frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 23 May 2025
DOI 10.3389/fonc.2025.1557064

OPEN ACCESS

EDITED BY
Daniele Loiacono,
Polytechnic University of Milan, Italy

REVIEWED BY
Chenbin Liu,
Chinese Academy of Medical Sciences and Peking Union Medical College
Eric Ehler,
University of Minnesota Medical Center, United States

*CORRESPONDENCE
Wei Liu
liu.wei@mayo.edu

RECEIVED 08 January 2025
ACCEPTED 29 April 2025
PUBLISHED 23 May 2025

CITATION
Wang P, Holmes J, Liu Z, Chen D, Liu T, Shen J and Liu W (2025) A recent evaluation on the performance of LLMs on radiation oncology physics using questions of randomly shuffled options.
Front. Oncol. 15:1557064.
doi: 10.3389/fonc.2025.1557064

A recent evaluation on the performance of LLMs on radiation oncology physics using questions of randomly shuffled options

Peilong Wang[1], Jason Holmes[1], Zhengliang Liu[2], Dequan Chen[3], Tianming Liu[2], Jiajian Shen[1] and Wei Liu[1]*

[1]Department of Radiation Oncology, Mayo Clinic, Phoenix, AZ, United States, [2]School of Computing, University of Georgia, Athens, GA, United States, [3]Department of Radiology, Mayo Clinic, Rochester, MN, United States

Purpose: We present an updated study evaluating the performance of large language models (LLMs) in answering radiation oncology physics questions, focusing on the recently released models.

Methods: A set of 100 multiple-choice radiation oncology physics questions, previously created by a well-experienced physicist, was used for this study. The answer options of the questions were randomly shuffled to create "new" exam sets. Five LLMs – OpenAI o1-preview, GPT-4o, LLaMA 3.1 (405B), Gemini 1.5 Pro, and Claude 3.5 Sonnet – with the versions released before September 30, 2024, were queried using these new exam sets. To evaluate their deductive reasoning ability, the correct answer options in the questions were replaced with "None of the above." Then, the explain-first and step-by-step instruction prompts were used to test if this strategy improved their reasoning ability. The performance of the LLMs was compared with the answers from medical physicists.

Results: All models demonstrated expert-level performance on these questions, with o1-preview even surpassing medical physicists with a majority vote. When replacing the correct answer options with 'None of the above', all models exhibited a considerable decline in performance, suggesting room for improvement. The explain-first and step-by-step instruction prompts helped enhance the reasoning ability of the LLaMA 3.1 (405B), Gemini 1.5 Pro, and Claude 3.5 Sonnet models.

Conclusion: These recently released LLMs demonstrated expert-level performance in answering radiation oncology physics questions.

KEYWORDS

radiation oncology, large language model (LLM), physics, evaluation, augmentation

1 Introduction

Large language models (LLMs) have advanced rapidly. On the one hand, the size of the data used for pre-training and the number of model parameters have grown a lot. For example, GPT-2 had 1.5 billion parameters (1), GPT-3 scaled up to 175 billion (2), and GPT-4 is estimated to have even more (3). On the other hand, the fine-tuning methods and prompt engineering strategies have advanced substantially (4, 5). Furthermore, agents and Retrieval-Augmented Generation (RAG) systems built on LLMs have seen considerable progress (6, 7). Notable recent developments as of September 2024 include OpenAI o1-preview (8), GPT-4o (9), LLaMA 3.1 (405B parameters) (10), Gemini 1.5 Pro (11), and Claude 3.5 Sonnet (12), demonstrating state-of-art performance in overall language processing, reasoning, and diverse downstream applications.

The rapid evolution of LLMs also renders prior performance evaluations outdated. As some LLMs cease providing services, new models are introduced, and existing versions are updated, studies published before may no longer accurately reflect the current state of LLM capabilities. A fresh evaluation is needed to address the dynamic landscape of LLM advancements.

In healthcare, LLMs have been explored for numerous potential applications (13–18). For their direct use in radiation oncology, unique challenges related to evaluation and validation arise due to the complexity and precision of treatment, which involves both clinical factors and physics considerations. Therefore, assessing the performance of LLMs in addressing questions related to radiation oncology physics is crucial. Such evaluations not only tell us how efficiently they process and reason about radiation oncology physics but also help us understand their limitations. In the past, several LLMs were evaluated on the 2021 American College of Radiology (ACR) Radiation Oncology In-Training Examination (TXIT), revealing that GPT-4-turbo achieved the highest score of 68.0%, outperforming some resident physicians (19). GPT-3.5 and GPT-4 were also assessed on Japan's medical physicist board examinations from 2018 to 2022, where GPT-4 demonstrated superior performance with an average accuracy of 72.7% (20). To offer insights into the recently released state-of-art LLMs and build on our prior work (21), we present here an updated study with refined methods evaluating their performance in radiation oncology physics.

