A multicentric study comparing a medical LLM's performance with clinical experts in radiation oncology

Fabio Dennstädt, Max Schmerder, Elena Riggenbach, 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, Hathal Haddad, Jan Peeken, Paul Martin Putora, Markus Glatzer, Florian Putz, Daniel Hoefler, Sebastian Christ, Irina Filchenko, Janna Hastings, Roberto Gaio, Lawrence Chiang, Daniel Aebersold, Nikola Cihoric

Submitted to: Journal of Medical Internet Research
on: December 10, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

A multicentric study comparing a medical LLM's performance with clinical experts in radiation oncology

Fabio Dennstädt[1]; Max Schmerder[1]; Elena Riggenbach[1]; Lucas Mose[1]; Katarina Bryjova[1]; Nicolas Bachmann[1]; Paul-Henry Mackeprang[1]; Maiwand Ahmadsei[1]; Dubravko Sinovcic[2]; Paul Windisch[2]; Daniel Zwahlen[2]; Susanne Rogers[3]; Oliver Riesterer[3]; Martin Maffei[4]; Eleni Gkika[5]; Hathal Haddad[6]; Jan Peeken[7, 8, 8]; Paul Martin Putora[1, 9]; Markus Glatzer[9]; Florian Putz[10]; Daniel Hoefler[10]; Sebastian Christ[11, 12]; Irina Filchenko[13]; Janna Hastings[14, 15, 16]; Roberto Gaio[1]; Lawrence Chiang[1]; Daniel Aebersold[1]; Nikola Cihoric[1]

[1]Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern Bern CH
[2]Department of Radiation Oncology, Cantonal Hospital Winterthur Winterthur CH
[3]Radiation Oncology Center Mittelland, Cantonal Hospital Aarau Aarau CH
[4]Department of Radiation Oncology, Hospital of Bolzano (SABES-ASDAA); Teaching Hospital of Paracelsus Medical University Bolzano IT
[5]Department of Radiation Oncology, University Hospital Bonn, University of Bonn Bonn DE
[6]Department of Radiation Oncology, University Hospital Tübingen Tübingen DE
[7]Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich (TUM) Munich DE
[8]Department of Radiation Oncology, Cantonal Hospital St. Gallen St. Gallen CH
[9]Department of Radiation Oncology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nürnberg Erlangen DE
[10]Department of Radiation Oncology, University Hospital Lausanne Lausanne CH
[11]Department of Radiation Oncology, University Hospital and University of Zurich Zurich CH
[12]Department of Neurology, Inselspital, Bern University Hospital and University of Bern Bern CH
[13]Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich Zurich CH
[14]School of Medicine, University of St. Gallen St. Gallen CH
[15]Swiss Institute of Bioinformatics Lausanne CH

Corresponding Author:
Nikola Cihoric
Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern
Freiburgstrasse 18
Bern
CH

Abstract

Background: Large Language Models (LLMs) hold promise for supporting clinical tasks, particularly in technical fields like radiation oncology. While prior evaluations have focused on exam-style settings, their performance in real-life clinical scenarios remains unclear.

Objective: This study aimed to assess a state-of-the-art medical LLM's ability to answer real-world clinical questions in radiation oncology compared to clinical experts.

Methods: Physicians from 10 departments collected routine clinical questions. Fifty of these questions were answered by three senior radiation oncology experts and the LLM Llama3-OpenBioLLM-70B. In a blinded review, physicians rated answer quality on a 5-point Likert scale, assessed safety, and determined if responses were from the LLM or an expert (recognizability). Comparisons were 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 M 4.00, interquartile range, IQR [3.00, 4.00] vs. M 3.67 IQR [3.33, 4.00]; p=0.263). The answers of the LLM were deemed potentially harmful in 16% of cases versus 13% for the clinical experts (p=0.633). Physicians correctly identified whether an answer was provided by an LLM or a clinician in 72% and 78% of cases, respectively.

Conclusions: The quality of the answers of the LLM seems similar to those of clinical experts. While great caution is recommended while using LLMs in clinical practice, their ability in answering real-life clinical questions is satisfactory, including highly specialized domains like radiation oncology.

