information

MDPI

Article

Performance Evaluation of Large Language Model Chatbots for Radiation Therapy Education

Jae-Hong Jung [1], Daegun Kim [1], Kyung-Bae Lee [2] and Youngjin Lee [3,*]

[1] Department of Radiation Oncology, College of Medicine, Soonchunhyang University Hospital Bucheon, 170, Jomaru-ro, Wonmi-gu, Bucheon-si 14584, Republic of Korea; rtjung@schmc.ac.kr (J.-H.J.); hidegun@schmc.ac.kr (D.K.)
[2] Department of Radiotechnology, Wonkwang Health Science University, 514, Iksan-daero, Iksan-si 54538, Republic of Korea; kblee@wu.ac.kr
[3] Department of Radiological Science, Gachon University, 191, Hambakmoe-ro, Yeonsu-gu, Incheon 21936, Republic of Korea
* Correspondence: yj20@gachon.ac.kr; Tel.: +82-32-820-4362

Abstract

This study aimed to develop a large language model (LLM) chatbot for radiation therapy education and compare the performance of portable document format (PDF)- and webpage-based question-and-answer (Q&A) chatbots. An LLM chatbot was created using the EmbedChain framework, OpenAI GPT-3.5-Turbo API, and Gradio UI. The performance of both chatbots was evaluated based on 10 questions and their corresponding answers, using the parameters of accuracy, semantic similarity, consistency, and response time. The accuracy scores were 0.672 and 0.675 for the PDF- and webpage-based Q&A chatbots, respectively. The semantic similarity between the two chatbots was 0.928 (92.8%). The consistency score was one for both chatbots. The average response time was 3.3 s and 2.38 s for the PDF- and webpage-based chatbots, respectively. The LLM chatbot developed in this study demonstrates the potential to provide reliable responses for radiation therapy education. However, its reliability and efficiency must be further optimized to be effectively utilized as an educational tool.

Keywords: radiation therapy education; professional education; large language model; chatbot

check for updates

Academic Editor: Evaggelos Karvounis

Received: 20 May 2025
Revised: 16 June 2025
Accepted: 20 June 2025
Published: 22 June 2025

Citation: Jung, J.-H.; Kim, D.; Lee, K.-B.; Lee, Y. Performance Evaluation of Large Language Model Chatbots for Radiation Therapy Education. *Information* **2025**, *16*, 521. https://doi.org/10.3390/info16070521

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Radiotherapy, along with surgery and chemotherapy, is an essential treatment for patients with cancer and is classified separately from diagnostic imaging in the medical field. The use of high-energy radiation in radiation therapy requires specialized expertise for professional execution. Radiologic science students or prospective radiologic technologists who wish to work in the field of radiation therapy must have a deep understanding of it. Radiotherapy involves complex concepts and the latest treatment techniques, making it challenging for students and learners. Educators must consider not only fundamental theories but also various learning methods and effective information delivery strategies to enhance radiation therapy science education. In addition, learners should approach this field with interest and strive to accumulate knowledge to develop the competencies required in clinical settings.

A large language model (LLM) is an artificial intelligence model that learns linguistic patterns from vast amounts of data, thereby enabling language comprehension and generation. It is designed to understand the context of natural language used by humans and

to generate coherent conversations and texts. Representative models include OpenAI's GPT [1], Meta's LLaMA [2], Google's PaLM [3], BARD (Gemini) [4], and BERT (Bidirectional Encoder Representations from Transformers) [5].

LLMs generally operate via three main processes: tokenization, transformers, and prompts. Tokenization is a natural language processing (NLP) technique that converts human language into sequences that machine-learning models can process. Second, the transformer model serves as the foundation for LLM-driven language generation, acting as a type of neural network that sequentially analyzes data and identifies word patterns. This process enables unsupervised learning, allowing the model to understand fundamental grammar, language structures, and knowledge. Finally, a prompt refers to the input provided to the LLM by developers to guide its processing and output based on the training data [6].

LLMs are widely used in various fields, including education, research, healthcare, and customer service, for tasks such as text generation, language translation, content summarization, code generation, and conversational interactions. Chatbots based on LLMs are particularly valuable in education. They allow students to receive instant responses to their questions while learning, thereby overcoming the time constraints of traditional teaching methods. By providing real-time feedback, learners can enhance their comprehension and improve their learning efficiency.

In the field of radiation therapy education, several studies have investigated the application of artificial intelligence (AI)-assisted chatbots to support knowledge transfer and improve learning outcomes. Chow et al. [7] developed an educational chatbot for radiotherapy that utilized dialogue tree structures and layered scenarios to simulate personalized conversations with cancer patients and their families, the general public, and radiation staff. This design incorporated emotionally supportive interactions, allowing users to select relevant topics based their needs while providing reassurance for individuals who may experience anxiety during information seeking related to cancer treatment [7]. In a subsequent study, Chow et al. [8] further enhanced the chatbot using the IBM Watson Assistant platform, employing a pre-designed flowchart structure to streamline dialogue management and incorporating NLP techniques to improve user comprehension and engagement. The chatbot delivered structured radiotherapy knowledge to both clinical professionals and individuals without prior medical expertise. Moreover, these studies have emphasized the importance of continuous evaluation and iterative refinement based on user feedback to optimize chatbot performance and educational effectiveness [9].