We utilized the 100-question radiation oncology physics exam we developed based on the American Board of Radiology exam style (22), and randomly shuffled the answer options to create "new" exam sets. We then queried the LLMs with these new exam sets and checked their ability to answer questions accurately. We also evaluated their deductive reasoning ability and tested whether the explain-first and step-by-step instruction prompts would improve their performance in reasoning tasks.

2 Methods

The 100-question multiple-choice examination on radiation oncology physics was created by our experienced medical physicist, following the official study guide of the American Board of Radiology. That exam includes 12 questions on basic physics, 10 questions on radiation measurements, 20 questions on treatment planning, 17 questions on imaging modalities and applications in radiotherapy, 13 questions on brachytherapy, 16 questions on advanced treatment planning and special procedures, and 12 questions on safety, quality assurance (QA), and radiation protection. 17 out of the 100 questions are math-based and require numerical calculation.

All the evaluated LLMs were queried with the exam questions through Application Programming Interface (API) services provided by their respective hosts, except LLaMA 3.1 (405B), an open-source LLM, which was hosted by us locally at our institution. All the LLMs used were the recently released version before September 30, 2024. The temperature was set to 0.1 for all LLMs to minimize variability in their responses[1], with the exception of the OpenAI o1-preview, whose temperature was fixed at 1 and could not be changed by the user.

2.1 Randomly shuffling the answer options

Since it was difficult to know whether any LLM had been pre-trained using our previously published 100-question multiple-choice exam, we wrote Python code to randomly shuffle the answer options for the 100 multiple-choice questions five times. For each shuffle, we obtained a "new" 100-question multiple-choice exam set. We then queried all the LLMs five times (Trial 1 - Trial 5), each with a new exam set. Each question of the new exam set was queried individually. We checked the distribution of the correct answers' locations for the five new exams where the options were shuffled and confirmed that the distribution of the correct options is fairly random among A, B, C, D, or E (only 2 questions offered option E), as shown in the Supplementary Material. The prompt we used for all the queries was as follows:

> "Please solve this radiation oncology physics problem: [radiation oncology physics problem]. "

This allowed the LLMs to reason and answer freely. Table 1 illustrates an example of the trials and how we queried the LLMs with the questions. For the responses generated by the LLMs, we utilized the LLaMA 3.1 (405B) model hosted locally to further extract the chosen answer options (letters A, B, C, D, or E) from free-form responses, thereby reducing some of the manual effort required to read and record them individually. We then conducted manual verification of the extracted options and obtained the final answer option sheet for all LLMs to compare with the ground truth answers. The accuracy of each LLM was reported as the mean score

1 The temperature was set to 0.1 rather than 0 due to the different meanings of a temperature of 0 across different LLMs. To avoid potential unexpected behaviors from models, we set the lower bound to 0.1 rather than 0.

TABLE 1 Illustration of the prompts and questions of randomly shuffled options to evaluate LLMs' performance on answering radiation oncology physics questions.

Trail	Prompt	Question
Trail 1	Please solve this radiation oncology physics problem:	Which of the following particles cannot be accelerated by an electric field? (a) Neutrons (b) Protons (c) Electrons (d) Positrons
Trail 2		Which of the following particles cannot be accelerated by an electric field? (a) Proton (b) Neutrons (c) Electrons (d) Positrons
Trail 3		Which of the following particles cannot be accelerated by an electric field? (a) Positrons (b) Protons (c) Electrons (d) Neutrons
Trail 4		Which of the following particles cannot be accelerated by an electric field? (a) Electrons (b) Neutrons (c) Protons (d) Positrons
Trail 5		Which of the following particles cannot be accelerated by an electric field? (a) Electrons (b) Positrons (c) Neutrons (d) Protons

across the five trials, and the measurement uncertainty was reported as the standard deviation of the five trials, as shown in the following equations:

$$\text{Mean}(\bar{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

where x_i represents each measurement.

The results of the LLMs' test scores were compared with the majority vote results from a group of medical physicists conducted in our previous study. The medical physicist group consisted of four experienced board-certified medical physicists, three medical physics residents, and two medical physics research fellows. For each question, the most common answer choice was selected as the group's answer. In case of a tie, one of the most common answer choices was chosen randomly.

2.2 Evaluating deductive reasoning ability

Deductive reasoning ability refers to the cognitive process of logically analyzing information to draw specific conclusions from general premises. The multiple-choice question with the answer option "None of the above" can effectively evaluate the test-taker's deductive reasoning ability, as it involves evaluating each option based on the information provided and ruling out incorrect choices contradicting known facts or logical outcomes to reach the correct answer. We therefore replaced the correct option in the exam with "None of the above." Since transformer-based LLMs predict the next word based on prior contexts, changing the correct option to "None of the above" removes a straightforward cue that might guide the model toward a known or patterned solution, thus forcing the LLMs to rely more on reasoning about the specific question and its options, rather than using surface-level lexical or statistical patterns that it may have learned.

