Talk Before You Retrieve: Agent-Led Discussions for Better RAG in Medical QA

Anonymous ACL submission

Abstract

Medical question answering (QA) is a reasoning-intensive task that remains challenging for large language models (LLMs) due to hallucinations and outdated domain knowledge. Retrieval-Augmented Generation (RAG) provides a promising post-training solution by leveraging external knowledge. However, existing medical RAG systems suffer from two key limitations: **(1)** a lack of modeling for human-like reasoning behaviors during information retrieval, and **(2)** reliance on suboptimal medical corpora, which often results in the retrieval of irrelevant or noisy snippets. To overcome these challenges, we propose *Discuss-RAG*, a plug-and-play module designed to enhance the medical QA RAG system through collaborative agent-based reasoning. Our method introduces a summarizer agent that orchestrates a team of medical experts to emulate multi-turn brainstorming, thereby improving the relevance of retrieved content. Additionally, a decision-making agent evaluates the retrieved snippets before their final integration. Experimental results on four benchmark medical QA datasets show that *Discuss-RAG* consistently outperforms MedRAG, especially significantly improving answer accuracy by up to 16.67% on BioASQ and 12.20% on PubMedQA. All code and prompt materials will be made publicly available.

1 Introduction

Large Language Models (LLMs) have significantly advanced a wide range of medical tasks (Singhal et al., 2023; Nori et al., 2023; Kim et al., 2024). However, their reliance on next-token prediction makes them susceptible to generating hallucinated responses (Ji et al., 2023). Additionally, once trained, LLMs operate with static parameters, meaning their internal knowledge remains fixed and cannot adapt to newly emerging research (Zhang et al., 2023). As a result, LLMs face notable limitations in dynamic, reasoning-intensive tasks (e.g., medical question answering (QA)), where both up-to-date knowledge and complex logical inference are essential.

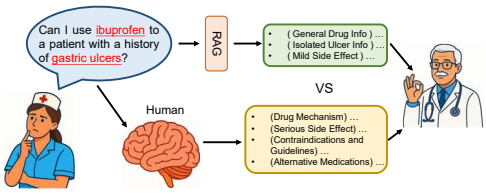

Figure 1: The illustration of difference between RAG and human for a medical query.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address the aforementioned limitations (Borgeaud et al., 2022; Guu et al., 2020; Izacard and Grave, 2020). By incorporating retrieved document snippets into the input prompt, RAG allows LLMs to generate responses that are grounded in up-to-date and trustworthy knowledge sources. Despite its success on several benchmarks, two concerns remain underexplored.

First, current medical RAG systems lack a human-like information retrieval process. They typically rely on statistical similarity metrics (e.g., cosine similarity) between the query(e.g., questions) and document embeddings to retrieve relevant content (Ke et al., 2024). This approach often fails to capture deeper contextual understanding, leading to the retrieval of superficially related but clinically irrelevant information. In contrast, as shown in Fig. 1, nurses in real-world clinical practice are more likely to recall and apply relevant clinical knowledge (e.g., drug contraindications) to guide decision-making, rather than relying solely on surface-level textual similarity. Second, existing systems often lack enough post-retrieval verification mechanisms (Barnett et al., 2024; He et al., 2024). Consequently, directly incorporating external knowledge may lead to overly cautious or

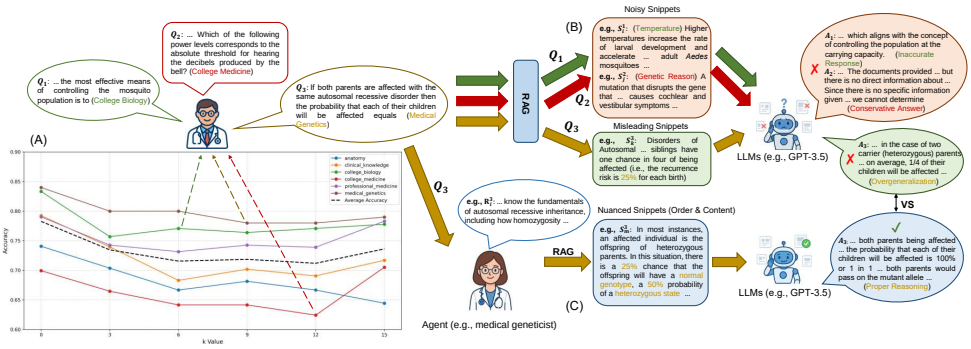

Figure 2: Preliminary experiments on the MMLU-Med benchmark. **(A)**. Accuracy trends as the number of retrieved documents k varies. Three representative questions (Q_1, Q_2 and Q_3) are selected to illustrate. **(B)**. Examples of retrieved snippets and the corresponding LLM (e.g., GPT-3.5) responses. **(C)**. Example of agent-led snippet selection and the resulting response for query Q_3. Additional details are discussed in Sec. 2.

outdated responses. In real-world settings, a judgmental role, such as a senior clinician reviewing a junior's recommendation (Fig. 1), is often necessary to assess the correlation between retrieved context and context before a final decision is made.

