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

Xiangru Tang[♡*], **Anni Zou**[♠*], **Zhuosheng Zhang**[♠], **Ziming Li**[♡],
Yilun Zhao[♡], **Xingyao Zhang**[♡], **Arman Cohan**[♡], **Mark Gerstein**[♡]
[♡]Yale University [♠]Shanghai Jiao Tong University
xiangru.tang@yale.edu, mark@gersteinlab.org

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).[1] (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.

[*] Equal contribution.

[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.

Findings of the Association for Computational Linguistics ACL 2024, pages 599–621
August 11-16, 2024

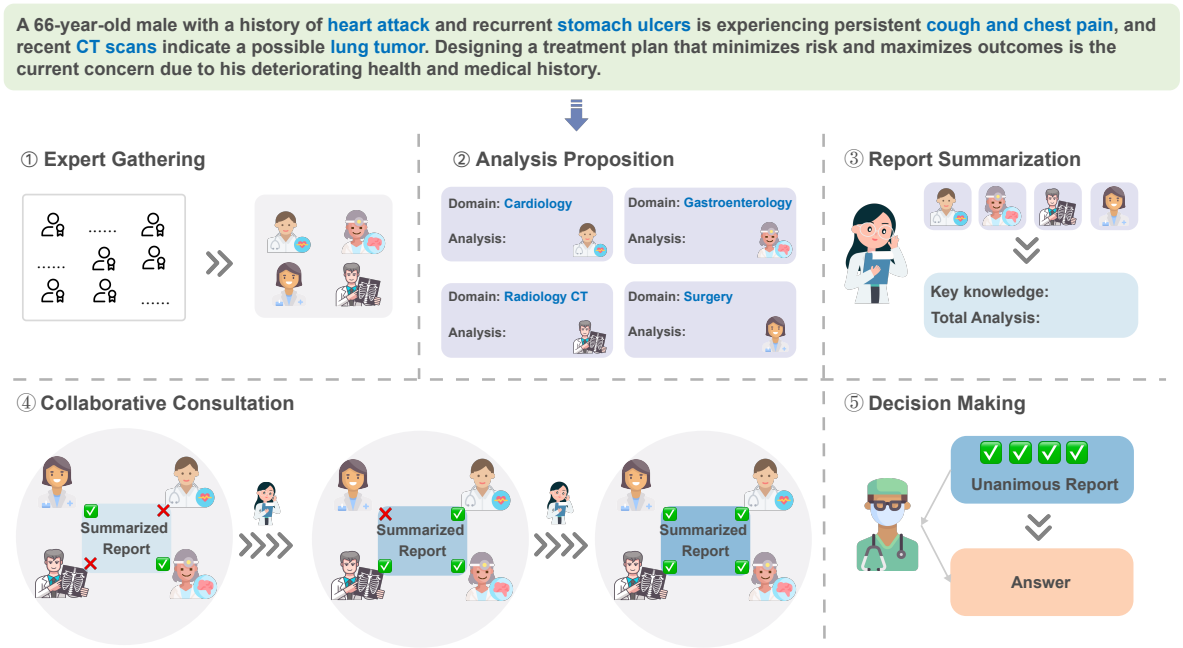

Figure 1: Diagram of our proposed MEDAGENTS framework. Given a medical question as input, the framework performs reasoning in five stages: (i) expert gathering, (ii) analysis proposition, (iii) report summarization, (iv) collaborative consultation, and (v) decision making.

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).[2] 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

[2]We follow the evaluation setting from Faln-PaLM (Singhal et al., 2023a).