2.2.1 Replacing the correct option with "None of the above"

We developed Python code to replace the correct option with "None of the above" for all questions in the five new exam sets derived by random shuffling. For each trial (Trial 1 - Trial 5), we queried the LLMs with a set of exams in which both the correct option was "None of the above," and its location was randomly shuffled. We used the same prompt as in Sec. 2.1 across all queries, and each question in an exam set was queried individually. This setup challenges the LLMs to avoid pattern-based answering and not rely on any single choice, but to process each answer option by reasoning.

As before, we utilized the previously described processes for answer-option extraction and manual verification, as outlined in Sec. 2.1. The performance accuracy and uncertainty of each LLM were reported as the average score and standard deviation across all five trials. Due to this setup, these exams were not used to test humans, as this pattern can be easily recognized by human test-takers[2].

2.2.2 Explain-first and step-by-step instruction

To further check if explicitly asking the LLMs to explain first and then develop answers step-by-step (chain-of-thought) would improve their deductive reasoning ability (23), we engineered the following prompt and queried the LLMs again with it:

> "Please solve this radiation oncology physics problem:
> [radiation oncology physics problem]
> Please first explain your reasoning, then solve the problem step by step, and lastly provide the correct answer (letter choice)."

We used the five exam sets and conducted the querying process both as described in Sec. 2.2.1. All five LLMs were evaluated using this prompting strategy. Accuracy and uncertainty were reported. The results from this strategy were compared with the test results

2 Since each question was queried through the API individually, it is assumed that LLMs would not notice this pattern.

10.3389/fonc.2025.1557064

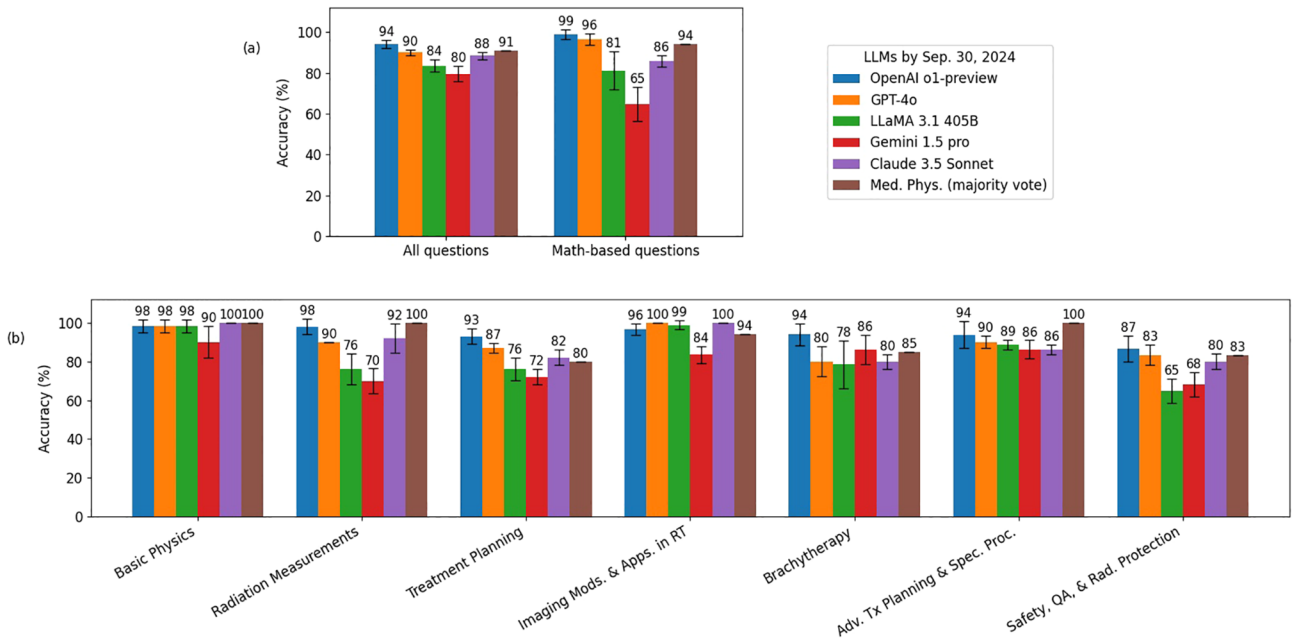

FIGURE 1
Evaluation results using five exam sets where the answer options of the questions were randomly shuffled. Panel **(a)** illustrates the evaluation results for all questions and math-based questions, while Panel **(b)** presents the evaluation results broken down by different topics.

from original prompts, where no explanation or step-by-step answering was required, as described in Sec. 2.2.1.

