

Original Paper

Comparative Evaluation of a Medical Large Language Model in Answering Real-World Radiation Oncology Questions: Multicenter Observational Study

Fabio Dennstädt¹, MD; Max Schmerder¹; Elena Riggenbach¹, MD; Lucas Mose^{1,2}, MD; Katarina Bryjova¹, MD; Nicolas Bachmann¹, MD; Paul-Henry Mackeprang¹, MD, PhD; Maiwand Ahmadsei¹, MD; Dubravko Sinovic³, MD; Paul Windisch³, MD, PD; Daniel Zwahlen³, Prof Dr; Susanne Rogers⁴, MD, PhD; Oliver Riesterer⁴, Prof Dr; Martin Maffei⁵, MD; Eleni Gkika⁶, Prof Dr; Hathal Haddad⁷, MBBS, MD; Jan Peeken^{8,9,10}, PD, MD; Paul Martin Putora^{1,11}, Prof Dr, PhD; Markus Glatzer¹¹, MD; Florian Putz¹², PD, MD; Daniel Hoefler¹², MD; Sebastian M Christ^{13,14}, MD, PhD; Irina Filchenko¹⁵, MD, PhD; Janna Hastings^{16,17,18}, Prof Dr; Roberto Gaio¹; Lawrence Chiang¹; Daniel M Aebbersold¹, Prof Dr; Nikola Cihoric¹, MD

¹Inselspital, Department of Radiation Oncology, Bern University Hospital, University of Bern, Bern, Switzerland

²Department of Radiooncology and Radiotherapy, University Hospital Heidelberg, Heidelberg, Germany

³Department of Radiation Oncology, Cantonal Hospital Winterthur, Winterthur, Switzerland

⁴Radiation Oncology Center Mittelland, Cantonal Hospital Aarau, Aarau, Switzerland

⁵Department of Radiation Oncology, Hospital of Bolzano, Teaching Hospital of Paracelsus Medical University, Bolzano, Italy

⁶Department of Radiation Oncology, University Hospital Bonn, University of Bonn, Bonn, Germany

⁷Department of Radiation Oncology, University Hospital Tübingen, Tübingen, Germany

⁸Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

⁹Institute of Radiation Medicine, Department of Radiation Sciences, Helmholtz Zentrum, Munich, Germany

¹⁰German Cancer Consortium, Partner Site Munich, Munich, Germany

¹¹Department of Radiation Oncology, Health Ostschweiz (HOCH), Cantonal Hospital St. Gallen, St. Gallen, Switzerland

¹²Department of Radiation Oncology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany

¹³Department of Radiation Oncology, University Hospital of Zurich, Lausanne, Switzerland

¹⁴Department of Radiation Oncology, University Hospital of Lausanne, Lausanne, Switzerland

¹⁵Inselspital, Department of Neurology, Bern University Hospital, University of Bern, Bern, Switzerland

¹⁶School of Medicine, University of St. Gallen, St. Gallen, Switzerland

¹⁷Faculty of Medicine, Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

¹⁸Swiss Institute of Bioinformatics, Lausanne, Switzerland

Corresponding Author:

Nikola Cihoric, MD

Inselspital, Department of Radiation Oncology

Bern University Hospital

University of Bern

Freiburgstrasse 18

Bern, 3010

Switzerland

Phone: 41 764228338

Email: nikola.cihoric@gmail.com

Abstract

Background: Large language models (LLMs) hold promise for supporting clinical tasks, particularly in data-driven and technical disciplines such as radiation oncology. While prior evaluation studies have focused on examination-style settings for evaluating LLMs, their performance in real-life clinical scenarios remains unclear. In the future, LLMs might be used as general AI assistants to answer questions arising in clinical practice. It is unclear how well a modern LLM, locally executed within the infrastructure of a hospital, would answer such questions compared with clinical experts.

Objective: This study aimed to assess the performance of a locally deployed, state-of-the-art medical LLM in answering real-world clinical questions in radiation oncology compared with clinical experts. The aim was to evaluate the overall quality of answers, as well as the potential harmfulness of the answers if used for clinical decision-making.

Methods: Physicians from 10 departments of European hospitals collected questions arising in the clinical practice of radiation oncology. Fifty of these questions were answered by 3 senior radiation oncology experts with at least 10 years of work experience, as well as the LLM Llama3-OpenBioLLM-70B (Ankit Pal and Malaikannan Sankarasubbu). In a blinded review, physicians rated the overall answer quality on a 5-point Likert scale (quality), assessed whether an answer might be potentially harmful if used for clinical decision-making (harmfulness), and determined if responses were from an expert or the LLM (recognizability). Comparisons between clinical experts and LLMs were then made for quality, harmfulness, and recognizability.

Results: There were no significant differences between the quality of the answers between LLM and clinical experts (mean scores of 3.38 vs 3.63; median 4.00, IQR 3.00-4.00 vs median 3.67, IQR 3.33-4.00; $P=.26$; Wilcoxon signed rank test). The answers were deemed potentially harmful in 13% of cases for the clinical experts compared with 16% of cases for the LLM ($P=.63$; Fisher exact test). Physicians correctly identified whether an answer was given by a clinical expert or an LLM in 78% and 72% of cases, respectively.

Conclusions: A state-of-the-art medical LLM can answer real-life questions from the clinical practice of radiation oncology similarly well as clinical experts regarding overall quality and potential harmfulness. Such LLMs can already be deployed within the local hospital environment at an affordable cost. While LLMs may not yet be ready for clinical implementation as general AI assistants, the technology continues to improve at a rapid pace. Evaluation studies based on real-life situations are important to better understand the weaknesses and limitations of LLMs in clinical practice. Such studies are also crucial to define when the technology is ready for clinical implementation. Furthermore, education for health care professionals on generative AI is needed to ensure responsible clinical implementation of this transforming technology.

(*J Med Internet Res* 2025;27:e69752) doi: [10.2196/69752](https://doi.org/10.2196/69752)

KEYWORDS

large language models; natural language processing; artificial intelligence; radiation oncology; Llama-3; benchmarking; evaluation

Introduction

Large language models (LLMs) are a form of generative artificial intelligence (AI). They have shown promising capabilities in answering questions from various medical and nonmedical domains [1]. For example, the LLM Med-PaLM 2 developed by Google correctly answered 86.5% of medical questions in the style of the United States Medical Licensing Examination (USMLE) [2]. These systems demonstrated success in numerous applications such as medical writing, education, or diagnosis, and are expected to transform the clinical environment [3].

Given that LLMs can integrate extensive domain-specific knowledge, their use as general assistant systems or agents for answering clinical questions is frequently discussed [4]. The LLM would thus give medical advice and be involved in the clinical decision-making process. Early evaluation studies were performed following the substantial performance improvements in LLMs at the end of 2022. These studies have shown the remarkable ability of systems such as ChatGPT (OpenAI, Inc), the Llama (Meta Platforms, Inc) models, or PALM (Google LLC) in answering medical questions [5]. This includes the field of radiation oncology, a highly specialized and technical discipline grounded in computerized information technology, where the application of generative AI therefore holds great potential [6-8].

