TALK BEFORE YOU RETRIEVE: AGENT-LED DISCUSSIONS FOR BETTER RAG IN MEDICAL QA

Xuanzhao Dong[1*], Wenhui Zhu[1*], Hao Wang[2*], Xiwen Chen[2,5*], Peijie Qiu[3], Rui Yin[1], Yi Su[4], Yalin Wang[1]
[1]Arizona State University, [2]Clemson University, [3]Washington University in St.Louis,
[4]Banner Alzheimer's Institute, [5]Morgan Stanley

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(QA) is a reasoning-intensive task that remains challenging for large language models (LLMs) due to hallucinations and outdated domain knowledge.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. The code is available at `https://github.com/LLM-VLM-GSL/Discuss-RAG`

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.

Retrieval-Augmented Generation (RAG) has emerged as a promising approach to address the aforementioned limitations Borgeaud et al. (2022); Guu et al. (2020); Izacard & 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

[*]These authors contributed equally to this paper.

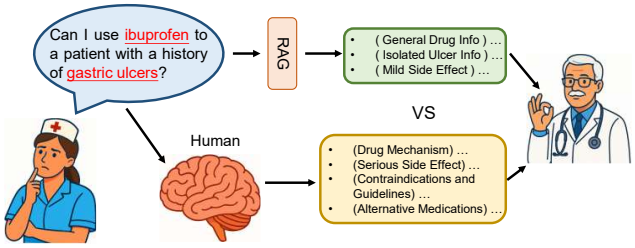

Figure 1: Illustration of the difference between standard RAG and human cognitive processes for a medical query. Specifically, conventional medical RAG systems (e.g., MedCPT Jin et al. (2023)) rely directly on the embedding similarity between the raw question and the text chunks. In contrast, a human (e.g., a nurse) first determines what specific knowledge is necessary to answer the question and uses these derived requirements to guide the retrieval process, leading to content like drug mechanism, serious side effect and alternative medications, etc. Additionally, a post-verification step (e.g., by a senior doctor) is typically required in practice.

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.

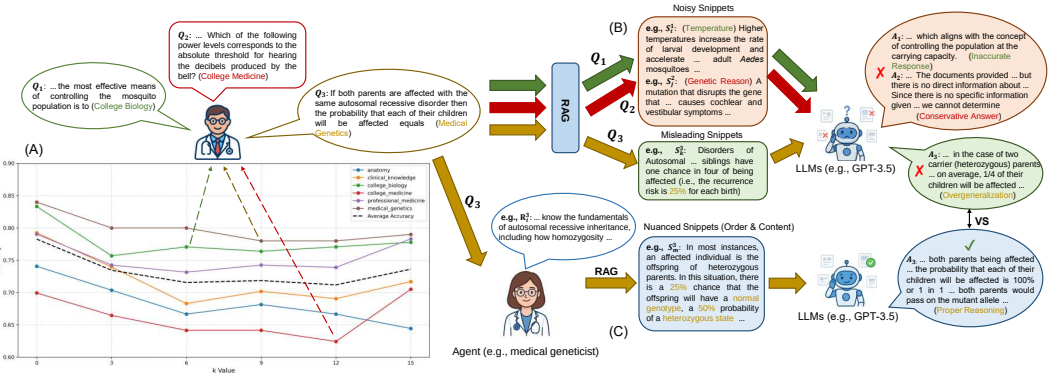

Figure 2: Preliminary experiments on the MMLU-Med benchmark. **(A)**. Accuracy trends as the number of retrieved documents k varies. Three representative questions (i.e., Q_1, Q_2 and Q_3) are selected for illustration. **(B)**. Examples of retrieved snippets (i.e., S_i^1, S_j^2 and S_k^3) and the corresponding LLM (e.g., GPT-3.5) responses (i.e., A_1, A_2 and A_3). Here, the responses reflect the model's reasoning process. The results demonstrate that naïve medical RAG systems may retrieve noisy snippets, causing LLMs to generate inaccurate responses (i.e., Q_1), overly conservative answers (i.e., Q_2), or overgeneralizations (i.e., Q_3). **(C)**. Example of agent-led snippet selection and the resulting response for query Q_3. We observed that simply prepending the agent's suggestions to the original question alters the content and ranking of retrieved snippets, leading to the correct answer. This finding inspired our pipeline design. Additional details are discussed in Sec. 3.

