Comparative Evaluation of a Citable RAG System and GPT-4o in Medical Education: Evidence Reliability and Usability Study

Suguru Murakami, Kanon Oumi, Kenji Hirata, Katsuhiko Ogasawara

Submitted to: JMIR Medical Education
on: August 29, 2025

Table of Contents

Original Manuscript **5**
Supplementary Files **22**
Multimedia Appendixes **23**
Multimedia Appendix 1 ..... **23**

Comparative Evaluation of a Citable RAG System and GPT-4o in Medical Education: Evidence Reliability and Usability Study

Suguru Murakami[1]; Kanon Oumi[2]; Kenji Hirata[3] MD, PhD; Katsuhiko Ogasawara[2] MBA, PhD

[1] Hokkaido University school of Medicine Sapporo JP
[2] Hokkaido University school of Health Science Sapporo JP
[3] Deaprtment of Diagnostic Imaging, Faculty of Medicine Hokkaido University school of Medicine Sapporo JP

Corresponding Author:
Katsuhiko Ogasawara MBA, PhD

Hokkaido University school of Health Science
N12-W5-Kita-ku
Sapporo
JP

Abstract

Background: Modern medical education requires efficient tools for knowledge acquisition, and large language models (LLMs) appear promising; however, they face challenges such as factual inaccuracies (“hallucinations”) and a lack of evidence transparency, particularly in critical fields such as medicine. Retrieval augmented generation (RAG) systems address these limitations by grounding LLM responses in external knowledge sources, which offers a potential solution for developing reliable educational tools.

Objective: This study aimed to examine the applicability of a citable RAG system in medical education, specifically to address the challenges of LLMs regarding hallucinations and unclear evidence, and to investigate its potential applications.

Methods: We designed and implemented a RAG system in Python, using the LangGraph for system construction. The system generated responses by converting input questions into search queries, retrieving relevant information from academic databases and web sources through dedicated search agents, and then integrating this information to generate comprehensive, citable answers. GPT-4o was used for both query generation and report generation within the RAG system. We used a data set of 103 medical questions for evaluation. We compared the RAG system's answers with those generated by stand-alone GPT-4o. Evaluation was conducted quantitatively by LLMs (GPT-4 and Gemini Flash 2.0) using the CLEAR reliability metric (Completeness, Lack of false information, Evidence, Appropriateness, Relevance). We also performed a subjective evaluation with 40 medical and nursing students, who used a Likert scale based on the CLEAR criteria for a subset of five randomly selected questions. Additionally, we assessed the consistency between generated text and its references using the SourceCheckup tool, which evaluated Citation URL Validity, Statement-level Support, and Response-level Support.

Results: In LLM evaluations, the RAG system consistently outperformed standalone GPT-4o in "Evidence." However, student evaluations showed a different trend, with GPT-4o receiving significantly higher scores in "Completeness," "Appropriateness," and "Relevance," while RAG excelled in "Evidence." For reference evaluation, all cited URLs were valid. However, Statement-level Support was 0.516 (95% CI: 0.460, 0.571), and Response-level Support was 0.228 (95% CI: 0.158, 0.317), indicating that not all claims or full responses were directly supported by the cited sources.

Conclusions: The RAG system effectively addressed the challenges of hallucination and unclear evidence in LLMs for medical education, consistently improving the evidence base of responses. However, discrepancies between LLM and human evaluations highlighted the need for further improvements in the overall structure and natural language flow of responses for practical educational implementation. Future work should focus on enhancing the consistency between referenced information and generated output, and on improving the overall coherence and clarity of the response structure, with Self-RAG emerging as a promising approach for self-verification and improved learning support.

(JMIR Preprints 29/08/2025:83247)
DOI: https://doi.org/10.2196/preprints.83247

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Comparative Evaluation of a Citable RAG System and GPT-4o in Medical Education: Evidence Reliability and Usability Study

Abstract

Background

Modern medical education requires efficient tools for knowledge acquisition, and large language models (LLMs) appear promising; however, they face challenges such as factual inaccuracies (“hallucinations”) and a lack of evidence transparency, particularly in critical fields such as medicine. Retrieval augmented generation (RAG) systems address these limitations by grounding LLM responses in external knowledge sources, which offers a potential solution for developing reliable educational tools.

Objective

This study aimed to examine the applicability of a citable RAG system in medical education, specifically to address the challenges of LLMs regarding hallucinations and unclear evidence, and to investigate its potential applications.

Methods

We designed and implemented a RAG system in Python, using the LangGraph for system construction. The system generated responses by converting input questions into search queries, retrieving relevant information from academic databases and web sources through dedicated search agents, and then integrating this information to generate comprehensive, citable answers. GPT-4o was used for both query generation and report generation within the RAG system. We used a data set of 103 medical questions for evaluation. We compared the RAG system's answers with those generated by stand-alone GPT-4o. Evaluation was conducted quantitatively by LLMs (GPT-4 and Gemini Flash 2.0) using the CLEAR reliability metric (Completeness, Lack of false information, Evidence, Appropriateness, Relevance). We also performed a subjective evaluation with 40 medical and nursing students, who used a Likert scale based on the CLEAR criteria for a subset of five randomly selected questions. Additionally, we assessed the consistency between generated text and its references using the SourceCheckup tool, which evaluated Citation URL Validity, Statement-level Support, and Response-level Support.

Results

In LLM evaluations, the RAG system consistently outperformed standalone GPT-4o in "Evidence." However, student evaluations showed a different trend, with GPT-4o receiving significantly higher scores in "Completeness," "Appropriateness," and "Relevance," while RAG excelled in "Evidence." For reference evaluation, all cited URLs were valid. However, Statement-level Support was 0.516 (95% CI: 0.460, 0.571), and Response-level Support was 0.228 (95% CI: 0.158, 0.317), indicating that not all claims or full responses were directly supported by the cited sources.

