Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions

Guangzhi Xiong[1,*], Qiao Jin[2,*], Xiao Wang[3], Minjia Zhang[3], Zhiyong Lu[2,†], Aidong Zhang[1,†]

[1]*Department of Computer Science, University of Virginia, VA 22904, USA*
[2]*National Library of Medicine, National Institutes of Health, MD 20892, USA*
[3]*Department of Computer Science, University of Illinois Urbana-Champaign, IL 61801, USA*
*Equal contribution. †Co-correspondence.
E-mail: hhu4zu@virginia.edu, qiao.jin@nih.gov, xiaow4@illinois.edu, minjiaz@illinois.edu,
zhiyong.lu@nih.gov, aidong@virginia.edu

The emergent abilities of large language models (LLMs) have demonstrated great potential in solving medical questions. They can possess considerable medical knowledge, but may still hallucinate and are inflexible in the knowledge updates. While Retrieval-Augmented Generation (RAG) has been proposed to enhance the medical question-answering capabilities of LLMs with external knowledge bases, it may still fail in complex cases where multiple rounds of information-seeking are required. To address such an issue, we propose iterative RAG for medicine (*i*-MedRAG), where LLMs can iteratively ask follow-up queries based on previous information-seeking attempts. In each iteration of *i*-MedRAG, the follow-up queries will be answered by a vanilla RAG system and they will be further used to guide the query generation in the next iteration. Our experiments show the improved performance of various LLMs brought by *i*-MedRAG compared with vanilla RAG on complex questions from clinical vignettes in the United States Medical Licensing Examination (USMLE), as well as various knowledge tests in the Massive Multitask Language Understanding (MMLU) dataset. Notably, our zero-shot *i*-MedRAG outperforms all existing prompt engineering and fine-tuning methods on GPT-3.5, achieving an accuracy of 69.68% on the MedQA dataset. In addition, we characterize the scaling properties of *i*-MedRAG with different iterations of follow-up queries and different numbers of queries per iteration. Our case studies show that *i*-MedRAG can flexibly ask follow-up queries to form reasoning chains, providing an in-depth analysis of medical questions. To the best of our knowledge, this is the first-of-its-kind study on incorporating follow-up queries into medical RAG.

Keywords: Large Language Models; Retrieval-Augmented Generation; Medical Question Answering; AI for Healthcare.

1. Introduction

Generative artificial intelligence (AI) technologies such as large language models (LLMs) have brought a wide variety of opportunities for biomedical applications.[1–4] For example, they have shown great potential for answering biomedical questions,[5–9] summarizing medical documents,[10–12] and matching patients to clinical trials.[13–16] However, LLMs often generate plausible-sounding but inaccurate content, an issue commonly known as "hallucination" in the literature.[17] They also possess outdated knowledge obtained from a fixed set of training data.[18] Retrieval-augmented generation (RAG) provides a lightweight post-training solution to these issues by providing LLMs with relevant documents retrieved from up-to-date and trustworthy sources.[19,20]

While there have been several medical applications of RAG, such as Almanac,[21] Clinfo.ai,[22] and MedRAG,[23] their RAG component is mainly beneficial to questions that have direct answers in a single document, such as those in the PubMedQA[24] and BioASQ[25] datasets. However, only marginal improvements are seen with RAG for questions that require multiple rounds of clinical reasoning like MedQA,[26] a dataset curated from medical license examinations. For example, to recommend a treatment for a patient with certain symptoms, a system needs to first infer the potential diagnosis from the symptoms and then find a suitable treatment for the diagnosis. Nevertheless, only one round of retrieval is conducted in the vanilla RAG architecture, prohibiting multiple rounds of information seeking that are required in complex clinical reasoning.

In this work, we propose i-MedRAG, a simple and effective framework for incorporating follow-up queries into RAG. Specifically, we prompt LLMs to iteratively generate follow-up queries to search for additional information from external medical corpora. The queries and the corresponding answers generated with RAG will be used to augment the answer generation of the original question. Empirical results demonstrate the effectiveness of i-MedRAG on both open- and close-source LLMs, which show improved performance on the United States Medical Licensing Examination (USMLE) subset of MedQA and medical questions from the Massive Multitask Language Understanding (MMLU) dataset. Our further analysis of the number of iterations and number of queries per iteration used in i-MedRAG reflects how its performance scales with different settings. Additionally, we present several case studies of i-MedRAG, showing how it overcomes the limitations in vanilla RAG to find the correct answers.

In summary, our contributions are three-fold:

- We introduce i-MedRAG, a novel RAG architecture that incorporates follow-up queries to solve complex reasoning tasks.
- We have conducted comprehensive experiments on medical question answering, and the results demonstrate that i-MedRAG not only outperforms vanilla RAG approaches but also surpasses all other prompt engineering approaches on MedQA with GPT-3.5, setting a new state-of-the-art performance of 69.68%.
- We also provide analyses to further characterize i-MedRAG, showing how its performance varies with the scaling of follow-up queries.

2. Related Work

2.1. *Retrieval-Augmented Generation for Medicine*

Retrieval-augmented generation (RAG) has been widely adopted in medicine. Here, we discuss several representative approaches. Almanac[21] is a system that augments LLMs with curated resources for medical guidelines and treatment recommendations, which shows improvements over the standard LLMs in six manually assessed metrics. Similarly, Low *et al.*[27] demonstrate the improvements of RAG-based systems for real-world clinical queries with manual evaluation. Clinfo.ai[22] is an open-source web application that answers clinical questions based on retrieved scientific literature from PubMed articles. Xiong *et al.*[23] conduct a benchmarking study with the MedRAG toolkit, and show the benefits of RAG in several medical multi-choice question answering datasets. There are also various biomedical literature search products[28] that use RAG to summarize the retrieved articles,[29] such as OpenEvidence[a] and ChatRWD[b]. However, most of the RAG studies in medicine use the vanilla architecture with only one round of retrieval. There have been several attempts to use iterative data refinement for LLM training[30] or RAG[31–33] in the general domain. Nevertheless, similar ideas have not yet been explored in the medical domain. To the best of our knowledge, our study presents the first approach and evaluations on incorporating follow-up queries in RAG for medicine.