To address these gaps between current medical RAG systems and real-world clinical decision-making processes, we proposed *Discuss-RAG*, an agent-led framework that enhances both the information retrieval and post-verification stages of medical RAG pipelines. Specifically, a summarizer agent collaborates with a team of specialized medical agents to generate progressively refined and context-rich background insights, which are incorporated into the retrieval process alongside the original query. Additionally, a decision-maker agent evaluates the relevance and coherence of the retrieved snippets and determines whether auxiliary components should be triggered. Notably, our framework is modular and can be seamlessly integrated into any existing training-free medical RAG pipeline. Experiments on four benchmark medical QA datasets demonstrate that *Discuss-RAG* consistently improves response accuracy compared to baseline systems.

In summary, this paper makes the following key contributions: ***(1)***. We propose *Discuss-RAG*, an agent-led RAG framework that simulates a human-like reference retrieval through multi-agent discussion and iterative summarization. ***(2)***. We introduce a post-retrieval verification agent that assesses the relevance and logical coherence of retrieved snippets before they are used in answer generation. ***(3)***. We conduct comprehensive experiments comparing *Discuss-RAG* with standard RAG systems, demonstrating its effectiveness in improving both answer accuracy and snippet quality.

2 Preliminary

In our empirical experiments, we found that limitations hinder the performance of medical RAG systems in medical QA tasks. As shown in Fig. 2(A), when the corpus is fixed (i.e., textbooks (Jin et al., 2021)), varying the number of retrieved documents k results in fluctuating accuracy across six medical subjects. To better understand the influence of document selection, we selected three representative questions (Q_1, Q_2, Q_3) across different k values and subject domains. A qualitative analysis reveals factors contributing to suboptimal model behavior.

First, snippets selected based solely on dense vector similarity with the query often retrieve content that is conceptually related but task-irrelevant. These snippets introduce excessive background information that may confuse the LLM. As shown in Fig. 2(B) for Q_1, high-scoring snippets focus on environmental factors such as climate and temperature in relation to mosquitoes, rather than addressing strategies for population control. This misalignment leads to noisy inputs, resulting in either inaccurate or overly cautious responses, as seen in Q_2. Second, even factually correct snippets can mislead the model. In the case of Q_3, retrieved snippets emphasize the 25% probability associated with autosomal inheritance, prompting the LLM to overgeneralize from heterozygous to homozygous cases. These findings further suggest that directly using retrieved snippets without verification can lead to reasoning errors.

To further examine the limitations of hard similarity-based retrieval, we conducted an exploratory experiment using the same query (Q_3). As shown in Fig. 2(C), we prompted a domain-specific agent (i.e., a medical geneticist) to identify the essential knowledge required to answer

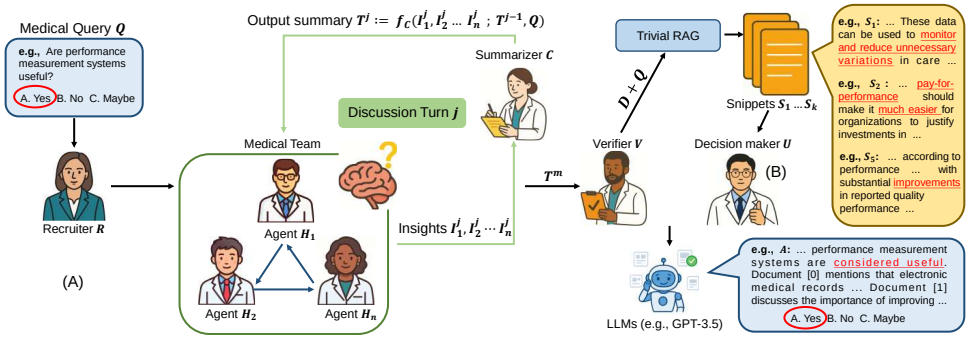

Figure 3: Illustration of the *Discuss-RAG* pipeline. **(A)**. depicts the multi-turn brainstorming and summarization process. **(B)**. presents the agent-led post-retrieval verification module. A medical query, the corresponding snippets, and the LLM's generated answer are used for illustration. Further details are provided in Sec 3.