EmbedChain is a framework for implementing LLM-based bots that utilizes embedding technology to collect information from various data sources {including webpages, videos, documents, research papers, and portable document format (PDF) files} and convert it into numerical representations (mathematical vectorization) for machine processing. It is an open-source, Python-based library [10,11]. Similar frameworks include LangChain, Haystack, Weaviate, Chroma, and Pinecone. Table 1 outlines the differences among these frameworks. EmbedChain is designed to facilitate the integration of diverse data sources, enabling seamless access to both structured and unstructured data [12]. Additionally, it offers an intuitive interface that simplifies the machine learning model deployment process, allowing users to quickly set up applications without extensive technical knowledge. This makes it more user-friendly than other frameworks, such as LangChain. Finally, EmbedChain supports application programming interface (API) integration with models, such as OpenAI's GPT, enabling developers to efficiently leverage advanced natural language processing capabilities. This functionality allows for the creation of educational applications tailored to user needs.

Table 1. Comparison of EmbedChain and similar frameworks in functionality and key differences.

Framework	Primary Functionality	Key Features
EmbedChain	Embedding and retrieval support for document data	Document and PDF embedding; flexible search capabilities
LangChain	Optimizes response workflows with LLMs	Robust data pipeline supports multiple data sources
Haystack	Builds NLP applications and supports RAG (Retrieval-Augmented Generation)	Multi-backend support, document search, and Q&A; integration with Pinecone
Weaviate	Vector database with AI-powered search	Real-time search integrates with knowledge graphs
Chroma	Embedding and search engine for vector data	High-performance real-time search and storage
Pinecone	Vector-based search and memory expansion	API-based management of largescale vector data; real-time search

NLP, natural language processing; LLM, large language model; Q&A, question-and-answer.

The EmbedChain framework combines ease of use, flexible data integration, and powerful AI capabilities, making it a viable tool for interactive education. In the education sector, an EmbedChain-powered LLM chatbot can retrieve relevant content from various learning materials, such as lecture notes and textbooks, to answer student queries. This enables the provision of personalized learning resources for each student and facilitates direct question-and-answer (Q&A) interactions based on textual and multimedia lecture materials. The framework operates by converting various sources into embedded representations, storing them in a vector database, and retrieving relevant information in real time when a user submits a query, thereby delivering accurate responses efficiently [13].

The purpose of this study was to develop an LLM chatbot for radiation therapy education and compare the performance of PDF- and webpage-based Q&A chatbots. Using the EmbedChain framework, relevant educational materials were collected and embedded from both formats. While previous studies have explored AI-based chatbots in healthcare education, this study offers a practical approach by applying dual-source integration in the context of radiation therapy learning. The findings are expected to inform future improvements in chatbot design and contribute to more effective AI-assisted learning environments.

2. Materials and Methods

2.1. Architecture Design for LLM Chatbot

The LLM chatbot used in this study was implemented using the Python-based EmbedChain framework, which enables the embedding of technology. Figure 1A shows a schematic diagram of the overall chatbot framework, and answer generation through a unified interface. The overall chatbot process is illustrated in the flowchart shown in Figure 1B, which outlines the step-by-step operation from user query input to answer generation and display. When the user submits a query, the chatbot retrieves the embedded information stored in EmbedChain and provides a textual response. To ensure intuitive responses to user queries, this study utilized OpenAI's GPT-3.5-Turbo model (San Francisco, CA, USA) integrated with an API key. The GPT-3.5-Turbo model by OpenAI was selected for this study due to its strong language comprehension capabilities, reliable performance, and compatibility with the EmbedChain framework. It also offers an efficient cost-to-performance ratio and is widely used in academic and industrial applications. For

user convenience, the chatbot was equipped with a Gradio-based user interface (UI) that allowed users to interact through questions and answers. Gradio is a tool designed to easily create web interfaces for machine learning models, offering ease of use and compatibility across various devices, including personal computers and smartphones. In addition, it provides flexible deployment options along with customizable input and output features, making it a highly versatile solution [14].

