ 12.0)} \\ Table 3: Ablation study for different processes on MedQA. Anal: Analysis proposition, Summ: Report summarization, Cons: Collaborative consultation.

Method & MedQA & MedMCQA \\ \hline MED\textsc{agents} (GPT-3.5) & 64.1 & 59.3 \\ MED\textsc{agents} (GPT-4) & 83.7 & 74.8 \\ MedAlpaca-7B & 55.2 & 45.8 \\ BioMedGPT-10B & 50.4 & 42.2 \\ BioMedLM-2.7B & 50.3 & - \\ BioBERT (large) & 36.7 & 37.1 \\ SciBERT (large) & - & 39.2 \\ BERT (large) & - & 33.6 \\ Table 4: Comparison with open-source medical models.

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.

4.2 Comparison with Open-source Medical Models

We conduct a comprehensive comparison between our proposed \textsc{medagents} framework with more baseline methods, including open-source domain-adapted models such as MedAlpaca-7B (Han et al., 2023), BioMedGPT-10B (Luo et al., 2023), BioMedLM-2.7B (Bolton et al., 2024), BioBERT (large) (Lee et al., 2020), SciBERT (large) (Beltagy et al., 2019) and BERT (large). We leverage them
in the early stages of our work for preliminary attempts. Results in Table 4 demonstrate that the open-source methods fell short of the baseline in Table 2, which leads us to focus on the more effective methods.

4.3 Number of agents
As our MEDAGENTS 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 and option agents while fixing other variables to observe the performance trends on the MedQA dataset. Figure 3 and Table 5 illustrate the corresponding trend and the optimal number of different agents. 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. 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.

4.4 Domain Variation Study
In order to investigate the influence of the changes in agent numbers, we perform additional studies where we manipulate agent numbers by selectively eliminating the most and least relevant domain experts based on domain relevance. Due to the manual evaluation involved in identifying the relevance of agent domains, our additional analysis was conducted on a limited set of 20 samples. Results in Table 6 depict minor variance for different sizes of data with random removing, reinforcing the notion that large-scale experiment performance remains largely robust against the effect of domain changes.

4.5 Agent Quantity Study
To further explore the effect of agent quantity without changes in domain representation, we conduct experiments with $k$ ($k = 6$) identical-domain agents, then with $k-1$ and $k-2$, to observe performance shifts. The process of selecting these domains is automated via prompting, and our manual inspection confirms the high relevance and quality of the selected domains. The experiment is conducted on a dataset of 300 samples.

4.6 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: 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: 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: the model may form and follow inaccurate rationales, leading to incorrect conclusions.

To illustrate the error examples intuitively, we select four typical samples from the four error categories, which can be shown in Figure 5: (i) The first error is due to a lack of domain knowledge regarding cutaneous larva migrans,
Lack of Domain Knowledge
Mis-retrieval of Domain Knowledge
Consistency Errors
CoT Errors

Figure 4: Ratio of different categories in error cases.

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.

Furthermore, we analyze the percentage of different categories by randomly selecting 40 error cases in MedQA and MedMCQA datasets. 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 there still exists a portion of domain knowledge that is explicitly beyond the intrinsic knowledge of LLMs, leading to a bottleneck of our proposed method. As a result, our analysis sheds light on future directions to mitigate the aforementioned drawbacks and further strengthen the model’s proficiency and reliability.

4.7 Correctional Capabilities and Interpretability

In our extensive examination of the MEDAGENTS framework, we discover the decent correctional capabilities of our framework. Please refer to Appendix B and Table B for an in-depth overview of instances where our approach successfully amends previous inaccuracies, steering the discussion towards more accurate outcomes. These corrections showcase the MEDAGENTS framework’s strength in collaborative synthesis; it distills and integrates diverse expert opinions into a cohesive and accurate conclusion. By interweaving a variety of perspectives, the collaborative consultation actively refines and rectifies initial analyses, thereby aligning the decision-making process closer to clinical accuracy. This iterative refinement serves as a practical demonstration of our model’s proficiency in rectifying errors, substantiating its interpretability and accuracy in complex medical reasoning tasks.

5 Related Work

5.1 LLMs in Medical Domains

Recent years have seen remarkable progress in the application of LLMs (Wu et al., 2023b; 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; 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). 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-promted data to obtain instruction datasets (Tu et al., 2023; Zhang et al., 2023a; 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., 2023a; 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., 2023). Unlike these methods, our work emphasizes harnessing latent medical knowledge intrinsic to LLMs and improving reasoning in a training-free setting.

5.2 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., 2023a; 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., 2023d; 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., 2023d; Hong et al., 2023) and communication (Wu et al., 2023a; Qian et al., 2023; Li et al., 2023b,c).