the question (mimicking the behavior of nurses, as illustrated in Fig. 1). When we used the agent's response, in conjunction with the original query, to guide retrieval, the resulting snippets were both more topically relevant and better organized. Under this setting, the LLM successfully distinguished between carriers and affected individuals and generated a well-reasoned response.

These findings motivate two key directions for better medical RAG: ***(1)***. While a single role-based agent can benefit retrieval quality, can a multi-agent setup, engaging diverse medical expertise in an iterative, self-refining discussion, yield a more comprehensive and contextually rich background? ***(2)***. Given that structured agent involvement benefits retrieval, can a similar structure be extended to the response stage? To address these questions, we propose an agent-led RAG paradigm, the details of which are presented in the following section.

3 Methodology

Multi-turn Discussion and summarization (MDS). This module simulates a collaborative brainstorming process between a team of medical experts and a summarizer (acting as a moderator). Specifically, given a medical query Q, a recruiter agent R assembles a team of medical domain experts H_i (for i in $1, 2 \dots n$), each contributing their domain-specific perspectives I_i^j at turn j (for j in $0, 1 \dots m$). A summarizer agent C is then prompted to extract key medical knowledge, background concepts, and reasoning steps from these inputs to generate a concise summary T^j. This iterative process is formally denoted as:

$$T^j := f_C(I_1^j, I_2^j, \dots, I_n^j ; T^{j-1}, Q) \quad (1)$$

Here $f_C(\cdot)$ denotes the summarization process performed by agent C, and T^j reflects the progressively refined understanding of the query, based on the current reflection I_i^j, previous summary T^{j-1} and the original query Q (with T^0 initialized as an empty summary). After the discussion concludes, a verifier agent V is introduced to evaluate the consistency and sufficiency of the final summary T^m. The verifier produces a distilled, verification-passed summary D, which is subsequently used for snippet retrieval, together with the original query Q.

As shown in Fig. 3(A), the recruiter R recruits a team consisting of three specialized agents (e.g., a health care quality specialist, a hospital administrator, and a health economist), who collaborate with the summarizer C to share their insights for the performance measurement system. The conversation terminates either when the maximum number of discussion rounds m is reached, or when all agents decline to contribute further. Notably, all agents in this module are explicitly instructed not to answer the original query or infer a final conclusion. This design ensures that the process remains focused on context construction for retrieval, rather than direct answer generation.

Post-retrieval Verification (PRV). This module leverages structured agent reasoning to mitigate the adverse effects of suboptimal retrieval. Specifically, given the distilled summary D and the medical query Q, a specialized decision-maker agent U is introduced to evaluate the top-k document chunks S_i retrieved by the underlying retrieval algorithm. If U returns a negative judgment, an alternative retrieval strategy is triggered (e.g., a CoT-based prompt (Wei et al., 2022) is used as a fallback in our implementation). Otherwise, the accepted snippets are incorporated into the context prompt for answer generation. As shown in Fig. 3(B), the verified snippets tend to be closely aligned with the

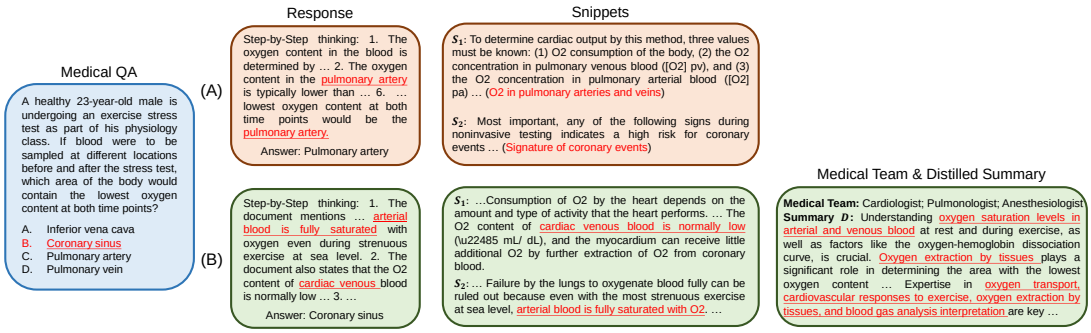

Figure 4: Example from the MedQA-US benchmark comparing MedRAG **(A)** and Discuss-RAG **(B)**. Answers and key phrases are highlighted in red.

intended focus of the query. In the shown example, the selected evidence explicitly highlights the effect (marked in red) of performance measurement systems, providing grounded support for a more accurate and contextually appropriate response.