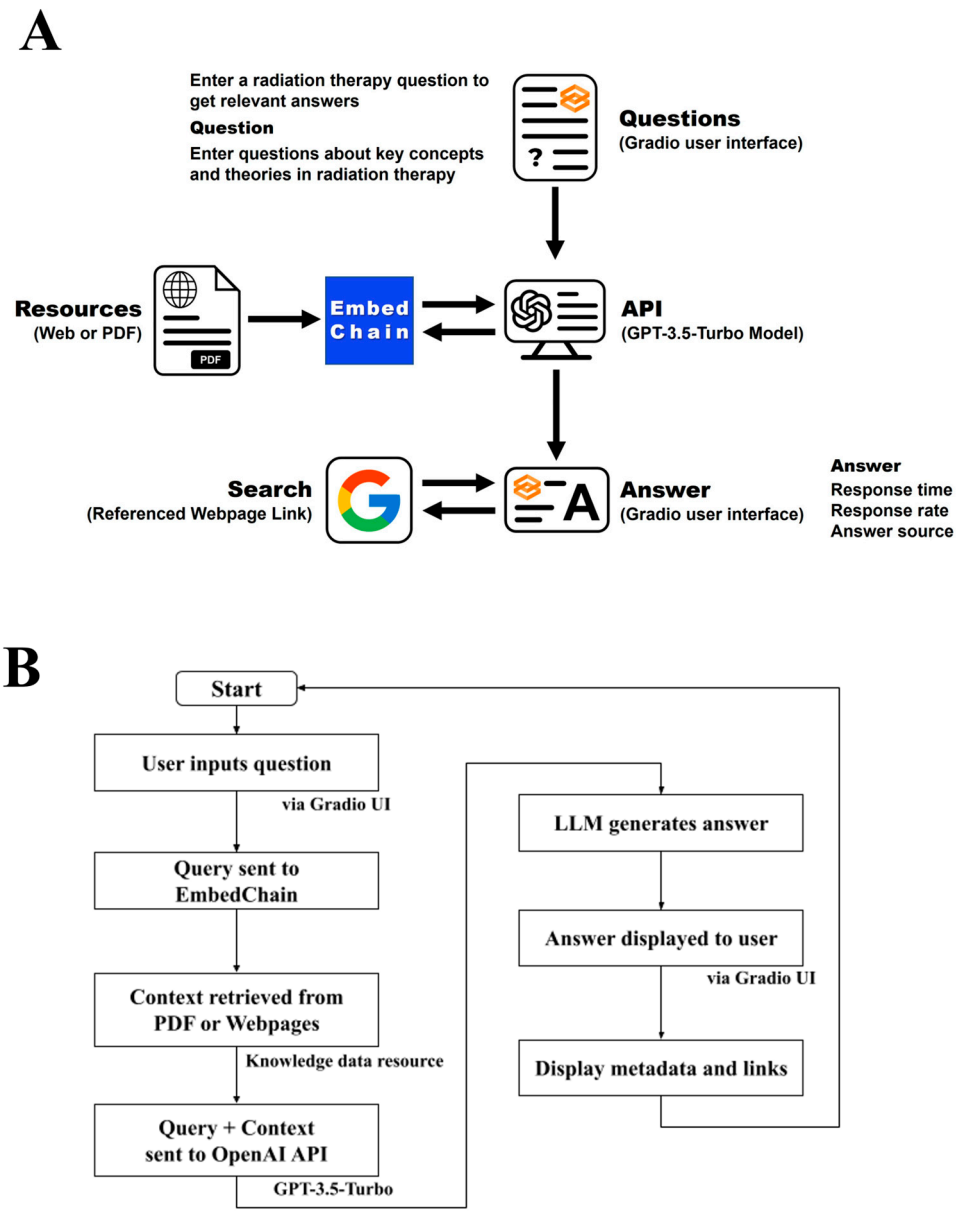


Figure 1. (**A**) Schematic diagram of the LLM chatbot framework for radiation therapy education, showing user interaction via Gradio UI, knowledge retrieval using the EmbedChain framework, and answer generation through the OpenAI GPT-3.5-Turbo API. (**B**) Standardized flowchart illustrating the sequential process of the chatbot operation from user question input, context retrieval from PDF or webpage sources, to LLM-based answer generation and final output with display and metadata.

The LLM chatbot program was developed using Python (version 3.10), with PyCharm (version 2024.03.05) (JetBrains, Prague, Czech Republic) as the programming editor [15]. The program utilizes the EmbedChain framework to store radiation therapy-related PDF documents or webpage data in a vector search database and incorporates a chatbot function for Q&A interactions. The chatbot operation began by setting the OpenAI GPT-3.5-Turbo API key as an environmental variable, followed by initializing the chatbot by creating an

EmbedChain instance. To enhance the embedding process for generating useful responses, relevant resources, including Q&A PDFs and webpage URLs, were applied. For Q&A processing, the {answer_query} function was used to generate responses using EmbedChain. The chatbot was designed to facilitate user interaction via a Gradio-based web interface (Table 2).

Table 2. Key features of the LLM chatbot code in this study.

Feature	Description
API Key setup	Environment variable, secure, sensitive information
Instance creation	App class, chatbot initialization
Resource addition	URL compilation, radiation therapy, error handling
Query handling	{answer_query} function, user input processing, answer retrieval
Interface setup	Gradio UI, input fields, submit button, visual elements
Chatbot execution	App accessibility

2.2. Knowledge Data Resource

The Q&A PDFs and webpage data were prepared to evaluate the functionality and performance of the LLM chatbot developed in this study. The Q&A PDF consisted of preexisting data used by the author for the research. It included a total of seven categories: the concept of radiation therapy, radiation therapy physics, linear accelerators, spatial dose distribution, simulation, treatment planning, and radiation therapy accessories. These categories were selected based on key learning objectives in radiation therapy education and reflect foundational concepts that learners must master for clinical application. The dataset contained 182 questions and their corresponding answers, comprising 16,899 words, and it was formatted as a text document that included Korean, English, numbers, and symbols [16].

Webpage data were selected to ensure easy access to real-time information, as they contain text-based content derived from hypertext markup language (HTML) structures. Ten webpages related to radiation therapy were selected from medical institutions and relevant organizations. These included the National Cancer Center (covering proton therapy and volumetric modulated arc therapy), the Samsung Medical Center (providing "50 Facts About Radiation Therapy"), and Chungbuk National University Hospital (covering radiation therapy). Other sources included Yeungnam University Medical Center (covering radiation therapy), Wikipedia (radiation therapy), medical articles (explaining why radiation therapy is administered in multiple sessions), the Seoul Asan Medical Center (surface guided radiation therapy), and the International Atomic Energy Agency (IAEA). Webpages were chosen based on the relevance of their content to the educational objectives, institutional credibility, and the accessibility of information for students.

Both the PDF and webpage data were embedded into a vector database using the EmbedChain framework, which converts text into numerical representations for real-time retrieval during chatbot interaction. To ensure content validity and professional relevance, the Q&A items were reviewed by two experts in radiation therapy, including a medical physicist (Ph.D.) and a senior radiation therapist (Ph.D.), each with over 20 years of clinical experience and extensive teaching experience in university-level education.

2.3. Q&A Data

The performance of the LLM chatbot developed in this study was evaluated by selecting 10 questions based on key concepts and theories related to radiation therapy (Table 3). Reference answers were prepared by the experts with reference to authoritative

sources, such as IAEA guidelines [17], the ICRU (International Commission on Radiation Units and Measurements) Report 83 [18], and AAPM (American Association of Physicists in Medicine) task group reports [19,20]. This ensured consistency and minimized bias in the evaluation process. For each question, a gold-standard answer was defined, and the chatbot's responses, generated using either PDF-based or webpage-based Q&A data, were compared with the gold-standard answers using evaluation metrics that included accuracy, semantic similarity, consistency, and response time.