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 Related Work

2.1 AI Agents

Large language models (LLMs) and, more recently, vision–language models (VLMs) have enabled natural-language interfaces for a broad range of tasks Wang et al. (2023a); Liu et al. (2024a). In particular, prompt engineering has emerged as a central method for adapting LLMs to specialized domains without additional training Wang et al. (2023a); Schmidgall et al. (2025). Early work improved downstream performance via template-based prompts that steer model behavior for question answering, captioning, and code generation Zhu et al. (2025); Dong et al. (2025); Chen et al. (2025). However, while effective in constrained settings, these approaches often suffer from brittleness, ambiguity, and limited generalization; minor wording changes can significantly alter outcomes, and scaling to heterogeneous tasks typically requires non-trivial prompt redesign Ngweta et al. (2025); Desmond & Brachman (2024); Chatterjee et al. (2024).

Consequently, as model capabilities and modalities have expanded, the field has shifted from static prompts to agentic workflows, where models function as autonomous agents capable of invoking tools and managing memory. For instance, systems such as AutoGPT Gravitas (2023) and Manus Manus (2023) introduced pipelines that dynamically plan and iteratively execute toward a goal superdesigndev (2023). Within this domain, two dominant design frameworks have shaped recent developments. Specifically, the ReAct Yao et al. (2022) agent design, popularized by libraries such as LangChain Chase (2022), interleaves reasoning with tool calls, allowing the agent to observe, reason, and act in an iterative loop. Alternatively, the Plan-and-Execute paradigm marked another agentic paradigm Wang et al. (2023b); it decouples a global planner (e.g., task decomposition) from an executor (e.g., stepwise completion). This separation improves structure and reliability over long horizons, making it particularly appealing for complex goals Liu et al. (2024b). Despite the success of these paradigms in general industry applications, the robust integration of agentic workflows into RAG pipelines, particularly within the biomedical domain, remains largely underexplored.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) Lewis et al. (2020) enables LLMs to access external knowledge bases without without the need for model retraining. Standard RAG systems typically utilize a retriever to identify the most relevant document snippets from a trusted corpus, which are then integrated into the original query as context to support the generation process. However, trivial RAG systems supper from two primary limitations. First, reliance on simple similarity search often retrieves noisy or irrelevant snippets. This noise can confuse the LLM, leading to hallucinations or factually incorrect responses Shi et al. (2023). This risk will be exacerbated when the external database contains outdated information Vu et al. (2024). Second, standard RAG operates as a single-step process, lacking the capability for multi-step reasoning Trivedi et al. (2023). Consequently, if the initial retrieval fails, the entire response is compromised Asai et al. (2024). These limitations pose significant risks in high-stakes domains such as biomedical field, where protocols evolve frequently. Furthermore, general-purpose retrievers often fail to accurately resolve specific medical acronyms, leading to semantic mismatches. Gao et al. (2023) Therefore, additional efforts are needed to transfer this success to the biomedical domain.

Recent developments in agentic RAG offer a promising alternative to address these limitations. By integrating specialized agents into the input processing, retrieval, and generation stages, agentic approaches transform the RAG pipeline from a static assembly line into a dynamic, adaptive workflow. For example, MA-RAG Nguyen et al. (2025) proposes leveraging a suite of specialized agents, such as Planners and Step Definers, to support information-dense tasks. Similarly, mRAG Salemi

et al. (2025) incorporates a reward-guided sampling process to optimize the collaborative dynamics among agents within the system. Despite these successes, current multi-agent RAG systems primarily focus on orchestrating the high-level pipeline rather than focusing on the retrieval step itself. *Discuss-RAG* represents a noteworthy exploration into using multi-agents to simulate brainstorming, thereby facilitating improved retrieval in RAG systems.

3 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 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.

4 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 $0, 1 \dots m$). As shown in Fig. 4, each expert is required to identify the essential medical knowledge and clinical background necessary to answer the question accurately. After each member provide their own reflection, 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:

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

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

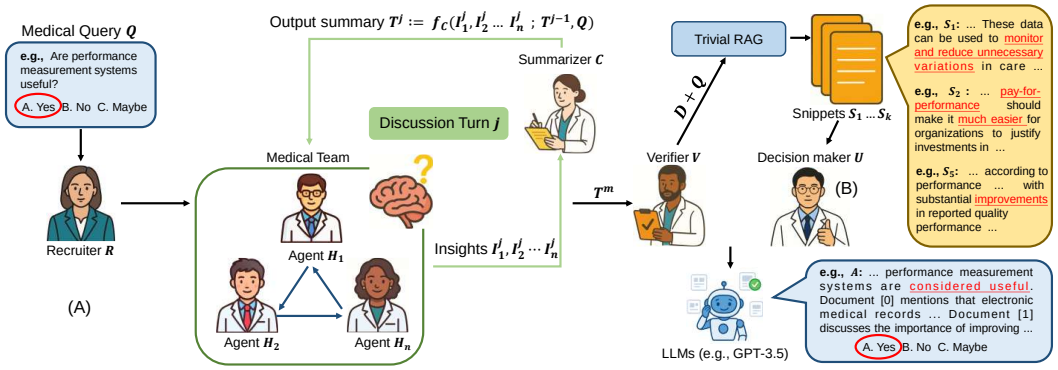

Figure 3: Illustration of the *Discuss-RAG* pipeline. Part **A** illustrate the Multi-turn Discussion and summarization (MDS) framework, including the recruiter R, medical team (e.g., experts H_1, H_2 and H_3), summarizer C and verifier V. This framework simulates the multi-turn brainstorming activity to provide necessary knowledge for medical query. Part **B** presents the agent-led post-retrieval verification module to simulate senior doctor's activity. The medical query, the corresponding snippets, and the LLM's generated answer are used for illustration. Further details are provided in Sec 4.

Prompt Template for Medical Expert H_i in Medical Team for MDS module

System Prompt: You are a **<role>** who **<expertise>**. Your job is to collaborate with other medical experts in a team. Given a medical question, your task is **not to answer it, but to share what kind of knowledge is necessary to answer the question correctly from the perspective of your own profession and expertise**. Focus on identifying the domains, facts, concepts, or reasoning approaches that would be essential for solving the question.

Input: Here is the medical question: **<question>**. As a domain-specific expert, remember your role and mission: **do not attempt to answer the question or infer a final conclusion**. Instead, **identify the essential medical knowledge, clinical background, or reasoning steps** that would be necessary to answer this question accurately. Please reflect from your area of expertise and think step-by-step. Please keep your response concise—ideally within 7 sentences. Focus on what is most essential. Begin your reflection below:

Figure 4: Illustration of the prompt template for the medical expert agent H_i. The **<role>** and **<expertise>** fields are dynamically assigned by the recruiter agent R based on the specific query (i.e., **<question>**). Key information is highlighted in bold. Additionally, all agents are strictly instructed to refrain from inferring a final answer; instead, they must provide essential medical knowledge that serves as guidance for retrieval. See Sec. 4 for more details.

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 (e.g., Fig. 4). 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. As shown in Fig. 5, if U returns a negative judgment (i.e., return No in answer part), an alternative retrieval strategy is triggered (e.g., a

Prompt Template for Decision-maker Agent U in for PRV module

System Prompt: System Prompt: You are an experienced medical assistant. You will be provided with a medical query and a paragraph of retrieved information. Your task is to **determine whether the paragraph contains sufficient information to reasonably support an answer to the medical query**, even if the answer is not explicitly stated, as long as the conclusion would be clear to a trained medical professional. Carefully review the entire paragraph and reason step by step before arriving at your conclusion. Strictly output your response in the following JSON format: {"step_by_step_thinking": **"<your explanation>"**, "answer": **"yes"** or **"no"**}.

Input: Here is the medical query:**<question>**. Here is a paragraph of retrieved information: **<information>**. Please reason step by step and produce your output in the following JSON format: {"step_by_step_thinking": **"<your explanation>"**, "answer": **"yes"** or **"no"**}.

Figure 5: Illustration of the prompt template for the specialized decision-maker agent U in PRV module. The **<question>** and **<information>** represent the medical query and distilled summary D from agent V shown in MDS module, respectively. Alternative RAG (e.g., CoT) method will be triggered if the response is No. Key information is highlighted in bold. See Sec. 4 for more details.

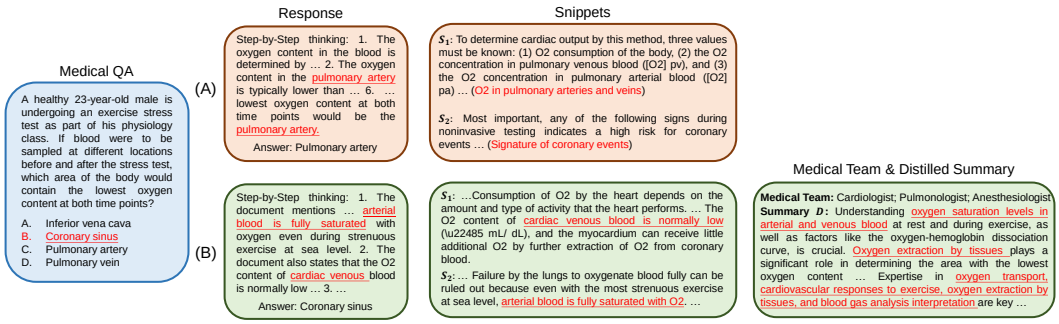

Figure 6: Example from the MedQA-US benchmark comparing MedRAG **(A)** and *Discuss-RAG* **(B)**. The **Response** column shows the output of the same LLM to the same question, while the **Snippets** column displays the top-2 retrieved snippets. Notably, incorporating *Discuss-RAG* alters both the content and the order of the retrieved snippets. **Medical Team & Distilled Summary** illustrates the specialized agent-led teams recruited by agent R, where summary D denotes the output from the summarizer agent C. Key factors are highlighted in red for better visualization. Additional details are provided in Sec. 5.

