Advances in Radiation Oncology (2026) **11**, 101929

advances in radiation oncology

www.advancesradonc.org

Research Letter

ChatGPT Versus DeepSeek: Assessing Artificial Intelligence Performance on Radiation Oncology Examination Questions

Ronald Chow, MD MS, MEng, FACE, FRSPH,[a,b,c] Ajay Zheng, BS,[a] Chenxi Gao, BS,[a] Milo Vermeulen, PhD,[a] Francis Yu, MS,[a] Irini Yacoub, MD,[a] Arpit M. Chhabra, MD,[a] J. Isabelle Choi, MD,[a] Haibo Lin, PhD,[a] Gilmer Valdes, PhD,[d] and Charles B. Simone, II, MD, FASTRO, FACRO[a,*]

[a]Department of Radiation Oncology, New York Proton Center, New York, New York; [b]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada; [c]Centre for Evidence-Based Medicine, University of Oxford, Oxford, United Kingdom; and [d]Department of Clinical Machine Learning Program, Moffitt Cancer Center, Tampa, Florida

Received 28 May 2025; accepted 9 October 2025

Purpose: Large language models have been assessed for their ability to receive and answer medical questions. Recently, there has been a new large language model named DeepSeek released, which has not been assessed for medical accuracy. This is the first study to assess DeepSeek for accuracy in responding to medical questions.

Methods and Materials: We prompted DeepSeek-R1 and several models of ChatGPT with 600 radiation oncology examination questions from national radiation oncology in-service multiple-choice examinations. These questions are used by medical residents in preparation for their certifying board examination and assess knowledge on anatomy, treatment planning, cancer epidemiology, and landmark trials. We recorded each model's accuracy, total prompt and completion tokens used, and total run time. Accuracy was compared across question categories and between models. Type I error was set at 0.05.

Results: DeepSeek-R1 answered 84.0% of questions correctly, requiring 59 seconds per question. DeepSeek-R1 demonstrated a significant difference in accuracy by question categories ($P = .012$) and was least accurate for questions about landmark studies (74.2% accuracy). ChatGPT o1 answered 89.0% of questions correctly, requiring 10 seconds per question. ChatGPT o1's accuracy did not significantly differ across question categories (93.5% accurate on questions about landmark studies). DeepSeek-R1 used 7.2% more tokens than ChatGPT o1. At February 2025 prices, DeepSeek-R1 costs up to $1.56, compared with ChatGPT's $37.96.

Conclusion: DeepSeek-R1 is less accurate and answers more slowly compared with ChatGPT o1, but is less costly at the time of this manuscript preparation. Careful analysis and consideration of the current landscape and performance of each model is needed before implementation of DeepSeek-R1 or ChatGPT o1 to determine if the added financial costs of ChatGPT o1 are within the intended goals of improved accuracy and efficiency.

Introduction

Advancements in artificial intelligence (AI) have enabled large language models (LLMs) to function as chatbots with the capability to receive and answer medical

Sources of support: This work had no specific funding.

Research data are stored in an institutional repository and will be shared uponrequest to the corresponding author.

*Corresponding author: Charles B. Simone, II, MD, FASTRO, FACRO; Email: csimone@nyproton.com

https://doi.org/10.1016/j.adro.2025.101929

questions. Much research has focused on OpenAI's ChatGPT, which has been evaluated in licensing,[1,2] primary care,[3] and specialty examinations across internal medicine,[4] medical subspecialties,[5-7] and surgical subspecialties.[8-10] Each iteration of ChatGPT has demonstrated improved accuracy.

Aside from ChatGPT, other LLMs exist.[11] In comparison with other models' performance in the setting of examination questions, the latest version of ChatGPT has been reported to be equal or superior to the latest versions of other models such as Gemini/Bard,[12] Claude,[13] LLaMA,[14] PaLM,[15] Perplexity,[16] and BingAI.[17]

In oncology, LLMs have likewise been evaluated. Akin to literature in other spaces, ChatGPT was found to be increasingly accurate compared with newer models when answering examination questions.[18] In comparisons with other models, including Gemini, Claude, LLaMA, and PaLM, ChatGPT has consistently demonstrated the greatest accuracy.[19]

Recently, a new LLM named DeepSeek-R1 was released in January 2025, which is similarly a large language model with chatbot capabilities. Initially trained for mathematical reasoning, it was reported to match or exceed ChatGPT's performance while being more cost-effective in training.[20,21] However, its accuracy in the medical context remains untested.