3 Results

3.1 Results of exam sets with randomly shuffled options

The evaluation results of the exam sets with options randomly shuffled are presented in Figure 1, where the height of each bar represents the mean test score, and the error bars indicate the standard deviations across five trials. All five LLMs exhibited strong performance, achieving mean test scores above 80%, which suggests their performance on these exams is comparable to that of human experts. When compared to the majority vote results from the medical physics group, the OpenAI o1-preview model outperformed the medical physicists with a majority vote. For math-based questions, both the o1-preview and GPT-4o models surpassed the medical physicists with a majority vote.

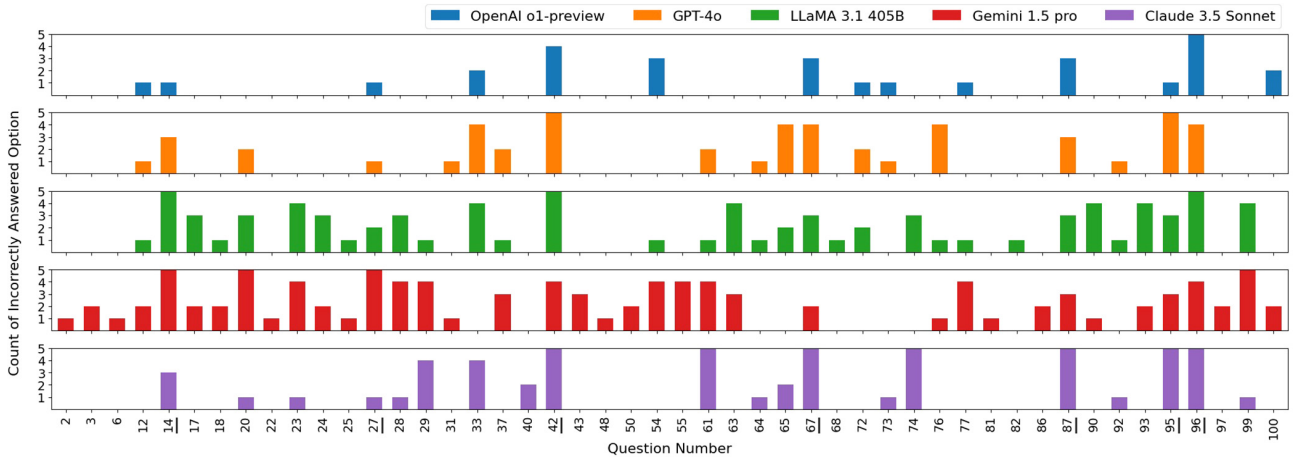

FIGURE 2
Distribution of incorrectly selected answer options by each LLM across five exam sets. All answer options in the five exam sets were randomly shuffled. Questions that were correctly answered by all LLMs in all five trials are not shown in this figure. Questions 14, 27, 42, 67, 87, 95, and 96, which were commonly answered incorrectly by all LLMs at least once, are underlined in the figure.

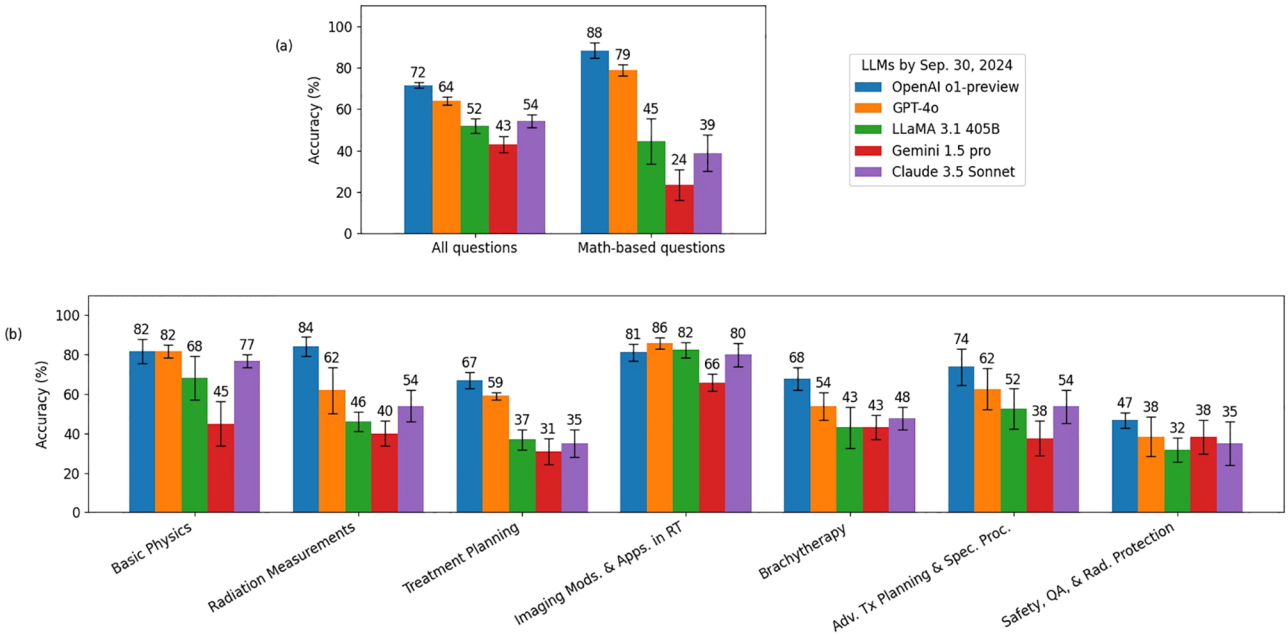

FIGURE 3
Deductive reasoning ability evaluation results for every LLM. The correct answer options were replaced with "None of the above" in all questions. Panel **(a)** illustrates the evaluation results for all questions and math-based questions, while Panel **(b)** presents the evaluation results broken down by different topics.

