Cureus
Part of SPRINGER NATURE

Open Access Original Article

Published via Contemporary Updates: Radiotherapy Innovation & Evidence (CURiE)

Putting ChatGPT to the Test: Accuracy and Error Patterns in Clinical and Physics-Based Radiation Oncology Questions

Received 10/14/2026
Review began 11/07/2026
Review ended 11/15/2026
Published 11/18/2026

© Copyright 2026
Fahmy et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.119879

Veronia Fahmy [1], Jane Georgas [2], James S. Welsh [3, 2]

1. Radiation Oncology, Loyola University Medical Center, Maywood, USA 2. Radiation Oncology, Loyola University Chicago Stritch School of Medicine, Maywood, USA 3. Radiation Therapy, Edward Hines Jr VA Hospital, Hines, USA

Corresponding author: Veronia Fahmy, veronia.fahmy@luhs.org

Abstract

Objective

Large language models (LLMs) are increasingly being used as revision tools. Radiation oncology encompasses a complex skill set, including clinical staging, trials, anatomy, radiobiology, physics, mathematical calculations, treatment planning, and quality assurance (QA). The accuracy, reasonability, and error types of two ChatGPT models were assessed across clinical and medical physics radiation oncology examination question banks.

Methodology

Two independent scoring studies were undertaken (n = 200). Cohort 1 assessed the accuracy of ChatGPT 5.5 across 100 single-best-answer clinical questions from the American College of Radiology (ACR) Radiation Oncology In-Training Examination (TXIT). Questions were selected at random using a computer-generated random sampling script (25 questions per year from the years 2021, 2024, 2025, and 2026). Cohort 2 assessed ChatGPT 5.6 (Luna build, v5.6) across 100 RAPHEX-style radiation oncology physics multiple-choice questions covering five domains of physics. Each question was asked once using a standardized zero-shot prompt requesting a single best answer with reasoning. Accuracy was scored against published answer keys. Response length, error types, and exploratory scoring of reasoning quality (0-2) were noted. Chi-square (χ^2), Fisher's exact, Welch's t-test, and Mann-Whitney U tests were used to compare results (P = 0.05).

Results

ChatGPT 5.5 had an overall accuracy of 82.0% (82/100) (95% confidence interval (CI): 72.8%-89.0%). Accuracy was 100% for radiobiology, pediatrics, and renal/bladder oncology. Accuracy was lowest for gynecological oncology (female) at 60.0% and benign/palliative radiation oncology at 66.7%. Accuracy varied between years of examination (2021-2026) from 76.0% to 88.0% (χ^2 = 2.44, df = 3, P = 0.49). Incorrect responses had a greater mean length compared to correct responses (227.657.5 vs. 202.144.3) (Welch's t = 1.77, P = 0.09) (Mann-Whitney U = 551.5, P = 0.10). Error types were mainly guideline/trial/numerical-specific recall errors (5/18, 27.8%).

ChatGPT 5.6 had an overall accuracy of 87.0% (87/100) (95% CI: 78.7%-93.2%). Accuracy was highest for brachytherapy (15/15, 100%) and imaging/planning (24/25, 96.0%). Accuracy was lowest for beam physics (21/27, 77.8%). Mean length of reasoning was 186.848.2 words (correct = 188.4, incorrect = 176.2) (Student's t = 0.81, P = 0.42). Error types were most commonly conceptual errors (5/13, 38%) and self-contradictory responses (4/13, 31%).

Conclusions

Both ChatGPT 5.5 and ChatGPT 5.6 had high agreement with the answer keys. However, error types differed. Clinical errors were mainly due to detailed guideline/trial recall and anatomical staging errors. Physics errors were often due to conceptual errors and mathematical contradictions. LLMs can be useful as revision tools; however, responses must be validated against authoritative literature.

Categories: Medical Education, Radiation Oncology, Medical Physics
Keywords: artificial intelligence, chatgpt, large language models, medical education, medical physics, radiation oncology, raphex

Introduction

Large language models (LLMs) are rapidly being applied to medical education, research, and knowledge dissemination [1-3]. LLMs can be used to answer complex questions, provide differential diagnoses, and generate detailed responses to complex clinical questions [4,5]. Their performance varies depending on the

medical specialty, question type, and software version being used. Furthermore, fluency does not necessarily indicate accuracy or logical reasoning [6].