2.2. *Medical Question Answering*

Question answering tasks such as MedQA,[26] PubMedQA,[24] MedMCQA,[34] BioASQ,[25] and Massive Multitask Language Understanding (MMLU)[35] are commonly used to benchmark the medical knowledge and reasoning capabilities of LLMs.[36] Most of these datasets focus on single-hop questions such as "what is the most common symptom of hypertension?", while only MedQA questions are longer patient vignettes where both medical knowledge and multi-step reasoning are required. As such, there have been many studies working on improving the GPT-3.5 performance on MedQA with prompt engineering. Figure 1 shows the comparison among different representative prompt engineering approaches on MedQA, including chain-of-thought (CoT) prompting,[37] self-consistency (SC) prompting,[38] multi-agent communication with MedAgents,[39] and RAG-based approaches such as Knowledge Solver (KSL),[40] LLMs Augmented with Medical Textbooks (LLM-AMT),[41] and MedRAG.[23] Much fewer studies focus on prompt engineering with GPT-4 on MedQA,[7,42] probably because the raw GPT-4 error rate[43] is close to the noise rate in MedQA annotations.[44] In this study, we focus on the zero-shot setting as it reflects realistic clinical scenarios. While not requiring any instances for training or few-shot learning, our approach surpasses all previous methods with GPT-3.5 on the MedQA dataset.

3. Methods

Figure 2 shows the overview of our *i*-MedRAG and its comparison to the vanilla Retrieval-Augmented Generation (RAG). Different from RAG, our *i*-MedRAG modifies its pipeline by

---

[a]https://www.openevidence.com/
[b]https://www.atroposhealth.com/chatrwd

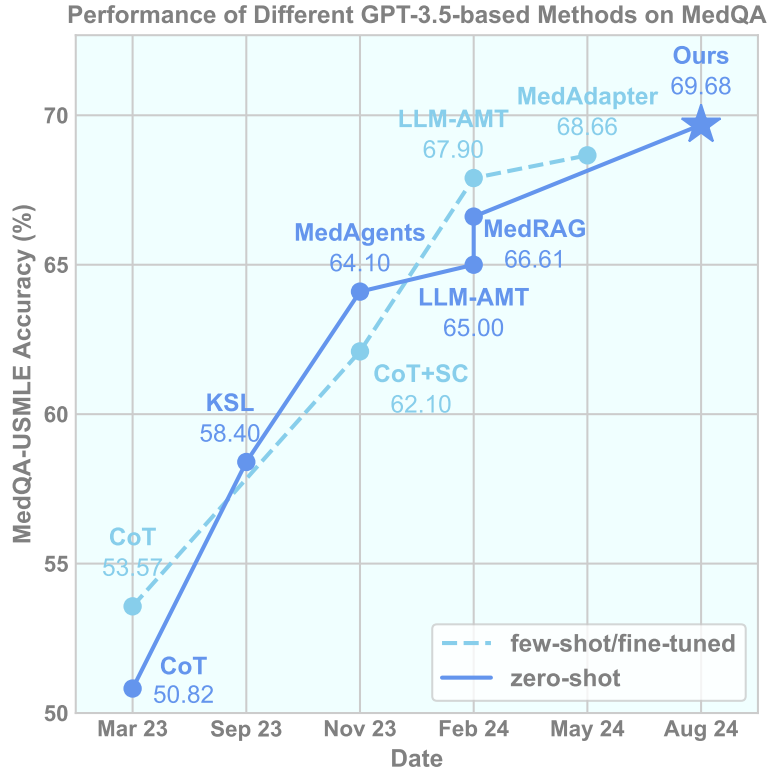**Performance of Different GPT-3.5-based Methods on MedQA**

Fig. 1. Comparison of various methods proposed to improve GPT-3.5 performance on MedQA. Our zero-shot i-MedRAG outperforms all previous prompt engineering and fine-tuning methods.

replacing the information retrieval step (Figure 2 left) with our proposed iterative question-answering step (Figure 2 middle and right). The settings of RAG are described in Section 3.1 and the pipeline of our new i-MedRAG is discussed in Section 3.2. The details of the iterative question answering are described in Section 3.3.

3.1. *Retrieval-Augmented Generation*

In the zero-shot setting of medical question answering, the task of LLM $\mathcal{M}$ is trying to find the correct answer $\mathcal{A}$ given the question $\mathcal{Q}$ only. The ideal answer prediction $\tilde{\mathcal{A}}$ can be provided by

$$\tilde{\mathcal{A}} = \arg\max_{\mathcal{A}} \mathbb{P}_{\mathcal{M}}(\mathcal{A} \mid \mathcal{Q}, \text{inst.}), \tag{1}$$

