Clinical and Translational Radiation Oncology 51 (2025) 100914

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

journal homepage: www.sciencedirect.com/journal/clinical-and-translational-radiation-oncology

Large language models for pretreatment education in pediatric radiation oncology: A comparative evaluation study

Dominik Wawrzuta [a,*], Aleksandra Napieralska [b,c,d], Katarzyna Ludwikowska [a], Laimonas Jaruševičius [e], Anastasija Trofimoviča-Krasnorucka [f,g], Gints Rausis [f], Agata Szulc [h], Katarzyna Pędziwiatr [a], Kateřina Poláchová [i,j], Justyna Klejdysz [k,l], Marzanna Chojnacka [a]

[a] Department of Radiation Oncology, Maria Sklodowska-Curie National Research Institute of Oncology, Wawelska 15B, 02-034 Warsaw, Poland
[b] Radiotherapy Department, Maria Sklodowska-Curie National Research Institute of Oncology, Wybrzeże Armii Krajowej 15, 44-100 Gliwice, Poland
[c] Department of Oncology, Maria Sklodowska-Curie National Research Institute of Oncology, Garncarska 11, 31-115 Cracow, Poland
[d] Faculty of Medicine & Health Sciences, Andrzej Frycz Modrzewski Krakow University, Gustawa Herlinga-Grudzińskiego 1, 30-705 Cracow, Poland
[e] Oncology Institute, Lithuanian University of Health Sciences, A. Mickevičiaus g. 9, LT-44307, Kaunas, Lithuania
[f] Department of Radiation Oncology, Riga East University Hospital, Hipokrāta iela 2, LV-1038 Riga, Latvia
[g] Department of Internal Diseases, Riga Stradiņš University, Dzirciema iela 16, LV-1007 Riga, Latvia
[h] Department of Radiation Oncology, Lower Silesian Center of Oncology, Pulmonology and Hematology, Hirszfelda 12, 53-413 Wroclaw, Poland
[i] Department of Radiation Oncology, Masaryk Memorial Cancer Institute, Žlutý kopec 7, 656 53 Brno, Czech Republic
[j] Department of Radiation Oncology, Faculty of Medicine, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[k] Department of Economics, Ludwig Maximilian University of Munich (LMU), Geschwister-Scholl-Platz 1, 80539 Munich, Germany
[l] ifo Institute, Poschinger Straße 5, 81679 Munich, Germany

A B S T R A C T

Background and purpose: Pediatric radiotherapy patients and their parents are usually aware of their need for radiotherapy early on, but they meet with a radiation oncologist later in their treatment. Consequently, they search for information online, often encountering unreliable sources. Large language models (LLMs) have the potential to serve as an educational pretreatment tool, providing reliable answers to their questions. We aimed to evaluate the responses provided by generative pre-trained transformers (GPT), the most popular subgroup of LLMs, to questions about pediatric radiation oncology.

Materials and methods: We collected pretreatment questions regarding radiotherapy from patients and parents. Responses were generated using GPT-3.5, GPT-4, and fine-tuned GPT-3.5, with fine-tuning based on pediatric radiotherapy guides from various institutions. Additionally, a radiation oncologist prepared answers to these questions. Finally, a multi-institutional group of nine pediatric radiotherapy experts conducted a blind review of responses, assessing reliability, concision, and comprehensibility.

Results: The radiation oncologist and GPT-4 provided the highest-quality responses, though GPT-4′s answers were often excessively verbose. While fine-tuned GPT-3.5 generally outperformed basic GPT-3.5, it often provided overly simplistic answers. Inadequate responses were rare, occurring in 4% of GPT-generated responses across all models, primarily due to GPT-3.5 generating excessively long responses.

Conclusions: LLMs can be valuable tools for educating patients and their families before treatment in pediatric radiation oncology. Among them, only GPT-4 provides information of a quality comparable to that of a radiation oncologist, although it still occasionally generates poor-quality responses. GPT-3.5 models should be used cautiously, as they are more likely to produce inadequate answers to patient questions.

Background

Approximately 33 % of pediatric cancer patients require radiation therapy as part of their treatment [1]. Although these patients and their parents are generally informed about the need for radiotherapy early on, they typically consult a radiation oncologist at a later stage of their treatment [2]. However, early access to detailed and reliable information about radiation therapy is an important aspect for them [3]. Due to limited access to credible educational resources before starting radiotherapy, patients often turn to the Internet for information about the treatment process, its effectiveness, and potential toxicity. Unfortunately, the reliability of this information is often questionable [4]. Inaccurate information about radiotherapy can lead to misunderstandings about its role and toxicity, resulting in increased anxiety

* Corresponding author.
E-mail address: dominik.wawrzuta@nio.gov.pl (D. Wawrzuta).

https://doi.org/10.1016/j.ctro.2025.100914
Received 17 December 2024; Accepted 5 January 2025
Available online 6 January 2025

and a desire to avoid treatment [5,6].