Table 3. Questions for measuring accuracy and semantic similarity.

Number	Question
1	What is the principle of radiation therapy?
2	What is the difference between IMRT and VMAT?
3	What are the advantages of proton therapy?
4	What are the side effects of radiation therapy?
5	What is the difference between diagnostic CT and therapeutic CT?
6	How is the radiation dose determined in radiation therapy?
7	What are the strategies for protecting normal tissues in radiation therapy?
8	What is the difference between SRS and SBRT?
9	What is the purpose of radiation therapy?
10	What is SGRT?

2.4. Performance Evaluation Methods

The accuracy of the chatbot responses was determined by combining the scores from the Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), SBERT (Sentence-BERT), and BERTScore. The final accuracy score was obtained by calculating a weighted average of the individual evaluation scores, with the following weight distribution: BLEU (15%), ROUGE (15%), SBERT (35%), and BERTScore (35%) [21]. BLEU measures sentence similarity by calculating the n-gram overlap between the chatbot-generated response and the gold-standard answer. This follows a simple word-matching approach: as shown in Equation (1), BP is an adjustment factor applied to penalize short sentences, where w_n represents the weight, and p_n denotes the n-gram precision. BLEU primarily evaluates exact word matches rather than the overall meaning of a sentence, focusing on n-gram correspondence [22].

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n\right) w_n log p_n \quad (1)$$

ROUGE evaluates the word-level similarity between the chatbot response and gold-standard answer by calculating ROUGE-1, ROUGE-2, and ROUGE-L. It measures the matching rate of consecutive words in a sentence. Equations (2) and (3) represent ROUGE-1, ROUGE-2, and ROUGE-L, respectively. In ROUGE-*n*, $n = 1$ refers to ROUGE-1 (unigram recall), and $n = 2$ refers to ROUGE-2 (bigram recall). ROUGE-L evaluates similarity based on the length of the longest common subsequence (LCS) within a sentence [23]. In Equation (3), C represents the candidate sentence (generated response), whereas R denotes the reference sentence (gold-standard answer). Essentially, ROUGE follows a recall-oriented evaluation approach that assesses the extent to which a summarized sentence retains important information from the original document.

$$ROUGE - N = \frac{\Sigma_{matched\ n-grams}}{\Sigma_{total\ n-grams\ in\ reference}} \quad (2)$$

$$ROUGE - L = \frac{LCS(C,R)}{|R|} \tag{3}$$

SBERT evaluates semantic similarity by calculating the cosine similarity between the chatbot's response and the gold-standard answer. Unlike simple word matching, SBERT assesses the contextual meanings of sentences. In Equation (4), A and B represent the embedding vector values of the two sentences transformed using the SBERT model. Terms $||A||$ and $||B||$ denote the Euclidean norms (lengths) of vectors A and B, respectively. By leveraging SBERT, the evaluation focused on semantic similarity rather than exact word matches, ensuring a more context-aware assessment of the chatbot's responses [24].

$$Cosine\ Similarity = \frac{A \cdot B}{||A||||B||} \tag{4}$$

BERTScore is a semantic similarity evaluation method that utilizes transformer models and serves as a key metric for assessing whether a chatbot generates semantically accurate responses. Equation (5) illustrates the word-embedding vector calculation process. Using BERT, the input sentence (X) and reference sentence (Y) were converted into embedding vectors X_i and Y_j, respectively. In this process, the evaluation involved computing precision (P), recall (R), and the F1 score, which is the harmonic mean of precision and recall. By employing BERTScore, the chatbot responses were precisely assessed in terms of semantic similarity, ensuring a more accurate evaluation of language understanding [25].

$$s_{i,j} = \frac{X_i \cdot Y_j}{||X_i||||Y_j||} \tag{5}$$

2.5. Semantic Similarity and Consistency

The semantic similarity score was evaluated using SBERT and BERTScore, and the final semantic similarity score was calculated as the average of the two metrics. BERTScore ranges from zero to one, with values closer to one indicating higher semantic similarity. BERTScore effectively evaluates sentences with different wording but identical meaning [26].

Consistency measures how reliably a chatbot provides consistent responses to identical questions. It was assessed using SBERT, which calculates the cosine similarity between responses to the same question. The final consistency score was obtained by averaging the measured cosine similarity values. A cosine similarity score approaching one indicates a high degree of response consistency [27].

2.6. Response Time

Response time was evaluated by analyzing the speed at which the chatbot generated answers and assessing its real-time Q&A performance. The time taken for the chatbot to generate a response to each question was measured, and the average response time was calculated.

3. Results

3.1. LLM Chatbot

In this study, two LLM chatbots were developed, one based on Q&A PDFs and the other on webpages. Figure 2 displays the user interface of the Q&A PDF-based LLM chatbot, including the input section for user queries, generated answers, category information, system response details, such as response source, response time, and response rate, and external webpage links offering supplementary reference materials. The entire chatbot interaction took place on a single screen, where all components were presented in an

integrated layout. Both LLM chatbots had the same UI. For user convenience, the chatbot was implemented with an intuitive Gradio UI layout, designed as a web-based chatbot interface. At the top, a banner image was displayed, along with a title emphasizing radiation therapy education and a brief description explaining the chatbot's purpose. For ease of identification, the title also indicated whether the chatbot was based on Q&A PDFs or webpages. Users could enter their questions in a text box labeled "Enter your question," and the chatbot's responses appeared in the output section labeled "Answer". Interaction began when users clicked the "Ask" button. The chatbot also provided the response time, response accuracy, and source information for each answer. In addition, to offer further reference material, users could click on the question content to access a relevant webpage for additional information.