where the "inst." is the task instruction the user provides that instructs the model to perform the task. As medical questions are knowledge-intensive,[36] it benefits from accessing large-scale external corpora to search for useful information.[21–23] A typical method to combine LLM reasoning with external corpora is RAG, which first retrieves relevant documents from the corpus for the given medical question and enters the retrieved documents along with the question into LLM to augment its answer generation. Formally, the RAG pipeline can be
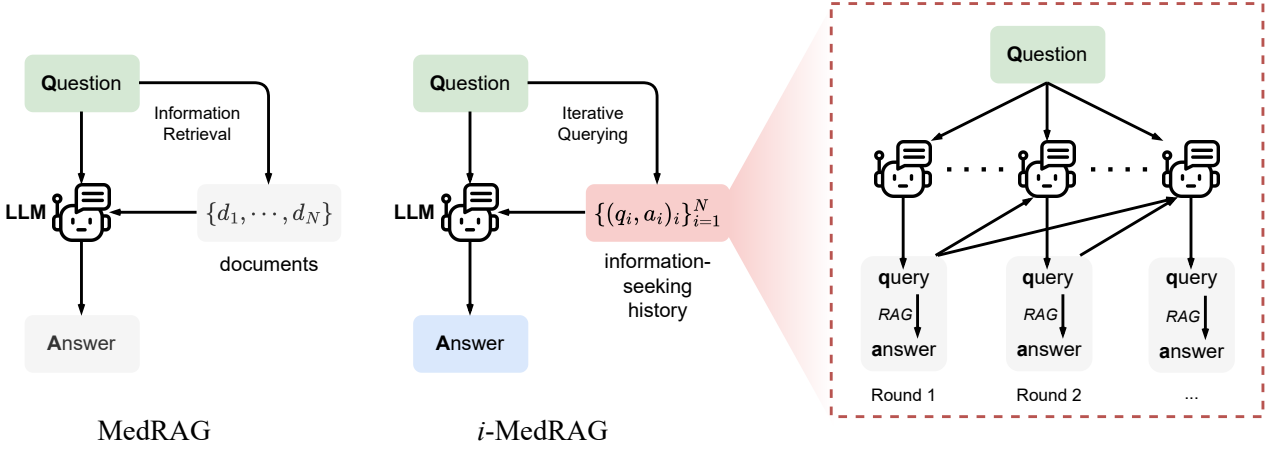
Fig. 2. Overview of i-MedRAG and its comparison to RAG (MedRAG). Left: the pipeline of Retrieval-Augmented Generation (RAG). Middle: the pipeline of our proposed i-MedRAG. Right: the iterative generation of question-specific medical query-answer (QA) pairs by asking follow-up queries.

described as

$$\tilde{\mathcal{A}} = RAG(\mathcal{Q}; \mathcal{M}, \mathcal{R}, \mathcal{D}) = \arg\max_{\mathcal{A}} \mathbb{P}_{\mathcal{M}}(\mathcal{A} \mid \mathcal{Q}, \text{inst.}, \{d_i\}_{i=1}^{N}), \tag{2}$$

where $\{d_i\}_{i=1}^{N}$ are the question-specific retrieved documents given by

$$\{d_i\}_{i=1}^{N} = \mathcal{R}(\mathcal{Q}; \mathcal{D}). \tag{3}$$

Here R is the text retriever and $\mathcal{D}$ is the corpus with a collection of documents.

3.2. Iterative Retrieval-Augmented Generation

While RAG exhibits promising performance in medical question answering,[23] it may be unable to handle certain complex medical questions in real-world cases. As text retrievers are typically trained to find relevant documents based on text similarity or lexicon overlap, they cannot break down a complex question and search for relevant information in a step-by-step manner. Thus, the inflexible retrieval step (Formula 3) in RAG may fail to analyze medical questions and find useful information to augment the answer generation, especially in complex clinical cases, where multiple rounds of information-seeking are required.

To address the issues mentioned, we propose to incorporate flexible information retrieval by prompting LLMs to iteratively generate follow-up queries based on the given medical question and previous information-seeking history. Moreover, as the context lengths of LLMs are limited, it can be impractical and infeasible to include all retrieved documents in the LLM context. Therefore, we prompt LLMs to directly answer the raised queries with relevant information and use such query-answer pairs as the information-seeking history. The pipeline of our proposed system can be formulated as

$$\tilde{\mathcal{A}} = i\text{-}MedRAG(\mathcal{Q}; \mathcal{M}, \mathcal{R}, \mathcal{D}) = \arg\max_{\mathcal{A}} \mathbb{P}_{\mathcal{M}}(\mathcal{A} \mid \mathcal{Q}, \text{inst.}, \{(q_i, a_i)\}_{i=1}^{N}), \tag{4}$$

where $\{(q_i, a_i)\}_{i=1}^{N}$ are the queries and the corresponding answers generated by LLMs with the help of RAG. The iterative process of query and answer generation will be detailed in Section 3.3.

3.3. *Iterative Generation of Follow-up Questions*

While the retrieved documents in RAG are determined by the question and the retrieval system, we propose to incorporate the reasoning capabilities of LLMs in i-MedRAG by prompting them to dynamically generate helpful queries in a step-by-step manner. Specifically, the LLM will be encouraged to generate n different queries to help find useful additional information for m iterations. In all iterations except for the first one, the model will be given the information-seeking history to generate context-specific follow-up queries. The queries $q_{i1}, \cdots, q_{in}$ generated in the i-th iteration can be formulated as

$$q_{i1}, \cdots, q_{in} = \begin{cases} \underset{q_{i1}, \cdots, q_{in}}{\arg\max} \, \mathbb{P}_{\mathcal{M}}(q_{i1}, \cdots, q_{in} \mid \mathcal{Q}, \text{inst.}'), & \text{if } i = 1, \\ \underset{q_{ic1}, \cdots, q_{in}}{\arg\max} \, \mathbb{P}_{\mathcal{M}}(q_{i1}, \cdots, q_{in} \mid \mathcal{Q}, \text{inst.}', \{(q_{jk}, a_{jk})\}_{j=1, \cdots, i-1}^{k=1, \cdots, n}), & \text{if } i > 1. \end{cases} \quad (5)$$

Different from the "inst." in Formula 2, the "inst.'" here is a modified instruction which focuses on generating follow-up queries instead of answering the medical question. For each query generation step, we prompt the LLM to analyze the existing information first and then generate new queries for additional knowledge. The step-by-step "reason-then-query" pipeline helps LLMs break down complex medical questions and find useful information from the external corpus. The answer to each generated query is given by a RAG system mentioned in Formula 2. This enables the system to leverage existing literature to provide grounded answers for generated queries.

The overall algorithm of i-MedRAG is presented in Algorithm 1.