Conclusion

The RAG system effectively addressed the challenges of hallucination and unclear evidence in LLMs for medical education, consistently improving the evidence base of responses. However, discrepancies between LLM and human evaluations highlighted the need for further improvements in the overall structure and natural language flow of responses for practical educational implementation. Future work should focus on enhancing the consistency between referenced information and generated output, and on improving the overall coherence and clarity of the response structure, with Self-RAG emerging as a promising approach for self-verification and improved learning support.

Keywords

Artificial Intelligence; Large Language Models; Medical Education; Retrieval Augmented Generation Question Answering System

Introduction

Modern medicine has advanced rapidly, and medical and nursing students, who will become doctors and nurses in the future, need to acquire a vast and constantly updated body of knowledge. Therefore, in addition to traditional textbooks and lectures, there has been a demand for the development of educational tools that enable students to actively explore and learn information [1]. In recent years, Large Language Models (LLMs), such as ChatGPT, have received significant attention for their potential applications in education because of their advanced natural language understanding and generation capabilities [2-7]. LLMs can generate quick, human-like responses to student questions and provide summaries on specific topics. However, in specialized fields such as medical education, where the accuracy of information directly affects human lives, the challenges inherent in conventional LLMs raise serious concerns. Traditional LLMs have fundamental issues, including the risk of generating factually incorrect answers ("hallucinations") and a lack of clarity regarding the basis of their responses [8, 9]. To address these challenges related to information reliability and clarity of evidence when applying such LLMs to medical education, Retrieval-Augmented Generation (RAG) has recently emerged as a promising technology [10]. RAG is a framework in which an LLM generates a response by first retrieving relevant information from external knowledge sources and then referencing that retrieved information. Therefore, we aim to examine the applicability of a RAG system in medical education.

Materials and Methods

Question Data Set

To evaluate the Retrieval-Augmented Generation (RAG) system, we created 120 medical questions using GPT-4 (model as of 2025-04-07). We removed duplicate questions and those with extremely similar content, which resulted in a data set of 103 questions. Table 1 shows some of the questions used.

Table 1. Question data set

No.	Question
1	What is ACP (Advance Care Planning)?
2	What are Bunina bodies?
3	What are the differences between COPD and asthma?
□	□
□	□
□	□
103	What are the common fall risks and countermeasures in older adults?

System

In this study, we designed and implemented a RAG system for question answering in the medical field using Python (version 3.13; Python Software Foundation)
. The main libraries included the ArXiv API for academic paper searches, the PubMed API for medical literature, and the Tavily API for web information retrieval. We used the LangChain and LangGraph frameworks to build the RAG system. As shown in Figure 2, the system collected and integrated information from multiple academic databases to generate comprehensive, citable answers. The system consisted of five main components. First, the Query Generation Agent converted the input question into search queries, generating one to five English words suitable for searching. Next, three search agents (ArXiv, PubMed, and Web) accessed their respective data sources to collect relevant information. Finally, the Report Generation Agent integrated this information and generated an answer report. Each search agent stored the retrieved information along with the source URL and managed it as a unified dictionary, which made it possible to clearly indicate the source of the information. This approach allowed for clear presentation of reference links that showed the basis for the descriptions in the final report. We used GPT-4o (model as of April 9, 2025) for query generation and report generation in the RAG system. To evaluate the pure effect of introducing RAG with citation capability, we unified conditions such as output structure (excluding citation features), question data set, and evaluation criteria between the GPT-4o and RAG systems.