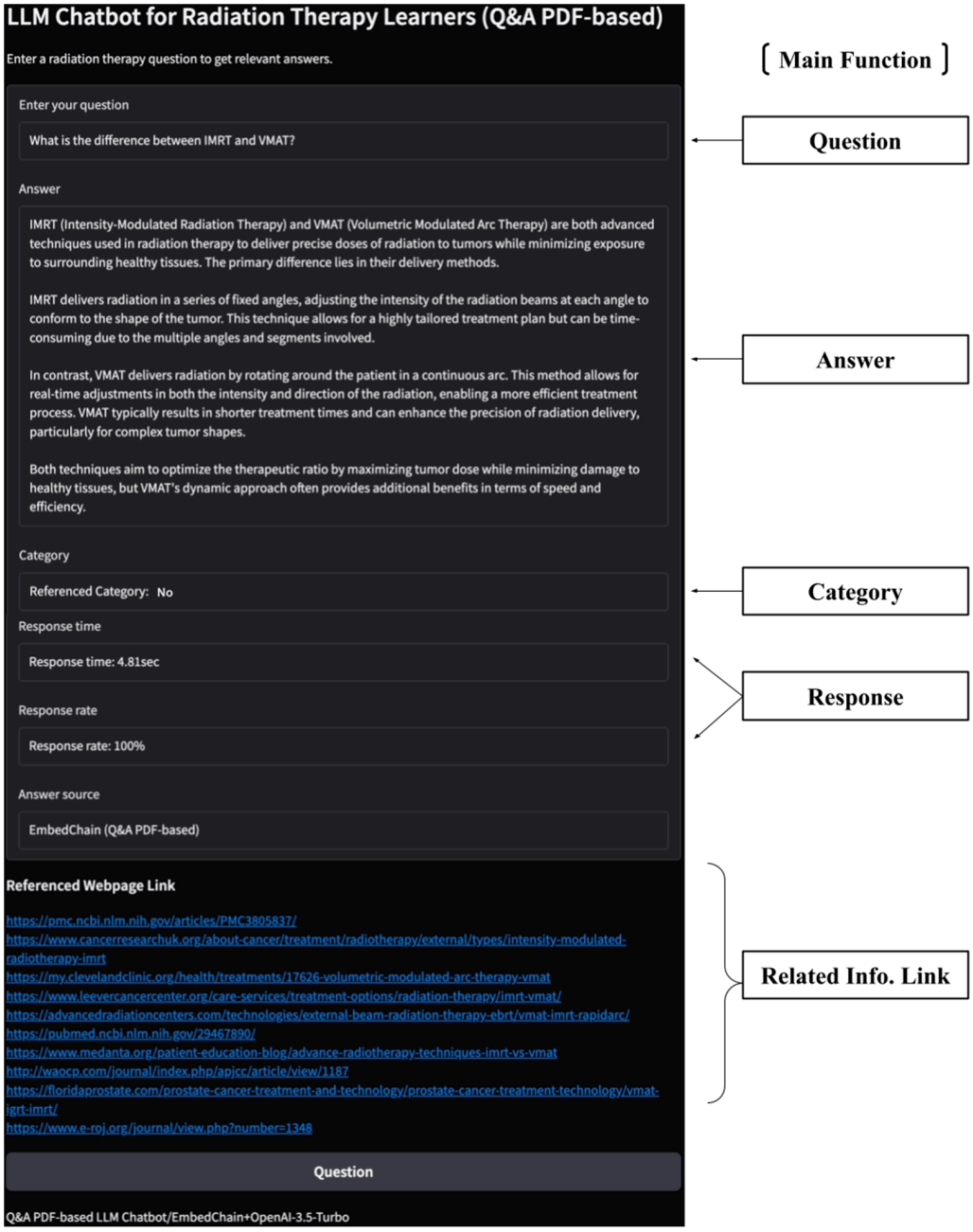


Figure 2. Example interface of the Q&A PDF-based chatbot developed using EmbedChain and the OpenAI API, showing the main functional components, such as user question input, generated answer, category, response details (answer source, response time, response rate), and referenced webpage links providing additional related information.

3.2. Performance Evaluation Results

The performance evaluation results of the developed LLM chatbots are listed in Table 4. The evaluation was conducted based on 10 questions, and the results represent the average performance of each chatbot. The Q&A PDF-based chatbot achieved an average accuracy score of 0.672, whereas the web page-based chatbot scored 0.675. Regarding semantic similarity, the SBERT score was 0.957 and 0.965 for the Q&A PDF- and web page-based chatbots, respectively. The average BERTScore was 0.9 and 0.901 for the Q&A PDF- and webpage-based chatbots, respectively. The overall semantic similarity between the chatbot responses and gold-standard answers was identical for both models, with a score of 0.928 (92.8%). Additionally, both chatbots demonstrated perfect consistency, scoring 1.0 in this category. The individual performance results for each of the 10 questions are illustrated in Figures 3 and 4, which show the evaluation outcomes for the Q&A PDF- and webpage-based chatbots, respectively. The Q&A PDF-based chatbot had an average response time of 3.3 s, whereas the webpage-based chatbot responded faster, with an average response time of 2.38 s.

Table 4. Performance comparison of Q&A PDF-based and webpage-based chatbots.

Performance	EmbedChain LLM Chatbot	
	Q&A PDF	**Webpage**
Accuracy	0.672	0.675
SBERT	0.957	0.965
BERTScore	0.9	0.901
Semantic similarity	0.928	0.928

LLM, large language model; Q&A, question-and-answer.