4. Experiments

4.1. *Evaluation settings*

To evaluate the performance of our proposed i-MedRAG on knowledge-intensive medical question-answering tasks and compare it with other approaches, we select MedQA[26] as the testbed, which contains medical questions collected from United States Medical Licensing Examination (USMLE). With complex clinical cases in the dataset, MedQA reflects the difficulty of decision-making in real-world clinical medicine. The approaches for comparison are prompt engineering or fine-tuning methods that try to improve the performance of GPT-3.5 on MedQA, including chain-of-thought (CoT) prompting,[43] self consistency (SC), knowledge solver (KSL),[40] medical agents (MedAgents),[39] LLMs augmented with medical textbooks (LLM-AMT),[41] medical retrieval-augmented generation (MedRAG),[23] and LLMs with test-time adaptations (MedAdapter).[45]

Additionally, we evaluate the generalizability of our i-MedRAG with more LLMs and medical datasets. Llama-3.1-8B is selected as the representative of open-source models, which has a context window of 128k tokens. We also include MMLU-Med, a set of six medical tasks (anatomy, clinical knowledge, professional medicine, human genetics, college medicine,

---

Algorithm 1 The algorithm of i-MedRAG for medical question answering

---

Input medical question $\mathcal{Q}$, large language model $\mathcal{M}$, text retriever $\mathcal{R}$, medical corpus $\mathcal{D}$, query instruction "inst.'", answer instruction "inst.", hyperparameters m, n, N

Output answer prediction $\tilde{\mathcal{A}}$

1: Initialize the information-seeking history $\mathcal{H} = \text{emptylist}()$
2: **for** i in $1, 2, \cdots, m$ **do**
3:     **if** $i = 1$ **then**
4:         generate n new queries $q_{i1}, \cdots, q_{in}$ using $\mathcal{M}$ given $\mathcal{Q}$
5:     **else if** $i > 1$ **then**
6:         generate n new queries $q_{i1}, \cdots, q_{in}$ using $\mathcal{M}$ given $\mathcal{Q}$ and $\mathcal{H}$
7:     **end if**
8:     **for** j in $1, 2, \cdots, n$ **do**
9:         retrieve N relevant documents $d_{ij}^1, \cdots, d_{ij}^N$ using $\mathcal{R}$ and $\mathcal{D}$ given q_{ij}
10:         generate the answer a_{ij} using $\mathcal{M}$ given q_{ij} and $d_{ij}^1, \cdots, d_{ij}^N$
11:         add the query-answer pair (q_{ij}, a_{ij}) to the list $\mathcal{H}$
12:     **end for**
13: **end for**
14: generate the predicted answer $\tilde{\mathcal{A}}$ using $\mathcal{M}$ given $\mathcal{Q}$ and $\mathcal{H}$
15: **return** $\tilde{\mathcal{A}}$

---

college biology) from Massive Multitask Language Understanding (MMLU), following previous studies.[8,23] MMLU-Med serves as a testbed to show the performance of i-MedRAG on a variety of different medical tasks.

Both MedQA and MMLU-Med are composed of multi-choice questions, whose evaluation metric is the accuracy of predicted answers chosen from given options. For the retrieval part in i-MedRAG, we select the Textbooks[26] and Statpearls[c] corpora introduced in MedRAG,[23] which are shown effective on medical examination questions. MedCPT[46] is chosen as the text retriever, which has been trained on domain-specific literature. For other baselines compared, the official settings described in their papers are used.

4.2. *Main results*

Table 1 shows the comparison results of i-MedRAG and other baseline approaches on MedQA using GPT-3.5. Official scores reported in previous research are used for a fair comparison. While methods with few-shot learning or model fine-tuning tend to perform better than the ones in a zero-shot setting, our i-MedRAG set a state-of-the-art performance of GPT-3.5 on MedQA without any training samples or parameter tuning. Among zero-shot approaches, i-MedRAG (69.68%) has a significant performance improvement ($p < 0.05$) compared to the previous best record achieved by MedRAG (66.61%).

The results of generalizing i-MedRAG to more LLMs and data are presented in Table 2. We compare i-MedRAG with our implemented CoT and MedRAG to see if i-MedRAG

---

[c]`https://www.statpearls.com/`

Table 1. Performance of GPT-3.5 with different prompt engineering / fine-tuning methods on MedQA. The "External Knowledge" column denotes if the method augments LLM generation with information retrieval of external knowledge.

Method	External Knowledge	Setting	Accuracy (%)
Chain of Thought[43]	No	zero-shot	50.82
Knowledge Solver[40]	Yes	zero-shot	58.40
Chain of Thought + Self Consistency[39]	No	zero-shot	61.30
MedAgents[39]	No	zero-shot	64.10
LLMs Augmented with Medical Textbook[41]	Yes	zero-shot	65.00
MedRAG[23]	Yes	zero-shot	66.61
Chain of Thought[43]	No	five-shot	53.57
Chain of Thought + Self Consistency[39]	No	five-shot	62.10
LLMs Augmented with Medical Textbook[41]	Yes	fine-tuned	67.90
MedAdapter[45]	No	fine-tuned	68.66
i-MedRAG (ours)	**Yes**	**zero-shot**	**69.68**

can bring a consistent improvement of LLM performance in medical question answering. For all experiments with i-MedRAG, we tune the hyperparameters on a validation set of 100 samples and then report its scores on the test set. Similar to the results on GPT-3.5, the open-source Llama-3.1-8B also shows improved performance on MedQA with the help of i-MedRAG. While Llama-3.1-8B shows a close performance to GPT-3.5 in CoT and MedRAG settings, its performance is significantly improved with i-MedRAG, achieving an accuracy of 75.02%. The improved performance of GPT-3.5 and Llama-3.1-8B on MMLU-Med also demonstrates the generalizability of i-MedRAG to more medical data. As medical questions in MMLU-Med are less complex than the USMLE questions in MedQA, follow-up queries may not be necessary to find relevant information for the given question. Thus, it can be observed that the improvement by i-MedRAG compared to MedRAG is less significant in MMLU-Med than in MedQA.

Table 2. Performance of i-MedRAG on different LLMs and datasets. "Acc." denotes the accuracy. "Δ" shows the relative performance improvement compared with CoT.