The raw counts of incorrect responses by the LLMs are shown in Figure 2, where each color represents the incorrect answers by an LLM across trials. As observed, each LLM exhibited variability in answering questions across trials. Notably, the models also showed similarities in incorrectly answering certain questions. We analyzed the questions that were commonly answered incorrectly by all LLMs at least once across all five trials – question numbers: 14, 27, 42, 67, 87, 95, and 96. Interestingly, only one of these questions was math-based, while the remaining seven were closely related to clinical medical physics knowledge, such as American Association of Physicists in Medicine (AAPM) Task Group (TG) reports and clinical experience. This observation suggests that current LLMs may still struggle with answering clinically focused radiation oncology physics questions. For example, question number 42 does not involve any calculations but instead focuses primarily on clinical hands-on experience.

3.2 Results of LLMs' deductive reasoning ability

Figure 3 shows the results of the deductive reasoning ability tests, where the correct answer options were replaced with "None of the above" in all questions. Overall, all LLMs performed much more poorly compared to the results in Sec. 3.1. Given that transformer-based LLMs (24) were designed to predict the next word in a sequence, replacing the correct answers with "None of the above" would likely disrupt their pattern recognition abilities, thereby reducing their overall scores performed on the exam sets. Nonetheless, the OpenAI o1-preview and GPT-4o still outperformed the others, especially on math-based questions, indicating the strong reasoning ability of these two models.

Figure 4 compares the performance of LLaMA 3.1 (405B), Gemini 1.5 Pro, and Claude 3.5 Sonnet models with the original simple prompts and with the explain-first, step-by-step instruction prompts. Overall, all three models demonstrated improved reasoning ability with the latter prompting strategy. Notably, Gemini 1.5 Pro showed significant gains on math-based questions, increasing its score from 24% to 68%. The o1-preview and GPT-4o showed only about a 1% overall difference, which was too small to be represented in this figure.

4 Discussion

4.1 Improvement of performance on answering radiation oncology physics questions of the state-of-art LLMs over the past two years

Over the past two years, our studies have observed a notable improvement in the performance of state-of-the-art LLMs on this highly specialized task – answering radiation oncology physics questions, as shown in Figure 5. Early versions of ChatGPT, like GPT-3.5 in late 2022 (25), scored around 54%, showing clear gaps in domain-specific knowledge. With the introduction of GPT-4 in early 2023, performance jumped to around 76%, reflecting improvements in accuracy and understanding. Subsequent