The rapid advancement of large language models (LLMs) opens new opportunities for patient education. The widespread availability and extensive information within these models can be a valuable resource for patients [7]. Initial investigations of the widely used LLM model, the generative pre-trained transformer (GPT) developed by OpenAI, demonstrated its utility as a chatbot to answer patient questions [8,9]. An avenue to improve the efficacy of LLMs in medical applications is to fine-tune them with high-quality medical data sources [10]. To date, no research has specifically examined GPT responses within the context of pediatric radiation oncology. Existing studies have focused on using GPT models in adult radiotherapy, but they have several limitations. They did not analyze original patient questions, and did not compare GPT responses to those of human experts, or explore the potential of fine-tuning models. Moreover, the results of these studies have been inconsistent; one study reported that only 6 % of the GPT-generated responses were inaccurate, while another found that 34 % contained inaccuracies, highlighting the significant risk of misinformation [11,12].

Aim of study

This study aimed to evaluate the utility of large language models (LLMs) as a tool for pretreatment education on radiation oncology for pediatric cancer patients and their families. We compared the responses generated by the GPT-3.5, GPT-4, and fine-tuned GPT-3.5 models with those provided by experienced pediatric radiation oncologists in terms of reliability, concision, and comprehensibility.

Materials and methods

Patient surveys

Between June 1, 2023, and December 31, 2023, we collected questions related to radiation oncology from pediatric radiotherapy patients and their parents. The inclusion criteria required participants to be ten years and older. We conducted anonymous and voluntary surveys with open-ended questions for all eligible patients and parents visiting the Department of Radiation Oncology at the Maria Sklodowska-Curie National Research Institute of Oncology in Warsaw during this period. Specifically, we asked participants to write down any questions they had before their initial visit to the radiation oncology department. These questions were then used to evaluate the quality of the responses provided by the GPT-3.5, fine-tuned GPT-3.5, and GPT-4 models. Fig. 1 illustrates the methodological steps followed in our study.

Fine-tuning procedure

To improve the performance of the GPT-3.5 model in the context of pediatric radiotherapy, we performed a systematic fine-tuning process. Initially, we identified high-quality educational guides for pediatric radiation therapy from various institutions (detailed list available in the Supplementary). Two researchers (DW and KL) searched for materials available in nine languages: English, German, French, Spanish, Portuguese, Italian, Polish, Dutch, and Romanian. We included only materials presented in a question-and-answer (Q&A) format to maintain consistency with the intended application. All non-English content was translated to ensure accurate medical terminology. Subsequently, we thoroughly reviewed all identified materials and excluded any incorrect content. The resulting dataset of Q&A pairs served as the basis for fine-tuning the GPT-3.5 model. The fine-tuning process was executed using the official OpenAI API [13]. For each Q&A pair, we structured the input with a system content parameter stating, “You are a radiation oncologist with medical expertise. You are responding to patient questions about pediatric radiotherapy”, the dataset question as user content, and the corresponding answer as assistant content. The fine-tuning was performed using default hyperparameters, with three training epochs, creating a specialized model accessible through the OpenAI API.