Model	Method	MedQA-USMLE		MMLU-Med		Average	
		Acc.	Δ	Acc.	Δ	Acc.	Δ
GPT-3.5-Turbo	CoT	65.04	+0.00%	72.91	+0.00%	68.98	+0.00%
GPT-3.5-Turbo	MedRAG	66.61	+2.41%	75.48	+3.52%	71.05	+3.00%
GPT-3.5-Turbo	i-MedRAG	**69.68**	**+7.13%**	**75.85**	**+4.03%**	**72.77**	**+5.49%**
Llama-3.1-8B	CoT	64.73	+0.00%	77.23	+0.00%	70.98	+0.00%
Llama-3.1-8B	MedRAG	66.54	+2.80%	78.05	+1.06%	72.30	+1.86%
Llama-3.1-8B	i-MedRAG	**73.61**	**+13.72%**	**78.42**	**+1.54%**	**76.02**	**+7.10%**

4.3. *Scaling with iterations and queries*

As we described in Section 3.3, the number of iterations to ask follow-up queries and the number of queries generated in each iteration are the two critical hyperparameters in our proposed iterative generation of follow-up queries. To explore how different selections of the hyperparameter values affect the model performance, we run i-MedRAG with different settings and compare their results. We test both GPT-3.5 and Llama-3.1-8B on MedQA and MMLU-Med to examine if there are model-specific or task-specific patterns.

Figure 3 shows the model performance with different hyperparameter settings. Generally, MedQA and MMLU-Med show distinct patterns in performance change with the increasing number of iterations. While the performance of both GPT-3.5 and Llama-3.1-8B on MedQA tends to improve with more iterations of follow-up queries, their performance on MMLU-Med converges or starts to drop with just one or two iterations, corresponding to the different complexities of these two tasks.

From the results on MedQA, it is also empirically shown that the number of generated queries per iteration determines the rate of performance improvement and convergence over multiple iterations. LLMs with more queries generated per iteration tend to have a larger improvement in accuracy but also converge more quickly. Such a result is intuitively reasonable as more information can be collected each iteration with more generated queries.

4.4. *Case studies*

Table 3 shows the predictions of GPT-3.5 on a MedQA question with different prompt engineering approaches. The question asks about the mechanism of the drug for transitional cell carcinoma of the bladder which causes hearing loss. To solve the problem, it is important to find the exact drug and then figure out how it causes the mentioned symptoms. However, the CoT result shows that GPT-3.5 does not inherently contain sufficient medical knowledge to solve this problem. Instead of inferring the described drug, GPT-3.5 with CoT directly hallucinates a wrong option as the answer. While free radicals are relevant to hearing loss, their connection to the disease of the patient is unclear and not discussed. Compared to CoT which solely relies on the internal knowledge of LLMs, MedRAG provides an opportunity for LLMs to augment their answer generation with external medical knowledge. Nevertheless, the model output shows that the MedRAG system fails to retrieve useful information about the drug from medical corpora. Given the complex problem description, it is difficult for text retrievers to find the asked mechanism without knowing the drug.

With iteratively generated follow-up queries, our i-MedRAG manages to identify the described drug and find information about its mechanism. From Table 3, it can be observed that GPT-3.5 starts with a general query about the asked mechanism. However, similar to the case in MedRAG, the RAG system fails to provide useful information about the query. With the information-seeking history, GPT-3.5 updates its actions with follow-up queries with respect to side effects especially hearing loss. With the updated queries, it manages to identify "cisplatin" as the drug which is not explicitly mentioned in the question. A query about the mechanism of action of cisplatin is further proposed to search for information about the answer to the original question. With several iterations of adaptive question answering, GPT-3.5
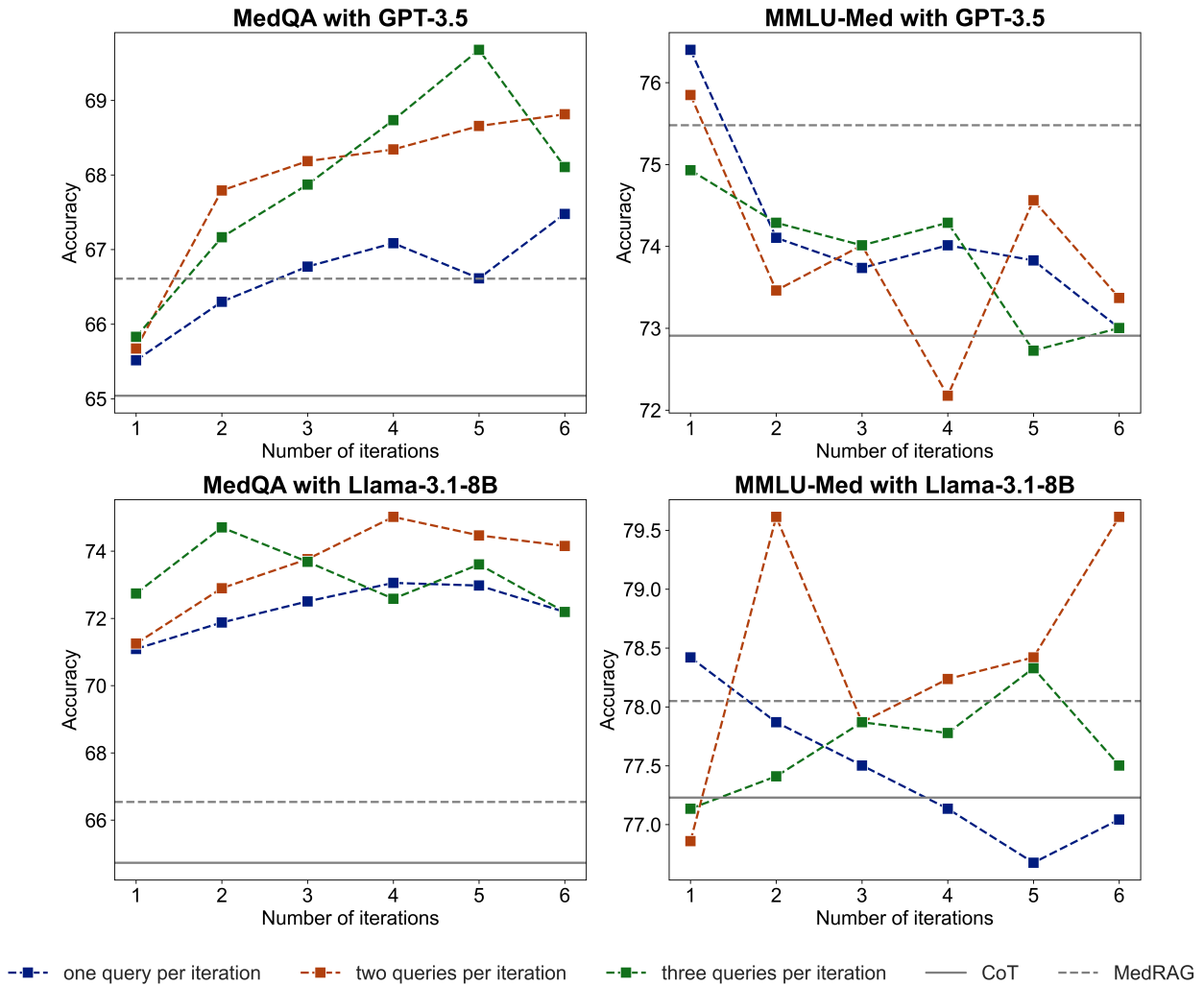
Fig. 3.   *i*-MedRAG performance on MedQA and MMLU-Med with different hyperparameters.