For instance, Solo Performance Prompting (SPP) (Wang et al., 2023d) managed to combine the strengths of multiple minds to improve performance by dynamically identifying and engaging multiple personas throughout 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; 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.

6 Conclusion

We present a novel medical QA framework that uses role-playing agents for multi-round discussions, offering greater reliability and clarity without prior training. Our method surpasses zero-shot baselines and matches few-shot baselines across nine datasets. Despite successes, human-based evaluations of errors have highlighted areas for refinement. Our approach differs from most existing methods by eliminating the dependency on knowledge bases, instead uniquely integrating medical knowledge through role-playing agents.
Limitation

The proposed MEDAGENTS framework has shown promising results, but there are still a few points that could be addressed in future studies. First, the parameterized knowledge within LLMs may need updating over time, and thus, continuous efforts are required to keep the framework up-to-date. Second, integrating diverse models at different stages of our framework might be an intriguing exploration. Third, the framework may have limited applicability in low-resource languages. Adapting this framework to a wider range of low-resource languages could meet their specific medical needs to some extent.

Ethics Statement

Although our work strictly adheres to well-established benchmarks in the field of medical question answering, it is possible that our approach introduces potential risks, e.g., some inherent biases of LLMs, when applying LLM reasoning to critical areas such as medicine.

Acknowledgments

Xiangru Tang and Mark Gerstein are supported by Schmidt Sciences. Zhuosheng Zhang is supported by CIPSC-SMP-Zhipu.AI Large Model Cross-Disciplinary Fund.

References


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


Emily Harris. 2023. Large language models answer medical questions accurately, but can’t match clinicians’ knowledge. *JAMA*.


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*.


A Analysis on the Addition of CoT

We provide an intriguing example that reveals a seemingly counter-intuitive observation: the addition of the CoT in a zero-shot setting led to a performance drop compared to the zero-shot one.

As demonstrated in the example, for specialist domains that demand considerable expert knowledge such as the medical domain knowledge, employing a CoT approach might sometimes lead to hallucination (Bubeck et al., 2023; Guerreiro et al., 2023; Ji et al., 2023; Maynez et al., 2020). Hallucinations refer to instances where the language model starts generating inaccurate or irrelevant information based on its insufficient understanding (Wei et al., 2022; Kojima et al., 2022; Shi et al., 2023). Consequently, in these instances, the use of the CoT method does not improve but hinders the overall performance.

This issue is particularly stated in the medical question-answering field by some recent work, where it has been demonstrated that the CoT’s step-by-step approach is unable to generate correct medical answers effectively. For example, the results from Liévin et al. (2022) demonstrate that CoT improvements are significantly limited.

Such failures in medical question-answering originate from a lack of domain knowledge (Harris, 2023; Kung et al., 2023; Tian et al., 2024) instead of reasoning rationale. This was also observed in our experiments, with a substantial 77% of errors related to domain knowledge, compared to a minor 8% due to CoT (as shown in Figure 4).

To address these issues, there has indeed been a recent shift toward utilizing RAG (Retrieval-Augmented Generation) for domain knowledge enhancement (Wang et al., 2023b; Lála et al., 2023) in medical QA. Despite these developments, our study underscores a novel approach within this field: our use of role-playing. Role-playing in our MEDAGENTS framework allows the model to explicitly reason with accurate knowledge, remarkably bypassing the need for RAG.
Table 8: Examples showing the influence of the addition of CoT

Question: Which of the following is the large bone found superior to the patella and inferior to the ischium? (A) Calcaneus (B) Femur (C) Symphysis pubis (D) Tibia

Zero-shot Answer: B ✓

Zero-shot-CoT Answer: D ✗

Zero-shot-CoT Rationale: The bone found superior to the patella and inferior to the ischium is most likely the tibia (✗). Let’s go through the options to confirm. Option A: Calcaneus - The calcaneus is the heel bone and is not located in the specified position. Option B: Femur - The femur is the thigh bone and is located superior to the patella, but it is not inferior to the ischium (✗). Option C: Symphysis pubis - The symphysis pubis is a joint located in the pelvis and is not a bone. Option D: Tibia - The tibia is the shin bone and is located superior to the patella and inferior to the ischium. This seems to be the correct option. Answer: D
B Interpretability: Case Study