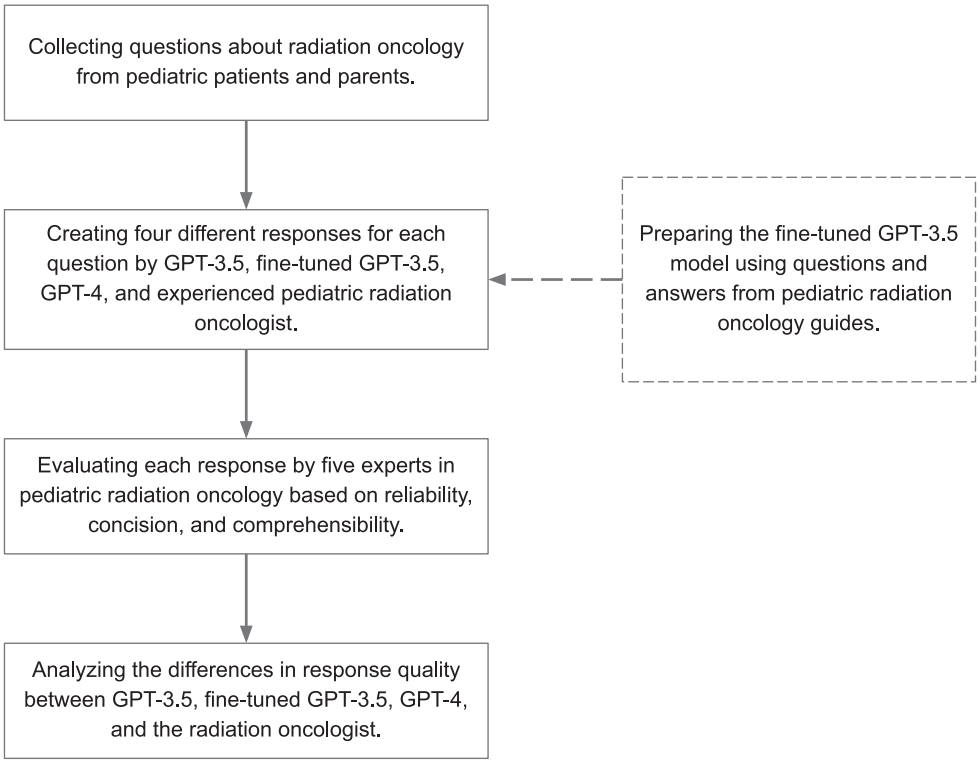

Fig. 1. Flowchart of the methodological steps.

Generating responses

We used prompt engineering [14] to instruct the GPT-3.5, GPT-4 and fine-tuned GPT-3.5 models to produce responses to patient inquiries related to pediatric radiation therapy. The following prompt was used: "You are a radiation oncologist with medical expertise. You are responding to patient questions about pediatric radiotherapy. Your responses are comprehensive and tailored to the patient's understanding.". Each patient question was presented to the models using zero-shot prompting, mirroring the typical interaction between patients and LLMs in real-world scenarios, where users ask direct questions without providing additional context or examples.

In addition, an experienced pediatric radiation oncologist (DW) authored an independent response to each question for comparative analysis. Each response from the radiation oncologist was then refined using the GPT-4 model with the prompt: "Refine the response to the patient's question about pediatric radiotherapy. Do not add new information or delete information; improve the structure of the response.". This step was designed to make expert responses stylistically more consistent with GPT-based responses and prevent easy differentiation during quality assessment. Following this process, the author validated all responses to ensure that the GPT-4 model only altered the overall appearance of the responses without modifying the information provided.

Evaluation process

Nine pediatric radiation oncology experts independently evaluated the quality of responses to patient questions. Each question was paired with four different responses: one from GPT-3.5, one from GPT-4, one from fine-tuned GPT-3.5, and one prepared by an experienced radiation oncologist. Experts evaluated blinded responses in three dimensions: reliability, concision, and comprehensibility. Each response received five independent ratings from different experts. We calculated Gwet's AC2 to estimate the inter-rater agreement [15]. The labeling process was carried out using the Label Studio platform [16].

Reliability was defined as the scientific quality of the responses, focusing on the accuracy and credibility of the content. This category ensures that the information is based on medical evidence and consistent with current best practices in healthcare. Experts rated reliability using a 5-point Likert scale:

1 – Very Poor (potentially harmful).
2 – Poor (incorrect but harmless).
3 – Acceptable (minor errors or significant information gaps).
4 – Good (correct but with some deficiencies).
5 – Very Good (entirely correct).

Concision was assessed based on the brevity of the responses, evaluating whether the information was presented clearly and without unnecessary elaboration. This category determines whether responses effectively convey essential information in a straightforward manner, addressing the needs of patients. Experts rated concision on a 3-point Likert scale:

1 – Poor (too lengthy, with unnecessary information or repetitions).
2 – Acceptable (contains some unnecessary information).
3 – Good (appropriate length).

Comprehensibility examined whether the patients could understand the language and terminology used in the responses. This category evaluated how well the information was presented in an accessible manner, avoiding medical jargon and complex explanations. Experts rated comprehensibility using a 3-point Likert scale:

1 – Poor (not suitable for patients).
2 – Acceptable (some minor issues).
3 – Good (well-adjusted for patients).

Finally, we created a composite score as a fourth dimension, calculated as the average of the reliability, conciseness, and comprehensibility scores, each standardized on a scale from 0 to 1.

