RAR^2: Retrieval-Augmented Medical Reasoning via Thought-Driven Retrieval

Kaishuai Xu[1], **Wenjun Hou**[1,2*], **Yi Cheng**[1*], **Wenjie Li**[1]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2]Research Institute of Trustworthy Autonomous Systems and
Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China
{kaishuaii.xu, alyssa.cheng}@connect.polyu.hk,
houwenjun060@gmail.com, cswjli@comp.polyu.edu.hk

Abstract

Large Language Models (LLMs) have shown promising performance on diverse medical benchmarks, highlighting their potential in supporting real-world clinical tasks. Retrieval-Augmented Generation (RAG) has emerged as a key approach for mitigating knowledge gaps and hallucinations by incorporating external medical information. However, RAG still struggles with complex medical questions that require intensive reasoning, as surface-level input often fails to reflect the true knowledge needs of the task. Existing methods typically focus on refining queries without explicitly modeling the reasoning process, limiting their ability to retrieve and integrate clinically relevant knowledge. In this work, we propose RAR^2, a joint learning framework that improves both Reasoning-Augmented Retrieval and Retrieval-Augmented Reasoning. RAR^2 constructs a thought process to uncover implicit knowledge requirements and uses it to guide retrieval and answer generation. We build a training dataset of mixed preference pairs and apply Direct Preference Optimization (DPO) to train the model. Moreover, we design two test-time scaling strategies to explore the boundaries of our framework. Experiments demonstrate the effectiveness of RAR^2 across several biomedical question answering datasets, outperforming RAG baselines with or without fine-tuning.

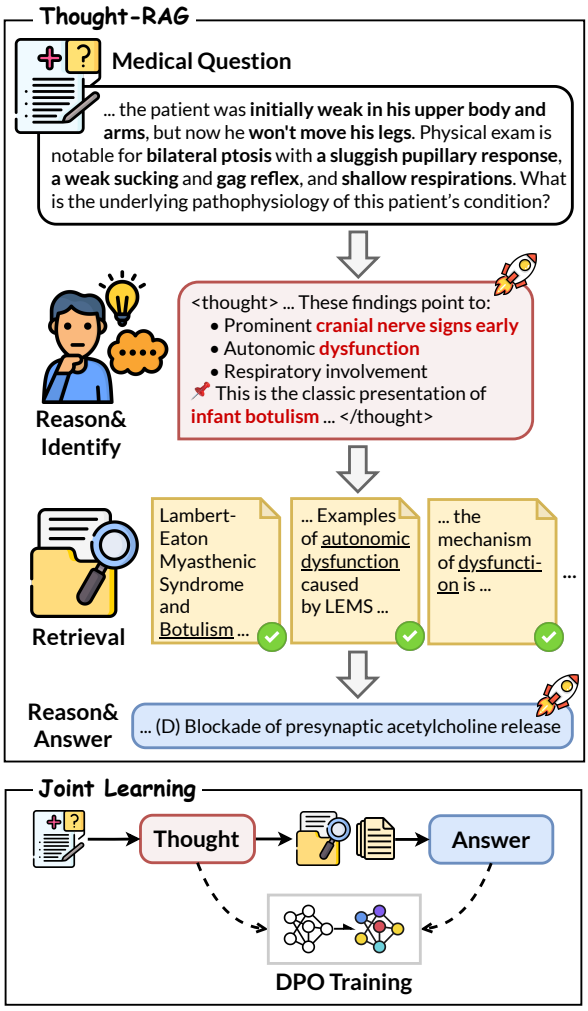

Figure 1: Thought-RAG for medical reasoning. Our method jointly optimizes the generation of thought processes and the subsequent retrieval-augmented answer generation.

1 Introduction

The capabilities of Large Language Models (LLMs) in medicine have long attracted substantial research attention (Singhal et al., 2023a,b; Wang et al., 2025). LLMs such as GPT-4o and Baichuan-M1 have demonstrated strong performance across diverse medical benchmarks, highlighting their potential to support real-world clinical tasks (Chen et al., 2024a; Wang et al., 2025; Xu et al., 2024a). More recently, given the vast scope and high-stakes nature of the medical domain, Retrieval-Augmented Generation (RAG) has emerged as a key approach for mitigating knowledge gaps and hallucinations by leveraging external medical knowledge (Lu et al., 2025; Wang et al., 2024; Wu et al., 2024; Yang et al., 2025).

While RAG performs well in a range of medical tasks, it still struggles with complex medical questions requiring intensive reasoning (Su et al., 2024).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20386–20396
November 4-9, 2025

This difficulty arises because the surface-level information in these questions does not reflect their actual knowledge requirements. For example, in Figure 1, although the question clearly describes the patient's symptoms, such as "weak in upper body and arms" and "won't move legs", the knowledge necessary to generate the correct answer lies in clinical concepts like "autonomic dysfunction" and "infant botulism". These concepts must be inferred through detailed analytical reasoning. Existing RAG methods for complex medical reasoning primarily focus on direct query refinement, but fail to construct and optimize a reasoning process that comprehensively uncovers the underlying knowledge requirements (Lu et al., 2025; Liang et al., 2025; Hu et al., 2024).