We analyze instances where our proposed MEDAGENTS framework manages to correct previous errors. Table B illustrates examples of our approach to turning previously incorrect answers into correct ones. These cases showcase the robustness of our MEDAGENTS framework. Different medical professionals offer depth and breadth in the analysis, transforming it into a comprehensive assessment. Furthermore, the collaborative consultation refines and corrects the initial interpretation, leading us toward the correct answer, thus demonstrating the effectiveness of our method in correcting errors and ensuring accuracy.
Question: A 23-year-old college student experiences monthly chest discomfort, shortness of breath, shakiness, and excessive sweating, likely due to stress. He fears having an episode in public, causing him to avoid leaving home. His medical history is unclear, and the physical exam is normal. Which of the following is the best medication for the long-term management of this patient’s condition? (A) Citalopram (B) Lithium (C) Lorazepam (D) Propranolol (E) Quetiapine

Experts: Cardiology, Psychiatry, Pulmonology, Endocrinology, Neurology

Analyses from various experts:

Cardiology Expert: ...These symptoms are not typical of a cardiac cause given the patient’s young age and normal physical exam...The best medication for long-term management of anxiety disorders is often selective serotonin reuptake inhibitors (SSRIs) like Citalopram.

Psychiatry Expert: ...Although quetiapine is commonly used in the treatment of schizophrenia and bipolar disorder, it is also effective in controlling the symptoms of rare endocrine disorders that present with anxiety-like symptoms.

Pulmonology Expert: ...this treatment based on the possibility that the symptoms could be due to a respiratory condition, such as hyperventilation syndrome, which Propranolol is thought to help by reducing the heart rate and perceived shortness of breath.

Endocrinology Expert: SSRIs like Citalopram can affect endocrine function by regulating serotonin levels, which in turn can modulate the hypothalamic-pituitary-adrenal (HPA) axis...

Neurology Expert: the focus would be on addressing the dysregulated ANS activity that is causing the physical symptoms of stress. The most appropriate medication would likely be Citalopram, an SSRI, which can help improve mood and reduce anxiety...

Expert analysis presents a multifaceted view of the problem. Based on the experts’ specialties, their analyses focus on different aspects, leading to the suggestion of two drugs, Citalopram and Propranolol.

Synthesized Report (v0): ... Based on the suggestions from the experts, the medication should be Citalopram along with Propranolol, which can help alleviate shortness of breath...

The first synthesis suggested a combination of Citalopram and Propranolol.

Vote Result:

Cardiology Expert: No
Psychiatry Expert: Yes
Pulmonology Expert: Yes
Endocrinology Expert: Yes
Neurology Expert: No

Synthesized Report (v1): ...the most appropriate medication for the long-term management of this patient’s condition appears to be Citalopram(SSRI).

The final report suggested Citalopram as the best medication for the patient’s condition.
Table 10: Case studies demonstrating how our proposed MEDAGENTS framework manages to correct previous errors.

**Question:** Eustachian tube function is best assessed by: (A) Tympanometry (B) VEMP (C) Rhinomanometry (D) Politzer tests

**Experts:** Otolaryngology, Audiology, Radiology, Allergy and Immunology, Pulmonology

**Analyses from various experts:**

**Otolaryngology Expert:** ...The focus is on Eustachian tube function, which is essential for equalizing ear pressure and facilitating normal hearing. Several tests are available, but tympanometry is a primary method for assessing function, indicating if there’s a blockage or proper movement of the eardrum.

**Audiology Expert:** ...Tympanometry is correct, VEMP is not related, Rhinomanometry is unrelated, and the Politzer test is appropriate for Eustachian tube assessment.

**Radiology Expert:** ...Proper function of the Eustachian tube is vital for hearing and balance. Tympanometry is one of the methods used to assess this function, by evaluating the compliance of the eardrum.

**Allergy and Immunology Expert:** Assessing Eustachian tube function is crucial for diagnosing related dysfunctions and infections. Tympanometry is a valuable test to determine if the tube is functioning correctly.

**Pulmonology Expert:** ...The assessment of the Eustachian tube function is important for diagnosing ear conditions that could affect breathing. Tympanometry helps evaluate eardrum mobility and middle ear pressure

Based on the experts’ specialties, their analyses focus on different aspects, leading to the suggestion of two methods, **Tympanometry and Politzer test**.

**Synthesized Report (v0):** ...Tympanometry is reaffirmed as a method to assess Eustachian tube function, with additional methods like sonotubometry and nasal endoscopy also important. The Politzer test is recognized as another method for such assessments.

The first synthesis suggested **Tympanometry and Politzer test**.

**Vote Result:**

**Otolaryngology Expert:** Yes
**Audiology Expert:** Yes
**Radiology Expert:** Yes
**Allergy and Immunology Expert:** Yes
**Pulmonology Expert:** Initially No, then Yes after revision.