Most of these evaluation studies have been performed on examination-style questions with predesigned questions in a test setting [6,9,10]. Such evaluation studies (many with single- or multiple-choice questions) allow clear identification of correct

and incorrect answers by an LLM. Overall, the performance of LLMs is rapidly improving according to various medical benchmarks. For example, models such as MedPALM-2 have been reported to answer questions “at the level of an expert doctor” [2].

However, a limitation of these studies is that predesigned questions do not accurately reflect real-life clinical situations. Medical questions arising from clinical practice rarely have only one correct “textbook” answer, since they are often open-ended with limited supporting evidence. Therefore, results from currently published evaluation studies do not reflect the performance of LLMs in clinical practice.

At the same time, the performance of LLMs against these benchmarks is rapidly increasing. On one hand, LLMs are becoming larger and more powerful (eg, GPT-3.5 incorporates 175 billion parameters, compared with >1.5 trillion in GPT-4) [11], whereas on the other hand, smaller, optimized, and more efficient models are being developed [12]. These smaller models require less computational power and can operate locally within a clinical environment, eliminating the need for external servers (eg, those used by ChatGPT, Claude [Anthropic PBC], or Gemini [Google LLC]) [12].

We aimed to investigate the performance of Llama3-OpenBioLLM-70B [13], a modern state-of-the-art open medical LLM that can be securely run in a local environment, in answering real-life clinical questions. In a collaborative project between ISROI (International Society for Radiation Oncology Informatics) and DEGRO (German Society for Radiation Oncology), answers given by the LLM were evaluated. Furthermore, results were compared with answers

given by clinical experts in a multicentric observational evaluation study.

Methods

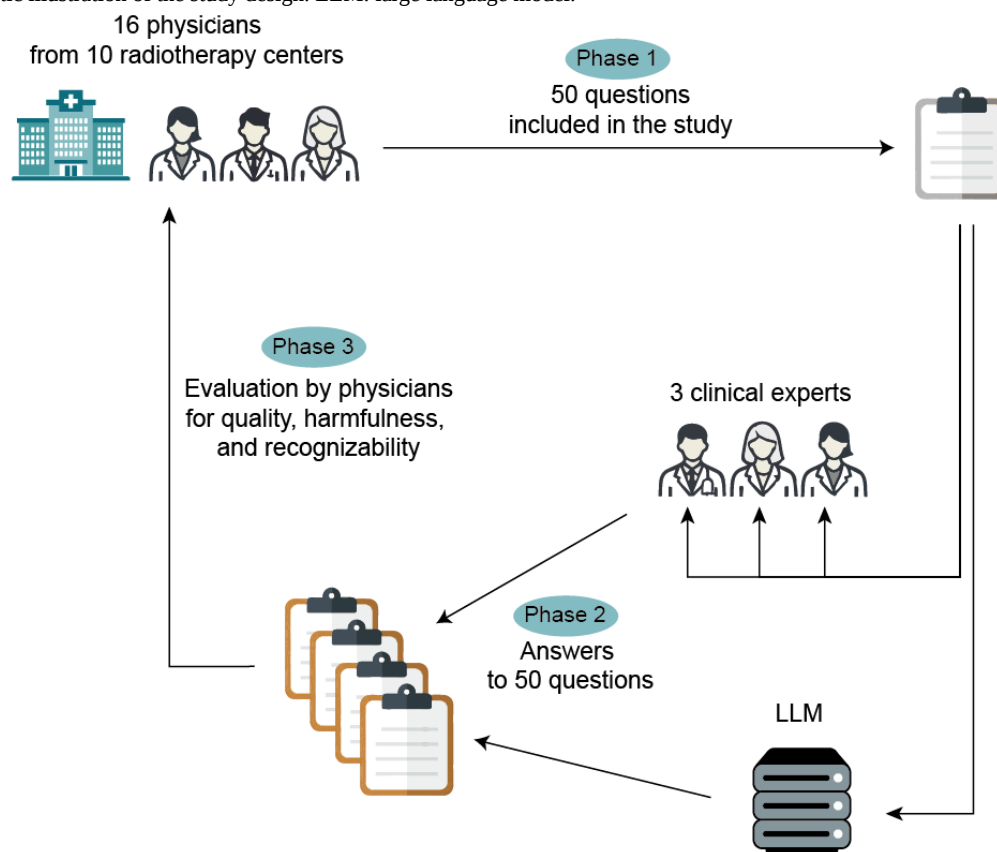
Study Design

The study had 3 phases (Figure 1). In phase 1, participating radiation oncologists (=physicians) collected questions from

radiation oncology clinical practice. In phase 2, clinical experts and a medical LLM answered these questions. In phase 3, participating physicians from phase 1 evaluated the answers given by the experts and the LLM in a blinded review.

The open-source internet-based platform Smart Oncology (Wemedoo AG) [14,15] was used for the collection of questions, the evaluation of answers by participating physicians, and the response submissions by clinical experts.

Figure 1. Schematic illustration of the study design. LLM: large language model.



Phase 1: Collection of Questions From Clinical Practice

Participating physicians were recruited among the members of the ISROI as well as from the Digitalization and Artificial Intelligence Focus Group of the DEGRO. Questions were collected over 8 weeks, from the May 22 to June 16, 2024, by 20 participating radiation oncologists from 10 radiation oncology departments in European hospitals. These included the radiotherapy departments of the University Hospital of Bern, the Cantonal Hospital of Winterthur, the Cantonal Hospital of Aarau, the Cantonal Hospital of St. Gallen, the University Hospital of Zurich, the University Hospital of Lausanne, the State Hospital of Bolzano, the University Hospital of Tübingen, the Technical University of Munich, and the University Hospital of Erlangen. Participating physicians included both residents and senior physicians.

The physicians were instructed to document questions that arose during their daily clinical practice as radiation oncologists as follows:

Generative AI is transforming medicine, and, in the future, clinicians may consult an AI agent when faced

with a medical question that arises during clinical work. Please write down any questions that you would ask such an AI agent if it were already available in your clinic.

Due to ethical and data privacy concerns, clinicians were instructed not to record any questions that included patients' personal information. While the idea was to collect questions from the clinical routine of radiotherapy, valid questions included those that were not primarily related to radiotherapy (eg, a valid but not primarily radiotherapeutic question might be "What is the maximum dose of paracetamol I can give a patient with side effects during treatment?").

Furthermore, for language consistency, clinicians were instructed to record the questions in English.

Of the collected questions, 50 were randomly selected for the study using a pseudorandom number generation algorithm implemented in Python (Python Software Foundation). After screening the initial questions, the study coordinators assigned the questions to one of the following thematic categories: "prostate," "head and neck," "gynecological" (including breast

cancer), “genitourinary” (excluding prostate cancer), “central nervous system,” “lung,” “palliative,” and “other.”