Radiation oncology is an important specialty to consider when evaluating the performance of generative artificial intelligence (AI), as it requires the integration of clinical knowledge, anatomy, and technical knowledge [7]. Clinical radiation oncology requires accurate knowledge of anatomy, tumor biology, tumor, node, and metastasis (TNM) staging, treatment selection, dose and fractionation, tolerance of normal tissues, and important clinical trials [7,8]. Radiation oncology also uses extensive medical imaging, including computed tomography (CT) and magnetic resonance imaging (MRI), to localize tumors and facilitate accurate contouring. This can be combined with AI to provide automated contouring, segmentation, and optimal treatment planning [7]. Radiation oncology physics requires additional knowledge of radiation interactions, dosimetry, treatment machine calibration, imaging, brachytherapy, quality assurance (QA), and complex calculations [7-10]. This makes radiation oncology a suitable specialty for assessing the accuracy of LLMs and understanding their limitations.

While there has been considerable focus on using AI in radiation oncology for automated contouring, segmentation, and optimal treatment planning [7], these applications use medical imaging and anatomy. Traditional LLMs, such as ChatGPT, use text-based information, including clinical scenarios, examination questions, and scientific literature [1,3]. They use the information provided in the prompt and apply previously learned clinical knowledge to generate an answer [1,3]. Newer multimodal models can also use visual information; however, their performance in interpreting medical imaging and specific radiotherapy data may vary [11,12]. It is also important to assess LLMs for knowledge and trainee education. Trainees and radiation oncologists are becoming increasingly involved in using AI tools for education [8,9]. Understanding their accuracy and limitations will help ensure their safe use in medical education.

Studies have shown that LLMs vary in their accuracy when answering general medical licensing examinations [4,5] and specialty board examinations [8,10]. ChatGPT 4 had an accuracy of ~73% when answering questions from the Japanese medical physicist examination, compared with ChatGPT 3.5, which had an overall accuracy of 42.2% [10]. Recently, Segura et al. assessed 6 LLMs on the 2024 RAPHEX Therapy Examination. Overall accuracies ranged from 47% (GPT-3.5-turbo and DeepSeek-R1) to 86% (Claude Sonnet 3.5). GPT-4o had an accuracy of 80%. The largest differences were observed in conceptual and radiation interaction questions [11]. Barbour and Barbour assessed ChatGPT 3.5 using a simulated oral board-style prostate cancer scenario [8]. They aimed to determine whether the model could demonstrate clinical knowledge and reasoning. Although the model could provide general diagnostic and treatment strategies, it had limitations in risk stratification and recommendations regarding the duration of hormone therapy [8].

Materials And Methods

Study design

A retrospective analysis of two cohorts of model responses to clinical and physics questions used during radiation oncology and medical physics board examination preparation was performed. Because radiation oncology board preparation encompasses two distinct domains - clinical oncology and radiation physics - two domain-specific question banks were selected (American College of Radiology (ACR) Radiation Oncology In-Training Examination (TXIT) for clinical knowledge and RAPHEX for medical physics). The study was designed to evaluate LLM performance within each educational domain independently rather than to conduct a head-to-head comparison between the two model versions [12]. The clinical and physics questions were analyzed separately to account for differences in model versions, question sources, subdomains, and analyses. A total of 200 question-analysis cycles were performed.

Clinical question selection (cohort 1)

The clinical question set consisted of 100 single-best-answer questions from the ACR TXIT. Permission was obtained to use the questions for research and publication. A random sampling script was generated using Python (v.3.11, Python Software Foundation, Wilmington, DE). A total of 25 unique questions were randomly selected from the 2021, 2024, 2025, and 2026 question sets. The questions were divided into 13 clinical subdomains: breast, prostate, gastrointestinal (GI), central nervous system (CNS), benign/palliative radiation oncology, female gynecological oncology, lung, head and neck, applied physics/other, radiobiology, lymphoma/hematological malignancies, pediatrics, and renal/bladder. The questions are used for formative revision during residency training and are not official American Board of Radiology (ABR) certifying board examinations. The model responses were scored against the published answer keys.