Additionally, few studies have explored enhancing the medical reasoning capabilities of LLMs within the RAG framework (Jeong et al., 2024). Reasoning with external knowledge is more challenging, as the retrieved content inevitably includes noise information. LLMs should learn to integrate relevant knowledge into the reasoning process while avoiding interference from irrelevant information. Therefore, optimizing retrieval-augmented reasoning is essential for improving both the accuracy and robustness of generation outcomes.

In this work, we construct a thought process to reason through medical questions and identify their implicit knowledge requirements. This thought process is directly used to retrieve relevant information, which is subsequently incorporated into reasoning to derive the final answer. To optimize both thought and answer generation, we propose a joint learning framework, **RAR**^{2}, that simultaneously improves **R**easoning-**A**ugmented **R**etrieval and **R**etrieval-**A**ugmented **R**easoning. Specifically, we first construct a training dataset consisting of mixed preference pairs. One type is thought pairs, in which a sampled thought process is annotated based on the outcome of subsequent retrieval-augmented generation. The other type is answer pairs, in which a sampled answer is annotated according to its correctness. We then apply Direct Preference Optimization (DPO) to fine-tune the LLM (Rafailov et al., 2023). These preference pairs enable supervised preference learning, allowing the model to identify relevant knowledge and reason effectively with external information. Extensive experiments across several biomedical question answering datasets demonstrate RAR^2's superiority over existing RAG baselines. Our further test-time scaling analysis validates the scalability of RAR^2.

In summary, our contributions are as follows:

- We propose a joint learning framework, RAR^2, that simultaneously improves reasoning-augmented retrieval and retrieval-augmented reasoning. Additionally, we design two test-time scaling strategies to explore the boundaries of our framework.
- We construct a mixed preference dataset to train the LLM to identify implicit knowledge needs and reason effectively with external information.
- Experimental results demonstrate the effectiveness of RAR^2 in six biomedical question answering datasets. Our framework outperforms the baseline under tuning-free and fine-tuned settings.

2 Preliminary

2.1 Problem Formulation

In this work, we focus on LLM-based medical reasoning for medical question answering. Given a question q and a large medical corpus $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, a set of top-relevant documents $\mathcal{D}_k$ is retrieved for q, where n is the total number of documents in the corpus and k is the number of retrieved ones. Then, the question and the retrieved documents are used as context for an LLM $\mathcal{M}$ to generate an answer y by step-by-step reasoning. Our goal is to jointly optimize document retrieval and retrieval-augmented medical reasoning.

2.2 Medical Corpus

A medical corpus is crucial for retrieval-augmented medical reasoning, as it provides the external knowledge necessary for LLMs to reason effectively about a given question. In this work, we adopt MedCorp, a large-scale medical corpus proposed by Xiong et al. (2024a). MedCorp integrates four major sources: PubMed[1], StatPearls[2], medical textbooks (Wang et al., 2024), and Wikipedia. It comprises 30.4M documents, including clinical guidelines, peer-reviewed research articles, and medical encyclopedic content. To process the documents, we reuse the original chunking strategy provided with the corpus and denote each document chunk as $d_i \in \mathcal{D}$, where $\mathcal{D}$ is the set of all chunks.

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.statpearls.com/

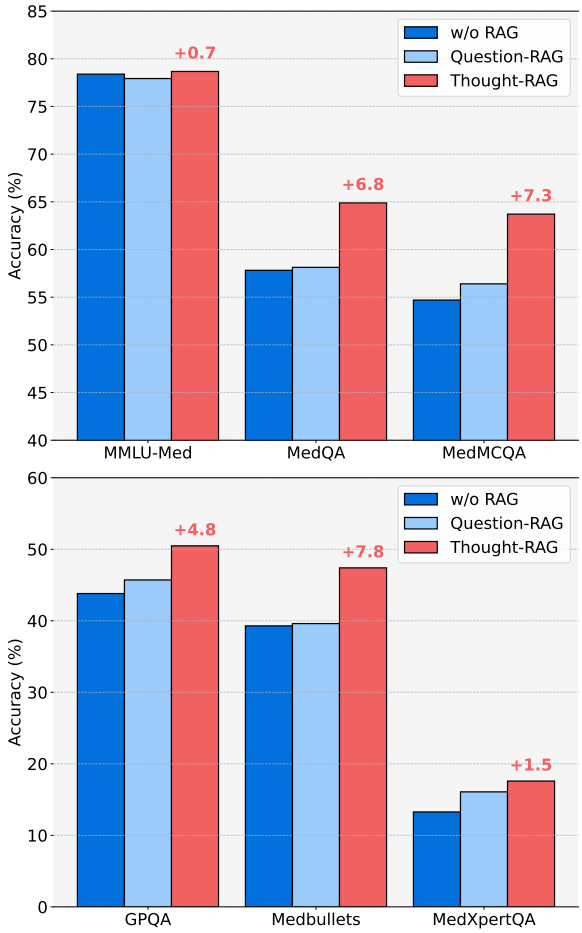

Figure 2: The performance comparison between Question-RAG and Thought-RAG on several medical question answering datasets.