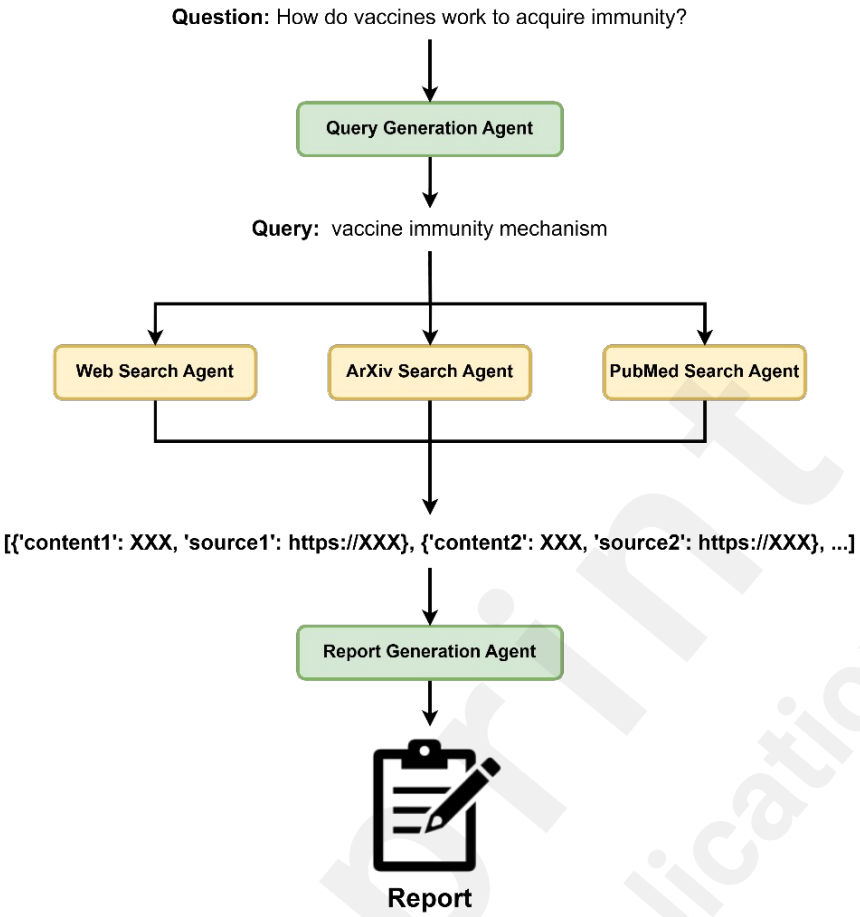


Fig 1. RAG System Overview

Assessment of RAG System Answers

Evaluation by LLMs

In this study, we used a RAG system to generate answers for 103 questions and compared these answers with those generated by GPT-4o for the same questions. We used CLEAR (C: Completeness, L: Lack of false information, E: Evidence, A: Appropriateness, R: Relevance) [11], a reliability evaluation metric for LLMs, to compare the responses. GPT-4 and Gemini Flash 2.0 evaluated the generated answers. We applied a paired t-test to compare the evaluation scores of the RAG system and GPT-4o for each CLEAR criterion, and we evaluated statistical significance at a significance level of 0.05.

- Completeness: Completeness refers to whether the generated content includes all the required information, ensuring that nothing essential is missing and that the response is neither excessive nor insufficient.

- Lack of false information: Lack of false information assesses the accuracy of the content, verifying that it does not contain any incorrect or misleading statements.

- Evidence: Evidence evaluates whether the content is supported by reliable scientific sources, indicating that the information is grounded in established research.

- Appropriateness: Appropriateness examines whether the content is clear, concise, unambiguous, and well-organized, which facilitates easy understanding for the reader.

- Relevance: Relevance determines whether the content directly addresses the question, avoiding the inclusion of unrelated or unnecessary information.

Evaluation by students

Based on the CLEAR criteria, we conducted a subjective evaluation with 40 students from the Faculty of Medicine and School of Nursing at Hokkaido University using a Likert scale. First, we randomly selected five questions from the 103-question data set. Each student received answers from both the RAG system and GPT-4o for these five questions. We then applied a paired t-test to compare the evaluation scores of the RAG system and GPT-4o for each CLEAR item, and we evaluated statistical significance at a significance level of 0.05.

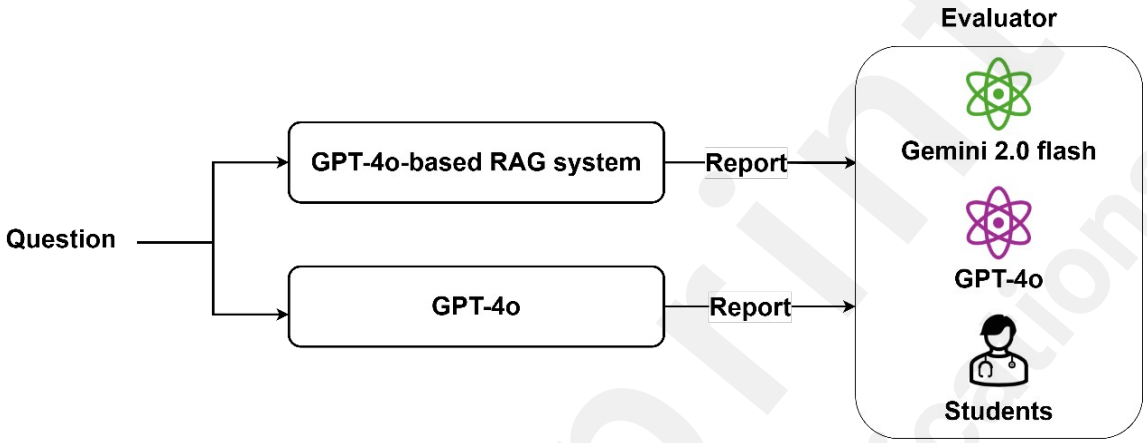


Fig 2. Evaluation Overview

Evaluation of references

In this study, we evaluated the extent to which LLMs supported their conclusions for medical-related questions with reliable references using SourceCheckup [12]. SourceCheckup is an automated, Python-based pipeline that divided LLM-generated answers into statement units and analyzed whether the claims matched the content of the URL sources. By inputting the LLM's answer and the reference URLs in a list format, the pipeline output the Statement-level Support and Response-level Support as described below. We downloaded SourceCheckup from its GitHub repository and used it on a local PC.

The evaluation was conducted based on the following three main criteria:

- Citation URL Validity: This measures the percentage of URLs generated by the model that return a valid webpage:

$$Source\ URL\ Validity = \frac{Number\ of\ URLs\ with\ status\ code\ 200}{Total\ Number\ of\ URLs}$$

- Statement-level Support: This measures the percentage of individual medical-related statements generated by the model that are supported by at least one source:

$$Statement\ Level\ Support = \frac{Number\ of\ Statements\ Supported\ by \geq 1\ Source}{Total\ Number\ of\ Statements}$$

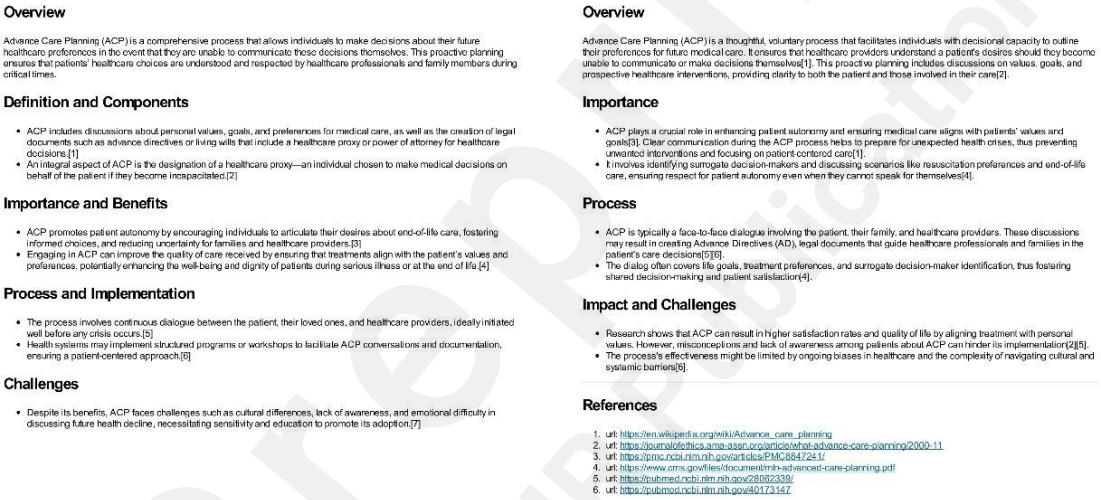
- Response-level Support: This measures the percentage of responses in which all statements are supported:

$$Response\ Level\ Support = \frac{Responses\ w/\ All\ Statements\ Supported}{Total\ Number\ of\ Responses}$$

Results

Responses to Questions from GPT-4o and RAG System

GPT-4o and the RAG system provided answers to all 103 questions presented in Table 1. Because Japanese students would perform the evaluation, the prompts instructed the systems to generate responses in Japanese, excluding references, with a body length of approximately 1,500 characters. Figure 3 shows example outputs from both the GPT-4o and RAG systems for the question related to Figure 3: "What is ACP (Advance Care Planning)?" The output language here has been changed to English for illustrative purposes.



Overview

Advance Care Planning (ACP) is a comprehensive process that allows individuals to make decisions about their future healthcare preferences in the event that they are unable to communicate those decisions themselves. This proactive planning ensures that patients' healthcare choices are understood and respected by healthcare professionals and family members during critical times.

Definition and Components

- ACP includes discussions about personal values, goals, and preferences for medical care, as well as the creation of legal documents such as advance directives or living wills that include a healthcare proxy or power of attorney for healthcare decisions.[1]
- An integral aspect of ACP is the designation of a healthcare proxy—an individual chosen to make medical decisions on behalf of the patient if they become incapacitated.[2]

Importance and Benefits

- ACP promotes patient autonomy by encouraging individuals to articulate their desires about end-of-life care, fostering informed choices, and reducing uncertainty for families and healthcare providers.[3]
- Engaging in ACP can improve the quality of care received by ensuring that treatments align with the patient's values and preferences, potentially enhancing the well-being and dignity of patients during serious illness or at the end of life.[4]

Process and Implementation

- The process involves continuous dialogue between the patient, their loved ones, and healthcare providers, ideally initiated well before any crisis occurs.[5]
- Health systems may implement structured programs or workshops to facilitate ACP conversations and documentation, ensuring a patient-centered approach.[6]

Challenges

- Despite its benefits, ACP faces challenges such as cultural differences, lack of awareness, and emotional difficulty in discussing future health decline, necessitating sensitivity and education to promote its adoption.[7]

Overview

Advance Care Planning (ACP) is a thoughtful, voluntary process that facilitates individuals with decisional capacity to outline their preferences for future medical care. It ensures that healthcare providers understand a patient's desires should they become unable to communicate or make decisions themselves[1]. This proactive planning includes discussions on values, goals, and prospective healthcare interventions, providing clarity to both the patient and those involved in their care[2].

Importance

- ACP plays a crucial role in enhancing patient autonomy and ensuring medical care aligns with patients' values and goals[3]. Clear communication during the ACP process helps to prepare for unexpected health crises, thus preventing unwanted interventions and focusing on patient-centered care[1].
- It involves identifying surrogate decision-makers and discussing scenarios like resuscitation preferences and end-of-life care, ensuring respect for patient autonomy even when they cannot speak for themselves[4].

Process

- ACP is typically a face-to-face dialogue involving the patient, their family, and healthcare providers. These discussions may result in creating Advance Directives (AD), legal documents that guide healthcare professionals and families in the patient's care decisions[5][6].
- The dialog often covers life goals, treatment preferences, and surrogate decision-maker identification, thus fostering shared decision-making and patient satisfaction[4].

Impact and Challenges

- Research shows that ACP can result in higher satisfaction rates and quality of life by aligning treatment with personal values. However, misconceptions and lack of awareness among patients about ACP can hinder its implementation[2][5].
- The process's effectiveness might be limited by ongoing biases in healthcare and the complexity of navigating cultural and systemic barriers[6].

References

1. url: https://en.wikipedia.org/wiki/Advance_care_planning
2. url: https://journalofethics.ama-assn.org/article/what-advance-care-planning/2000-11
3. url: https://pmc.ncbi.nlm.nih.gov/articles/PMC8847241/
4. url: https://www.cms.gov/files/document/mln-advanced-care-planning.pdf
5. url: https://pubmed.ncbi.nlm.nih.gov/29092339/
6. url: https://pubmed.ncbi.nlm.nih.gov/40173147

Fig 3. Example Outputs of GPT-4o and RAG System

Assessment of RAG System Answers

Evaluation by LLMs

While the RAG system consistently outperformed GPT-4o in terms of Evidence, we did not observe a consistent trend for the other items. Additionally, GPT-4o tended to assign higher scores across all items compared to Gemini Flash 2.0. This tendency is likely attributable to the answer-generating AI being GPT-based. When the evaluator model and the generation model belong to the same family of models, there may be a tendency for them to perceive higher consistency or persuasiveness in GPT-series outputs, possibly because of similarities in their internal evaluation criteria and output patterns.