Physics question selection (cohort 2)

The physics question set consisted of 100 multiple-choice questions that mirrored the RAPHEX examination. The questions were systematically sampled across the five physics subdomains: beam physics, dosimetry and units, brachytherapy, imaging and treatment planning, and safety and QA. Questions requiring image interpretation without sufficient context were excluded. Institutional permission was obtained for use of the questionnaire.

Model analysis and prompts

Each question was entered individually into the model. Because radiation oncology board preparation encompasses two distinct domains - clinical oncology and radiation physics - two domain-specific question banks were selected (ACR TXIT for clinical knowledge and RAPHEX for medical physics). The study was designed to evaluate LLM performance within each educational domain independently rather than to conduct a head-to-head comparison between the two model versions. ChatGPT 5.5 was used for the clinical questions and ChatGPT 5.6 (Luna build v.5.6) for the physics questions, reflecting the default flagship model available on the standard web platform at the time each respective cohort was queried. Both models were accessed using a standard web platform [12]. A standard zero-shot prompt was used: "Please answer the following multiple-choice question designed for radiation oncology board preparation and provide the single best answer choice followed by a comprehensive explanation justifying your selection." Each question was queried only once. No additional clinical context, multi-prompting, or resampling occurred.

Outcome measures and quality scoring

The primary outcome measure was agreement with the published answer. Overall accuracy and accuracy per subdomain were calculated with 95% confidence intervals (CIs). Secondary outcomes included the number of words in explanations, accuracy per subdomain, and analysis of disagreements. An exploratory scoring system was used to assess the quality of model explanations (scored from 0-2):

Clinical rationale: Scored based on correct answers. 0 = Poor/no rationale. 1 = Partially coherent rationale. 2 = Structured and clinically appropriate rationale.

Physics rationale: Scored based on incorrect answers. 2 = Coherent and structured rationale. 1 = Partially coherent rationale with confusion. 0 = Large inconsistencies in rationale.

Disagreement

Disagreement was analyzed by one reviewer. Clinical disagreements were split into the following categories: guideline-specific, trial-specific, numerical-specific, ambiguous answer key, anatomical/staging-specific, image-specific, self-contradictory, criteria confusion, and multi-step quantitative/physics. Physics disagreements were split into the following categories: knowledge/conceptual error, calculation error, self-contradictory, misapplication of guidelines, and overgeneralization.

Clinical disagreements were categorized as follows: guideline-specific (failure to apply current National Comprehensive Cancer Network (NCCN)/American Society for Radiation Oncology (ASTRO) treatment recommendations), trial-specific (misstating outcomes or inclusion criteria of landmark clinical trials), numerical-specific (incorrect dose, fractionation, or constraint values), ambiguous answer key (questions with debatable or outdated reference answers), anatomical/staging-specific (errors in American Joint Committee on Cancer (AJCC) staging or anatomical boundaries), image-specific (misinterpretation of imaging or field setup descriptions), self-contradictory (internal inconsistency between the explanation and chosen option), criteria confusion (conflating diagnostic or toxicity grading criteria), and multi-step quantitative/physics (errors requiring sequential clinical-dosimetric calculations). Physics disagreements were categorized as follows: knowledge/conceptual error (misunderstanding fundamental radiation physics principles), calculation error (arithmetic or formula-execution mistakes), self-contradictory (explanation contradicting the selected answer), misapplication of guidelines (incorrect application of American Association of Physicists in Medicine (AAPM) Task Group protocols or regulatory limits), and overgeneralization (applying broad rules to specific edge cases where they do not hold).

Statistical analysis

Categorical data were reported as frequencies and percentages (*n*, %). CIs were calculated using Wilson scores for clinical accuracy. Chi-square (χ^2) tests were used to assess accuracy per examination year. Welch's t-tests and Mann-Whitney U tests were performed to compare response and explanation lengths between correct and incorrect answers. Chi-square (χ^2) tests or Fisher's exact tests were used to assess accuracy per subdomain. All statistical analyses were performed using SPSS Statistics for Windows, Version 28.0 (IBM Corp., Armonk, NY) and R (v.4.3.2) (R Foundation for Statistical Computing, Vienna, Austria) [13]. A two-sided significance level of $P = 0.05$ was used.