3 Reasoning Before Retrieval

We begin with a preliminary study on the impact of reasoning prior to retrieval. Rather than fine-tuning a model to generate or refine a query, we instead sample a thought process that could guide the model toward a final answer and use it directly as a query to retrieve relevant documents. The thought process is sampled in a zero-shot manner by prompting the LLM with: "*Please reason step by step to identify the relevant knowledge that may be involved.*" For retrieval, we use the BM25 algorithm (Robertson and Zaragoza, 2009) to retrieve the top-k documents from the corpus. These documents are then provided as context to the LLM for answer generation. We evaluate this approach—referred to as Thought-RAG—against standard question-based retrieval (Question-RAG) using the Qwen2.5-7B-Instruct model (Yang et al., 2024) with $k = 32$.

Figure 2 shows the performance comparison on several medical question answering datasets. We can observe that Thought-RAG consistently outperforms Question-RAG, with an average improvement of 4.82% across all datasets. Especially in challenging datasets like Medbullets and GPQA, the gains are more significant. Furthermore, for most datasets, the accuracy remains unchanged or even declines using Question-RAG. A similar phenomenon can be found in the MIRAGE benchmark (Xiong et al., 2024a). These observations suggest that reasoning-intensive medical questions often do not explicitly state the specific information needs required for successful retrieval. In contrast, a thought process generated through prior reasoning more accurately captures the underlying knowledge requirements, thereby improving overall RAG performance. Reasoning first and then using the resulting thought process for retrieval is a straightforward yet powerful approach to collect relevant knowledge, but it has received limited attention within the RAG community (Sohn et al., 2024). Therefore, further optimizing the generation of the thought process holds promise for enhancing retrieval quality and improving the RAG performance.

4 Method

In this section, we introduce the RAR^2 framework, which constructs a mixed preference dataset consisting of thought and answer pairs, and applies DPO to jointly enhance both reasoning-augmented retrieval and retrieval-augmented reasoning for LLMs. As shown in Figure 3, we first sample a set of thought processes and answers. The thought processes are generated based on the medical question, while the answers are generated using the question along with the corresponding retrieved documents. All samples are annotated based on the correctness of the answer (§4.1). We collect samples to form a mixed preference dataset and apply DPO to jointly optimize two types of reasoning processes (§4.2). Besides, we design two test-time scaling strategies to examine the boundaries of RAR^2 (§4.3).

4.1 Construction of Mixed Preference Pairs

Our goal is to formulate and optimize two types of reasoning processes: one for retrieval and the other for retrieval-augmented generation. Previous work often overlooks the reasoning process required to investigate the underlying knowledge requirements, and few studies focus on jointly optimizing both

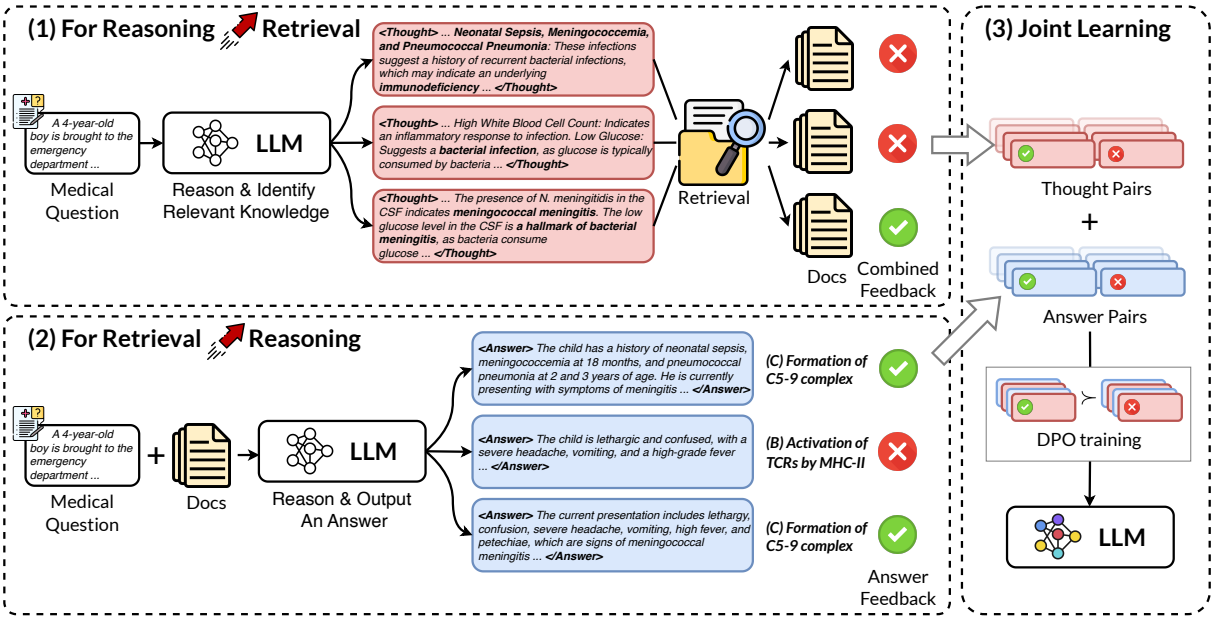

Figure 3: Construction of mixed preference pairs. (1) We sample several thought processes and append each to the question to form a query for retrieval. Each thought is then annotated based on whether it leads to a correct answer in subsequent RAG. (2) We sample several answers and annotate each one according to its correctness.

retrieval and generation. To address this, we design distinct preference pairs tailored to each type of reasoning process.