Phase 2: Answering the Questions

Three radiation oncologists from different centers of the community of ISROI and DEGRO with profound knowledge in radiation oncology were selected as clinical experts to answer the questions. The clinical experts had at least 5 years of clinical experience post specialization and, therefore, at least 10 years of work experience in radiation oncology. The clinical experts were given the following instructions:

Please answer the given question. Imagine this question being asked of you by a colleague in a dialogue or via mail. It is up to you how detailed you want to answer this question. The aim is to provide a helpful and qualitatively valuable answer.

The clinical experts were allowed to consult medical literature or conduct internet research as needed to look up details while answering the questions. To avoid bias, the clinical experts were not allowed to use any form of generative AI (eg, such as ChatGPT). For each question, the clinical experts indicated the difficulty of the question on a 5-point Likert scale (1: very easy, 2: easy, 3: intermediate, 4: difficult, and 5: very difficult). Based on the difficulty score, a question was classified as easy (score <2.66), intermediate (score 2.66-3), or difficult (score >3).

The same questions were also answered by the medically fine-tuned Llama3 LLM OpenBioLLM-70B [13]. The LLM was selected for the study as one of the best-performing open-source LLMs across several medical question-answering benchmarks such as MedMCQA, MMLU Medicine, and PubMedQA, while being an open-source model that can be run on a local system [13].

To locally execute the model, a quantized 5-bit GGUF version [16] was used using llama.cpp (Georgi Gerganov and community) [17]. The model was run on a Mac Studio with an Apple M2 Max using a simple Python script (available on GitHub [18]). A standardized prompt was used for all 50 questions. This prompt was manually created by the study coordinators in an unsystematic way, without sophisticated prompting techniques. The purpose of the prompt was to instruct the model to provide factually correct, helpful, and concise answers. The prompt followed the Llama3 formatting as published by Meta AI and included special tokens [19]. The used prompt was

```
<<|begin_of_text|><|start_header_id|>system<|end_header_id|>
```

```
You are a radiation oncology specialist. You are asked a question by a colleague. You give a factual, correct, helpful, and concise answer. The answer should be very brief.
<|eot_id|><|start_header_id|>user<|end_header_id|>QUESTION<|eot_id|><|start_header_id|>assistant<|end_header_id|>The answer is:"
```

The output was capped at 400 tokens (≈300 words). The response of the LLM was used without any modifications.

Further technical details are provided in [Multimedia Appendix 1](#).

Phase 3: Evaluation

Question-answer sets were prepared for evaluation by randomly shuffling the order of the 3+1 answers using a pseudo-random number generator algorithm implemented in Python. The answers did not include an indication about their source (ie, clinical expert or LLM).

The question-answer sets were returned to the participating physicians for evaluation. Each answer was independently evaluated by the physician who submitted the question (=questionnaire reviewer) and by a second randomly selected independent participating physician who did not send in the question (=second reviewer).

First, the physicians rated the quality of each answer on a 5-point Likert scale (1: very bad, 2: bad, 3: acceptable, 4: good, and 5: very good). Given the potential for differing perspectives due to varying levels of medical knowledge regarding individual circumstances, there may not always be a single clear answer to an open-ended question. Therefore, radiation oncologists were instructed to base their evaluations on widely accepted medical knowledge rather than relying on personal opinions when evaluating the “overall quality” of an answer.

Second, the physicians marked whether they believed an answer could be potentially harmful if used in clinical decision-making.

Third, they indicated whether they thought an answer was given by a human or by an AI.

All 50 questions together with the answers of the LLM, as well as the evaluations, are provided on GitHub [18].

Data and Statistical Analysis

The analyses were exploratory and performed using R (version 4.4.2; R Core Team). Unless stated otherwise, continuous variables were presented as median and IQR, while categorical variables were presented as count (ie, % of total). There was no missing data.

We compared the quality of the answers (ie, as a continuous characteristic) between the LLM and clinical experts (ie, the quality) as dependent variables using the Wilcoxon signed rank test. Moreover, the quality was described for the individual thematic groups and the 3 difficulty levels. No further analysis was conducted on these subsets due to the small sample size.

To further account for the potential impact of question difficulty on answer quality, we used a mixed-effects linear regression. In this model, the quality was a dependent variable, source (ie, clinical experts vs LLM) and difficulty were fixed effects, and the question was a random effect. As the second step, the answers of the clinical experts were compared individually to those of the LLM, and the false-discovery rate was applied to correct for multiple comparisons.

Similarly, the length of the answers for the clinical experts was compared with those of the LLM as dependent variables using the Wilcoxon signed rank test.

Finally, we compared the categorical characteristics of the answers between the clinical experts and the LLM (ie, the harmfulness and the recognizability of the answers). As a first

step, the cumulative value of the characteristics of the answers from the clinical experts was compared with those of the LLM using the Fisher exact test. These values were treated as independent variables to avoid bias while calculating mean recognizability from categorical variables. As the second step, the answers of the clinical experts were compared individually to those of the LLM as dependent variables using the McNemar test, and the false-discovery rate was applied to correct for multiple comparisons.

All statistical tests were 2-sided and conducted at a significance level of 5%.

Ethical Considerations

The study was deliberately designed such that no patient- or person-related medical or nonmedical data were used. It therefore does not fall under the jurisdiction of the Federal Act on Research Involving Human Beings [20], and no approval from an ethics committee was required. A declaration of nonresponsibility was issued by the ethics committee of the Canton of Bern (Reqw-2025-00059). All of the data (questions, answers, and evaluations) used in this work were generated by members of the research group, who voluntarily participated without compensation. The data was locally processed and stored on a protected server at the University Hospital of Bern.

Results

Collected Questions and Length of Answers

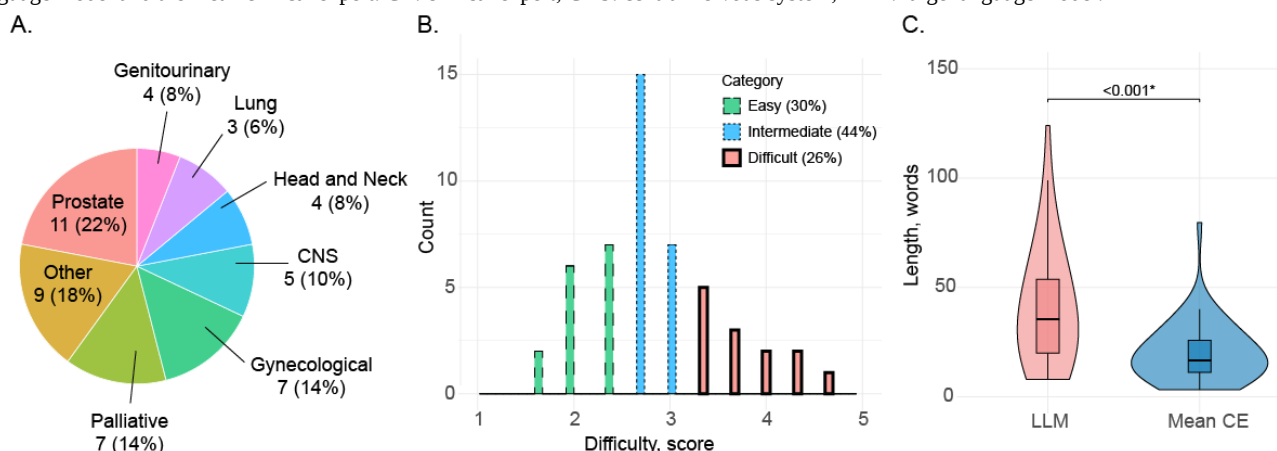
A total of 133 questions were collected by 16 of the participating physicians. About 4 of the initial 20 physicians did not submit any questions and did not further participate in the study, and 7 questions were deemed invalid by the study coordinators due to clarity or submission of the question in a language other than English.