10.3389/fonc.2025.1557064

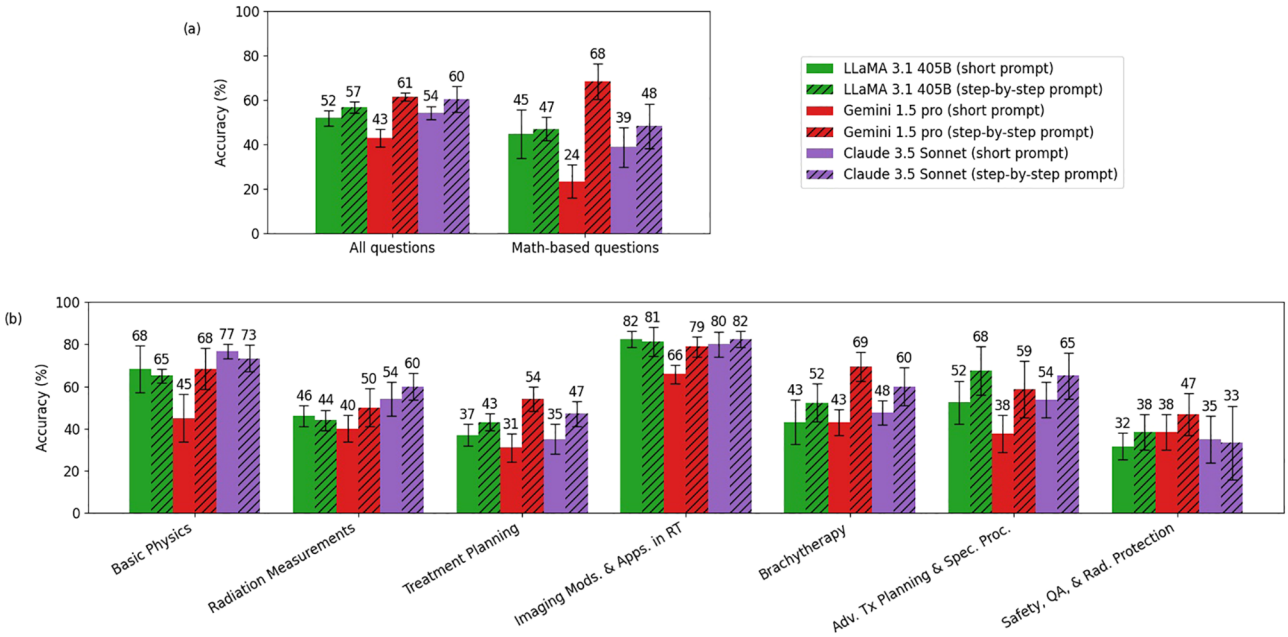

FIGURE 4
Comparison of accuracy between the short prompt and the explain-first, step-by-step instruction prompt (chain-of-thought) for LLaMA 3.1 (405B), Gemini 1.5 Pro, and Claude 3.5 Sonnet. Panel **(a)** shows the comparison between the short prompt and the explain-first, step-by-step instruction prompt for all questions and math-based questions, while Panel **(b)** breaks down the comparison by topic. (o1-preview and GPT-4o showed only about a 1% overall difference, which was too small to be represented in this figure.).

releases of the GPT-4o model and more recently the o1-preview (both in 2024), pushed scores even higher to 90% and 94% respectively, indicating increasing capabilities in radiation oncology physics. This steady improvement can be attributed to more extensive domain pre-training, increase of number of parameters, refined architectural updates, and enhanced fine-tuning techniques (26, 27), all of which have led to improved understanding, stronger reasoning skills, and better alignment with expert-level knowledge. The evolution of these models over the last two years underscores the rapid growth of LLMs, suggesting their potential as useful tools in areas such as radiation oncology physics education and training.

4.2 Potential applications of LLMs in radiation oncology physics

Recent advancements in exploring potential applications of LLMs in radiation oncology physics have focused on auto-contouring, dose prediction and treatment planning. For auto-

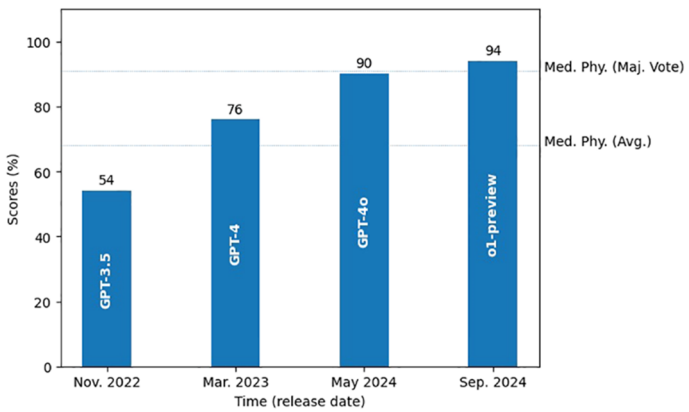

FIGURE 5
Growth of the state-of-art LLMs' performance in radiation oncology physics over the past two years. Two dotted lines mark the average score and the majority vote score of the medical physicists.

contouring, LLMs have been utilized to extract electronic medical records (EMR) text data and align them with the image embeddings of the mixture-of-experts model to improve the performance of the target volume contouring for radiation therapy (28). In addition, LLMs have also been used to extract text-based features and incorporated them into vision transformer to help improve the target delineation results (29). In dose prediction, LLM have been explored to encode knowledge from prescriptions and interactive instructions from clinicians into neural networks to enhace the prediction of dose-volume histograms (DVH) from medical images (30). Regarding treatment planning, GPT-4V has been investigated for evaluating dose distribution and DVH and assisting with the optimization of the treatment planning (18). Furthermore, an LLM-based multi-agent system has also been developed to mimick the workflow of dosimetrists and medical physicists to generate text-based treatment plans (31). Collectively, these advancements highlight the transformative potential of LLMs in radiation oncology physics, offering potential improvements in efficiency and outcomes.

4.3 Possible further improvement of LLMs in radiation oncology physics

The performance of LLMs on radiation oncology physics, although encouraging, still requires further improvement due to two primary factors. First, radiation oncology physics represents a very specialized domain characterized by both the complexity of physics concepts and specific clinical contexts, neither of which was extensively represented in the general datasets used during the initial pre-training of these models. Second, existing LLMs still encounter difficulties with reasoning tasks specific to radiation oncology physics, indicating a need for enhanced general reasoning capabilities. To address these limitations, further studies could explore strategies of fine-tuning existing LLMs using specialized medical physics domain datasets with clinical contexts. Such fine-tuning would likely enable the models to better capture the complexities and contextual details of the domain, enhancing their accuracy and practical clinical utility in medical physics tasks. Additionally, to improve reasoning capabilities, techniques such as chain-of-thought, which encourages models to articulate intermediate reasoning steps explicitly, and reinforcement learning, which optimizes model responses in desired patterns, could be investigated (32).