Thought Pairs. We use an instruction-tuned LLM as the base model to sample thought processes. Given a medical question q, we first prompt the model with "*Please reason step by step to identify the relevant knowledge that may be involved.*" to sample several thought processes:

$$y^t \sim \mathcal{M}(q, \text{Prompt}_t). \tag{1}$$

We then annotate them as chosen and rejected based on two criteria. The first criterion appends the prompt "*The answer is:* " to the end of each thought process and lets the LLM complete the answer. The correctness of the result is denoted as $g^{\text{direct}} \in \{0, 1\}$. The second criterion uses the thought process as a query to retrieve the top-k documents, which are then used to generate an answer. The correctness of this result is denoted as $g^{\text{rag}} \in \{0, 1\}$. To ensure deterministic outputs, we set the temperature to 0 during annotation. The thought samples are labeled as chosen only if both $g^{\text{direct}} = 1$ and $g^{\text{rag}} = 1$. Such an annotation ensures that the thought process can lead to a correct answer. We collect one chosen and one rejected thought process to form a preference pair, denoted as (y^{t+}, y^{t-}).

Answer Pairs. We use the same base model to sample answer candidates. Given a medical question q and a thought process y^t, we first retrieve top-k documents $\mathcal{D}_k$ from the corpus. Then, we sample a group of answers based on the question and each retrieved document set:

$$y^a \sim \mathcal{M}(q, \mathcal{D}_k, \text{Prompt}_a), \tag{2}$$

where Prompt_a is "*Please reason step by step and choose one option from the above*". The sample is labeled as chosen if its answer is correct. We collect one chosen and one rejected answer to form a preference pair, denoted as (y^{a+}, y^{a-}).

4.2 Joint Learning

After collecting the thought and answer pairs, we construct a mixed preference dataset $\mathcal{Y} = \{(y^{t+}, y^{t-})\} \cup \{(y^{a+}, y^{a-})\}$. We then apply DPO to jointly optimize both types of reasoning processes. Joint training is adopted because the two processes are complementary and can benefit from each other: reasoning to identify relevant knowledge facilitates answer generation, while answer generation, in turn, helps refine the analytical quality of the thought process. Previous studies fail to optimize these reasoning processes or consider their complementary relationship.

For DPO training, we thoroughly shuffle all pref-

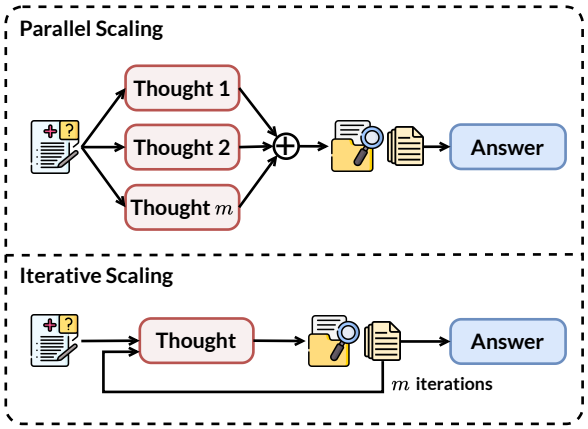

Figure 4: Frameworks of test-time scaling. Parallel scaling concatenates multiple thought processes to retrieve documents, while iterative scaling uses the retrieved documents in sequence to enhance thought generation.

erence pairs and employ the DPO loss as follows:

$$\mathcal{L} = -\mathbb{E}_{(q,y^+,y^-)\sim\mathcal{Y}}\big[\log\sigma\big(\beta\log\frac{\pi_\theta(y^+|q)}{\pi_\mathcal{M}(y^+|q)} - \beta\log\frac{\pi_\theta(y^-|q)}{\pi_\mathcal{M}(y^-|q)}\big)\big], \quad (3)$$

where σ is the sigmoid function, π_θ is the policy model, $\pi_\mathcal{M}$ is the reference model, and β is the hyperparameter that regulates the extent of deviation from the reference model.

4.3 Test-Time Scaling

To investigate the test-time scaling effect of RAR^2, we design two scaling strategies, as illustrated in Figure 4. Previous studies have paid little attention to the continual improvement of medical reasoning within the RAG framework.

The first strategy (**Paralleling Scaling**) generates m thought processes in parallel and concatenates them to form a single extended thought process. This combined thought is then used to retrieve documents, based on which an answer is generated. The second strategy (**Iterative Scaling**) generates a thought process and retrieves documents, which are then appended to the prompt for the next round of thought generation. This iterative process is repeated m times, after which an answer is generated. We design these two strategies to explore how RAR^2 scales with the number of generated thought processes. The first strategy merges more useful information to retrieve documents, while the second exploits the retrieved documents to generate more accurate thought processes.

5 Experiments

5.1 Evaluation Datasets

We evaluate our method on six biomedical question answering datasets: MedQA, MedMCQA (Pal et al., 2022), MMLU-Med (MMLU.) (Hendrycks et al., 2021), Medbullets (Chen et al., 2024a), GPQA (Rein et al., 2023), and MedXpertQA (Zuo et al., 2025). Among these, MedQA is the in-domain dataset, as our training data is derived from it. The remaining datasets are considered out-of-domain. The number of answer options across datasets ranges from 4 to 10. Notably, Medbullets, GPQA, and MedXpertQA are more recent and consist of graduate- or expert-level problems. Overall, the diversity of these datasets enables a robust evaluation of the model's medical reasoning capabilities.