A

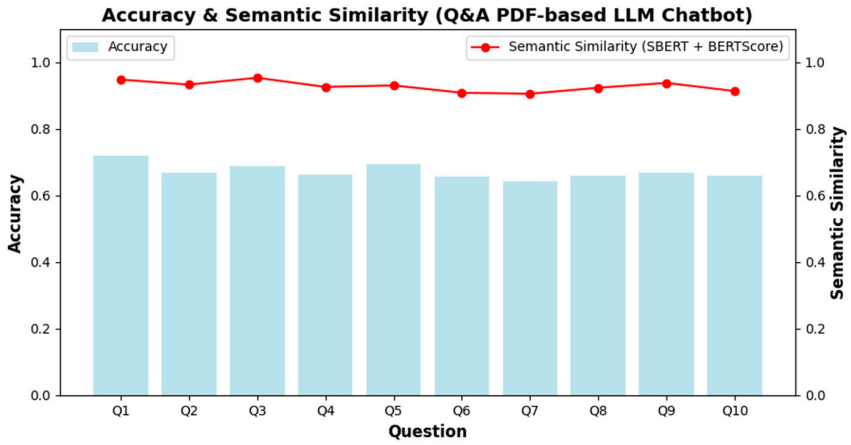


Figure 3. *Cont.*

B

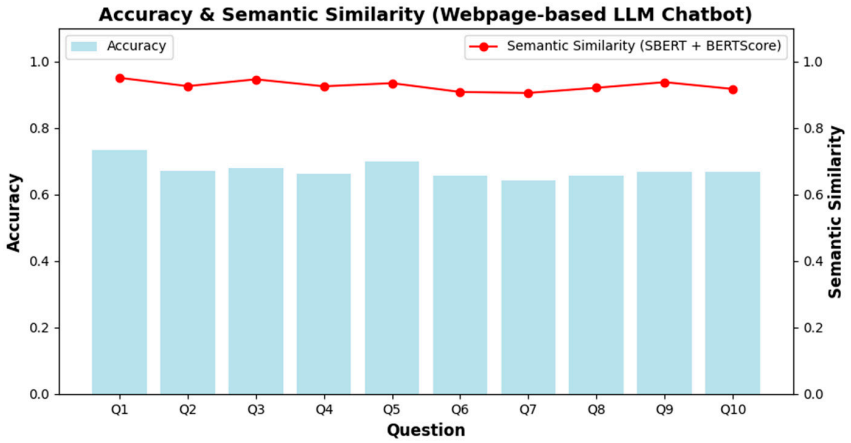
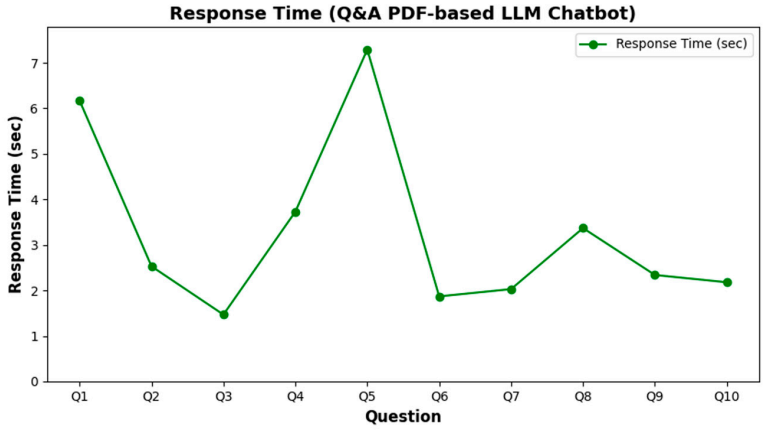


Figure 3. (**A**) Accuracy and semantic similarity of the Q&A PDF-based chatbot for 10 different questions. (**B**) Corresponding results for the webpage-based chatbot. In both cases, higher scores indicate better alignment with the gold-standard answers.

A



B

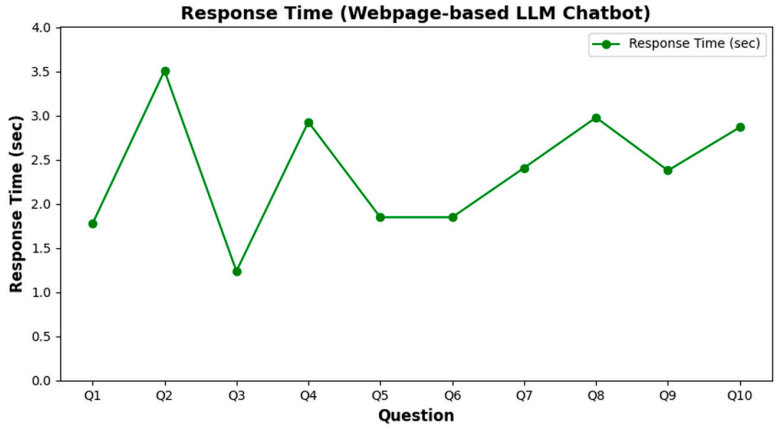


Figure 4. (**A**) Response time for each question using the Q&A PDF-based chatbot. (**B**) Response time for each question using the webpage-based chatbot. In both cases, shorter times reflect higher efficiency.

4. Discussion

4.1. System Implementation and User Interface Design

The LLM chatbot developed in this study was designed for radiation therapy education by incorporating Q&A PDF- and webpage-based systems. Both chatbots were implemented using a Gradio UI to ensure an intuitive UI. For user convenience, the system was designed not only to facilitate question input and response generation but also to provide meta-information about each response, including response time, response accuracy, and the source of the retrieved document. The chatbot was further designed to support learning by allowing users to explore additional resources through webpage links. Furthermore, the Gradio UI enables LLM chatbots to be accessed and used on smartphones, ensuring greater flexibility and ease of use.