4.4 Limitations

Although the LLMs evaluated in this study exhibit expert-level performance on radiation oncology physics questions, such results do not directly translate to effectiveness in practical clinical tasks like treatment planning and delivery. This limitation arises from differences between theoretical examinations and practical clinical applications. Clinical scenarios encountered in radiation oncology are inherently more complex, context-dependent, and require integrating multiple sources of clinical and patient-specific data, whereas theoretical examinations often have clearly defined questions and objective answers. Consequently, strong performance in controlled question-answering tasks may not effectively transfer to real-world contexts, which frequently involve ambiguity, uncertainty, and nuanced clinical judgment. Additionally, clinical decision-making encompasses not only physics-based calculations but also multidisciplinary collaboration, patient safety considerations, regulatory compliance, and human factors in clinical workflows. Therefore, although the evaluated models demonstrate promise in foundational physics knowledge, caution must be exercised when inferring their direct clinical utility.

5 Conclusion

We evaluated recently released LLMs using a method that randomly shuffled the answer options of radiation oncology physics questions. Our results demonstrated that these models achieved expertlevel performance on these questions, with some even surpassing human experts with a majority vote. However, when the correct answer options were replaced with "None of the above," all models exhibited a steep decline in performance, suggesting room for improvement. Employing the technique of explain-first and step-by-step instruction prompt enhanced the reasoning abilities of LLaMA 3.1 (405B), Gemini 1.5 Pro, and Claude 3.5 Sonnet.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/Mayo-Clinic-RadOnc-Foundation-Models/Radiation-Oncology-NLP-Database.

Author contributions

PW: Data curation, Formal Analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. JH: Conceptualization, Methodology, Visualization, Writing – review & editing, Formal Analysis. ZL: Resources, Software, Writing – review & editing. DC: Resources, Software, Writing – review & editing. TL: Funding acquisition, Resources, Writing – review & editing. JS: Data curation, Formal Analysis, Methodology, Supervision, Visualization, Writing – review & editing. WL: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the National Cancer Institute (NCI) R01CA280134, the Eric & Wendy Schmidt Fund for AI Research & Innovation, The

Fred C. and Katherine B. Anderson Foundation, and the Kemper Marley Foundation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. Generative AI was used to correct the grammar errors.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2025.1557064/full#supplementary-material

References

1. OpenAI. Better language models and their implications(2019). Available online at: https://openai.com/index/betterlanguage-models/ (Accessed February 14, 2019).

2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., (United States: Neural Information Processing Systems Foundation, Inc.) (2020) 33:1877–901.

3. Open AI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. *arXiv: 2303.08774*. (2024). https://arxiv.org/abs/2303.08774.

4. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., (United States: Neural Information Processing Systems Foundation, Inc.) (2022), 27730–44.

5. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: low-rank adaptation of large language models. *arXiv: 2106.09685*. (2021). https://arxiv.org/abs/2106.09685.

6. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, et al. The rise and potential of large language model based agents: A survey. In: *Science China Information Sciences*. (Beijing, China: Science China Press) 68.2 (2025), 121101.

7. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Advances in Neural Information Processing Systems. Curran Associates, Inc.*, (United States: Neural Information Processing Systems Foundation, Inc.) (2020), 9459–74.

8. OpenAI. Introducing OpenAI o1-preview(2024). Available online at: https://openai.com/index/introducing-openaio1-preview/ (Accessed September 12, 2024).

9. Open AI, Hurst A, Lerer A, Goucher A, Perelman A, et al. GPT-4o system card. *arXiv: 2410.21276*. (2024). https://arxiv.org/abs/2410.21276.

10. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, et al. The llama 3 herd of models. *arXiv: 2407.21783*. (2024). https://arxiv.org/abs/2407.21783.

11. Gemini Team, Georgiev P, Lei V, Burnell R, Bai L, Gulati A, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv: 2403.05530*. (2024). https://arxiv.org/abs/2403.05530.

12. Anthropic. Introducing claude 3.5 sonnet(2024). Available online at: https://www.anthropic.com/news/claude-3-5sonnet (Accessed June 20, 2024).

13. Liu Z, Zhang L, Wu Z, Yu X, Cao C, Dai H, et al. Surviving chatGPT in healthcare. In: *Frontiers in Radiology*, (Lausanne, Switzerland: Frontiers Media S.A.) vol. 3. (2024). issn: 2673-8740. doi: 10.3389/fradi.2023.1224682

14. Liu C, Liu Z, Holmes J, Zhang L, Zhang L, Ding Y, et al. Artificial general intelligence for radiation oncology. In: *Meta-Radiology*, (Beijing, China: KeAi Communications Co., Ltd.) vol. 1. (2023). p. 100045. issn: 2950-1628. doi: 10.1016/j.metrad.2023.100045

15. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. In: *Nature Medicine* (New York, United States; London, United Kingdom: Nature Portfolio) (2024). p. 1–13.