5.2 Baseline Methods

We compare our framework with six strong baselines in medical reasoning. Qwen2.5-7B-Instruct (Yang et al., 2024) is used as our base model. Two recent models enhanced for medical reasoning that do not use RAG are included: m1-7B-23K (Huang et al., 2025) and HuatuoGPT-o1 (Chen et al., 2024b). For a fair comparison, we use HuatuoGPT-o1-7B with the same model size. Additionally, we include three tuning-free RAG methods that have been applied to medical tasks: MedRAG (Xiong et al., 2024a), i-MedRAG (Xiong et al., 2024b), and Med-R^2 (Lu et al., 2025). Med-R^2 is developed through a few-shot generation strategy. Finally, we include four RAG methods that involve fine-tuning on specific components: Self-BioRAG (Jeong et al., 2024), SimRAG-8B (Xu et al., 2024b), RAG^2-7B (Lu et al., 2025), and SPO (Chen et al., 2025). Baselines with publicly released models are reimplemented for comparison.

5.3 Implementation Details

We utilize the MedQA training set (Jin et al., 2020) to implement our method, which comprises 10K medical problems derived from professional medical board exams. Each problem requires selecting the correct answer from four options. Our base model for sampling preference pairs is Qwen2.5-7B-Instruct (Yang et al., 2024). We set the number of thought sampling attempts to 15 and the number of answer sampling attempts to 5. The temperature of sampling is set to 0.2, and the top-p is set to 0.9. For each question, we retrieve the top 32 document

Methods	MedQA†	MedMCQA	MMLU.	GPQA	Medbullets	MedXpertQA	Avg.
No RAG							
Qwen2.5-7B	57.82	54.70	78.39	43.81	39.29	13.27	47.20
m1-7B	64.34	59.34	78.02	36.19	48.38	16.29	50.43
HuatuoGPT-o1	68.81	64.95	79.87	45.71	50.65	14.98	54.17
Tuning-Free RAG							
MedRAG	54.20	52.35	75.81	42.86	39.29	14.20	46.45
i-MedRAG	62.84	55.18	**79.87**	51.43	44.81	16.37	**51.75**
Med-R^2	**81.06**	49.27	72.39	-	-	-	-
RAR^2 (w/o train)	64.89	**63.71**	78.67	50.48	**47.40**	**17.59**	**54.27**

Table 1: Comparisons of RAR^2 with other medical large language models and tuning-free RAG methods. † denotes the in-domain dataset.

Methods	MedQA†	MedMCQA	MMLU.
Self-BioRAG	43.60	42.15	53.92
SimRAG-8B	62.92	67.51	75.57
RAG^2-7B	75.64	63.04	78.67
SPO	**76.98**	**71.08**	85.49
RAR^2 (w/ train)	76.43	65.69	**86.32**

Table 2: Comparisons of RAR^2 with other RAG methods that involve fine-tuned components. † denotes the in-domain dataset.

chunks using the BM25 algorithm (Robertson and Zaragoza, 2009) with parameters $k_1 = 1.2$ and $b = 0.75$ as context for sampling solutions. A total of 12K preference pairs are collected after filtering and selection, where 4K are thought pairs and 8K are answer pairs. The model is then trained for 4 epochs with a global batch size of 64 and a learning rate of 1e-6, while the parameter β for the DPO loss is set to 0.2. All experiments are conducted on eight A100 GPUs, and we employ DeepSpeed ZeRO3 to optimize memory usage.

5.4 Main Results

We report the main results on six biomedical datasets shown in Table 1 and Table 2. We calculate the average performance across several datasets as **Avg.** in the final column. We present the results of RAR^2 under two settings: **w/o train** and **w/ train**. The former uses the base model for inference, while the latter uses the DPO-tuned model.

Table 1 presents the results in comparison with medically enhanced LLMs and tuning-free RAG methods. Our framework significantly improves the medical reasoning capabilities of the LLM over its backbone model, Qwen2.5-7B. It outperforms the backbone on all six datasets, achieving an average accuracy gain of 7.07%. Additionally, RAR^2 performs competitively with state-of-the-art medical language models: it surpasses m1-7B and achieves comparable average accuracy to HuatuoGPT-o1. Notably, on challenging MedXpertQA, RAR^2 achieves accuracy gains of 1.3% and 2.61% over m1-7B and HuatuoGPT-o1, respectively. Compared to other tuning-free RAG methods, RAR^2 also achieves higher average accuracy. Specifically, it outperforms i-MedRAG, which uses iterative query refinement, and Med-R^2, which employs a more complex retrieval pipeline.

Table 2 presents the results in comparison with other RAG methods that are applied to medical tasks and involve fine-tuned components. Our RAR^2 outperforms most baseline methods. Notably, on MMLU-Med, it achieves the highest accuracy among all models. SPO performs well on MedQA and MedMCQA, which may be attributed to its use of larger and open-source medical knowledge sources and additional test sample selection. Overall, the results on both in-domain and out-of-domain datasets demonstrate that our framework can help general LLMs consistently improve their medical reasoning abilities for both retrieval and answer generation.