For each rated response, we calculated the mean score in each dimension from the five experts' ratings. To compare the quality differences among the different models across dimensions, we used the Kruskal-Wallis test, a non-parametric method, to determine if the samples originated from the same distribution [17]. Additionally, we used the Wilcoxon signed-rank test with Bonferroni correction to perform pairwise comparisons of the median ratings between different sources of answers [18,19].

Results

We collected 40 surveys, including 12 from pediatric patients and 28 from parents. These surveys contained 80 different questions covering ten different topics, as described in Table 1. Respondents provided between one and eight questions spanning various categories. The most frequently addressed topics were toxicity, the impact of radiation therapy on daily life, skincare, effectiveness of irradiation, and the treatment course. Less frequently, respondents sought information on supportive care during treatment, the risk of hair loss, the duration of treatment, pain associated with irradiation, and dietary considerations.

For the fine-tuning of GPT-3.5, we used data from 18 sources, detailed in the Supplementary Materials. Table 2 outlines the distribution of Q&A guides in different languages: five in English, three in French, two in German, Spanish and Dutch, and one in Portuguese, Italian, Polish, and Romanian. In total, we collected 145 pairs of questions and the corresponding answers on pediatric radiotherapy. The majority were formulated in English (54 %), followed by German (12 %) and French (9 %).

Nine pediatric radiation oncology experts from six institutions (three from Warsaw, Poland; two from Riga, Latvia; one from Brno, Czech Republic; one from Gliwice, Poland; one from Kaunas, Lithuania; and one from Wroclaw, Poland) evaluated 320 responses to 80 patient questions. Each response was reviewed by five experts in three dimensions (reliability, concision, and comprehensibility), resulting in a total of 4,800 labels. The inter-rater agreement was 0.66 (95 % CI 0.63–0.69) for reliability, 0.68 (95 % CI 0.64–0.72) for concision, and 0.77 (95 % CI 0.74–0.80) for comprehensibility.

According to the composite score (Fig. 2A), the radiation oncologist (RO) produced the highest quality responses, with a median score of 0.90 (Q1 0.84, Q3 0.94). The GPT-4 model followed closely, with a median score of 0.86 (Q1 0.76, Q3 0.93). Although RO responses were rated higher than those of GPT-4, the difference was not statistically significant ($p = 0.26$). The fine-tuned GPT-3.5 model achieved a median score of 0.81 (Q1 0.70, Q3 0.90), outperforming the baseline GPT-3.5 model, which had a median score of 0.74 (Q1 0.60, Q3 0.85) ($p =$

Table 1
Categories of questions.

Question category	Frequency in children	Frequency in parents
Toxicity	6 (27 %)	17 (29 %)
Everyday life impact	5 (23 %)	4 (7 %)
Skincare	2 (9 %)	7 (12 %)
Efficacy of treatment	1 (5 %)	8 (14 %)
Treatment course description	4 (18 %)	4 (7 %)
Supportive care	0 (0 %)	7 (12 %)
Hair loss	1 (5 %)	4 (7 %)
Duration of treatment	2 (9 %)	2 (3 %)
Pain	1 (5 %)	3 (5 %)
Diet	0 (0 %)	2 (3 %)
Sum	**22 (100 %)**	**58 (100 %)**

Table 2
Data sources for fine-tuning.

Language	Number of sources	Number of questions
English	5 (28 %)	78 (54 %)
French	3 (17 %)	13 (9 %)
German	2 (11 %)	17 (12 %)
Spanish	2 (11 %)	10 (7 %)
Dutch	2 (11 %)	9 (6 %)
Portuguese	1 (6 %)	4 (3 %)
Italian	1 (6 %)	2 (1 %)
Polish	1 (6 %)	6 (4 %)
Romanian	1 (6 %)	6 (4 %)
Sum	**18 (100 %)**	**145 (100 %)**

0.04). While the fine-tuned GPT-3.5 model was significantly inferior to RO responses ($p < 0.001$), it was not statistically different from GPT-4 responses ($p = 0.47$).

In the reliability dimension (Fig. 2B), the fine-tuned GPT-3.5 model produced the weakest responses, with a median score of 4.20 (Q1 4.00, Q3 4.60). On the contrary, the RO (median 4.50, Q1 4.20, Q3 4.60), GPT-4 (median 4.60, Q1 4.40, Q3 4.80), and GPT-3.5 (median 4.50, Q1 4.20, Q3 4.80) provided similar reliability scores, without significant differences between any pair (all $p > 0.76$).