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 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.

5 Experiments

Experimental details. To demonstrate the generalizability of the *Discuss-RAG* pipeline, we selected four diverse 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). As these benchmarks consist of closed-form questions, we employ accuracy as our evaluation metric. We adopt MedRAG Xiong et al. (2024) as the baseline pipeline. To ensure a fair comparison, we utilize the same medical textbook corpus Jin et al. (2021) and MedCPT retriever Jin et al. (2023) across all ex-

periments. We select GPT-3.5 (`gpt-3.5-turbo-0125` OpenAI (2024)) as the backbone LLM. For other necessary parameters, we set $n = 3$, $k = 9$, and $m = 2$.

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 1: Results on benchmark datasets using accuracy as the evaluation metric. The *Discuss-RAG* framework improves answer accuracy across all four benchmarks. Notably, it achieves gains of 12.2% on PubMedQA and 16.67% on BioASQ. See Sec. 5 for more information.

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. 6, 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. For example, it emphasizes the importance of oxygen saturation levels, oxygen extraction by tissues, which are reflected by the *Discuss-RAG*'s retrieved snippets.

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%

Ablation Study. The *Discuss-RAG* framework comprises two primary components. Specifically, the MDS module simulates collaborative brainstorming to enable agent-led information retrieval. The PRV module serves as a post-verification mechanism, determining whether to trigger a backup plan based on the retrieved snippets. To investigate the orthogonal contribution of these two components, we conduct an ablation study on the MMLU-Med benchmark. Here, we still use accuracy as our metrics. As shown in Tab. 2, 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.

6 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.

7 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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. 2024.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 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*, pp. 194–199, 2024.

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

Harrison Chase. Langchain. `https://github.com/langchain-ai/langchain`, 2022.

Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. Posix: A prompt sensitivity index for large language models. *arXiv preprint arXiv:2410.02185*, 2024.

Yamei Chen, Haoquan Zhang, Yangyi Huang, Zeju Qiu, Kaipeng Zhang, Yandong Wen, and Weiyang Liu. Symbolic graphics programming with large language models. *arXiv preprint arXiv:2509.05208*, 2025.

Michael Desmond and Michelle Brachman. Exploring prompt engineering practices in the enterprise. *arXiv preprint arXiv:2403.08950*, 2024.

Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. Llada-medv: Exploring large language diffusion models for biomedical image understanding. *arXiv preprint arXiv:2508.01617*, 2025.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Significant Gravitas. Autogpt. `https://github.com/Significant-Gravitas/AutoGPT`, 2023.

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

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

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

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

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

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

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

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. Development and testing of retrieval augmented generation in large language models–a case study report. *arXiv preprint arXiv:2402.01733*, 2024.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024b.

Manus. Manus. `https://manus.im/`, 2023. Accessed: 2025-09-20.

Thang Nguyen, Peter Chin, and Yu-Wing Tai. Ma-rag: Multi-agent retrieval-augmented generation via collaborative chain-of-thought reasoning. *arXiv preprint arXiv:2505.20096*, 2025.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. Towards llms robustness to changes in prompt format styles. *arXiv preprint arXiv:2504.06969*, 2025.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

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

Alireza Salemi, Mukta Maddipatla, and Hamed Zamani. Ciir@ liverag 2025: Optimizing multi-agent retrieval augmented generation through self-training. *arXiv preprint arXiv:2506.10844*, 2025.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

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

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

superdesigndev. superdesign. https://github.com/superdesigndev/superdesign, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish 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)*, pp. 10014–10037, 2023.

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

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*, 2023b.

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

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 6233–6251, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

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

Wenhui Zhu, Xin Li, Xiwen Chen, Peijie Qiu, Vamsi Krishna Vasa, Xuanzhao Dong, Yanxi Chen, Natasha Lepore, Oana Dumitrascu, Yi Su, et al. Retinalgpt: A retinal clinical preference conversational assistant powered by large vision-language models. *arXiv preprint arXiv:2503.03987*, 2025.