Results

Clinical cohort

ChatGPT 5.5 had an overall accuracy of 82.0% (95% CI: 72.8%-89.0%) when answering 82/100 clinical questions correctly. The model performed best in the radiobiology, pediatrics, and renal/bladder oncology categories, with an accuracy of 100%. It performed worst in the female gynecological oncology (60.0%) and benign/palliative radiation oncology (66.7%) categories. For each examination year (2021, 2024, 2025, and 2026), accuracy ranged from 76.0% to 88.0%, with no significant difference between years ($\chi^2 = 2.44$, df = 3, $P = 0.49$) (Figure *1*). The mean response length across all 100 clinical questions was 206.7 words. The mean

response length was 202.1 ± 44.3 words for correct answers and 227.6 ± 57.5 words for incorrect answers. There was a nonsignificant trend toward longer explanations for incorrect answers (Welch's t = 1.77, $P = 0.09$; Mann-Whitney $U = 551.5$, $P = 0.10$). For correct clinical answers, the mean exploratory reasoning quality score was high at 1.85 ± 0.36 (scale, 0-2) (Table 1).

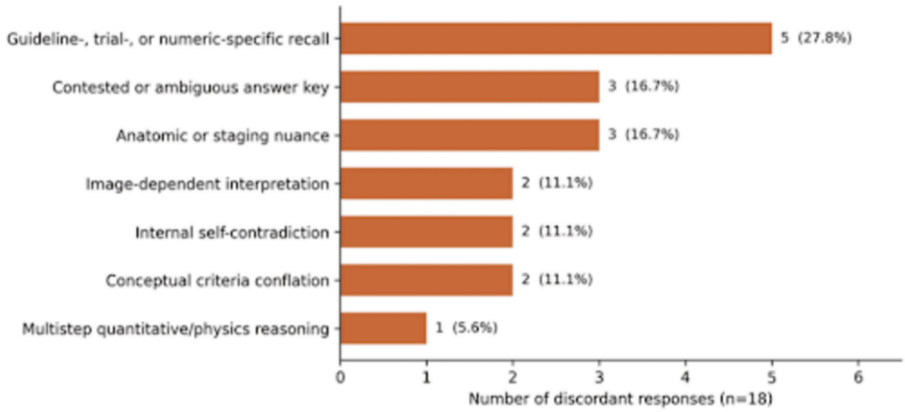


FIGURE 1: Patterns of ChatGPT 5.5 responses diverging from the TXIT answer key.

Primary classification of the 18 clinical responses generated by ChatGPT 5.5 that differed from the TXIT answer key, based on the evaluated TXIT question set. Categories describe the primary observed response pattern, were assigned by a single reviewer, and do not necessarily indicate an unequivocal clinical error.

TXIT, Radiation Oncology In-Training Examination

Variable	Correct (*n* = 82)	Incorrect (*n* = 18)	Test statistic	*P*-value
Response length (words), mean ± SD	202.1 ± 44.3	227.6 ± 57.5	t = 1.77	0.09

TABLE 1: Response length of ChatGPT 5.5 across answer-key concordance groups.

Clinical cohort response length for ChatGPT 5.5 across TXIT answer-key concordance groups. Data are presented as mean ± SD where available. Statistical comparisons were performed using the Mann-Whitney U test, with a two-sided significance threshold of $P < 0.05$. The U statistic was not available in the source materials.

TXIT, Radiation Oncology In-Training Examination; SD, standard deviation

SD, standard deviation

For the 18 incorrect clinical answers, the most common primary error type was guideline/trial/numerical-specific recall errors (5/18, 27.8%). Disputed/ambiguous answer keys (3/18, 16.7%), anatomical/staging specifics (3/18, 16.7%), image-dependent limitations (2/18, 11.1%), internal contradictions (2/18, 11.1%), conceptual criteria mix-ups (2/18, 11.1%), and multi-step numerical reasoning (1/18, 5.6%) were also common (Table 2).