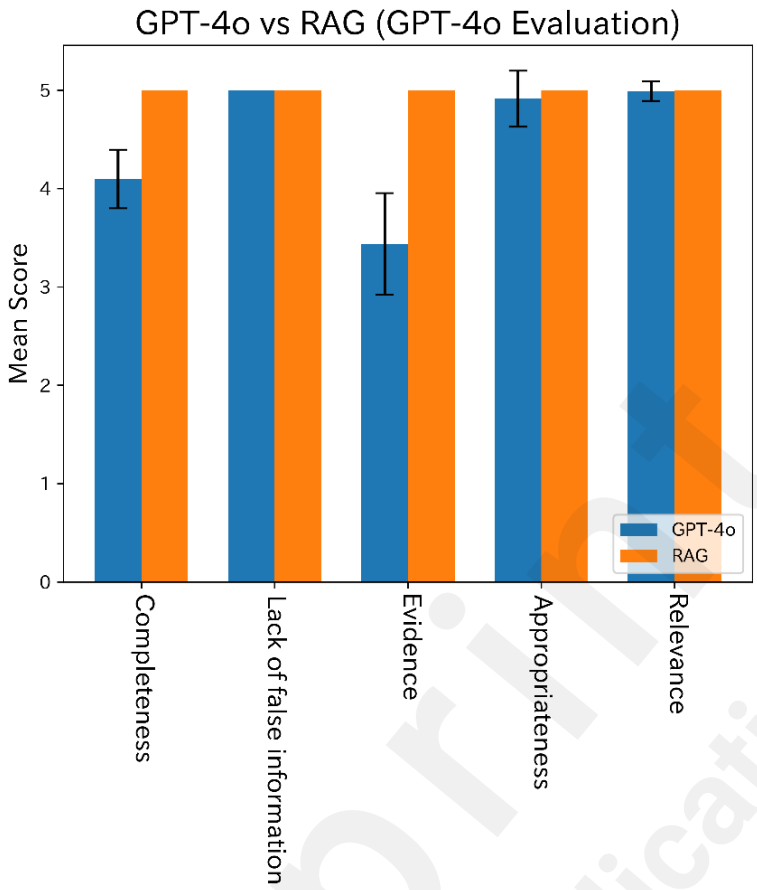


Fig 4. Comparison of Average Scores between GPT-4o and RAG (Evaluator: GPT-4o)

Figure 4 presents a comparison of the average scores for GPT-4o and RAG on the CLEAR evaluation metrics, as assessed by GPT-4. The error bars indicate the standard deviation. We conducted a paired t-test to evaluate whether the differences in average scores were statistically significant.

As a result, we found that GPT-4o achieved significantly higher average scores than RAG for Completeness ($p < .001$), Evidence ($p < .001$), and Appropriateness ($p = .002$). In contrast, there was no statistically significant difference for Relevance ($p = .32$), and the p-value for Lack of false information was not computable.

Table 2. Quantitative Assessment of GPT-4o and RAG by GPT-4

	GPT-4o (Mean ± SD)	RAG (Mean ± SD)	*p* Value
Completeness	4.10 ± 0.30	5.00 ± 0.00	<.001
Lack of false information	5.00 ± 0.00	5.00 ± 0.00	-
Evidence	3.44 ± 0.52	5.00 ± 0.00	<.001
Appropriateness	4.91 ± 0.28	5.00 ± 0.00	.002
Relevance	4.99 ± 0.10	5.00 ± 0.00	.320

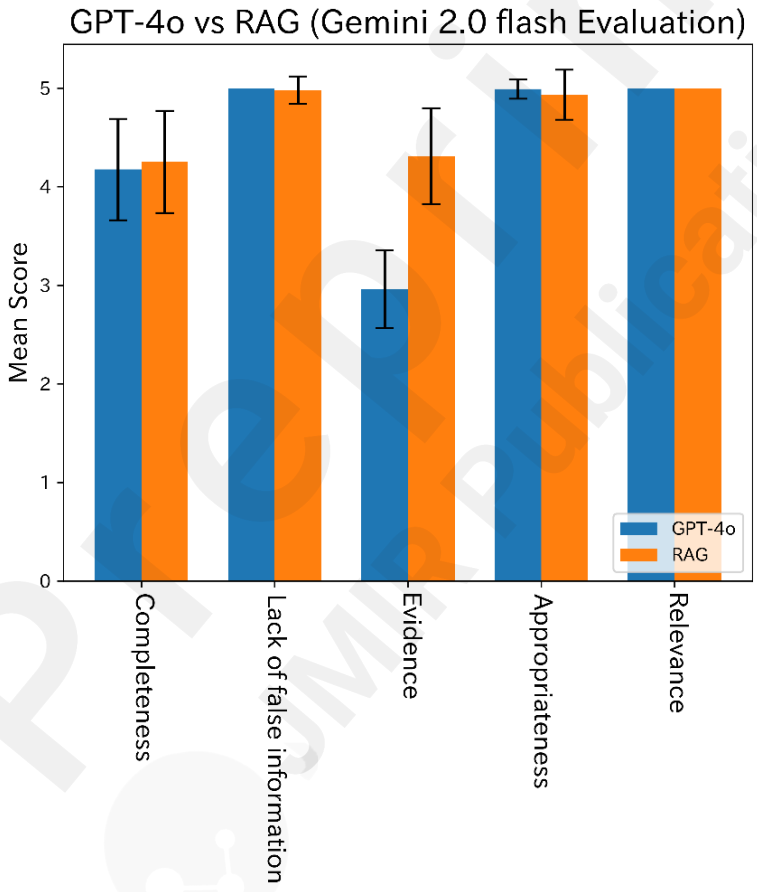


Fig 5. Comparison of Average Scores between GPT-4o and RAG (Evaluator: Gemini Flash 2.0)

Figure 5 shows a comparison of the average scores for GPT-4o and RAG on the CLEAR evaluation metrics, as assessed by Gemini Flash 2.0. The error bars represent the standard deviation. We conducted a paired t-test to evaluate whether the differences in average scores were statistically significant.