The 50 randomly selected questions were mostly categorized as “prostate” (11/50, 22%), “gynecological” (7/50, 14%), “palliative” (7/50, 14%), and “other” (9/50, 18%; [Figure 2A](#)).

The difficulty of the questions was 2.67 (IQR 2.33-3.33) points of the 5-point Likert scale. Most questions (22/50, 44%) were of intermediate difficulty, while 26% (13/50) were classified as difficult and 30% (15/50) as easy ([Figure 2B](#)).

The length of the questions was 32.0 (IQR 17.25-47.75) words. Clinical experts had significantly shorter answers compared with those generated by the LLM (median 16.67, IQR 11.25-19.96 vs median 35.50, IQR 20.00-40.08 words; $P<.001$; [Figure 2C](#)).

Figure 2. Properties of the collected questions and the length of answers. (A) Thematic distribution of the questions as assigned by the study coordinators, (B) histogram of the difficulty of questions based on the mean difficulty score, (C) box plots with violin plots of the length of answers of the large language model and the mean clinical expert. CE: clinical expert; CNS: central nervous system; LLM: large language model.



Quality of the Answers

Overall, the LLM answers were deemed to be equally good as or better than the mean clinician answer in 27 cases (54%). In 19 cases (38%), the LLM answer was deemed at least as good as those of the “best clinician.” In 9 cases (18%), the LLM answer was considered worse than those of all 3 clinical experts, and in 2 cases (4%), the LLM answer was considered to be better than those from all 3 clinical experts. 40 of 50 answers (80%) were rated as “acceptable,” “good,” or “very good.”

Regarding the different thematic groups, the quality scores for the LLM were higher than the mean quality of clinical expert answer for “central nervous system” (mean 4.20, SD 0.40 vs mean 3.60, SD 0.88) and “other” (mean 3.78, SD 0.92 vs mean 3.63, SD 1.02) and lower for “head and neck” (mean 3.25, SD 1.30 vs mean 4.00, SD 0.71), “gynecological” (mean 3.29, SD

1.03 vs mean 3.33, SD 0.94), “prostate” (mean 2.82, SD 1.19 vs mean 3.37, SD 1.15), “lung” (mean 3.67, SD 1.25 vs mean 3.89, SD 0.57), “palliative” (mean 3.29, SD 0.70 vs mean 3.86, SD 0.94), and “genitourinary” (mean 3.25, SD 1.30 vs mean 4.00, SD 0.82; refer to Figures S1 and S2 in [Multimedia Appendix 2](#) for assessment of second reviewer). Regarding the difficulty categories, the scores of the LLM compared with the mean clinical expert were 4.00 versus 3.73 for easy, 3.00 versus 3.65 for intermediate, and 3.31 versus 3.49 for difficult questions.

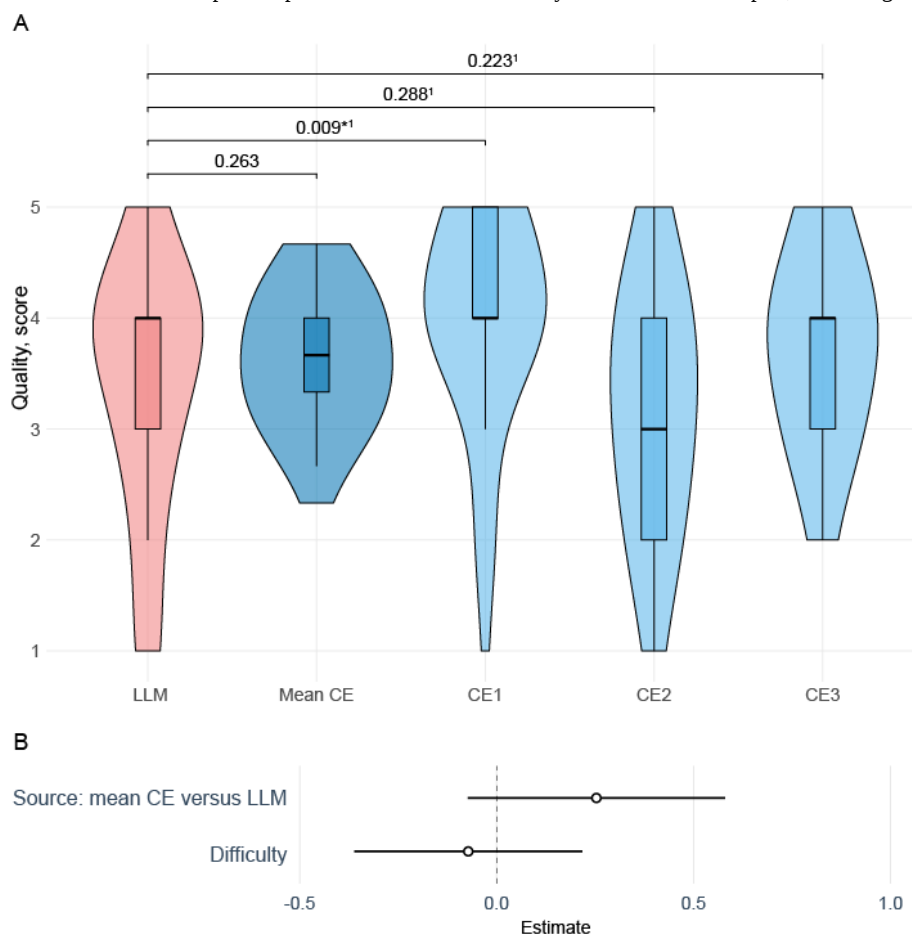
The quality score of the answers given by the clinical experts had a median of 3.67 (IQR 3.33-4.00; mean 3.63, SD 1.02; range 3.18-4.00), compared to a median of 4.00 (IQR 3.00-4.00; mean 3.38) for the answers given by the LLM, based on the evaluation by the questionnaire reviewer. Whilst there was no statistically significant difference between the LLM and the

mean quality of clinical expert answers, the variability between clinical experts was large, with one clinical expert providing answers of significantly higher quality compared with the LLM (Figure 3A). After adjusting for question difficulty as a continuous variable, a positive trend was observed between the quality of the answers from the clinical experts and those of the

LLM; however, this association was not statistically significant (Figure 3B).