For the concision dimension (Fig. 2C), RO responses were the most concise, with a median score of 3.00 (Q1 2.80, Q3 3.00). GPT models tended to produce more extended responses, with GPT-3.5 showing the lowest concision (median 2.40, Q1 2.20, Q3 2.60) performing worse than fine-tuned GPT-3.5 (median 2.80, Q1 2.60, Q3 3.00) and GPT-4 (median 2.80, Q1 2.60, Q3 2.80) (both $p < 0.0001$).

In the dimension of comprehensibility (Fig. 2D), the GPT-3.5 model had the lowest performance, with a median score of 2.80 (Q1 2.60, Q3 3.00). This was significantly lower than the scores for RO (median 2.80, Q1 2.80, Q3 3.00, $p = 0.03$) and GPT-4 (median 2.80, Q1 2.80, Q3 3.00, $p = 0.05$) and similar to fine-tuned GPT-3.5 (median 2.80, Q1 2.60, Q3 3.00, $p = 1.00$).

A subanalysis of inadequate responses (defined as a mean rating of less than 2 for concision and comprehensibility or less than 3 for reliability) identified ten poor-quality responses, accounting for 4 % of all GPT-generated responses across all models. Each inadequate response received low rating in only one dimension. In particular, none of the responses from the radiation oncologist was rated as poor. Out of the ten inadequate responses, seven were rated poorly for concision, six from GPT-3.5, and one from fine-tuned GPT-3.5. Two responses were rated poorly for reliability, one from the fine-tuned GPT-3.5 and one from GPT-4. One response was rated poorly for comprehensibility, generated by the fine-tuned GPT-3.5.

Discussion

Data sources

The original patient questions were used as a data source to evaluate the GPT models. Previous studies mainly analyzed questions prepared by experts, which may have had different structures and vocabulary compared to those used by patients [11,12,20]. This is particularly important considering the technical nature of radiation oncology, where patients often have questions not only about clinical aspects but also related to physics, radiobiology, and medical engineering, formulating them non-professionally. Taking into account the patient's perspective is crucial, as their emotions and behaviors can influence the form and content of their questions [21].

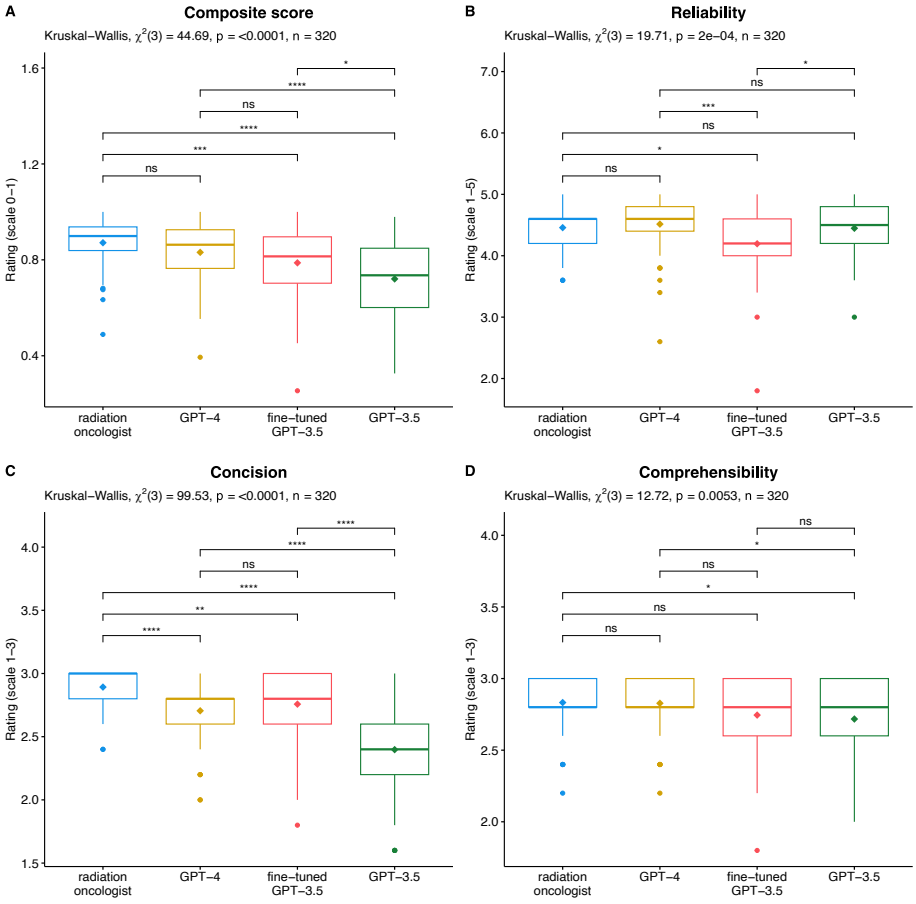

Fig. 2. Summary of the ratings for the responses of the GPT models (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$).