4 Experiments

Experimental details. We selected four medical QA benchmarks: MMLU-Med (Hendrycks et al., 2020), MedQA-US (Jin et al., 2021), BioASQ (Tsatsaronis et al., 2015), and PubMedQA (Jin et al., 2019). We adopt MedRAG (Xiong et al., 2024) as our baseline RAG pipeline. To ensure a fair comparison, we employed the same medical textbooks (Jin et al., 2021) as the corpus and MedCPT (Jin et al., 2023) as the retriever. For LLM, GPT-3.5 (i.e., gpt-3.5-turbo-0125 (OpenAI, 2024)) was selected. For other necessary parameters, we chose $n = 3$, $k = 9$, and $m = 2$.

Table 1: Benchmark dataset results. Answer accuracy was used as the evaluation metric.

Dataset	MedRAG	+ Discuss-RAG	Δ
MMLU-Med	71.53%	77.23%	+5.70%
MedQA-US	62.45%	66.85%	+4.40%
BioASQ	58.61%	75.28%	+16.67%
PubMedQA	35.60%	47.80%	+12.20%

Table 2: Ablation study over MMLU-Med. We keep use the same setting as main experiment.

	MedRAG	+ MDS	MDS + PRV
Accuracy%	71.53%	73.74%	77.23%

Experimental results and analysis. *Discuss-RAG* can enrich the background information available and mitigates the impact of suboptimal retrieval. As shown in Tab. 1, integrating our method consistently improves MedRAG performance across all four benchmarks, especially achieving gains of up to 16.67% on the BioASQ dataset and 12.20% on PubMedQA. Further, as illustrated in Fig. 4, for the same query, the top-2 snippets retrieved by *Disscuss-RAG* provide more grounded and factual support for correctly answering the question. Specifically, snippets S_1 explicitly mention the low oxygen (O_2) content in cardiac venous blood, while snippets S_2 support the reasoning process from a contrasting perspective. Additionally, the final distilled summary D generated by the medical team highlights the essential knowledge required to focus the retrieval process, leading to more reliable and contextually appropriate evidence selection. Table 2 presents the ablation study on the MMLU-Med benchmark. Incorporating the multi-turn discussion and summarization modules increases accuracy from 71.53% to 73.74%. Further adding the post-retrieval verification module yields an additional 3.49% performance gain. These results demonstrate the complementary contributions of the two modules in improving accuracy. Finally, deploying *Discuss-RAG* on MMLU-Med incurs a cost of approximately $12, which translates to an additional $0.01 per question, which is an acceptable trade-off given the substantial accuracy improvements.

5 Conclusion

In this work, we propose *Discuss-RAG*, an agent-led framework designed to enhance the response accuracy of LLMs in medical QA. Specifically, we introduce a multi-turn discussion and summarization module to facilitate context-rich and self-refined document retrieval, and a post-retrieval verification agent to make the final judgment on the retrieved content. Experiments conducted on four medical QA benchmark datasets demonstrate that *Discuss-RAG* consistently improves both response accuracy and snippet quality.

6 Limitation

We acknowledge that *Discuss-RAG* is hindered by two primary limitations. **(1)**. Limited interaction among team members. The specialized medical agents H_i do not communicate directly with one another, but interact through the summary from the previous round. Direct peer-to-peer interaction may facilitate deeper and more dynamic reasoning. **(2)**. Increased computational overhead. Our framework involves prompting multiple LLM-based agents, each requiring careful instruction design to perform their respective roles effectively. This introduces additional computational and engineering costs.

References

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 194–199.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, and 1 others. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. *arXiv preprint arXiv:2410.05801*.

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*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

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, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, and Daniel Shu Wei Ting. 2024. Development and testing of retrieval augmented generation in large language models–a case study report. *arXiv preprint arXiv:2402.01733*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, Hae Park, and 1 others. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

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. 2024. Gpt-3.5 turbo. https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2025-04-27.

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented

generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*.