As a result, we observed statistically significant differences between GPT-4o and RAG in several key metrics. While RAG outperformed GPT-4o in Evidence (p < .001), GPT-4o achieved a significantly higher average score than RAG in Appropriateness (p = .033). In

contrast, we found no statistically significant differences for Completeness (p = .312), Lack of false information (p = .158), and Relevance (p = not computable).

Table 3. Quantitative Assessment of GPT-4o and RAG by Gemini 2.0 flash

	GPT-4o (Mean ± SD)	RAG (Mean ± SD)	*p* Value
Completeness	4.17 ± 0.51	4.25 ± 0.52	.312
Lack of false information	5.00 ± 0.00	4.98 ± 0.14	.158
Evidence	2.96 ± 0.39	4.31 ± 0.49	<.001
Appropriateness	4.99 ± 0.10	4.93 ± 0.25	.033
Relevance	5.00 ± 0.00	5.00 ± 0.00	-

Evaluation by Students

In the student evaluation, although the RAG system outperformed GPT-4o in terms of Evidence, GPT-4o achieved significantly higher average scores than RAG on other items, particularly Completeness, Appropriateness, and Relevance.

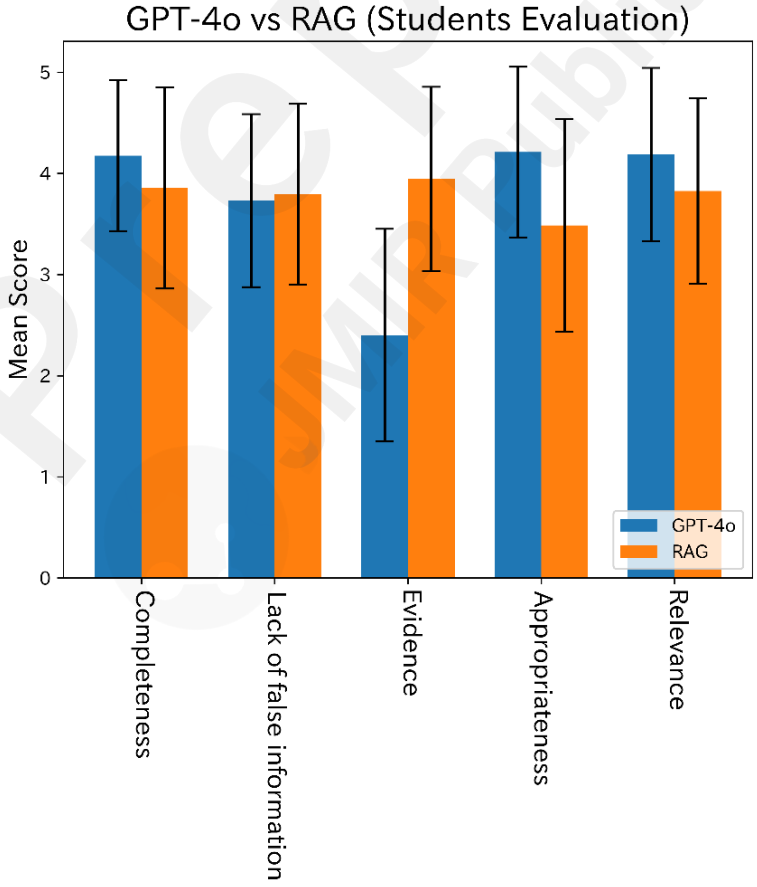


Fig 6. Comparison of Average Scores between GPT-4o and RAG (Evaluator: Students)

Figure 6 presents a comparison of the average scores for GPT-4o and RAG on the CLEAR

evaluation metrics, as assessed by students. The error bars indicate the standard deviation. We conducted a paired t-test to examine the statistical significance of the differences in average scores.

As a result, while RAG outperformed in Evidence ($p < 0.001$), GPT-4o achieved significantly higher average scores than RAG in Completeness ($p < 0.001$), Appropriateness ($p < 0.001$), and Relevance ($p < 0.001$). For Lack of false information ($p = .226$), we observed no statistically significant difference.

Table 4. Quantitative Assessment of GPT-4o and RAG by Students

	GPT-4o (Mean ± SD)	RAG (Mean ± SD)	*p* Value
Completeness	4.17 ± 0.75	3.85 ± 0.99	<.001
Lack of false information	3.73 ± 0.85	3.79 ± 0.89	.226
Evidence	2.40 ± 1.05	3.94 ± 0.91	<.001
Appropriateness	4.21 ± 0.84	3.48 ± 1.05	<.001
Relevance	4.18 ± 0.86	3.83 ± 0.92	<.001

Evaluation of References

In the citation evaluation, all cited URLs returned valid responses to HTTP requests. However, according to SourceCheckup's evaluation metrics, Statement-level Support was 0.516, which shows that approximately half of the claims were supported by appropriate references. The 95% confidence interval was (0.460, 0.571). Response-level Support was 0.228, indicating that only about 20% of the entire responses were fully supported. The 95% confidence interval was (0.158, 0.317). These results suggest that individual statements within LLM responses were not always fully substantiated by the cited sources, and that the entire response was not sufficiently supported by the references. In particular, we found a discrepancy between the information presented by the model and the content of the references it relied upon.