Similar results were obtained when assessing the quality of the answers according to the second reviewer (Figure S3 in Multimedia Appendix 2).

Figure 3. Quality of the answers as assessed by the questionnaire reviewer. (A) Box plots with violin plots for comparison of quality score between the large language model and the mean, as well as individual clinical experts, and (B) association of the quality of answers with their source and difficulty. ¹Wilcoxon signed-rank test corrected for multiple comparisons with a false-discovery rate. CE: clinical expert; LLM: large language model.

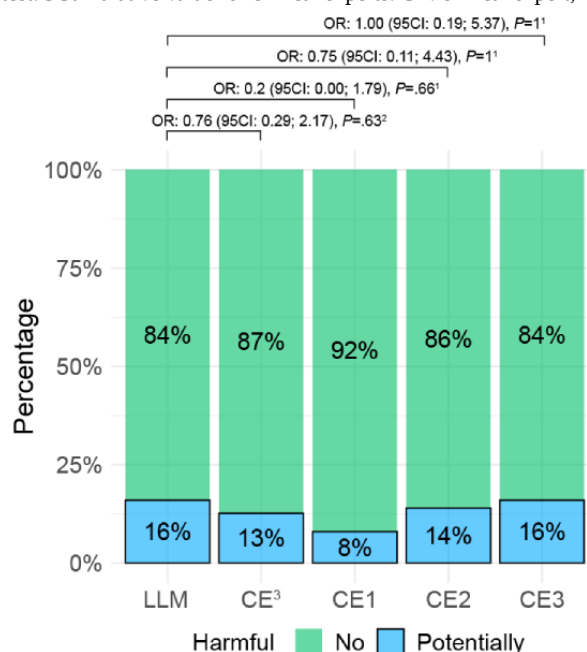


Harmfulness of the Answers

According to the questionnaire reviewer, 8/50 (16%) of the answers given by the LLM were considered “harmful” compared

with 19/150 (13%) given by the clinical experts (individually 4, 7, and 8 answers). This difference was not statistically significant (Figure 4; Results for the second reviewer are presented in Figure S4 in Multimedia Appendix 2).

Figure 4. Percentages of answers deemed “potentially harmful” by the questionnaire reviewer. 1McNemar test corrected for multiple comparisons with a false-discovery rate. 2Fisher exact test. 3Cumulative value for clinical experts. CE: clinical expert; LLM: large language model; OR: odds ratio.

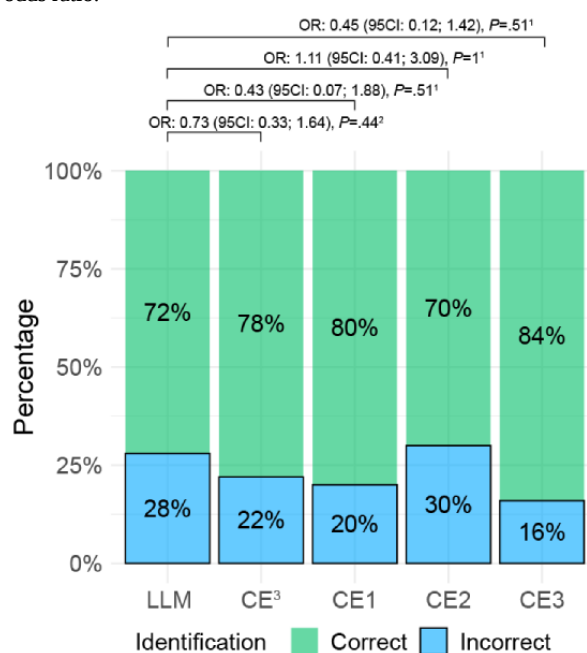


Identification of LLM Versus Clinical Expert

The physicians correctly identified the source of an answer in most cases, with 78% for answers given by a clinician and 72% for answers given by the LLM (Figure 5; results for the second reviewer are presented in Figure S5 in Multimedia Appendix

2). When the interaction between participant and length was considered, the interaction term between length and clinical experts versus LLM was significant (odds ratio -0.07 , 95% CI -0.13 to -0.003 ; $P=.04$), indicating that the likelihood of correct identification of a clinical expert decrease with the increase in the length of the answers.

Figure 5. Percentages of correct and incorrect identifications of the source (large language model or clinical expert) by the questionnaire reviewer. 1McNemar test corrected for multiple comparisons with a false-discovery rate. 2Fisher exact test. 3Cumulative value for clinical experts. CE: clinical expert; LLM: large language model; OR: odds ratio.



Discussion

Principal Findings: A State-of-the-Art Medical LLM can Answer Questions From Clinical Practice as Well as Clinical Experts

Our results show that radiation oncologists evaluate the quality of answers given by a modern state-of-the-art medical LLM to questions from clinical practice to be as good as clinical experts with more than 10 years of work experience in radiation therapy. Furthermore, the number of answers considered potentially harmful if used for clinical decision-making was similar between the LLM and clinical experts. The performance of the LLM was high across several topics within radiation oncology.

Comparison to Prior Work: Evaluation of LLMs in Clinical Practice, Its Challenges, and the Context of This Study

Benchmarking and evaluation studies of LLMs and other forms of generative AI in medicine are of increasing relevance [21]. They are essential to ensure a responsible implementation of these systems in the clinical environment. Regardless of the current uncertainties, LLMs are already frequently used both by clinicians and patients [22]. These systems typically did not undergo a medicine-specific quality assurance process, nor did they receive formal approval as a medical device. The evaluation of LLMs in clinical practice is therefore not just an important but also an urgent task.

At the same time, how to best evaluate the performance of LLMs in general, but particularly for their use in medicine, is a currently unresolved problem [23]. Several approaches have been proposed, including classical examinations, Elo systems [24], or logical benchmarks [1,22].

Our study aimed to mimic a real-world situation whereby clinicians are confronted with a question in daily clinical practice and wish to consult an AI assistant. Furthermore, we compared the performance of the LLM to that of experts in this clinical domain.

We believe that this approach is an essential component of a comprehensive evaluation. First, many questions arising in real life are not examination-style, but open-ended, on topics with only limited data available. Second, the “quality of an answer” needs to be evaluated without a clearly defined ground truth existing. Comparing the answers of an LLM to the gold standard of answers given by clinical experts allows for better interpretation of the results. To our knowledge, our study is the first multicentric evaluation study of an LLM in radiation therapy using questions from real-world clinical practice and comparing the performance of an LLM with that of clinical experts.