Analyzed models

Our study evaluated the quality of responses from the most popular LLMs, the GPT models developed by OpenAI. We selected these models due to their widespread use among the general public and patients [22,23]. Specifically, we examined the performance of GPT-3.5, the most accessible free option for patients; GPT-4, a more advanced model requiring a subscription; and the fine-tuned GPT-3.5, which allows for the creation of custom models tailored to specific applications using one's own data and theoretically best suited for specific tasks [10]. Despite the potential benefits of fine-tuning in medical education, the performance of such models has only been analyzed in a few medical studies and never in the context of radiation oncology.

Previous analyses of GPT applications in radiation therapy have focused primarily on evaluating AI responses without comparing them with human experts [11,20,24]. We decided to compare the quality of the GPT model responses with those prepared by an experienced radiation oncologist, as these responses can serve as a reference point due to their expected higher quality [25]. Additionally, to prevent GPT answers from being easily distinguishable from human responses due to stylistic differences, we refined the human responses using the GPT-4 model to give them a GPT-style. This process did not impact any of the dimensions analyzed, as the original author reviewed all refined responses to ensure that only the style was slightly altered.

Quality of models

We assembled a multinational panel of experts to evaluate the quality of responses generated by GPT models and a human expert. This diverse panel was designed to minimize bias from the experiences and subjective opinions of a single institution or cultural background [26]. Five different experts evaluated each response, reducing the impact of outlier ratings. Despite this diversity, the inter-rater agreement was high and consistent across the dimensions analyzed, demonstrating the potential to generate objectively correct answers to patient questions.

The best responses to the patients' and their parents' questions were provided by the radiation oncologist and the GPT-4 model. The only shortcoming of GPT-4 was in the dimension of concision, as it often generated responses that were too long, which is a common characteristic of LLMs [27]. In general, GPT-4 outperformed both fine-tuned GPT-3.5 and GPT-3.5, due to its larger model size and enhanced training data [28]. Although fine-tuned GPT-3.5 is expected to outperform GPT-3.5 in all dimensions, GPT-3.5 was superior in reliability [10]. This discrepancy arose from the content structure of the training guides, which often provide brief and concise answers that are correct but may omit some information. It is worth noting that, even when the GPT responses were generally of good quality, they occasionally generated inadequate answers. It happened even in the best-performing GPT-4, showing that despite its generally high-quality answers, comparable to a physician's, potentially harmful outlier content can still occur.

Research in context

The application of LLMs in radiation oncology remains relatively unexplored, with only a handful of published studies. While there is extensive research on LLMs in general medicine, the highly technical nature of radiation oncology presents unique challenges that require specific evaluation. Current research in patient education consists of a small-scale analysis examining responses to radiosurgery questions [9] and two broader investigations of general radiation therapy inquiries [11,12]. However, these studies differ significantly from our approach. First, they analyzed pre-formulated questions rather than authentic patient inquiries, potentially missing the nuanced language and concerns that characterize real patient communication. Second, they evaluated single LLM models without exploring the potential of fine-tuned models specifically adapted for radiation oncology applications.

Yalamanchili et al. [12] similarly employed expert responses as a "gold standard" for comparison, but their methodology differed in a crucial aspect. Their expert answers were sourced from official Q&A websites, which could have made them easily distinguishable from LLM-generated responses. In contrast, our study refined expert responses using GPT-4 to maintain stylistic consistency while preserving the original medical content, thus minimizing potential evaluation bias. Beyond patient education, LLMs have demonstrated utility in various aspects of radiation oncology, including patient symptom summarization [24], media content classification [29], supporting with scientific tasks [30], medical questions responses in a professional context [20,31,32], and insurance documentation preparation [33].

Limitations and future directions

Our study focused solely on the GPT family of models because they are considered the best for general applications and are most accessible to average Internet users. We excluded specific LLMs designed for medical applications, such as Med-PaLM [34] or Clinical Camel [35]. Despite their theoretically high-quality responses to general medical queries, these models are currently not widely available to patients.