**Synthesized Report (v1):** ...Tympanometry is a key method among several to assess Eustachian tube function, crucial for diagnosing Eustachian tube dysfunction, infections, and hearing issues....

The final report suggested **Tympanometry** as the best medication for the patient’s condition.
Table 11: Case studies demonstrating how our proposed MEDAGENTS framework manages to correct previous errors.

**Question:** To prevent desiccation and injury, the embryos of terrestrial vertebrates are encased within a fluid secreted by the? (A) amnion (B) chorion (C) allantois (D) yolk sac

**Experts:** Embryology, Physiology, Dermatology, Endocrinology, Reproductive Medicine

**Analyses from various experts:**
- **Embryology Expert:** ... amnion is the primary structure responsible for secreting amniotic fluid, with the chorion also contributing to this process.
- **Physiology Expert:** ...the amniotic sac and amniotic fluid, produced by the amnion, is important in creating a stable and protected environment for the embryo. Besides, its cushioning effect and role in the exchange of nutrients and gases cannot be ignored either.
- **Dermatology Expert:** ...the importance of amniotic fluid in fetal development, particularly for the skin...
- **Endocrinology Expert:** the question relates to embryological development rather than an endocrinological condition, ... the amniotic sac plays a vital role in providing a protective environment.
- **Reproductive Medicine Expert:** ...amniotic fluid is of crucial importance for the normal development and survival of the embryo.

Based on the experts’ specialties, their analyses focus on different aspects, leading to the suggestion of two methods, amnion and chorion.

**Synthesized Report (v0):** ... Both the amnion and chorion play vital roles in this process.
- The first synthesis suggested amnion and chorion.

**Vote Result:**
- **Embryology Expert:** Yes
- **Physiology Expert:** Yes
- **Dermatology Expert:** No
- **Endocrinology Expert:** No
- **Reproductive Medicine Expert:** Yes

**Synthesized Report (v1):** ... the amnion as the primary source of the amniotic fluid that protects the developing embryos of terrestrial vertebrates, with an understanding of the contributions from other structures like the chorion and the allantois.

Dermatology and Endocrinology experts have revised their initial focus to align with the consensus on the role of the amnion in secreting amniotic fluid.
C  Dataset Information

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).
D Prompt Templates

Prompt templates involved in the experiments are presented in Table 12.
Table 12: Prompt templates and role descriptions employed in our MEDAGENTS framework.

**rqd**: You are a medical expert who specializes in categorizing a specific medical scenario into specific areas of medicine.

**prompt**: 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 the same format as: Medical Field: | .

**rqa**: 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.

**prompt**: You need to complete the following steps: 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 the same format as: Medical Field: | .

**rs**: You are a medical assistant who excels at summarizing and synthesizing based on multiple experts from various domain experts.

**prompt**: 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:

**vote**: You are a medical expert specialized in the domain domain. 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.

**prompt**: Here is a medical report: synthesized_report. As a medical expert specialized in domain, please carefully read the report and decide whether your opinions are consistent with this report. Please respond only with: [YES or NO].

**prompt**: Here is advice from a medical expert specialized in domain: advice. Based on the above advice, output the revised analysis in the same format as: Key Knowledge:; Total Analysis:

**prompt**: Here is a synthesized report: syn_report. Based on the above report, select the optimal choice to answer the question. Points to note: 1. The analyses provided should guide you towards the correct response. 2. Any option containing incorrect information inherently cannot be the correct choice. 3. Please respond only with the selected option’s letter, like A, B, C, D, or E, using the following format: "'Option: [Selected Option’s Letter]'". Remember, it’s the letter we need, not the full content of the option.
E Multiple Runs

we have performed multiple runs on a set of 300 samples for GPT-4 and GPT-3.5 to account for variability.

The preliminary tests involved the average score of 5 repetitions for each sample, and the results indicated a decent small variance between runs. In summary, the consistency observed in this set provides confidence in the stability of our reported results.

<table>
<thead>
<tr>
<th></th>
<th>MedQA</th>
<th>MedMCQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC framework GPT-3.5 (single try)</td>
<td>64.1</td>
<td>59.3</td>
</tr>
<tr>
<td>MC framework GPT-3.5 (5 repetitions)</td>
<td>64.3</td>
<td>59.2</td>
</tr>
<tr>
<td>MC framework GPT-4 (single try)</td>
<td>83.7</td>
<td>74.8</td>
</tr>
<tr>
<td>MC framework GPT-4 (5 repetitions)</td>
<td>83.5</td>
<td>74.9</td>
</tr>
</tbody>
</table>

Table 13: GPT-4 versus GPT-3.5 based on multiple runs of 300 samples.