5.5 Ablation Studies

We demonstrate the effectiveness of RAR^2 under different training settings, as detailed below: (1) **w/o training**, which adopts the thought-based RAG framework without any fine-tuning; (2) **w/o answer**, which removes answer pairs and optimizes only the generation of thought processes; and (3) **w/o thought**, which removes thought pairs and op-

Methods	MedQA	MedMCQA	MMLU.	GPQA	Medbullets	MedXpertQA	Avg.
RAR^2	76.43	65.69	86.32	56.19	57.14	20.98	60.46
- w/o training	64.89	63.71	78.67	50.48	47.40	17.59	54.27
- w/o answer	74.39	63.88	83.75	53.33	53.57	18.90	57.97
- w/o thought	74.86	64.79	84.85	54.29	55.19	19.51	58.92

Table 3: Ablation study on several medical question answering datasets.

timizes only the generation of answers. The results are shown in Table 3.

As shown in the table, RAR^2 achieves the best performance across all three selected datasets, demonstrating the effectiveness of our proposed method. Compared to the setting without training, both *w/o answer* and *w/o thought* yield significant performance improvements, indicating that optimizing the generation of either thought processes or answers is crucial for RAR^2's effectiveness. More importantly, optimizing both types of preference pairs yields complementary gains, enhancing the performance of both reasoning-augmented retrieval (i.e., thought generation) and retrieval-augmented reasoning (i.e., answer generation).

We also investigate the impact of the number of retrieved documents on the performance of RAR^2. As shown in Figure 5, RAR^2 achieves its best overall performance when retrieving 32 documents, although the optimal number varies slightly across different datasets.

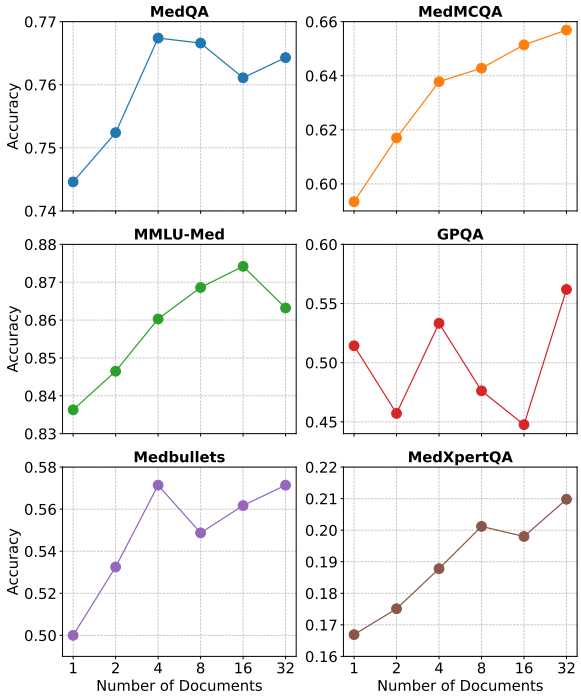

Figure 5: The performance comparison with a different number of retrieved documents.

5.6 Impact of Test-Time Scaling

We investigate the impact of test-time scaling on RAR^2, as shown in Figure 6. The total number of generated thought processes is scaled from 1 to 8. For parallel scaling, the sampling temperature is set to 1.0, and the top-p value is also set to 1.0. For iterative scaling, the number of thoughts depends on the number of iterations.

Figure 6 shows the results with increasing iterations on the MedQA and Medbullets datasets, where one is the in-domain dataset and the other is a challenging out-of-domain dataset. For Parallel Scaling, accuracy shows an upward trend as the number of thought processes increases. The improvement is particularly stable on MedQA. A similar effect is observed in Iterative Scaling, where accuracy generally increases with the number of iterations, achieving a maximum improvement of approximately 2% on both datasets.

6 Related Work

6.1 Medical Reasoning

Medical reasoning is a critical component of clinical decision-making (Ledley and Lusted, 1959). It entails integrating medical knowledge, patient information, and contextual factors to develop accurate diagnoses and effective management plans. Recent advancements in LLMs have drawn increasing attention to their use in medical reasoning (OpenAI, 2024; Singhal et al., 2023b; Wang et al., 2025). A large body of research has focused on enhancing the medical reasoning capabilities of LLMs through further pre-training with additional medical knowledge or instruction tuning on question-answering datasets, such as MEDITRON (Chen et al., 2023), Meerkat (Kim et al., 2024), and MedAdapter (Shi et al., 2024). More recently, reinforcement learning has been applied to improve