The two most commonly used methods for creating personalized LLMs are fine-tuning and retrieval-augmented generation (RAG) [36]. We used fine-tuning instead of RAG because RAG's effectiveness heavily depends on the similarity between queries and stored content, which could be particularly challenging given patient questions' varied and colloquial nature. Fine-tuning, although limited by the static nature of the training data, provides a more streamlined and computationally efficient solution. Future research should compare the performance of fine-tuned models against RAG-based systems to determine the optimal approach for handling patient inquiries in radiation oncology. It would also be worthwhile to investigate the performance of fine-tuned GPT-4 models, as this option was not available at the time of the study but was released to the public in July 2024 [37].

Conclusions

GPT models can be a valuable tool to educate patients about pediatric radiation oncology before starting treatment. In our study, only the most advanced model, GPT-4, responded to patient questions with a quality comparable to that of a radiation oncologist. However, even GPT-4 occasionally generated poor responses. GPT-3.5-based models should be used with caution, as their responses are generally inferior to those of radiation oncologists. Although GPT models can be beneficial in educating pediatric cancer patients and their parents, especially when access to medical experts is limited, it is crucial to recognize their potential to produce low-quality responses.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2025.100914.

References

[1] Zaghloul M. Pediatric radiotherapy utilization rate (PUR) in a large pediatric oncology center in Low-/Middle-Income country (LIMC): Report on 14473 patients. Int J Radiat Oncol 2022;114(5):1067–8. https://doi.org/10.1016/j.ijrobp.2022.09.024.