Clinical discordance category	*N* (*n* = 18)	Percentage (%)
Guideline-, trial-, or numeric-specific recall	5	27.8
Contested or ambiguous answer key	3	16.7
Anatomic or staging classification nuance	3	16.7
Image-dependent interpretation	2	11.1
Internal self-contradiction	2	11.1
Conceptual criteria conflation	2	11.1
Multistep quantitative or physics reasoning	1	5.6

TABLE 2: Classification of ChatGPT 5.5 clinical responses diverging from the TXIT answer key.

Primary classification of clinical responses generated by ChatGPT 5.5 that differed from the TXIT answer key in the evaluated question set. Data are presented as number (*N*) and percentage (%). Categories describe the primary observed response pattern and do not necessarily indicate an unequivocal clinical error. Analysis is descriptive only.

TXIT, Radiation Oncology In-Training Examination

Physics cohort

ChatGPT 5.6 had an overall accuracy of 87.0% (95% CI: 78.7%-93.2%) when answering 87/100 physics questions correctly. The model had different accuracies depending on the physics category ($\chi^2 = 7.42$, df = 4, $P = 0.116$). It performed best in the brachytherapy category (15/15, 100%) and imaging and treatment planning (24/25, 96.0%). It performed next best in the dosimetry and units (11/13, 84.6%), safety and QA (16/20, 80.0%), and beam physics (21/27, 77.8%) categories. Beam physics had the most errors (6/13, 46.2%) (Figure *2*, Table *3*).

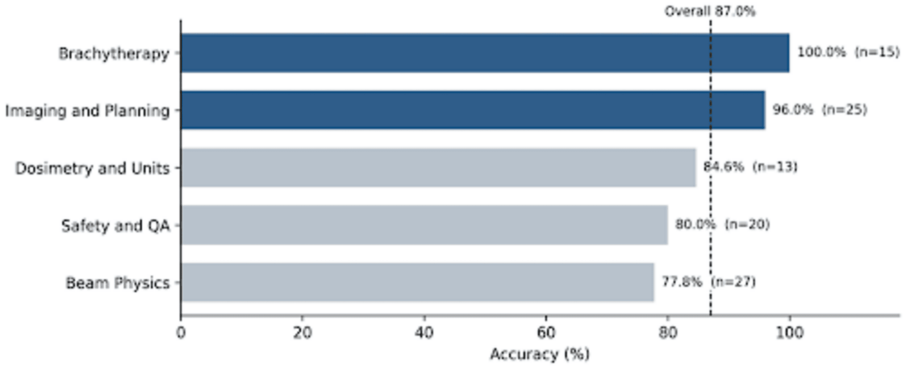


FIGURE 2: Topic-specific accuracy of ChatGPT 5.6 on RAPHEX-style physics questions.

Accuracy of ChatGPT 5.6 across topics in a 100-question RAPHEX-style physics assessment. The dashed horizontal line indicates the overall accuracy of 87.0%. All results reflect the evaluation of the 100-question physics question set using ChatGPT 5.6.

Physics domain	*N*	Correct	Incorrect	Accuracy (%)
Safety and Quality Assurance	20	16	4	80.0
Dosimetry and Units	13	11	2	84.6
Imaging and Planning	25	24	1	96.0
Beam Physics	27	21	6	77.8
Brachytherapy	15	15	0	100.0
Total	100	87	13	87.0

TABLE 3: Topic-specific accuracy of ChatGPT 5.6 across RAPHEX-style physics questions.

Accuracy of ChatGPT 5.6 across RAPHEX-style physics topics in the 100-question physics assessment. Data are presented as number (*N*) and percentage (%). Topic-specific chi-square or Fisher's exact test statistics and *P*-values were not available in the source results; therefore, no inferential comparisons are reported. The two-sided significance threshold specified in the source methods was $P < 0.05$.