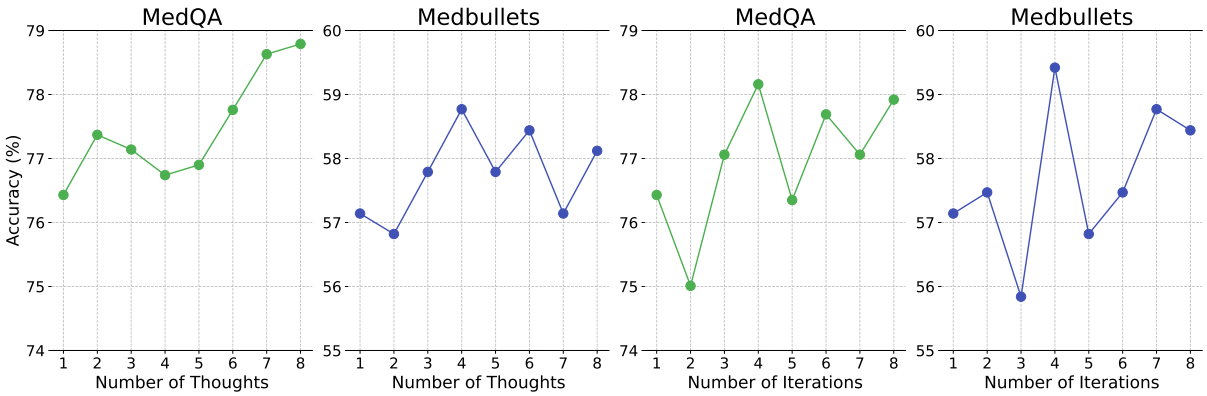

Figure 6: Accuracies with different scaling strategies on the MedQA and Medbullets datasets. The left two present Parallel Scaling, and the right two present Iterative Scaling.

test-time scaling performance in medical reasoning tasks (Chen et al., 2024b; Zhang et al., 2025; Huang et al., 2025; Yu et al., 2025).

6.2 Medical RAG

Retrieval-Augmented Generation (RAG) plays a crucial role in medical applications due to the knowledge-intensive and high-stakes nature of the domain (Xiong et al., 2024a). Recent research efforts focus on refining queries to access more relevant information. Specifically, Med-R^2 (Lu et al., 2025) and RGAR (Liang et al., 2025) enhance retrieval by iteratively modifying medical queries based on generation outcomes. SeRTS (Hu et al., 2024) optimizes query generation using Monte Carlo Tree Search with document-relevance feedback. Additionally, some work focuses on improving knowledge construction and retrieval-based knowledge usage. Wang et al. (2024) builds a RAG pipeline with query enhancement and knowledge filtering. MedGraphRAG (Wu et al., 2024) constructs a local Knowledge Graph (KG) from medical queries and efficiently retrieves relevant subgraphs. KARE (Jiang et al., 2024) constructs a multi-source KG by integrating a medical corpus and LLM-generated insights. SPO (Chen et al., 2025) investigates source planning and multi-source utilization. However, few studies have explicitly constructed and optimized the reasoning process to improve retrieval.

6.3 Reasoning-augmented Retrieval

Reasoning-intensive tasks present greater challenges to retrieval, primarily because critical knowledge requirements are often concealed within surface information and require reasoning to uncover (Su et al., 2024). For example, in medical diagnosis tasks, surface information (e.g., age, gender, and examination results) requires further analysis to establish standardized knowledge requirements (e.g., symptoms of hypertension during pregnancy). Although critical, improving retrieval for reasoning-intensive tasks has attracted relatively little research attention. RAG^2 (Sohn et al., 2024) applies RAG to medical question answering, where queries are augmented with LLM-generated rationales. JudgeRank (Niu et al., 2024) designs query and document analysis modules to enhance relevance judgment and improve document reranking. Search-R1 (Jin et al., 2025) learns to generate a series of search queries during step-by-step reasoning with real-time retrieval. RARE (Tran et al., 2025) applies MCTS to generate queries. To the best of our knowledge, we are the first to jointly optimize reasoning-augmented retrieval and retrieval-augmented reasoning, and to perform a single retrieval instead of multiple time-consuming retrieval steps.

7 Conclusion

In this work, we propose RAR^2, a novel joint learning framework designed to enhance LLM reasoning capabilities within the RAG framework, specifically targeting complex medical questions. Unlike existing RAG approaches, which rely primarily on direct query refinement, RAR^2 explicitly constructs and optimizes a thought-based reasoning process to uncover implicit knowledge requirements. We introduce a mixed preference dataset comprising thought pairs and answer pairs and leverage DPO for joint training. Extensive experiments on multiple biomedical question answering datasets demonstrate that RAR^2 outperforms state-of-the-art RAG

baselines, both with and without fine-tuning. Moreover, our analysis of test-time scaling strategies validates the scalability and robustness of RAR^2, highlighting its potential to significantly improve medical reasoning and decision-making in clinical scenarios.

Limitations

Our work has several limitations. First, the medical corpus we use does not incorporate knowledge graphs. As an important source of structured medical knowledge, knowledge graphs could help address the lack of structure in our current corpus by introducing explicit entity relationships and semantic hierarchies, thereby improving the retrieval of clinically relevant information and supporting more accurate reasoning. Second, our optimization of thought process and answer generation does not involve step-level supervision. Step-level DPO has shown promising results in various domains and represents a worthwhile direction for future exploration. Lastly, our method does not aim to improve the retrieval model itself. Enhancing the retrieval model's reasoning and understanding capabilities could further unlock the potential of reasoning-augmented retrieval.