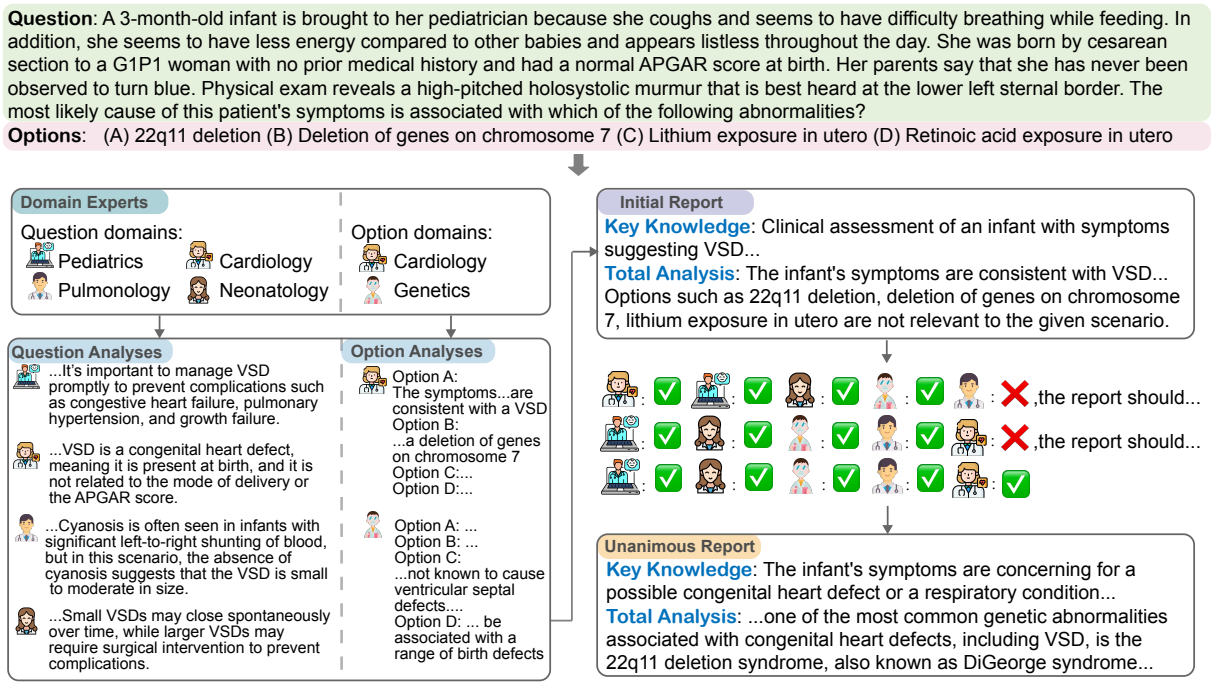

Figure 2: Illustrative example of our proposed MedAgents, a multi-disciplinary collaboration framework. The questions and options are first presented, with domain experts subsequently gathered. The recruited experts conduct thorough Question and Option analyses based on their respective fields. An initial 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.

demonstrate the general effectiveness of our proposed MEDAGENTS framework. Besides, we identify and categorize common error types in our approach through rigorous human evaluation to shed light on future studies.

2 Method

This section presents the details of our proposed multi-disciplinary collaboration MEDAGENTS framework. Figure 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]

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 $\mathcal{QD} = \{qd_1, qd_2, \ldots, qd_m\}$ and option domain experts $\mathcal{OD} = \{od_1, od_2, \ldots, od_n\}$ 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:

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

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

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

[4] We design domain experts for the question and options in order to leverage diverse agents to elicit multifaceted and comprehensive knowledge.

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

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: $\mathcal{QA} = \{qa_1, qa_2, \ldots, qa_m\}$ and $\mathcal{OA} = \{oa_1, oa_2, \ldots, oa_n\}$.

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

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

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

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

2.3 Report Summarization

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

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

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*): $vote = \text{LLM}\left(Repo, \texttt{r}_{\texttt{vote}}, \texttt{prompt}_{\texttt{vote}}\right)$, 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 $Repo_j = \text{LLM}\left(Repo_{j-1}, Mod_j, \texttt{prompt}_{\texttt{mod}}\right)$. In this way, the discussions are held iteratively until all experts vote *yes* for the final report $Repo_f$ or the discussion number attains the maximum attempts threshold.

Algorithm 1: Collaborative Consultation

Input: Domain experts $D = \{d_1, ..., d_n\}$, initial report R_0, Model $\mathcal{M}$, maximum attempts t, prompts $\{p_{vote}, p_{mod}, p_{rev}\}$
Output: Final report R_f

```
// Initialize variables
nocon_flag ← True, n_try ← 0
R_cur ← R_0, Mods ← ∅
// Iterative review
while nocon_flag is True and n_try < t do
    n_try ← n_try + 1
    nocon_flag ← False
    // vote for the report
    for i in 1, ..., n do
        vote_i ← M(R_cur, d_i, p_vote)
        // propose modifications
        if vote_i is no then
            Mod_i ← M(R_cur, d_i, p_mod)
            Update Mods with Mod_i
            nocon_flag ← True
        end
    end
    // modify the report
    if nocon_flag is True then
        R_cur ← M(R_cur, Mods, p_rev)
    end
end
return R_f ← R_cur
```

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

report $Repo_f$:

$$ans = \text{LLM}\left(q, op, Repo_f, \texttt{prompt}_{\texttt{dm}}\right). \quad (5)$$

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.[5] 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 k of options is 4 except for PubMedQA (3). The numbers of domain experts for the question and options are set as: $m = 5, n = 2$ except for PubMedQA ($m = 4, n = 2$). 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 **¢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) appends the prompt /emphA: The answer is to a given question and utilizes it as the input fed into LLMs. Few-shot (Brown et al., 2020) introduces several manually templated demonstrations, structured as [Q: q, A: The answer is a], preceding the input question.

• **Setting w/ CoT**: Zero-shot CoT (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

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

[6] We acknowledge that our proposed method requires more cost compared with CoT or direct prompting. However, our approach is relatively cost-effective and the improved performance potentially leads to better health outcomes.

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

Table 2: Main results (Acc). SC denotes the self-consistency prompting method. Results in **bold** are the best performances.

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

Table 3: Ablation study for different processes on MedQA. Anal: Analysis proposition, Summ: Report summarization, Cons: Collaborative consultation.