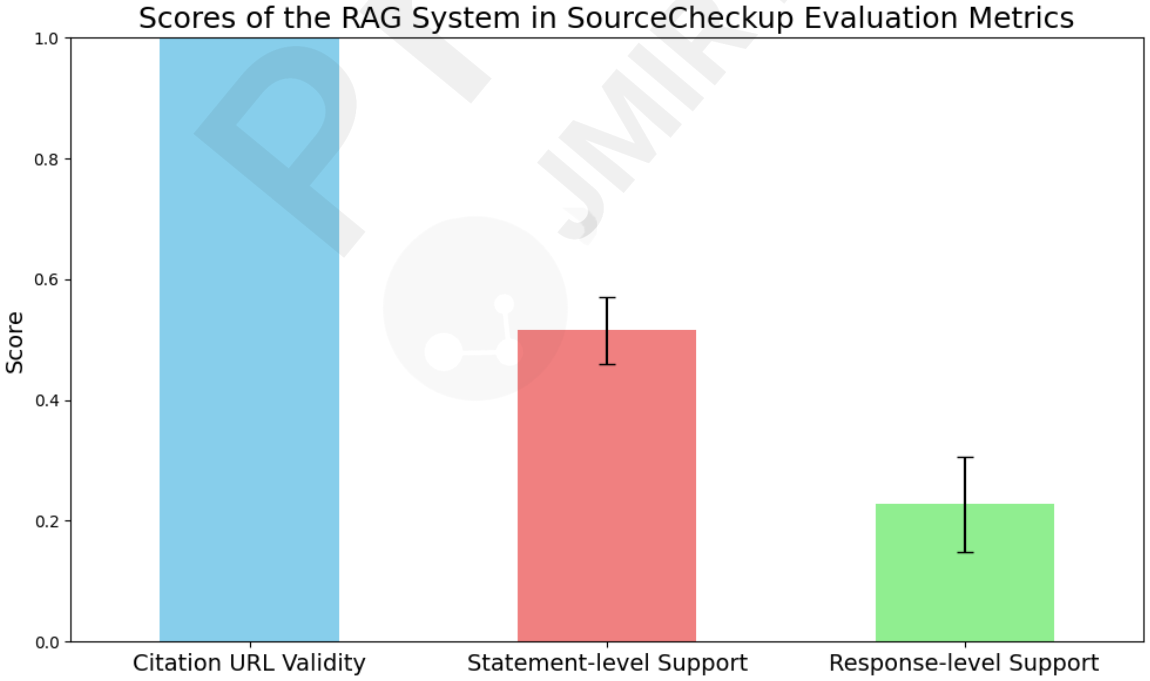


Fig 7. Calculated Scores and 95% Confidence Intervals (Bootstrap Method)

This bar graph presents three key scores calculated from the data set: Citation URL Validity, Statement-level Support, and Response-level Support. Error bars indicate the respective 95% confidence intervals, which we calculated using the bootstrap method.

Discussion

This study showed that the RAG system was a promising approach to address the challenges faced by LLMs in medical education, particularly regarding hallucinations and ambiguous evidence. In evaluations conducted by LLMs, the RAG system consistently achieved higher scores than standalone GPT-4o, especially in the “Evidence” category. These results suggest that RAG may increase the reliability and validity of answers by using external knowledge sources as evidence. This finding supports the idea that RAG’s characteristic of generating LLM responses by referencing retrieved information contributes to strengthening “Evidence.”

However, a different trend appeared in the students’ subjective evaluations (Figure 6). While the RAG system showed superiority in "Evidence," students rated GPT-4o higher in "Completeness," "Appropriateness," and "Relevance." This result suggests a discrepancy in the evaluation criteria for response "quality" between LLMs and human evaluators (students). In medical education, factors such as the overall response structure, naturalness of language, and directness in addressing the question are likely considered more important than only presenting accurate evidence. In this context, although the RAG system contributed to informational accuracy and explicit evidence by generating responses based on referenced information, it sometimes compromised the overall structure and naturalness of expression, which led to an impression of reduced consistency or clarity. A useful perspective for interpreting this discrepancy is the concept of “Incorrect Specificity” identified by Barnett et al. [13] as one of the common failure points in RAG systems. Previous studies have highlighted that this issue arises when responses are either too general or too narrowly tailored to be helpful. This effect may occur because the emphasis on fragmented integration of referenced documents can sacrifice the overall flow and coherence of the response.
The Citation URL Validity of 1.0 (100%) in the reference evaluation indicated that the RAG system could reliably provide valid URLs, which is an important foundation for students to verify information sources. This result demonstrated RAG's effectiveness in addressing the hallucination problem inherent in traditional LLMs. However, the Statement-level Support of 0.516 (approximately half) and Response-level Support of 0.228 (approximately 20%) highlighted a remaining challenge for the RAG system: even if the provided URLs were valid, their content did not necessarily directly support all statements in the generated response. When LLMs interpreted reference information, they sometimes added their own inferences or combined fragmented information from multiple sources to generate new statements, which made it possible for individual claims not to be fully supported by a single source.

A promising approach to address the challenges identified in this study is the introduction of Self-RAG (Self-Retrieval-Augmented Generation) [14]. Unlike conventional RAG, Self-RAG uses a structure in which the model self-reflects on the generated response and then re-searches and re-generates sections where the evidence is ambiguous or insufficient. This process allows the model to autonomously verify whether each claim is supported by explicit information sources and to make corrections as needed. This mechanism is expected to compensate for the shortcomings of the RAG system used in this study,

specifically the low statement-level support (0.516) and response-level support (0.228). Furthermore, Self-RAG showed good results in human evaluations. In research by Asai et al. [14], human evaluators judged Self-RAG's outputs for short question answering (PopQA) as valid and based on the provided information. Additionally, the self-evaluation metrics automatically indicated by the model closely matched human judgment, which demonstrates that the model itself can confirm the accuracy and appropriateness of its outputs to some extent. These results show that Self-RAG not only references external information but also functions as a mechanism to self-verify the consistency between the response content and the evidence, which leads to more accurate responses. For the challenges observed in this study, such as insufficient evidence and differences in evaluation between humans and models, Self-RAG could serve as a concrete improvement method.

Acknowledgments

We sincerely thank Kohki Yamada, from the Faculty of Medicine, Hokkaido University, for their cooperation in data collection and invaluable advice throughout all stages of this research.