This study evaluated the accuracy of DeepSeek-R1 in answering radiation oncology examination questions and compared its performance to that of different models of ChatGPT.

Methods and Materials

We used a similar methodology to our previously published study comparing different versions of ChatGPT's accuracy in the setting of radiation oncology examination questions.[18] In brief, we sourced 600 radiation oncology examination questions from consecutive years of a national radiation oncology in-service multiple-choice examination. These questions are used by medical residents in preparation for their certifying board examination and assess knowledge on anatomy, treatment planning, cancer epidemiology, and landmark trials. Approximately 1 quarter (26.8%) of questions pertain to radiation biology/pathology/anatomy, another 1 quarter (23.4%) to treatment recommendations/planning, one-fifth (19.1%) to medical physics, one-seventh (15.7%) to cancer epidemiology, and another one-seventh (15.0%) to knowledge of landmark studies.

Questions were input into DeepSeek-R1 using 1-shot learning, with this setting chosen over in-context learning or chain-of-thought reasoning as it best approximates pragmatic use of DeepSeek-R1. The Kluster.ai API and Azure AI Foundry's API were used for the DeepSeek-R1 model. Answers from DeepSeek-R1 were recorded and compared with the answer keys. The same questions were also input into prior models of ChatGPT, including ChatGPT 3.5 and ChatGPT 4o, as well as the most recent version of ChatGPT, ChatGPT o1, indicated to excel at reasoning tasks. One-shot learning was similarly used for each model.

To assess the efficiency and cost of each model, the total prompt and completion tokens used, as well as the total run time (performed on the same server) for each model and for all 600 consecutive questions, were recorded.

Descriptive statistics were used to present the total number of correct and incorrect answers across all questions and by question categories of (1) radiation biology/pathology/anatomy, (2) treatment recommendations/planning, (3) medical physics, (4) cancer epidemiology, and (5) knowledge of landmark studies. Pearson's χ^2 test and pairwise Z-test for proportions were used to assess for differences in accuracy by question categories and to compare DeepSeek-R1 to the different models of ChatGPT. Type I error was set at 0.05. All analyses were conducted using StataBE 18.0 (StataCorp).

Results

All 600 questions were answered by DeepSeek-R1 and ChatGPT o1. The total run time for DeepSeek-R1 was 9 hours, 47 minutes, and 35 seconds, costing 897,193 tokens (249,353 prompt tokens and 647,840 completion tokens); on average, it required 59 seconds per question. The total run time for ChatGPT o1 was 1 hour, 38 minutes, and 35 seconds, costing 837,235 tokens (272,808 prompt tokens and 564,427 completion tokens); on average, it required 10 seconds per question (Table 1).

DeepSeek-R1 answered 504 (84.0%) questions correctly. There was a significant difference in accuracy by question categories (P = .012) (Fig. 1). DeepSeek-R1 was

Table 1 Comparison of DeepSeek-R1 and ChatGPT o1, by resources

Model	Total run time (h:min:s)	Total token required
DeepSeek-R1	09:47:35	Prompt: 249,353 Completion: 647,840
ChatGPT o1	01:38:35	Prompt: 272,808 Completion: 564,427