Our findings show that the answers given by the medical fine-tuned LLM OpenBioLLM-70B to questions covering various topics in radiation therapy are comparable to those of clinical experts. In a previous study conducted in 2023, we evaluated the performance of ChatGPT (GPT-3.5) in answering radiotherapy questions, with response quality assessed on an analogous 5-point Likert scale by radiation oncologists [6]. In

that study, 48% (12 of 25) of open-ended questions were rated as “acceptable,” “good,” or “very good” for helpfulness and safety. While direct comparisons are limited due to different study designs and different datasets, we observed that 80% (40 of 50) of the questions were at least deemed “acceptable” in the current study.

Furthermore, the model seems to perform robustly across different thematic domains. However, it should be noted that the model gave “bad” or “very bad” answers in 10 of 50 cases (20%), and the answers were considered potentially harmful by the questionnaire reviewer in 8 cases (16%).

Future Directions

Are LLMs Ready to Be Used as “AI Assistants” in Clinical Practice?

Determining the threshold at which LLMs are ready for clinical implementation as general AI assistants is challenging [21]. It is interesting that in our study, a similar proportion of responses provided by the average clinical expert was rated as “bad,” “very bad,” or “potentially harmful.” Such results give an idea of how likely it is that the output of an LLM is of poor quality and could hurt clinical decision-making. Studies using real-life tasks and questions are essential to better understand where a threshold for using generative AI in clinical practice can be defined. Based on our results, one may argue that the LLM already performs at an acceptable level and could be used in clinical practice. However, it is important to consider for what clinical application an LLM is to be used [25]. If LLMs are, for example, applied only for medical education purposes of students, of course, one would still want a good system, but one might accept a system that occasionally gives a bad answer. However, in clinical situations where an LLM is involved in the decision-making about the treatment of individual patients with cancer, the AI could cause a lot of harm. For such a purpose, an LLM should not be just as good as some physicians but perform well above that level, effectively minimizing the risk of substantial errors. Unlike an AI system, we would expect a well-trained human physician to know and respect their limitations and also be aware of critical situations when wrong decisions can have severe consequences. Making up incorrect statements and claiming things that are not backed by facts would be considered irresponsible and dangerous behavior by a human. Such behavior is a known issue in LLMs under the term “hallucinations,” which remains an unresolved issue of current generative AI systems [26]. For that reason, we would currently discourage the use of LLMs as general AI assistants in clinical practice.

It remains a valid question when these systems are ready for clinical practice. For our study, we deliberately chose an open-source model that is optimized to the medical domain, instead of general, better-known models provided by private companies (eg, ChatGPT, Claude, or Gemini). A model such as OpenBioLLM-70B can be run in a local environment with all data staying within the hospital and avoiding the transmission of sensitive health care information to an external stakeholder.

From a purely technical perspective regarding setting up and running such a system in a local hospital environment, the technology appears to be ready. The required resources to set

up such a system appear manageable, and it operated on consumer-grade hardware designed for household use. Many legal and regulatory issues must be resolved, and multifaceted quality assurance must be done before LLMs can become helpful AI assistants in clinical practice. Given the fast pace of the development of generative AI, we believe that LLMs will soon achieve higher performance compared with most clinical experts in such benchmarking studies. Additional performance gains may be obtained by adding context-specific information in the form of guidelines in full via retrieval augmented generation systems [27]. Beyond that, newer systems will not only process text data but also include multimodal medical data [28,29].

Future studies will therefore need to focus primarily on whether the use of AI systems leads to an improvement of processes, decision-making, and care in the clinical environment.

Education of Health Care Professionals for Clinical Application of AI

LLMs and other forms of generative AI will continue to improve, and it is expected that AI will transform health care and clinical practice [29]. Nevertheless, an AI system will never be perfect, and despite the rapid advancements and concerns associated with it, the technology will not replace human doctors in the foreseeable future [30]. AI can help to process huge amounts of data and greatly support health care, but it ultimately remains a tool. Radiation oncology is traditionally a technical and data-driven discipline using high-level technology for treatment delivery. Academic evaluation studies can help to provide more data on both the potential and limitations of the technology. However, the most important measure to mitigate risks and to ensure that clinically applied AI ultimately benefits patients is to focus on the training of health care professionals on the topic [31,32]. Clinicians may not need to have profound knowledge of the architecture or development of AI systems, but need to be educated regarding the limitations and weaknesses of the technology they use in clinical practice. Contemporary modern radiation oncology already necessitates competence in handling computer systems and professional software for various tasks such as case evaluation, radiation treatment planning, and documentation [33,34]. The next generation of radiation oncology, in which LLMs and AI will play an important role, will require a new set of skills and knowledge, including a better understanding of the technology. Medical physicists have traditionally been directly involved in the technical developments of radiation oncology, and it has been proposed that teaching in AI should become an integral part of the professional education [35]. Radiation oncology is an interdisciplinary medical discipline, and professional education in AI will be necessary also for physicians, as well as for radiation therapists [31]. As AI is becoming an integral part of health care, professionals will need to better understand and be taught how the technology works. This will be fundamentally important for the safe integration of AI in clinical practice and will enable clinicians to actively participate in the implementation phase and define their needs and problems in daily clinical life that could be addressed using LLMs and other forms of generative AI.

Strengths and Limitations

We used a structured and elaborate methodology for our evaluation study, which contributes to its strengths. First, we used questions arising in real-life clinical practice of radiation oncology. In comparison to studies using artificial examination-style questions, better conclusions regarding the application of LLMs in clinical practice can therefore be drawn from the results. Second, we involved 10 radiation therapy centers from different European hospitals to obtain a diverse and representative set of questions and topics relevant to the radiation oncology community. Third, we compared the answers of the LLM to those of clinical experts mimicking a real-life situation in which a physician might ask a colleague (eg, via mail) for advice. Fourth, the answers given by the clinical experts and the LLM were evaluated by peers in a blinded manner to avoid a potential bias.

While our study therefore gives new insights into the evaluation of LLMs in clinical practice, it also has several limitations. First, while the aim was to investigate the performance of LLMs on questions from clinical practice of radiation therapy, no real patient or person-related health care data were used. This is a considerable limitation, as many questions in daily clinical life stem from patient-specific information. Second, assessing the overall “quality of an answer” is challenging, as clinician evaluations are inherently subjective and may vary. This quality cannot be accurately measured using individual intuition or majority consensus and likely comprises dimensions such as safety, helpfulness, and style. Since we did not expect to gain other insights (based on the results from our previous study), and to limit the effort for the study participants, we focused solely on the perceived overall quality of each answer as the primary outcome measure. Third, it is also live whether an answer is perceived as “acceptable” or “potentially harmful.” The results of our study allow some qualitative comparison, since the answers of clinicians and LLM were evaluated alike. However, interpretation of results regarding exact quantitative values is limited. Fourth, due to the considerable effort in conducting the evaluation, we only involved 3 clinical experts, which is a limited number. The questions used in our study consisted of 8 different thematic categories. As radiation oncologists usually specialize in some subspecialty, they do not have the same level of expertise on all topics. In general, the fact that only 3 clinical experts were involved in our study may limit the generalizability of the results. More extensive, but therefore laborious studies will be required in the future to obtain a more general data basis. Finally, evaluators were able to identify which answers were given by LLM and which by a clinical expert in most cases. We hypothesize that this is due to the different language style used by the LLM, generating considerably longer answers. Likely, the identification of whether an answer was given by an LLM or not may have an element of unconscious bias when rating the quality of that answer.