successfully finds the correct answer for the given clinical medical question.

Table 4 shows another case from the MMLU-Med dataset by Llama-3.1-8B. The case is selected from the "professional medicine" subtask of MMLU-Med, which contains complex clinical cases similar to those in MedQA. In the question, LLMs are asked to choose the next step in diagnosis given the described symptoms, which requires the initial analysis of potential diseases. Similar to GPT-3.5, the open-source Llama-3.1-8B may hallucinate wrong answers with just CoT or MedRAG, as shown by the highlighted incorrect rationales in Table 4. With iterative follow-up queries, *i*-MedRAG enables the LLM to find out the correct answer by making specific queries given existing information. For example, Llama-3.1-8B asks about the symptoms of *Clostridium difficile* infection (CDI) and diverticulitis in its first iteration which helps it to identify diverticulitis as the potential disease for the diagnosis. In the second iteration, the model queries about the CT scan for diverticulitis, where the answer provides key information that helps it identify the correct next step.

Table 3.   An example of how *i*-MedRAG helps GPT-3.5 to find the correct answer. The <mark>incorrect rationale</mark> and <mark>useful evidence</mark> are highlighted.

MedQA Question
0001: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions? **A.** Inhibition of proteasome, **B.** Hyperstabilization of microtubules, **C.** Generation of free radicals, **D.** Cross-linking of DNA

CoT
Analysis: The patient is experiencing a ringing sensation in his ear, which is a symptom of tinnitus. [...] Among the given options, the most likely mechanism of action that could cause sensorineural hearing loss is the generation of free radicals. Free radicals can cause oxidative damage to the hair cells and auditory nerve, leading to hearing loss. Therefore, the answer choice is C. Generation of free radicals. **Answer**: C ✗

MedRAG
Retrieved documents: [...] **Analysis**: [...] To determine the mechanism of action of the drug causing the symptoms, we need to identify the drug used for neoadjuvant chemotherapy. However, the drug is not mentioned in the question stem or the provided documents. Without this information, it is not possible to definitively determine the mechanism of action. Therefore, the answer cannot be determined based on the given information. **Answer**: Cannot be determined. ✗

***i*-MedRAG**
Query (Iteration 1): What is the mechanism of action of the neoadjuvant chemotherapy drug used in transitional cell carcinoma of the bladder? **Retrieved documents**: [...]   ***Answer***: Based on the provided documents, the mechanism of action of the neoadjuvant chemotherapy drug used in transitional cell carcinoma of the bladder is not explicitly mentioned [...]   ***Query (Iteration 2)***: What are the potential side effects of neoadjuvant chemotherapy drugs used in transitional cell carcinoma of the bladder? **Retrieved documents**: [...]   ***Answer***: The potential side effects [...] include myelosuppression, dysuria, urinary frequency, contact dermatitis, and systemic illness associated with granulomatous infections in multiple sites.   ***Query (Iteration 3)***: What is the specific neoadjuvant chemotherapy drug [...] that can cause sensorineural hearing loss? **Retrieved documents**: [...]   ***Answer***: <mark>The specific neoadjuvant chemotherapy drug used in transitional cell carcinoma of the bladder that can cause sensorineural hearing loss is **cisplatin**.</mark>   ***Query (Iteration 4)***: What is the <mark>mechanism of action of **cisplatin**</mark>? **Retrieved documents**: [...]   ***Answer***: The mechanism of action of cisplatin is achieved through the covalent binding of platinum to the purine bases guanine and adenine in DNA. This covalent binding leads to <mark>intra-strand and inter-strand crosslinks</mark> causing subsequent strand breaks [...]   **Final analysis**: [...] The drug that is most likely responsible for the patient's symptoms is cisplatin, which is known to cause sensorineural hearing loss. The mechanism of action of cisplatin involves cross-linking of DNA, leading to subsequent strand breaks and cell death. Therefore, the correct answer is D. **Answer**: D ✔

Table 4.   An example of how *i*-MedRAG helps Llama-3.1-8B to find the correct answer. The incorrect rationale and useful evidence are highlighted.