The mean number of words given for all 100 physics answers was 186.8 words. The mean number of words for correct answers was 188.4 words and for incorrect answers was 176.2 words (Student's t = 0.81, $P = 0.42$). For the 13 incorrect physics answers, the most common error was knowledge and conceptual (5/13, 38.5%). Internal contradictions (4/13, 30.8%), incorrect guideline application (2/13, 15.4%), calculation errors (1/13, 7.7%), and overgeneralization (1/13, 7.7%) were also common (Figure *3*, Table *4*).

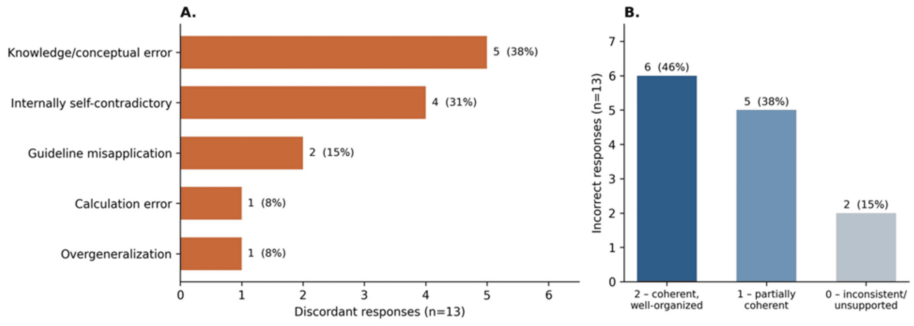


FIGURE 3: Characterization of ChatGPT 5.6 responses diverging from the reference answers.

Characterization of the 13 ChatGPT 5.6 responses that differed from the reference answers in the 100-question RAPHEX-style physics assessment. (A) Primary error category. (B) Exploratory rationale-quality score for the written explanation (2, coherent and well-organized despite an incorrect answer; 1, partially coherent; 0, substantially inconsistent or unsupported). Scores reflect the quality of the written explanations rather than hidden reasoning.

Physics discordance category	*N* (*n* = 13)	Percentage (%)
Knowledge or conceptual error	5	38
Internally self-contradictory response	4	31
Guideline misapplication	2	15
Calculation error	1	8
Overgeneralization	1	8

TABLE 4: Classification of ChatGPT 5.6 physics responses diverging from the reference answers.

Primary classification of ChatGPT 5.6 responses that differed from the reference answers in the 100-question RAPHEX-style physics assessment. Data are presented as numbers (*N*) and percentages (%). Categories are descriptive and mutually exclusive, as assigned by the original reviewer.

For incorrect physics answers, the exploratory rationale-quality scores were 2 (6/13, 46.2%), 1 (5/13, 38.5%), and 0 (2/13, 15.4%) (Table *5*). Both answers were internally contradictory, with correct calculations but a different option selected.

Rationale-quality score among incorrect physics responses	*N* (*n* = 13)	Percentage (%)
2: coherent and well-organized	6	46
1: partially coherent	5	38
0: substantially inconsistent or unsupported	2	15

TABLE 5: Rationale quality of ChatGPT 5.6 explanations among incorrect physics responses.

Exploratory rationale-quality scores among incorrect ChatGPT 5.6 responses in the 100-question RAPHEX-style physics assessment. Data are presented as numbers (*N*) and percentages (%). Scores characterize the coherence and organization of the written explanations, not the model's hidden reasoning.

Discussion

We found that ChatGPT 5.5 had an overall accuracy of 82.0% in answering 100 board-style questions, and ChatGPT 5.6 had an overall accuracy of 87.0% in answering 100 RAPHEX-style questions. While neither the ACR TXIT nor RAPHEX enforces a formal pass/fail threshold - and the American Board of Radiology (ABR) utilizes criterion-referenced scaled scoring rather than a fixed raw percentage - scores of 82.0% and 87.0% comfortably exceed typical resident mean percentages on in-training examinations (which generally range from ~60% to 75%, depending on postgraduate training year) and would generally correspond to passing-level performance on standardized board examinations [2]. Similarly, Singhal et al. demonstrated that LLMs can be trained to contain a large amount of medical knowledge [3]. However, due to variations in the versions of ChatGPT used, question banks, and the criteria used to assess ChatGPT, these results cannot be directly compared.