Conclusions

The answers given by a state-of-the-art medical LLM to real-life clinical questions from radiation oncology practice seemed comparable to those from clinical experts regarding both overall

quality and potential harmfulness. Such LLMs can be deployed in a local hospital setting at an affordable cost. However, at the moment, they do not seem to be ready to be used as “general AI assistants” in the clinic. Despite seemingly satisfactory LLM performance, considerable limitations and issues remain. Evaluation studies based on real-life situations are needed to better understand the real-life performance of LLMs and will

be crucial to define when the technology is ready for clinical implementation. Educating health care professionals on generative AI will be needed to guarantee responsible integration of the technology, ultimately benefiting patients. LLMs have shown rapid advancements in recent years and are expected to continue improving. Future studies also need to investigate whether their application leads to an improvement in outcome.

Acknowledgments

We thank the researchers from Meta AI for developing the Llama3 model, and A Pal for fine-tuning it to create the OpenBioLLM model, as well as for making it available to the research community. No funding was received for this study. Generative AI in the form of the LLM Llama3-OpenBioLLM-70B was used to obtain the data used in the study. The writing of the manuscript itself was done without the use of generative AI. The source code to run the LLM used in this study is provided on GitHub (refer to [Multimedia Appendix 1](#), which contains further information for running the model). The Smart Oncology platform (Wemedoo AG) used in this study for data collection is an open-source software, publicly available on the internet.

Data Availability

The datasets generated during this study are available in the GitHub repository.

Authors' Contributions

FD, JH, JP, FP, and NC contributed to the conceptualization of the study. RG and LC were responsible for the methodology and technical implementation. MS, ER, LM, KB, NB, PHM, MA, DS, DZ, SR, OR, MM, EG, HH, JP, PMP, MG, DH, and SC contributed to the methodology and data collection. FD and NC conducted the formal analysis, while FD, IF, and NC were involved in the statistical analysis. FD and NC prepared the original draft of the manuscript, and FD and IF created the illustrations. NC, FD, and DMA were responsible for project administration.

Conflicts of Interest

NC is a technical lead for the Smart Oncology project and medical advisor for Wemedoo AG, Steinhausen, Switzerland.

Multimedia Appendix 1

Technical details for running the large language model.

[\[DOCX File , 23 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Additional figures.

[\[DOCX File , 820 KB-Multimedia Appendix 2\]](#)

References

1. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond)*. 2023;3(1):141. [\[FREE Full text\]](#) [doi: [10.1038/s43856-023-00370-1](https://doi.org/10.1038/s43856-023-00370-1)] [Medline: [37816837](#)]
2. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards expert-level medical question answering with large language models. *Arxiv*. Preprint posted online on May 16, 2023. [\[FREE Full text\]](#)
3. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. *iScience*. 2024;27(5):109713. [\[FREE Full text\]](#) [doi: [10.1016/j.isci.2024.109713](https://doi.org/10.1016/j.isci.2024.109713)] [Medline: [38746668](#)]
4. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nat Commun*. 2024;15(1):2050. [\[FREE Full text\]](#) [doi: [10.1038/s41467-024-46411-8](https://doi.org/10.1038/s41467-024-46411-8)] [Medline: [38448475](#)]
5. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30(9):2613-2622. [doi: [10.1038/s41591-024-03097-1](https://doi.org/10.1038/s41591-024-03097-1)] [Medline: [38965432](#)]
6. 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. [\[FREE Full text\]](#) [doi: [10.1016/j.adro.2023.101400](https://doi.org/10.1016/j.adro.2023.101400)] [Medline: [38304112](#)]
7. Huang Y, Gomaa A, Semrau S, Haderlein M, Lettmaier S, Weissmann T, et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and red journal gray zone cases: potentials and challenges for ai-assisted medical education and