The LLM chatbot developed in this study was implemented using a Gradio UI, enabling its use across various devices, including smartphones and PCs. However, to ensure sustainable use in educational settings, several operational and improvement measures must be considered. First, a cost-reduction strategy for servers and APIs must be established. Currently, the Gradio server has a 72 h usage limitation, and the token usage cost of the OpenAI API could become a financial burden for long-term operations. To address this issue, building a local server-based LLM model or exploring open-source alternatives may be beneficial. For example, utilizing open-source LLMs, such as LLaMA 2, Mistral AI, or GPT-4-All, could reduce API costs while providing advantages in terms of data protection. Second, the role of LLM chatbots in radiation therapy education must be clearly defined. The chatbot can serve as a teaching assistant to support instructors or as a self-learning tool for students. Therefore, it is necessary to establish guidelines on how educators can use chatbots effectively. In addition, a process to verify the reliability of chatbot responses should be implemented to ensure their effectiveness in education. Third, continuous improvements should be made by incorporating user feedback from educational settings. This study primarily conducted a quantitative performance evaluation of the chatbots' responses. However, a regular review process based on student and instructor feedback is essential. To achieve this, a user evaluation system for chatbot responses should be integrated, allowing for the analysis of user satisfaction and identification of areas for improvement.

4.2. Comparative Performance Evaluation of Chatbots

The performance of the LLM chatbots was evaluated using accuracy, semantic similarity, and consistency metrics, allowing for a comparative analysis between the two systems. The difference in accuracy between the Q&A PDF- and webpage-based chatbots was only 0.3%, indicating that both systems provided similar response quality. Several factors may explain this small accuracy difference. The first possible reason is the difference in expression between the questions and answers. The gold-standard answers used for evaluation were formulated at the sentence level based on reference books and relevant materials. Because the chatbot generates responses based on the LLM's learning process, its word choices and sentence structures are likely to differ from those of predefined answers rather than match at the sentence level.

Therefore, it is important to consider that n-gram-based accuracy measurements, such as BLEU and ROUGE scores, may be lower because of variations in phrasing. The second reason is the limitation of the evaluation method. BLEU and ROUGE primarily use n-gram-based quantitative evaluation metrics, meaning that if a chatbot's response lacks exact word matches, the accuracy score may decrease. In other words, the evaluation results may be influenced by how closely the chatbot's response aligns with the gold-standard answer. Since this study considered semantic similarity as an accuracy indicator

rather than exact word matching, this methodological limitation was taken into account. Future research should consider adjusting the weighting of BLEU and ROUGE scores while incorporating semantic similarity-based evaluation methods for a more balanced assessment. The third factor is the limitation of the LLM model's training data. A chatbot's response quality improves when it has access to high-quality, comprehensive information from the Q&A PDFs and webpages. However, the Q&A- and webpage-based data available in this study were relatively limited. Expanding the Q&A dataset with a larger and more precise collection of reference materials could significantly enhance accuracy. Additionally, it is essential to consider model uncertainties and potential inaccuracies in the responses, which should be addressed in future studies.

This high similarity score suggests that the LLM chatbot generates reliable responses for radiation therapy education. However, LLM-generated responses are not always accurate. Since LLMs generate the most probable response based on their training data, they are susceptible to the hallucination problem, where the model produces factually incorrect or misleading information. Therefore, it is essential to explore methods for enhancing the reliability of chatbot responses. One proposed solution is the retrieval-augmented generation technique. This approach enables the LLM to retrieve relevant information from a trusted external database before generating a response. By integrating retrieval-augmented generation, chatbots can provide more accurate and trustworthy information, thereby reducing the risk of hallucinations.

4.3. Study Limitations

There are several limitations in this study. First, user group testing was not conducted, which limits the evaluation of user satisfaction, usability, and actual learning outcomes. Second, the system-level performance was assessed using only ten questions, which may not sufficiently capture the chatbot's ability to handle a wider range of clinical or educational scenarios. Third, the evaluation relied on a fixed set of expert-generated reference answers, and inter-rater variability in defining gold-standard responses was not assessed. Lastly, potential biases in the embedded data, stemming from the selection of source materials, may have influenced the chatbot's responses. These limitations should be addressed in future studies to enhance the robustness and generalizability of the findings.

5. Conclusions

This study developed two types of LLM chatbots for radiation therapy education using the EmbedChain framework, one utilizing instructor-generated Q&A PDF datasets and the other using web-based radiation therapy information. The developed chatbots demonstrated reliable response capabilities, with both achieving a high semantic similarity score of 0.928 (92.8%) and a perfect consistency score of 1.0. In particular, the integration of dual data sources highlighted the versatility of the EmbedChain framework in handling diverse educational materials, facilitating a robust comparative analysis.

However, further enhancements are needed to improve response accuracy and optimize educational effectiveness. Future research will involve comprehensive evaluations with actual learners, expansion of radiation therapy-specific datasets, reinforcement of response consistency, and adoption of open-source LLM models to ensure long-term cost efficiency and practical applicability in clinical education. In addition, feedback from various educators and students will help refine the chatbot for use as an effective educational tool.