Our findings from the clinical radiation oncology questions are consistent with previous research. Barbour and Barbour used radiation oncology board questions to assess ChatGPT and showed that it was capable of answering a large proportion of questions correctly; however, ChatGPT struggled with more complex scenarios [8]. Nori et al. demonstrated that GPT-4 can be used to answer complex medical knowledge questions [2]. Kung et al. demonstrated that ChatGPT could be used as an educational tool by using it to answer questions from the United States Medical Licensing Examination (USMLE) [4]. Similarly, Gilson et al. used ChatGPT to answer questions from the USMLE [5]. Overall, this shows that LLMs could be used as supplementary tools in medical education. However, it does not show that they are capable of making clinical decisions [1].

From the clinical questions, we see that the inaccurate questions tended to be those that required detailed clinical knowledge, such as recommendations from guidelines, inclusion/exclusion criteria for trials, staging criteria, and numerical values. Radiation oncology often requires knowledge of staging, guidelines, trial eligibility, dose/fractionation, and sequencing of multidisciplinary treatments [7]. Barbour and Barbour showed that ChatGPT struggled with more complex radiation oncology questions [8]. Overall, Thirunavukarasu et al. discussed the limitations of using LLMs in medicine [1]. Therefore, high overall

scores do not mean that we can trust the model to make clinical recommendations without verifying them against guidelines.

The medical physics questions found that ChatGPT 5.6 had an overall accuracy of 87.0%. This adds to the growing body of literature discussing the use of LLMs in medical physics education. Kadoya et al. used ChatGPT to answer questions from the Japanese medical physicist exam and found an accuracy of ~73% [10]. Segura et al. used LLMs to answer questions from the 2024 RAPHEX exam [11]. Our work expands on this by using a newer version of ChatGPT and a specific question bank of 100 questions. We also look at how well ChatGPT performed on individual topics of medical physics and the types of errors it made.

Overall, we saw that ChatGPT had relatively high accuracy in answering the questions. However, we saw that there were variations in accuracy between topics, with beam physics having the lowest accuracy at 77.8%. This shows that even if a model performs well overall, there may be variations in accuracy between different topics. The questions that required multiple steps and multiple concepts, such as monitor unit calculations, radioactive decay, and field matching, were more challenging. Kadoya et al. showed that using generative AI to answer specialized medical physics questions can be challenging [10]. Similarly, Segura et al. looked at LLM performance using RAPHEX questions [11]. This shows that LLMs should be assessed on specific topics.

We found that nearly a third of the physics questions for which ChatGPT did not agree with the answer had internal contradiction errors. This means that ChatGPT would provide an explanation for the question with what appeared to be a correct calculation or reasoning but provide an incorrect answer. This shows a limitation of using LLMs based on how well they explain their answers. Nori et al. showed that GPT-4 has capabilities in answering medical challenge questions. However, it is important to assess how well they perform when answering challenging questions [2]. Similarly, Thirunavukarasu et al. discussed the limitations of LLMs in medicine [1]. This is important to consider for trainees using LLMs to assess themselves. Just because they provide a detailed explanation does not mean the answer is correct. We have conducted an exploratory analysis to look at how well ChatGPT explains the reasoning behind its answers.

Overall, our findings should be considered in the context of the increasing use of artificial intelligence in radiation oncology education. Zafar et al. performed a review of current and emerging uses of artificial intelligence in radiation oncology [7]. However, appropriate governance will be required to implement these tools into radiation oncology training. Malone et al. looked at artificial intelligence literacy in radiation oncology and showed the importance of preparing the workforce [9]. Our findings show that LLMs may be useful as supplementary resources in both clinical radiation oncology and medical physics. However, they should not be used as a replacement for existing resources, teaching from faculty, and clinical judgment.