Method	**MedQA**	**MedMCQA**
MEDAGENTS (GPT-3.5)	64.1	59.3
MEDAGENTS (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.

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

Table 5: Optimal number of agents on MedQA, MedMCQA, PubMedQA, and MMLU.

Method	**MedQA**	**MedMCQA**
MEDAGENTS	63.8	58.9
Remove most relevant	60.5	55.4
Remove least relevant	66.2	61.5
Remove randomly	62.2	56.3

Table 6: Domain variation study. The results are based on GPT-3.5.

4.2 Comparison with Open-source Medical Models

We conduct a comprehensive comparison between our proposed 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

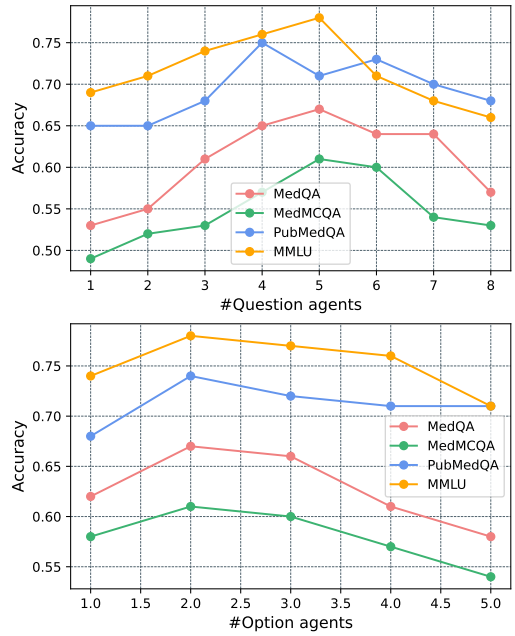

Figure 3: Influence of the number of question and option agents on various datasets.

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.

Method	MedQA	MedMCQA
MEDAGENTS		
w/ 6 different domains	64.1	59.3
w/ 6 same domains	59.2	58.1
w/ 5 same domains	57.5	57.3
w/ 4 same domains	55.9	57.0

Table 7: Agent quantity study. The results are based on GPT-3.5.

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*,

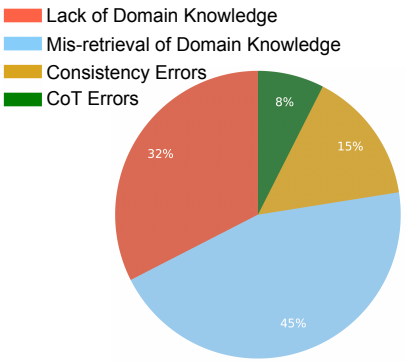

Figure 4: Ratio of different categories in error cases.

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; Rosoł et al., 2023). Although LLMs have demonstrated their potential in distinct medical applications encompassing diagnostics (Singhal et al., 2023a; Han et al., 2023), genetics (Duong and Solomon, 2023; Jin et al., 2023), pharmacist (Liu et al., 2023), and medical evidence summarization (Tang et al., 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-prompted data to obtain instruction datasets (Tu et al., 2023; Zhang et al., 2023a; Singhal et al., 2023b; Tang et al., 2023c). These

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

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

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., 2023b). 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

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. 2024. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*.

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*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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*.

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.

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

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.

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

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

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

Emily Harris. 2023. Large language models answer medical questions accurately, but canâĂŹt match cliniciansâĂŹ knowledge. *JAMA*.

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*.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

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

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*.

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*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Chunyuan 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*.

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*.

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

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

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.

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

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*.

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*.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

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.

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)*.

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*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

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.

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.

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*.

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.

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.

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.

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

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

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

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*.

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*.

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.

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.

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

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

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.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, 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*.

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

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

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

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023b. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint arXiv:2309.02233*.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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

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

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*.

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*.

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.

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

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*.

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*.

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*.

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

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.

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 hindrances 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.

Table 9: Case studies demonstrating how our proposed MEDAGENTS framework manages to correct previous errors.

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 middle ear.

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.

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

$prompt_{qd}$: 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: | .`

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.

$prompt_{od}$: 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 the same format as: `Medical Field: | .`

r_{qa}: You are a medical expert in the domain of `question_domain`. From your area of specialization, you will scrutinize and diagnose the symptoms presented by patients in specific medical scenarios.

$prompt_{qa}$: 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.

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

$prompt_{oa}$: 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 at first glance but could potentially be misleading in reality.

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

$prompt_{rs}$: 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:`

r_{vote}: You are a medical expert specialized in the `domain` domain.

$prompt_{vote}$: 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]`.

$prompt_{mod}$: 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:`

$prompt_{dm}$: 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.

	MedQA	MedMCQA
MC framework GPT-3.5 (single try)	64.1	59.3
MC framework GPT-3.5 (5 repetitions)	64.3	59.2
MC framework GPT-4 (single try)	83.7	74.8
MC framework GPT-4 (5 repetitions)	83.5	74.9

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