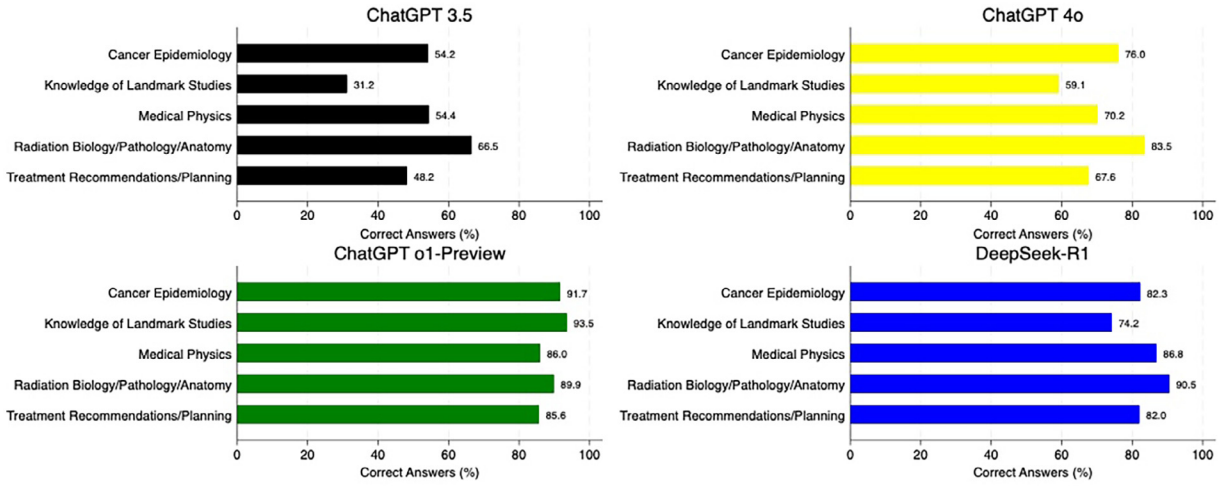

Figure 1 Comparison of the accuracy of AI models.

less accurate for questions about landmark studies (74.2% accuracy) compared with other categories. DeepSeek-R1 correctly answered over 80% of questions in all other categories; specifically, it answered correctly 90.5% of questions relating to radiation biology/pathology/anatomy, 86.8% relating to medical physics, 82.3% relating to cancer epidemiology, and 82.0% relating to treatment recommendations/planning.

In comparison, ChatGPT o1 answered 534 (89.0%) questions correctly. There was no significant difference in accuracy by question category (P = .241). ChatGPT o1 answered with 86.0% to 93.5% accuracy across all categories. In contrast to DeepSeek-R1, ChatGPT o1 achieved the highest accuracy in questions about landmark studies.

DeepSeek-R1 was significantly less accurate than ChatGPT o1 (P = .011), but it was significantly more accurate than preceding models of both ChatGPT 4o (72.2% accuracy; $P < .001$) and 3.5 (53.8% accuracy; $P < .001$) (Fig. 1).

Discussion

This is the first study to assess the accuracy of DeepSeek-R1 in medical knowledge as assessed using examination questions and to compare the accuracy of DeepSeek-R1 in this setting with that of ChatGPT. DeepSeek-R1 answered 84.0% of questions correctly, which is less accurate than the latest version of ChatGPT o1 but notably superior to preceding models of ChatGPT 4o and 3.5. DeepSeek-R1 was the least accurate in questions assessing knowledge of landmark studies, where the newest ChatGPT model excelled (93.5% accuracy for ChatGPT o1).

When comparing DeepSeek-R1 to ChatGPT's iteratively improved models, DeepSeek-R1's model falls short of the current model but outperforms predecessors. In isolation, DeepSeek-R1 may be considered an inferior model to ChatGPT. However, we would be remiss if we were to definitively characterize 84.0% accuracy as poor accuracy. In an examination setting, this 84.0% would be considered satisfactory competence for practicing specialized physicians.

Furthermore, when assessing the key differences in accuracy between DeepSeek-R1 and ChatGPT o1, DeepSeek-R1 continues to perform worse on questions related to landmark studies (74.2%), which is in keeping with the poor performance of prior ChatGPT models 4o (59.1%) and 3.5 (31.2%); ChatGPT o1 performs similarly well on questions related to landmark studies and other biology/physics/clinical practice questions. The difference between DeepSeek-R1 and ChatGPT o1's performance may perhaps be underpinned by a difference in knowledge of landmark trials. We have previously commented that the difference in knowledge is unlikely to be related to the cutoff dates for the models but rather related to knowledge synthesis, which has been improving with each iteration of ChatGPT.[18] Therefore, the current level of performance of DeepSeek-R1 relative to ChatGPT o1 is not unexpected, given DeepSeek-R1's focused development and its initial training for mathematical reasoning.