Limitations

There are several limitations to this study. First, there were only 100 questions in each cohort, meaning that some subdomains had a small number of questions. Second, each question was only provided with one zero-shot prompt. We did not look at how well the model performed when using prompting techniques, multiple turns, or multiple samples. This may affect how consistent and accurate the LLMs are [1]. Third, we used different versions of ChatGPT (5.5 and 5.6) and different question banks. Therefore, we cannot make any direct comparisons between the two cohorts or attribute any differences in accuracy to the version of the model. Fourth, we used published answer keys. Some questions or answers may have been ambiguous. Finally, the qualitative errors and exploratory scoring of how well ChatGPT explained its answers were only performed by one reviewer. This means there could be bias in how they were categorized. Future work should use larger question banks, standardized prompting, multiple independent reviewers, and direct comparisons using the same instrument. This will allow us to better understand how reliable LLMs are and how useful they are in radiation oncology and medical physics [1,10]. LLMs can also create hallucinations, whereby they produce answers that are coherent yet contain false, unfounded, or made-up information [14]. We identified incorrect responses during our evaluation; however, hallucinations cannot be independently assessed. This is an important consideration when using LLMs in medical education.

Conclusions

ChatGPT 5.5 and ChatGPT 5.6 have shown strong agreement with answer keys in clinical radiation oncology and RAPHEX-style physics questions. However, the types of errors differ between question topics. Clinical questions mainly demonstrate errors involving detailed guidelines, trials, and staging. Physics questions often show a lack of understanding and inconsistencies in calculations. LLMs can be useful resources; however, their responses should be carefully checked against the literature and expert faculty.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Veronia Fahmy, James S. Welsh

Acquisition, analysis, or interpretation of data: Veronia Fahmy, Jane Georgas

Drafting of the manuscript: Veronia Fahmy

Critical review of the manuscript for important intellectual content: Veronia Fahmy, Jane Georgas, James S. Welsh

Supervision: James S. Welsh

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS: Large language models in medicine. Nat Med. 2023, 29:1930-40. 10.1038/s41591-023-02448-8
2. Nori H, King N, McKinney SM, et al.: Capabilities of GPT-4 on medical challenge problems . arXiv. 2023, 10.48550/arXiv.2303.13375
3. Singhal K, Azizi S, Tu T, et al.: Large language models encode clinical knowledge . Nature. 2023, 620:172-80. 10.1038/s41586-023-06291-2
4. Kung TH, Cheatham M, Medenilla A, et al.: Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023, 2:e0000198. 10.1371/journal.pdig.0000198
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D: How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023, 9:e45312. 10.2196/45312
6. Wang H, Wu W, Dou Z, He L, Yang L: Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform. 2023, 177:105173. 10.1016/j.ijmedinf.2023.105173
7. Zafar F, Vilsan J, Mani S, et al.: AI in radiation oncology: a comprehensive review of current applications and future directions. Cureus. 2025, 17:e92964. 10.7759/cureus.92964
8. Barbour AB, Barbour TA: A radiation oncology board exam of ChatGPT . Cureus. 2023, 15:e44541. 10.7759/cureus.44541
9. Malone C, Callens D, Forde E, et al.: Quantifying the AI readiness gap: an international, multidisciplinary assessment of artificial intelligence literacy in the radiation oncology community. Clin Transl Radiat Oncol. 2026, 59:101210. 10.1016/j.ctro.2026.101210
10. Kadoya N, Arai K, Tanaka S, et al.: Assessing knowledge about medical physics in language-generative AI with large language model: using the medical physicist exam. Radiol Phys Technol. 2024, 17:929-37. 10.1007/s12194-024-00838-2
11. Segura AC, Perkins D, Dumane VA, et al.: Large language models in radiation oncology physics education: performance assessment using the 2024 RAPHEX exam. Int J Radiat Oncol Biol Phys. 2025, 123:779. 10.1016/j.ijrobp.2025.06.3253
12. ChatGPT-5.5 and ChatGPT-5.6 Language Model System Card and Technical Documentation . (2026). https://openai.com.
13. IBM Corp. IBM SPSS Statistics for Windows, Version 28.0. Armonk, NY: IBM Corp . (2021). https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-28.
14. Heydari S, Alavi ZS: Artificial intelligence hallucinations in large language models for medical education and healthcare: a scoping review of detection, mitigation, and governance. Med Educ Online. 2026, 31:2756806. 10.1080/10872981.2026.2756806