16. Holmes J, Zhang L, Ding Y, Feng H, Liu Z, Liu T, et al. Benchmarking a foundation large language model on its ability to relabel structure names in accordance with the american association of physicists in medicine task group-263 report. In: *Practical Radiation Oncology*, (United States: Elsevier Inc.) vol. 14. (2024). p. e515–21. issn: 1879-8500. doi: 10.1016/j.prro.2024.04.017

17. Li X, Zhao L, Zhang L, Wu Z, Liu Z, Jiang H, et al. Artificial general intelligence for medical imaging analysis. In: *IEEE Reviews in Biomedical Engineering*. (New York, United States / Piscataway, New Jersey, United States: IEEE) (2024). p. 1–18. doi: 10.1109/RBME.2024.3493775

18. Liu S, Pastor-Serrano O, Chen Y, Gopaulchan M, Liang W, Buyyounouski M, et al. Automated radiotherapy treatment planning guided by GPT-4Vision. *arXiv: 2406.15609*. (2024). https://arxiv.org/abs/2406.15609.

19. Thaker N, Redjal N, Loaiza-Bonilla A, Penberthy D, Showalter T, Choudhri A, et al. Large language models encode radiation oncology domain knowledge: performance on the american college of radiology standardized examination. In: *AI in Precision Oncology*, (New Rochelle, New York, United States: Mary Ann Liebert, Inc.) vol. 1. (2024). p. 43–50. doi: 10.1089/aipo.2023.0007

20. Kadoya N, Arai K, Tanaka S, Kimura Y, Tozuka R, Yasui K, et al. Assessing knowledge about medical physics in language-generative AI with large language model: using the medical physicist exam. *Radiol Phys Technol*. (Tokyo, Japan: Springer Japan) (2024) 17:929–37. doi: 10.1007/s12194-024-00838-2

21. Holmes J, Liu Z, Zhang L, Ding Y, Sio T, McGee L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. In: *Frontiers in Oncology*, (Lausanne, Switzerland: Frontiers Media S.A.) vol. 13. (2023). issn: 2234-943X. doi: 10.3389/fonc.2023.1219326

22. Liu Z, Holmes J, Liao W, Liu C, Zhang L, Feng H, et al. The radiation oncology NLP database. *arXiv: 2401.10995*. (2024). https://arxiv.org/abs/2401.10995.

23. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv: 2201.11903*. (2023). https://arxiv.org/abs/2201.11903.

24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., (2017).

25. OpenAI. Introducing chatGPT(2022). Available online at: https://openai.com/index/chatgpt/ (Accessed November 30, 2022).

26. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. In: Jurafsky D, et al. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, Pennsylvania, United States: Association for Computational Linguistics (2020). p. 8342–60. doi: 10.18653/v1/2020.acl-main.740

27. Kaplan J, McCandlish S, Henighan T, Brown T, Chess B, Child R, et al. Scaling laws for neural language models. *arXiv: 2001.08361*. (2020). https://arxiv.org/abs/2001.08361.

28. Rajendran P, Chen Y, Qiu L, Niedermayr T, Liu W, Buyyounouski M, et al. Auto-delineation of treatment target volume for radiation therapy using large language model-aided multimodal learning. In: *International Journal of Radiation Oncology*Biology*Physics*, (United States: Elsevier Inc.) vol. 121. (2025). p. 230–40. issn: 0360-3016. doi: 10.1016/j.ijrobp.2024.07.2149

29. Oh Y, Park S, Li X, Wang Y, Paly J, Efstathiou J, et al. Mixture of multicenter experts in multimodal generative AI for advanced radiotherapy target delineation. *arXiv: 2410.00046*. (2024). https://arxiv.org/abs/2410.00046.

30. Dong Z, Chen Y, Gay H, Hao Y, Hugo G, Samson P, et al. Large-language-model empowered 3D dose prediction for intensity-modulated radiotherapy. In: *Medical Physics*, (Hoboken, New Jersey, United States: Wiley) vol. 52. (2025). p. 619–32. doi: 10.1002/mp.17416

31. Wang Q, Wang Z, Li M, Ni X, Tan R, Zhang W, et al. A feasibility study of automating radiotherapy planning with large language model agents. In: *Physics in Medicine & Biology*, (Bristol, United Kingdom: IOP Publishing Ltd.) vol. 70. (2025). p. 075007. doi: 10.1088/1361-6560/adbff1

32. Shao Z, Wang P, Zhu Q, Xu R, Song J, Bi X, et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models. *arXiv: 2402.03300*. (2024). https://arxiv.org/abs/2402.03300.