Of interest, DeepSeek-R1 requires significantly more time and tokens to answer the questions compared to ChatGPT o1. From a time perspective, DeepSeek-R1's nearly 10-hour run time is nearly 6 times longer than ChatGPT o1's 100-minute run time. Furthermore, ChatGPT used 7.2% fewer tokens. However, DeepSeek-R1 is financially less costly. When comparing the cost in US dollars at prices in January 2025, DeepSeek-R1 costs between $1.45 and $1.56 (at a rate of $0.14 per 1 million prompt tokens that are cache hit or $0.55 per 1 million prompt tokens that are cache miss and $2.19 per 1 million completion tokens), compared with ChatGPT o1's cost of

$37.96 (at a rate of $15 per 1 million prompt tokens and $60 per 1 million completion tokens).

When interpreted as a whole, DeepSeek-R1 is less accurate than ChatGPT o1, but it is financially less expensive. DeepSeek-R1, in contrast, requires substantially more time to run than ChatGPT o1. Ultimately, this creates a scenario where, borrowing from the health economics literature, neither model is dominant over the other (ie, there is not one model that is cheaper and more effective). Rather, careful analysis and consideration of the current performance and cost of each LLM before implementation is needed to achieve optimal alignment with the intended goals of improved accuracy and efficiency.

Disclosures

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.

References

1. 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.
2. Penny P, Bane R, Riddle V. Advancements in AI medical education: assessing ChatGPT's performance on USMLE-style questions across topics and difficulty levels. *Cureus*. 2024;16:e76309.
3. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's family medicine board exam. *J Chin Med Assoc*. 2023;86:762-766.
4. Kaneda Y, Tayuinosho A, Tomoyose R, et al. Evaluating ChatGPT's effectiveness and tendencies in Japanese internal medicine. *J Eval Clin Pract*. 2024;30:1017-1023.
5. Madrid-García A, Rosales-Rosado Z, Freites-Nuñez D, et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. *Sci Rep*. 2023;13:22129.
6. Miao J, Thongprayoon C, Garcia Valencia OA, et al. Performance of ChatGPT on nephrology test questions. *Clin J Am Soc Nephrol*. 2024;19:35-43.
7. Milutinovic S, Petrovic M, Begosh-Mayne D, et al. Evaluating performance of ChatGPT on MKSAP cardiology board review questions. *Int J Cardiol*. 2024;417:132576.
8. Guerra GA, Hofmann H, Sobhani S, et al. GPT-4 artificial intelligence model outperforms ChatGPT, medical students, and neurosurgery residents on neurosurgery written board-like questions. *World Neurosurg*. 2023;179:e160-e165.
9. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT −3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg*. 2023;31:1173-1179.
10. Patel J, Robinson P, Illing E, Anthony B. Is ChatGPT 3.5 smarter than Otolaryngology trainees? A comparison study of board style exam questions. *PLoS One*. 2024;19:e0306233.
11. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930-1940.
12. Du W, Jin X, Harris JC, et al. Large language models in pathology: a comparative study of ChatGPT and Bard with pathology trainees on multiple-choice questions. *Ann Diagn Pathol*. 2024;73:152392.
13. Deng L, Wang T, Yangzhang, et al., et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. *Int J Surg*. 2024;110:1941-1950.
14. Li DJ, Kao YC, Tsai SJ, et al. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan psychiatric licensing examination and in differential diagnosis with multi-center psychiatrists. *Psychiatry Clin Neurosci*. 2024;78:347-352.
15. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: a pilot study. *J Dent*. 2024;144:104938.
16. Sadeq MA, Ghorab RMF, Ashry MH, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Sci Rep*. 2024;14:18859.
17. Mete U. Evaluating the performance of ChatGPT, Gemini, and Bing compared with resident surgeons in the otorhinolaryngology in-service training examination. *Turk Arch Otorhinolaryngol*. 2024;62:48-57.
18. Chow R, Hasan S, Zheng A, et al. The accuracy of artificial intelligence ChatGPT in oncology examination questions. *J Am Coll Radiol*. 2024;21:1800-1804.
19. Rydzewski NR, Dinakaran D, Zhao SG, et al. Comparative evaluation of LLMs in clinical oncology. *NEJM AI*. 2024;1. https://doi.org/10.1056/aioa2300151.
20. Liu A, Feng B, Xue B, et al. DeepSeek-v3 technical report. Preprint. Posted online December 27, 2024. arXiv:2412.19437. doi: 10.48550/arXiv.2412.
21. Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. Preprint. Posted online January 22, 2025. arXiv:2501.12948. doi: 10.48550/arXiv.2501.12948.