Ethic Statement

Our proposed framework aims to improve retrieval-augmented medical reasoning and focuses on medical question answering. All datasets used for training and evaluation have been anonymized, and there is no risk of privacy exposure. However, when using LLMs for medical tasks, it is important to be aware that LLMs are prone to hallucinations, and their suggestions should not be considered definitive diagnostic conclusions. Medical advice generated by LLMs must be reviewed by qualified healthcare professionals. Therefore, we do not recommend the direct use of LLMs for medical diagnosis or decision-making at this stage. Furthermore, the scientific artifacts that we used are freely available for research, including Transformers, PyTorch and other GitHub codes. And this paper's use of these artifacts is consistent with their intended use.

References

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2024a. Benchmarking large language models on answering and explaining challenging medical questions. *CoRR*, abs/2402.18060.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024b. Huatuogpt-o1, towards medical complex reasoning with llms. *CoRR*, abs/2412.18925.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. 2025. Towards omni-rag: Comprehensive retrieval-augmented generation for large language models in medical applications. *CoRR*, abs/2501.02460.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. Serts: Self-rewarding tree search for biomedical retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1321–1335. Association for Computational Linguistics.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. m1: Unleash the potential of test-time scaling for medical reasoning with large language models. *Preprint*, arXiv:2504.00869.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinform.*, 40(Supplement_1):i119–i129.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha A. Kass-Hout, Jimeng Sun, and Jiawei Han. 2024. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *CoRR*, abs/2410.04585.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. 2024. Small language models learn enhanced reasoning skills from medical textbooks. *CoRR*, abs/2404.00376.

Robert S Ledley and Lee B Lusted. 1959. Reasoning foundations of medical diagnosis: symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, 130(3366):9–21.

Sichu Liang, Linhai Zhang, Hongyu Zhu, Wenwen Wang, Yulan He, and Deyu Zhou. 2025. Rgar: Recurrence generation-augmented retrieval for factual-aware medical question answering. *Preprint*, arXiv:2502.13361.

Keer Lu, Zheng Liang, Da Pan, Shusen Zhang, Xin Wu, Weipeng Chen, Zenan Zhou, Guosheng Dong, Bin Cui, and Wentao Zhang. 2025. Med-r^2: Crafting trustworthy LLM physicians through retrieval and reasoning of evidence-based medicine. *CoRR*, abs/2501.11885.

Tong Niu, Shafiq Joty, Ye Liu, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Judgerank: Leveraging large language models for reasoning-intensive reranking. *CoRR*, abs/2411.00142.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

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, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Haotian Sun, Hang Wu, Carl Yang, and May Dongmei Wang. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 22294–22314. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, and 12 others. 2023b. Towards expert-level medical question answering with large language models. *CoRR*, abs/2305.09617.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. 2024. Rationale-guided retrieval augmented generation for medical question answering. *CoRR*, abs/2411.00300.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Ö. Arik, Danqi Chen, and Tao Yu. 2024. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *CoRR*, abs/2407.12883.

Hieu Tran, Zonghai Yao, Zhichao Yang, Junda Wang, Yifan Zhang, Shuo Han, Feiyun Ouyang, and Hong Yu. 2025. RARE: retrieval-augmented reasoning enhancement for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18305–18330. Association for Computational Linguistics.

Bingning Wang, Haizhou Zhao, Huozhi Zhou, Liang Song, Mingyu Xu, Wei Cheng, Xiangrong Zeng, Yupeng Zhang, Yuqi Huo, Zecheng Wang, Zhengyun Zhao, Da Pan, Fei Kou, Fei Li, Fuzhong Chen, Guosheng Dong, Han Liu, Hongda Zhang, Jin He, and 23 others. 2025. Baichuan-m1: Pushing the medical capability of large language models. *Preprint*, arXiv:2502.12671.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024. Augmenting black-box llms with medical textbooks for biomedical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 1754–1770. Association for Computational Linguistics.

Junde Wu, Jiayuan Zhu, and Yunli Qi. 2024. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. *CoRR*, abs/2408.04187.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024a. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6233–6251. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024b. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *CoRR*, abs/2408.00727.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024a. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6796–6814. Association for Computational Linguistics.

Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C. Ho, Carl Yang, and Qi He. 2024b. Simrag: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. *CoRR*, abs/2410.17952.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems*, 2(1):2.

Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. 2025. Finemedlm-o1: Enhancing the medical reasoning ability of LLM from supervised fine-tuning to test-time training. *CoRR*, abs/2501.09213.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *Preprint*, arXiv:2502.19655.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *CoRR*, abs/2501.18362.

Datasets	Number
MedQA	1273
MedMCQA	4183
MMLU-Med	1089
GPQA	105
Medbullets	308
MedXpertQA	2450

Table 4: The number of samples for each evaluation dataset

A Appendix

A.1 Package Details

We used the following Python packages and their corresponding versions: transformers 4.44.2, pytorch 2.4.0, and xformers 0.0.27.post2.

A.2 Baseline Implementations

We adopt the officially released model checkpoints for baseline methods: m1-7B-23K[3], HuatuoGPT-o1-7B[4], MedRAG (i-MedRAG)[5], Self-BioRAG[6].

A.3 Details of Evaluation Datasets

We present the statistics of evaluation datasets in the Table.

[3] https://huggingface.co/UCSC-VLAA/m1-7B-23K

[4] https://huggingface.co/FreedomIntelligence/HuatuoGPT-o1-7B

[5] https://github.com/Teddy-XiongGZ/MedRAG

[6] https://github.com/dmis-lab/self-biorag