MMLU Question
professional_medicine-020: A 52-year-old woman comes to the emergency department because of a 1-week history of low-grade fever and increasing abdominal cramps that are exacerbated by bowel movements [...] Her temperature is 38.1°C (100.6°F), pulse is 75/min, and blood pressure is 150/80 mm Hg [...] there is tenderness to palpation of the left lower quadrant with guarding but no rebound [...] leukocyte count is 15,000/mm3 [...] Which of the following is the most appropriate next step in diagnosis? **A.** Test of the stool for Clostridium difficile toxin **B.** Endoscopic retrograde cholangiopancreatography **C.** Pelvic ultrasonography **D.** CT scan of the abdomen with contrast

CoT
Analysis: [...] the differential diagnosis includes diverticulitis, appendicitis, and bowel obstruction. However, the absence of bloody stools and the patient's response to antibiotics makes diverticulitis less likely. Therefore, the most appropriate next step in diagnosis is to test for Clostridium difficile toxin in the stool. **Answer**: A ✗

MedRAG
Retrieved documents: [...] **Analysis**: [...] Given the patient's symptoms and history, the most likely diagnosis is a pelvic abscess or a complication related to the previous surgery. The most appropriate next step in diagnosis would be to investigate the pelvic area for any potential complications or abscesses. **Answer**: C ✗

i-MedRAG
Query 1 (Iteration 1): What are the common symptoms and diagnostic criteria for Clostridium difficile infection (CDI)? **Retrieved documents**: [...] **Answer**: [...] ***Query 2 (Iteration 1)***: What are the typical findings on physical examination and laboratory tests for diverticulitis? **Retrieved documents**: [...] **Answer**: Based on the provided documents, the typical findings on physical examination and laboratory tests for diverticulitis are: *Physical Examination*: Tenderness to palpation over the area of inflammation (almost always present) [...] Fever is almost always present, but hypotension and shock are uncommon. *Laboratory Tests*: Leukocytosis (elevated white blood cell count) [...] ***Query 1 (Iteration 2)***: What are the typical imaging findings for diverticulitis on a CT scan of the abdomen with contrast? **Retrieved documents**: [...] **Answer**: [...] These findings are consistent with the diagnosis of diverticulitis, and are often used to differentiate it from other conditions that may present with similar symptoms. ***Query 2 (Iteration 2)***: What are the indications for surgical intervention in a patient with diverticulitis? **Retrieved documents**: [...] **Answer**: [...] **Final Analysis**: [...] The patient's laboratory results show a leukocyte count of 15,000/mm^3, indicating leukocytosis [...] The patient's symptoms of abdominal cramps, fever, and leukocytosis, combined with tenderness to palpation of the left lower quadrant, are consistent with diverticulitis [...] the most appropriate next step in diagnosis is to perform a CT scan of the abdomen with contrast to evaluate for diverticulitis and rule out other causes of her symptoms. **Answer**: D ✔

5. Discussion

Overall, our proposed i-MedRAG effectively improves the performance of LLMs on complex medical questions by prompting them to iteratively ask follow-up queries. The experimental results show that our approach is better than previously proposed prompt engineering and fine-tuning methods, and is generalizable to various LLMs and medical question-answering datasets. Nevertheless, our approach has certain limitations which need to be discussed. It is also worthwhile to discuss the future work of this study to analyze how it can be further improved to facilitate real-world medical assistance.

5.1. *Limitations*

The first limitation of i-MedRAG is its high cost. While generating more follow-up queries tends to provide LLMs with more comprehensive and focused information about the given medical question, the cost also grows linearly with the number of queries generated. The time cost can be further increased if more documents are used to help answer the generated queries with RAG. While the cost is comparable to approaches using multiple LLM agents[39] or self consistency[47] which also prompt LLMs multiple times for each question, it is much more costly than baseline prompting methods such as CoT.[37]

Another limitation is the selection of hyperparameter values for optimal performance. As shown in Figure 3, different LLMs can have different hyperparameter settings for their optimal performance. Even for the same LLM, its optimal hyperparameters can vary based on the medical questions being processed. Thus, it is non-trivial to find the optimal hyperparameters of i-MedRAG for a new medical task, which may be inefficient for real-world deployments.

5.2. *Future work*

Given the limitations of i-MedRAG, we consider several potential future directions that could further improve the performance of retrieval-augmented generation for medicine. The first direction is the automation of hyperparameter selection in i-MedRAG. To reduce the laborious process of hyperparameter selection, one may use an LLM agent to dynamically determine how many follow-up queries should be asked each iteration. This can improve the efficiency and flexibility of the hyperparameter selection process. Another future direction is to improve the performance of i-MedRAG with few-shot demonstrations. While few-shot CoT prompting is demonstrated to perform better than the zero-shot counterpart,[43] it is not easy to adapt such strategies to i-MedRAG as the reasoning process can be dynamically affected by the use of external corpora and retrievers. Investigating how i-MedRAG can benefit from one or few-shot samples could be a potential direction to further enhance its performance on medical question answering. More quantitative analysis can also be performed to examine the error types of i-MedRAG compared to existing methods.

6. Acknowledgments

References

1. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, Large language models in medicine, *Nature medicine* **29**, 1930 (2023).

2. H. Zhou, B. Gu, X. Zou, Y. Li, S. S. Chen, P. Zhou, J. Liu, Y. Hua, C. Mao, X. Wu *et al.*, A survey of large language models in medicine: Progress, application, and challenge, *arXiv preprint arXiv:2311.05112* (2023).

3. J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou and R. Daneshjou, Large language models in medicine: the potentials and pitfalls: a narrative review, *Annals of Internal Medicine* **177**, 210 (2024).

4. S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, X. Chen, Y. Yang, Q. Chen, W. Kim, D. C. Comeau *et al.*, Opportunities and challenges for chatgpt and large language models in biomedicine and health, *Briefings in Bioinformatics* **25**, p. bbad493 (2024).

5. R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon and T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in bioinformatics* **23**, p. bbac409 (2022).

6. V. Liévin, C. E. Hother, A. G. Motzfeldt and O. Winther, Can large language models reason about medical questions?, *Patterns* **5** (2024).

7. H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu *et al.*, Can generalist foundation models outcompete special-purpose tuning? case study in medicine, *arXiv preprint arXiv:2311.16452* (2023).

8. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, Large language models encode clinical knowledge, *Nature* **620**, 172 (2023).