Author Contributions: Conceptualization, J.-H.J. and D.K.; methodology, J.-H.J. and Y.L.; software, J.-H.J.; validation, K.-B.L. and Y.L.; formal analysis, J.-H.J. and K.-B.L.; investigation, D.K. and K.-B.L.; writing—original draft preparation, J.-H.J.; writing—review and editing, D.K., K.-B.L. and Y.L.; project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xin, Q.; Na, Q. Enhancing inference accuracy of Llama LLM using reversely computed dynamic temporary weights. *TechRxiv* **2024**. [CrossRef]
2. Hadi, M.U.; Al-Tashi, Q.; Qureshi, R.; Shah, A.; Muneer, A.; Irfan, M.; Zafar, A.; Shaikh, M.B.; Akhtar, N.; Wu, J.; et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv* **2024**. [CrossRef]
3. Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PaLM-E: An embodied multimodal language model. In Proceedings of the ICML'23: Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023. [CrossRef]
4. Patnaik, S.S.; Hoffmann, U. Comparison of ChatGPT vs. bard to Anesthesia-related Queries. *medRxiv* **2023**. [CrossRef]
5. Koroteev, M.V. BERT: A review of applications in natural language processing and understanding. *arXiv* **2021**, arXiv:2103.11943. [CrossRef]
6. Rajaraman, N.; Jiao, J.; Ramchandran, K. An analysis of tokenization: Transformers under markov data. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 9–15 December 2024.
7. Chow, J.C.L.; Sanders, L.; Li, K. Design of an educational chatbot using artificial intelligence in radiotherapy. *AI* **2023**, *4*, 319–332. [CrossRef]
8. Chow, J.C.L.; Wong, V.; Sanders, L.; Li, K. Developing an AI-Assisted Educational Chatbot for Radiotherapy Using the IBM Watson Assistant Platform. *Healthcare* **2023**, *11*, 2417. [CrossRef] [PubMed]
9. Chow, J.C.L.; Li, K. Developing Effective Frameworks for Large Language Model-Based Medical Chatbots: Insights From Radiotherapy Education With ChatGPT. *JMIR Cancer* **2025**, *11*, e66633. [CrossRef] [PubMed]
10. Eickhoff, T.; Hakoua, A.N.; Gobel, J.C. AI-assisted engineering data integration for small and medium-sized enterprises. In Proceedings of the NordDesign, Reykjavik, Iceland, 12–14 August 2024. [CrossRef]
11. Kaplan, A.; Sayan, I.U.; Sahan, H.; Begen, E.; Bayrak, A.T. Response performance evaluations of ChatGPT models on large language model frameworks. In Proceedings of the 32nd Signal Processing and Communications Applications Conference (SIU), Mersin, Turkiye, 15–18 May 2024. [CrossRef]
12. Deepchecks. EmbedChain Modeling. Available online: https://www.deepchecks.com/llm-tools/embedchain/ (accessed on 1 May 2025).
13. Saxena, R.R. Beyond flashcards: Designing an intelligent assistant for USMLE mastery and virtual tutoring in medical education (A study on harnessing chatbot technology for personalized Step 1 prep). *arXiv* **2024**, arXiv:2409.10540. [CrossRef]
14. Ferreira, R.; Canesche, M.; Jamieson, P.; Vilela Neto, O.P.; Nacif, J.A.M. Examples and tutorials on using Google Colab and Gradio to create online interactive student-learning modules. *Comput. Appl. Eng. Educ.* **2024**, *32*, e22729. [CrossRef]
15. Jetbrains. PyCharm Version 2024.1 [Computer Software]. Available online: https://www.jetbrains.com/pycharm (accessed on 1 May 2025).
16. Jung, J.H.; Lee, K.B. Research of BERT-Based Q&A System for Radiation Therapy Education. *J. Radiol. Sci. Technology* **2025**, *48*, 171–178. [CrossRef]
17. Podgorsak, E.B. (Ed.) *Radiation Oncology Physics: A Handbook for Teachers and Students*; International Atomic Energy Agency: Vienna, Austria, 2005.
18. International Commission on Radiation Units and Measurements. *Prescribing, Recording, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT)*; ICRU Report 83; Oxford University Press: Oxford, UK, 2010.
19. Huq, M.S.; Fraass, B.A.; Dunscombe, P.B.; Gibbons, J.P., Jr.; Ibbott, G.S.; Mundt, A.J.; Mutic, S.; Palta, J.R.; Rath, F.; Thomadsen, B.R.; et al. The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. *Med. Phys.* **2016**, *43*, 4209–44262. [CrossRef] [PubMed]

20. Siochi, R.A.; Peter, B.; Charles, D.B.; Santanam, L.; Blodgett, K.; Curran, B.H.; Engelsman, M.; Feng, W.; Mechalakos, J.; Pavord, D.; et al. A rapid communication from AAPM Task Group 201: Recommendations for the QA of external beam radiotherapy data transfer. AAPM TG 201: Quality assurance of external beam radiotherapy data transfer. *J. Appl. Clin. Med. Phys.* **2011**, *12*, 170–181. [CrossRef] [PubMed]
21. Chen, Y.; Belouadi, J.; Eger, S. Reproducibility issues for BERT-based evaluation metrics. *arXiv* **2022**, arXiv:2204.00004. [CrossRef]
22. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. ACL 2002. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002. [CrossRef]
23. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain; 2024. Available online: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/07/was2004.pdf (accessed on 1 May 2025).
24. Sharma, K.V.; Ayiluri, P.R.; Betala, R.; Kumar, P.J.; Reddy, K.S. Enhancing query relevance: Leveraging SBERT and cosine similarity for optimal information retrieval. *Int. J. Speech Technol.* **2024**, *27*, 753–763. [CrossRef]
25. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating text generation with BERT. *arXiv* **2020**, arXiv:1904.09675. [CrossRef]
26. Ye, Y.; Simpson, E.; Rodriguez, R.S. Using similarity to evaluate factual consistency in summaries. *arXiv* **2024**, arXiv:2409.15090. [CrossRef]
27. Zheng, X.; Wu, H. Autoregressive linguistic steganography based on BERT and consistency coding. *Secur. Commun. Netw.* **2022**, *2022*, 9092785. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.