The question dataset was generated using the generative AI tool ChatGPT by OpenAI and subsequently revised by the authors. Both the original ChatGPT transcripts and the complete finalized dataset are available in the Multimedia Appendix.

We would like to thank Editage (www.editage.jp) for English language editing.

Limitations

Scale and Diversity of the Evaluation Dataset

The question data set used to evaluate the RAG system consisted of 103 questions, which may represent a limited sample size. To cover the broad knowledge domain of medicine and draw more general conclusions, validation with a larger and more diverse set of questions is necessary. Furthermore, because the questions were generated by an LLM (GPT-4), we must consider the potential divergence from questions that students would actually have.

LLM Evaluator Bias

In the evaluation of responses by LLMs (GPT-4o and Gemini Flash 2.0), there was a potential for self-evaluation bias because the evaluating model, GPT-4, belonged to the same family as the response-generating model, GPT-4o. When the evaluator and generator models share the same lineage, their internal evaluation criteria and output patterns may be similar. This similarity could lead to a tendency to perceive greater consistency or persuasiveness in GPT-series outputs. This potential bias might have hindered objective evaluation. We believe that this bias could be reduced by increasing evaluations from models with different architectures or by including more assessments from entirely independent human experts.

Limitations of Student Subjective Evaluation

While the subjective evaluation by 40 students from the Faculty of Medicine and School of

Nursing at Hokkaido University provided valuable insights, the number of evaluators was limited. Increasing the number of participating students, perhaps from multiple universities, could increase the reliability of the evaluation.

Conclusion

This study showed that the RAG system could effectively address challenges faced by LLMs in medical education, particularly the risks of hallucinations and unclear answer sourcing. Although the RAG system consistently improved the evidence supporting LLM responses—a critical factor for accuracy in medicine—subjective evaluations by students indicated a tendency to rate GPT-4o higher than RAG in areas such as completeness and appropriateness. This finding suggests that, for learners, elements such as response structure and natural language flow are also important, and further improvements are needed for practical implementation in educational settings. Additionally, while RAG provided valid URLs, these references did not always directly support every claim in the generated answers. Future enhancements should focus on integrating mechanisms to improve consistency between referenced information and the generated output. Self-RAG (Self-Retrieval Augmented Generation), which includes self-verification capabilities, appears to be a promising approach for this purpose, supporting the development of educational tools that offer both higher accuracy and clearer evidence. In conclusion, RAG serves as a valuable foundational technology for increasing the reliability of LLM applications in medical education, and advanced methods such as Self-RAG may enable even higher-quality learning support in the future.

Conflicts of Interest

None declared.

Ethical Considerations

Upon submitting an ethics review application to the Hokkaido University Hospital Institutional Review Board, we received a response stating that this research would be exempt from the ethics guidelines as it is being conducted for educational purposes.

Data Availability

The code supporting the findings of this study, including the RAG system implementation and evaluation scripts, are openly available in the GitHub repository at [https://github.com/S-murakami1/MedRAG.git].

Multimedia Appendix

The original ChatGPT transcripts and complete version of Table 1.

References

1. Howley LD, Whelan AJ. From the World Wide Web to AI: Why We Must Learn From Our Past to Transform the Future of Medical Education. Acad Med. Published online May 30, 2025. doi: 10.1097/ACM.0000000000006103

2. Lieb, A., & Goel, T. (2024). Student Interaction with NewtBot: An LLM-as-tutor Chatbot for Secondary Physics Education. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.doi: 10.1145/3613905.3647957

3. Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare, 11. doi: 10.3390/healthcare11060887

4. Lucas, H.C., Upperman, J.S., & Robinson, J.R. (2024). A systematic review of large language models and their implications in medical education. Medical Education, 58, 1276 - 1285. doi: 10.1111/medu.15402

5. Xu, X., Chen, Y., & Miao, J. (2024). Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. Journal of Educational Evaluation for Health Professions, 21. doi: 10.3352/jeehp.2024.21.6

6. Benítez, T.M., Xu, Y., Boudreau, J.D., Kow, A.W., Bello, F., Phuoc, L.V., Wang, X., Sun, X., Leung, G.K., Lan, Y., Wang, Y., Cheng, D., Tham, Y., Wong, T.Y., & Chung, K.C. (2024). Harnessing the potential of large language models in medical education: promise and pitfalls. Journal of the American Medical Informatics Association : JAMIA. doi: 10.1093/jamia/ocad252

7. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. doi: 10.1016/j.lindif.2023.102274

8. Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. ACM Trans Inf Syst. 2025;43(2):Article 42. doi: 10.1145/3703155

9. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. Healthcare, 13. doi: 10.3390/healthcare13060603

10. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv, abs/2005.11401. doi: 11.48550/arXiv.2005.11401

11. Sallam M, Alrahabi M, Alabbad A, et al. Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. Cureus. 2023;15(11):e49373. doi: 10.7759/cureus.49373

12. Wu K, Wu E, Wei K, et al. An automated framework for assessing how well LLMs cite relevant medical references. Nat Commun. 2025;16:3615. doi: 10.1038/s41467-025-58551-

13. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., & Abdelrazek, M. (2024). Seven Failure Points When Engineering a Retrieval Augmented Generation System. 2024 IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN), 194-199. doi: 10.1145/3644815.3644945

14. Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to

Retrieve, Generate, and Critique through Self-Reflection. ArXiv, abs/2310.11511. doi: 10.48550/arXiv.2310.11511

Abbreviations

CLEAR: Completeness, Lack of false information, Evidence, Appropriateness, Relevance
LLM: Large Language Model
RAG: Retrieval Augmented Generation

Supplementary Files

Multimedia Appendixes

The original ChatGPT transcripts for the question dataset and the complete version of Table 1 are provided.
URL: http://asset.jmir.pub/assets/8620987fa3e1cd81e380d4fe7eed4014.pdf