9. E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin *et al.*, Biomedlm: A 2.7 b parameter language model trained on biomedical text, *arXiv preprint arXiv:2403.18421* (2024).

10. C. Shaib, M. Li, S. Joseph, I. Marshall, J. J. Li and B. C. Wallace, Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success), in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

11. L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau *et al.*, Evaluating large language models on medical evidence summarization, *npj Digital Medicine* **6**, p. 158 (2023).

12. D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová *et al.*, Adapted large language models can outperform medical experts in clinical text summarization, *Nature Medicine* , 1 (2024).

13. Q. Jin, Z. Wang, C. S. Floudas, F. Chen, C. Gong, D. Bracken-Clarke, E. Xue, Y. Yang, J. Sun and Z. Lu, Matching patients to clinical trials with large language models, *ArXiv* (2023).

14. C. Wong, S. Zhang, Y. Gu, C. Moung, J. Abel, N. Usuyama, R. Weerasinghe, B. Piening, T. Naumann, C. Bifulco *et al.*, Scaling clinical trial matching using large language models: A case study in oncology, in *Machine Learning for Healthcare Conference*, 2023.

15. M. Wornow, A. Lozano, D. Dash, J. Jindal, K. W. Mahaffey and N. H. Shah, Zero-shot clinical trial patient matching with llms, *arXiv preprint arXiv:2402.05125* (2024).

16. S. Zhuang, B. Koopman and G. Zuccon, Team ielab at trec clinical trial track 2023: Enhancing clinical trial retrieval with neural rankers and large language models, *arXiv preprint arXiv:2401.01566* (2024).

17. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto and P. Fung, Survey of hallucination in natural language generation, *ACM Computing Surveys* **55**, 1 (2023).

18. T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu and G. Haffari, Continual learning for large language models: A survey, *arXiv preprint arXiv:2402.01364* (2024).

19. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems* **33**, 9459 (2020).

20. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun and H. Wang, Retrieval-augmented generation for large language models: A survey, *arXiv preprint arXiv:2312.10997* (2023).

21. C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley *et al.*, Almanac—retrieval-augmented language models for clinical medicine, *NEJM AI* **1**, p. AIoa2300068 (2024).

22. A. Lozano, S. L. Fleming, C.-C. Chiang and N. Shah, Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature, in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, 2023.

23. G. Xiong, Q. Jin, Z. Lu and A. Zhang, Benchmarking retrieval-augmented generation for medicine, *arXiv preprint arXiv:2402.13178* (2024).

24. Q. Jin, B. Dhingra, Z. Liu, W. Cohen and X. Lu, Pubmedqa: A dataset for biomedical research question answering, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

25. G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos *et al.*, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, *BMC bioinformatics* **16**, 1 (2015).

26. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang and P. Szolovits, What disease does this patient have? a large-scale open domain question answering dataset from medical exams, *Applied Sciences* **11**, p. 6421 (2021).

27. Y. S. Low, M. L. Jackson, R. J. Hyde, R. E. Brown, N. M. Sanghavi, J. D. Baldwin, C. W. Pike, J. Muralidharan, G. Hui, N. Alexander *et al.*, Answering real-world clinical questions using large language model based systems, *arXiv preprint arXiv:2407.00541* (2024).

28. Q. Jin, R. Leaman and Z. Lu, Pubmed and beyond: biomedical literature search in the age of artificial intelligence, *EBioMedicine* **100** (2024).

29. Q. Jin, R. Leaman and Z. Lu, Retrieve, summarize, and verify: how will chatgpt affect information seeking from the medical literature?, *Journal of the American Society of Nephrology* **34**, 1302 (2023).

30. X. Li, P. Yu, C. Zhou, T. Schick, O. Levy, L. Zettlemoyer, J. E. Weston and M. Lewis, Self-alignment with instruction backtranslation, in *The Twelfth International Conference on Learning Representations*, 2024.

31. Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan and W. Chen, Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy, in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

32. H. Trivedi, N. Balasubramanian, T. Khot and A. Sabharwal, Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

33. Z. Jiang, M. Sun, L. Liang and Z. Zhang, Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach, *arXiv preprint arXiv:2407.13101* (2024).

34. A. Pal, L. K. Umapathi and M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in *Conference on Health, Inference, and Learning*, 2022.

35. D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song and J. Steinhardt, Measuring massive multitask language understanding, *arXiv preprint arXiv:2009.03300* (2020).

36. Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu and S. Yu, Biomedical question answering: a survey of approaches and challenges, *ACM Computing Surveys (CSUR)* **55**, 1 (2022).

37. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* **35**, 24824 (2022).

38. X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, *arXiv preprint arXiv:2203.11171* (2022).

39. X. Tang, A. Zou, Z. Zhang, Y. Zhao, X. Zhang, A. Cohan and M. Gerstein, Medagents: Large language models as collaborators for zero-shot medical reasoning, *arXiv preprint arXiv:2311.10537* (2023).

40. C. Feng, X. Zhang and Z. Fei, Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs, *arXiv preprint arXiv:2309.03118* (2023).

41. Y. Wang, X. Ma and W. Chen, Augmenting black-box llms with medical textbooks for clinical question answering, *arXiv preprint arXiv:2309.02233* (2023).

42. T. Savage, A. Nayak, R. Gallo, E. Rangan and J. H. Chen, Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine, *NPJ Digital Medicine* **7**, p. 20 (2024).

43. H. Nori, N. King, S. M. McKinney, D. Carignan and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* (2023).

44. K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi *et al.*, Capabilities of gemini models in medicine, *arXiv preprint arXiv:2404.18416* (2024).

45. W. Shi, R. Xu, Y. Zhuang, Y. Yu, H. Wu, C. Yang and M. D. Wang, Medadapter: Efficient test-time adaptation of large language models towards medical reasoning, *arXiv preprint arXiv:2405.03000* (2024).

46. Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur and Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, *Bioinformatics* **39**, p. btad651 (2023).

47. X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-consistency improves chain of thought reasoning in language models, in *The Eleventh International Conference on Learning Representations*,