- decision making in radiation oncology. *Front Oncol*. 2023;13:1265024. [FREE Full text] [doi: [10.3389/fonc.2023.1265024](https://doi.org/10.3389/fonc.2023.1265024)] [Medline: [37790756](https://pubmed.ncbi.nlm.nih.gov/37790756/)]
8. Putz F, Haderlein M, Lettmaier S, Semrau S, Fietkau R, Huang Y. Exploring the capabilities and limitations of large language models for radiation oncology decision support. *Int J Radiat Oncol Biol Phys*. 2024;118(4):900-904. [doi: [10.1016/j.ijrobp.2023.11.062](https://doi.org/10.1016/j.ijrobp.2023.11.062)] [Medline: [38401978](https://pubmed.ncbi.nlm.nih.gov/38401978/)]
 9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. [FREE Full text] [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
 10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. [FREE Full text] [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
 11. Dentella V, Guenther F, Leivada E. Language in vivo vs. in silico: size matters but larger language models still do not comprehend language on a par with humans. *ArXiv*. Preprint posted online on Apr 23, 2024. [FREE Full text]
 12. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. *ArXiv*. Preprint posted online on Jul 12, 2023. [FREE Full text]
 13. OpenBioLLM-70B. Hugging Face. URL: <https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B> [accessed 2025-05-10]
 14. SmartOncology. GitHub. URL: <https://github.com/wemedoo/smartoncology> [accessed 2025-05-10]
 15. SmartOncology. Wemedoo. URL: <https://wemedoo.com/smart-oncology/> [accessed 2025-05-10]
 16. OpenBioLLM-Llama3-70B-GGUF. Hugging Face. URL: <https://huggingface.co/mradermacher/OpenBioLLM-Llama3-70B-GGUF> [accessed 2025-05-10]
 17. Llama.cpp. GitHub. URL: <https://github.com/ggerganov/llama.cpp> [accessed 2025-05-10]
 18. LLM-evaluation-in-RO. GitHub. URL: <https://github.com/med-data-tools/LLM-evaluation-in-RO> [accessed 2025-05-10]
 19. Llama3 Model Card. Meta AI. URL: <https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3/> [accessed 2025-05-10]
 20. Federal Act on Research Involving Human Beings. Fedlex. URL: <https://www.fedlex.admin.ch/eli/cc/2013/617/en> [accessed 2025-05-10]
 21. Thirunavukarasu AJ. How can the clinical aptitude of AI assistants be assayed? *J Med Internet Res*. 2023;25:e51603. [FREE Full text] [doi: [10.2196/51603](https://doi.org/10.2196/51603)] [Medline: [38051572](https://pubmed.ncbi.nlm.nih.gov/38051572/)]
 22. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak*. 2024;24(1):72. [FREE Full text] [doi: [10.1186/s12911-024-02459-6](https://doi.org/10.1186/s12911-024-02459-6)] [Medline: [38475802](https://pubmed.ncbi.nlm.nih.gov/38475802/)]
 23. Kanithi PK, Christophe C, Pimentel MAF, Raha T, Saadi N, Javed H, et al. MEDIC: towards a comprehensive framework for evaluating LLMs in clinical applications. *ArXiv*. Preprint posted online on Sep 11, 2024. [FREE Full text]
 24. Elo AE. *The Rating of Chessplayers, Past and Present*. New York, NY: Arco Pub; 1978.
 25. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review. *J Med Internet Res*. 2024;26:e22769. [FREE Full text] [doi: [10.2196/22769](https://doi.org/10.2196/22769)] [Medline: [39509695](https://pubmed.ncbi.nlm.nih.gov/39509695/)]
 26. Guerreiro NM, Alves D, Waldendorf J, Haddow B, Birch A, Colombo P, et al. Hallucinations in Large Multilingual Translation Models. *ArXiv*. Preprint posted online on Mar 28, 2023. [FREE Full text] [doi: [10.1162/tacl_a_00615](https://doi.org/10.1162/tacl_a_00615)]
 27. Kresevic S, Giuffrè M, Ajcevic M, Accardo A, Crocè LS, Shung DL. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ Digit Med*. 2024;7(1):102. [FREE Full text] [doi: [10.1038/s41746-024-01091-y](https://doi.org/10.1038/s41746-024-01091-y)] [Medline: [38654102](https://pubmed.ncbi.nlm.nih.gov/38654102/)]
 28. Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. *NPJ Digit Med*. 2024;7(1):205. [FREE Full text] [doi: [10.1038/s41746-024-01208-3](https://doi.org/10.1038/s41746-024-01208-3)] [Medline: [39112822](https://pubmed.ncbi.nlm.nih.gov/39112822/)]
 29. Sezgin E. Redefining virtual assistants in health care: the future with large language models. *J Med Internet Res*. 2024;26:e53225. [FREE Full text] [doi: [10.2196/53225](https://doi.org/10.2196/53225)] [Medline: [38241074](https://pubmed.ncbi.nlm.nih.gov/38241074/)]
 30. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. *Digit Health*. 2023;9:20552076231186520. [FREE Full text] [doi: [10.1177/20552076231186520](https://doi.org/10.1177/20552076231186520)] [Medline: [37426593](https://pubmed.ncbi.nlm.nih.gov/37426593/)]
 31. Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M. AI education for clinicians. *EClinicalMedicine*. 2025;79:102968. [FREE Full text] [doi: [10.1016/j.eclinm.2024.102968](https://doi.org/10.1016/j.eclinm.2024.102968)] [Medline: [39720600](https://pubmed.ncbi.nlm.nih.gov/39720600/)]
 32. Wosny M, Strasser LM, Kraehenmann S, Hastings J. Practical recommendations for navigating digital tools in hospitals: qualitative interview study. *JMIR Med Educ*. 2024;10:e60031. [FREE Full text] [doi: [10.2196/60031](https://doi.org/10.2196/60031)] [Medline: [39602211](https://pubmed.ncbi.nlm.nih.gov/39602211/)]
 33. Barbosa B, Oliveira C, Bravo I, Couto JG, Antunes L, McFadden S, et al. An investigation of digital skills of therapeutic radiographers/radiation therapists: a european survey of proficiency level and future educational needs. *Radiography (Lond)*. 2023;29(3):479-488. [FREE Full text] [doi: [10.1016/j.radi.2023.02.009](https://doi.org/10.1016/j.radi.2023.02.009)] [Medline: [36878157](https://pubmed.ncbi.nlm.nih.gov/36878157/)]
 34. Siochi RA, Balter P, Bloch CD, Bushe HS, Mayo CS, Curran BH, et al. Information technology resource management in radiation oncology. *J Appl Clin Med Phys*. 2009;10(4):16-35. [FREE Full text] [doi: [10.1120/jacmp.v10i4.3116](https://doi.org/10.1120/jacmp.v10i4.3116)] [Medline: [19918239](https://pubmed.ncbi.nlm.nih.gov/19918239/)]

35. Zanca F, Hernandez-Giron I, Avanzo M, Guidi G, Crijns W, Diaz O, et al. Expanding the medical physicist curricular and professional programme to include artificial intelligence. *Phys Med*. 2021;83:174-183. [doi: [10.1016/j.ejmp.2021.01.069](https://doi.org/10.1016/j.ejmp.2021.01.069)] [Medline: [33798903](https://pubmed.ncbi.nlm.nih.gov/33798903/)]

Abbreviations

AI: artificial intelligence

DEGRO: German Society for Radiation Oncology

ISROI: International Society for Radiation Oncology Informatics

LLM: large language model

USMLE: United States Medical Licensing Examination

Edited by N Cahill; submitted 10.Dec.2024; peer-reviewed by M Meinschad, L Gascon; comments to author 13.Jan.2025; revised version received 22.Feb.2025; accepted 31.Mar.2025; published 23.Sep.2025

Please cite as:

Dennstädt F, Schmerder M, Riggerbach E, Mose L, Bryjova K, Bachmann N, Mackeprang P-H, Ahmadsei M, Sinovcic D, Windisch P, Zwahlen D, Rogers S, Riesterer O, Maffei M, Gkika E, Haddad H, Peecken J, Putora PM, Glatzer M, Putz F, Hoefler D, Christ SM, Filchenko I, Hastings J, Gaio R, Chiang L, Aebersold DM, Cihoric N

Comparative Evaluation of a Medical Large Language Model in Answering Real-World Radiation Oncology Questions: Multicenter Observational Study

J Med Internet Res 2025;27:e69752

URL: <https://www.jmir.org/2025/1/e69752>

doi: [10.2196/69752](https://doi.org/10.2196/69752)

PMID:

©Fabio Dennstädt, Max Schmerder, Elena Riggerbach, Lucas Mose, Katarina Bryjova, Nicolas Bachmann, Paul-Henry Mackeprang, Maiwand Ahmadsei, Dubravko Sinovcic, Paul Windisch, Daniel Zwahlen, Susanne Rogers, Oliver Riesterer, Martin Maffei, Eleni Gkika, Hatal Haddad, Jan Peecken, Paul Martin Putora, Markus Glatzer, Florian Putz, Daniel Hoefler, Sebastian M Christ, Irina Filchenko, Janna Hastings, Roberto Gaio, Lawrence Chiang, Daniel M Aebersold, Nikola Cihoric. Originally published in the *Journal of Medical Internet Research* (<https://www.jmir.org>), 23.Sep.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the *Journal of Medical Internet Research* (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.