[2] Shen CJ, Terezakis SA. The evolving role of radiotherapy for pediatric cancers with advancements in molecular tumor characterization and targeted therapies. Front Oncol 2021;16(11):679701. https://doi.org/10.3389/fonc.2021.679701.
[3] Ångström-Brännström C, Engvall G, Mullaney T, Nilsson K, Wickart-Johansson G, Svärd A-M, Nyholm T, Lindh J, Lindh V. Children undergoing radiotherapy: Swedish parents' experiences and suggestions for improvement. PLOS ONE 2015; 10(10). https://doi.org/10.1371/journal.pone.0141086. Glod JW, editor.
[4] Ogasawara R, Katsumata N, Toyooka T, Akaishi Y, Yokoyama T, Kadokura G. Reliability of cancer treatment information on the internet: observational study. JMIR Cancer 2018;4(2):e10031. PMID:30559090.
[5] Fuji H, Fujibuchi T, Tanaka H, Ogawa Y, Noda C, Hayakawa M, et al. Changes in satisfaction and anxiety about radiotherapy for pediatric cancer by two-step audio-visual instruction. Tech Innov Patient Support Radiat Oncol 2023;27:100214. https://doi.org/10.1016/j.tipsro.2023.100214.
[6] Lazard AJ, Nicolla S, Vereen RN, Pendleton S, Charlot M, Tan H-J, et al. Exposure and reactions to cancer treatment misinformation and advice: survey study. JMIR Cancer 2023;28(9):e43749. https://doi.org/10.2196/43749.
[7] Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med 2023;3(1): 141. https://doi.org/10.1038/s43856-023-00370-1.
[8] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589. https://doi.org/10.1001/jamainternmed.2023.1838.
[9] Dayawansa S, Mantziaris G, Sheehan J. Chat GPT versus human touch in stereotactic radiosurgery. J Neurooncol 2023;163(2):481–3. https://doi.org/10.1007/s11060-023-04353-z.
[10] Yang Q, Wang R, Chen J, Su R, Tan T. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise. 10.48550/ARXIV.2407.11536 arXiv 2024.
[11] Floyd W, Kleber T, Pasli M, Qazi JJ, Huang CC, Leng JX, et al. Evaluating the reliability of Chat-GPT model responses for radiation oncology patient inquiries. Int J Radiat Oncol 2023;117(2):e383.
[12] Yalamanchili A, Sengupta B, Song J, Lim S, Thomas TO, Mittal BB, et al. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open 2024;7(4):e244630. https://doi.org/10.1001/jamanetworkopen.2024.4630.
[13] OpenAI. GPT. OpenAI; 2023.
[14] Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. Npj Digit Med Nat Publ Group 2024;7(1):1–9. https://doi.org/10.1038/s41746-024-01029-4.
[15] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(1):29–48. https://doi.org/10.1348/000711006X126600.
[16] Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. Label Studio: Data labeling software. 2020.
[17] McKight PE, Najab J. Kruskal-Wallis Test. In: Weiner IB, Craighead WE, editors. Corsini Encycl Psychol. 1st ed. Wiley; 2010. . p. 1–1. 10.1002/9780470479216.corpsy0491ISBN:978-0-470-17024-3.
[18] R.F. Woolson Wilcoxon signed-rank test. R.B. D'Agostino L. Sullivan J. Massaro editors. Wiley Encycl Clin Trials 1st ed Wiley; 2008. p. 1–3. 10.1002/9780471462422.eoct979ISBN:978-0-471-35203-7.
[19] Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014;34(5):502–8. https://doi.org/10.1111/opo.12131.
[20] Dennstädt F, Hastings J, Putora PM, Vu E, Fischer GF, Süveg K, et al. Exploring capabilities of large language models such as ChatGPT in radiation oncology. Adv Radiat Oncol 2024;9(3):101400. https://doi.org/10.1016/j.adro.2023.101400.
[21] Beisecker AE, Beisecker TD. Patient information-seeking behaviors when communicating with doctors. Med Care 1990;28(1):19.
[22] Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. Health Promot Perspect 2023;13(3):183–91. PMID:37808939.
[23] chat.openai.com Traffic Analytics, Ranking & Audience [June 2024]. Similarweb. Available from: https://www.similarweb.com/website/chat.openai.com/ [accessed Jul 25, 2024].
[24] Wu DJ, Bibault J-E. Pilot applications of GPT-4 in radiation oncology: Summarizing patient symptom intake and targeted chatbot applications. Radiother Oncol Elsevier 2024. Jan 1;190. PMID:37913954.
[25] Reis M, Reis F, Kunde W. Influence of believed AI involvement on the perception of digital medical advice. Nat Med 2024. https://doi.org/10.1038/s41591-024-03180-7.
[26] Dovidio JF, Glick P, Hewstone M. The SAGE handbook of prejudice, stereotyping and discrimination. Sage; 2010. p. 1–672.
[27] Saito K, Wachi A, Wataoka K, Akimoto Y. Verbosity Bias in Preference Labeling by Large Language Models. arXiv 2023;10.48550/arXiv:2310.10076.
[28] Chang EY. Examining gpt-4: Capabilities, implications and future directions. 2023.
[29] Wawrzuta D, Klejdysz J, Chojnacka M. The rise of negative portrayals of radiation oncology: A textual analysis of media news. Radiother Oncol 2024;190:110008. https://doi.org/10.1016/j.radonc.2023.110008.
[30] Guckenberger M, Andratschke N, Ahmadsei M, Christ SM, Heusel AE, Kamal S, et al. Potential of ChatGPT in facilitating research in radiation oncology? Radiother Oncol 2023;188:109894. https://doi.org/10.1016/j.radonc.2023.109894.
[31] Duggan R, Tsuruda KM. ChatGPT performance on radiation technologist and therapist entry to practice exams. J Med Imaging Radiat Sci 2024;55(4):101426. https://doi.org/10.1016/j.jmir.2024.04.019.
[32] Holmes J, Liu Z, Zhang L, Ding Y, Sio TT, McGee LA, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. Front Oncol 2023;17(13):1219326. https://doi.org/10.3389/fonc.2023.1219326.
[33] Kiser KJ, Waters M, Reckford J, Lundeberg C, Abraham CD. Large language models to help appeal denied radiotherapy services. JCO Clin Cancer Inform 2024;8: e2400129. https://doi.org/10.1200/CCI.24.00129.
[34] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P, Seneviratne M, Gamble P, Kelly C, Babiker A, Schärli N, Chowdhery A, Mansfield P, Demner-Fushman D, Agüera Y, Arcas B, Webster D, Corrado GS, Matias Y, Chou K, Gottweis J, Tomasev N, Liu Y, Rajkomar A, Barral J, Semturs C, Karthikesalingam A, Natarajan V. Large language models encode clinical knowledge. Nature 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2.
[35] Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical camel: an open expert-level medical language model with dialogue-based knowledge encoding. arXiv 2023. https://doi.org/10.48550/arXiv.2305.12031.
[36] Soudani H, Kanoulas E, Fine HF. Tuning vs. Retrieval augmented generation for less popular knowledge. In: Proc 2024 Annu Int ACM SIGIR Conf Res Dev Inf Retr Asia Pac Reg Tokyo Japan. ACM; 2024. p. 12–22. 10.1145/3673791.3698415.
[37] GPT-4o mini: advancing cost-efficient intelligence. Available from: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ [accessed Jul 25, 2024].