(JMIR Preprints 10/12/2024:69752)
DOI: https://doi.org/10.2196/preprints.69752

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

A multicentric study comparing a medical LLM's performance with clinical experts in radiation oncology

AUTHORS: Fabio Dennstädt[1], Max Schmerder[1], Elena Riggenbach[1], Lucas Mose[1], Katarina Bryjova[1], Nicolas Bachmann[1], Paul-Henry Mackeprang[1], Maiwand Ahmadsei[1], Dubravko Sinovcic[2], Paul Windisch[2], Daniel R Zwahlen[2], Susanne J Rogers[3], Oliver Riesterer[3], Martin Maffei[4], Eleni Gkika[5], Hathal Haddad[6], Jan C Peeken[7,8,9], Paul Martin Putora[1,10], Markus Glatzer[10], Florian Putz[11], Daniel Hoefler[11], Sebastian M Christ[12,13], Irina Filchenko[14], Janna Hastings[15,16,17], Roberto Gaio[1], Lawrence Chiang[18], Daniel Aebersold[1], Nikola Cihoric*[1]

Affiliations

1 – Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland.

2 – Department of Radiation Oncology, Cantonal Hospital Winterthur, Winterthur, Switzerland.

3 – Radiation Oncology Center Mittelland, Cantonal Hospital Aarau, Aarau, Switzerland.

4 – Department of Radiation Oncology, Hospital of Bolzano (SABES-ASDAA), Bolzano-Bozen, Italy; Teaching Hospital of Paracelsus Medical University.

5 – Department of Radiation Oncology, University Hospital Bonn, University of Bonn, Bonn, Germany.

6 – Department of Radiation Oncology, University Hospital Tübingen, Tübingen, Germany.

7 – Department of Radiation Oncology, Klinikum rechts der Isar, Technical University of Munich (TUM), Munich, Germany.

8 – Institute of Radiation Medicine (IRM), Department of Radiation Sciences (DRS), Helmholtz Zentrum, Munich, Germany.

9 – German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany.

10 – Department of Radiation Oncology, Cantonal Hospital St. Gallen, St. Gallen, Switzerland.

11 – Department of Radiation Oncology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany.

12 – Department of Radiation Oncology, University Hospital Lausanne, Lausanne, Switzerland.

13 – Department of Radiation Oncology, University Hospital and University of Zurich, Zurich, Switzerland.

14 – Department of Neurology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland.

15 – School of Medicine, University of St. Gallen, St. Gallen, Switzerland.

16 – Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich, Zurich, Switzerland.

17 – Swiss Institute of Bioinformatics, Lausanne, Switzerland.

18 – Department of Radiology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland.

*corresponding author

E-mail: nikola.cihoric@insel.ch

ABSTRACT

Background: Large Language Models (LLMs) hold promise for supporting clinical tasks, particularly in technical fields like radiation oncology. While prior evaluations have focused on exam-style settings, their performance in real-life clinical scenarios remains unclear. This study aimed to assess a state-of-the-art medical LLM's ability to answer real-world clinical questions in radiation oncology compared to clinical experts.

Methods: Physicians from 10 departments collected routine clinical questions. Fifty of these questions were answered by three senior radiation oncology experts and the LLM Llama3-OpenBioLLM-70B. In a blinded review, physicians rated answer quality on a 5-point Likert scale, assessed safety, and determined if responses were from the LLM or an expert (recognizability). Comparisons were 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 M 4.00, interquartile range, IQR [3.00, 4.00] vs. M 3.67 IQR [3.33, 4.00]; *P*=.263; Wilcoxon signed-rank test). The answers of the LLM were deemed potentially harmful in 16% of cases versus 13% for the clinical experts (*P*=.633; Fisher Exact test). Physicians correctly identified whether an answer was provided by an LLM or a clinician in 72% and 78% of cases, respectively.

Conclusion: The quality of the answers of the LLM seems similar to those of clinical experts. While great caution is recommended while using LLMs in clinical practice, their ability in answering real-life clinical questions is satisfactory, including highly specialized domains like radiation oncology.

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 non-medical 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 Exam (USMLE) [2]. These systems demonstrated success in numerous applications like 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 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 remarkable ability of systems like ChatGPT, the Llama models or PALM 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] [7] [8].

Most of these evaluation studies have been performed on exam-style questions with pre-designed 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 like MedPALM-2 have been reported to answer questions "at the level of an expert doctor" [2]. However, a limitation of these studies is that pre-designed 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 (e.g., Chat-GPT 3.5 incorporates 175 billion parameters, compared to >1.5 trillion in Chat-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 (e.g., those used by ChatGPT, Claude or Gemini) [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 to answers given by clinical experts in a multicentric evaluation study.

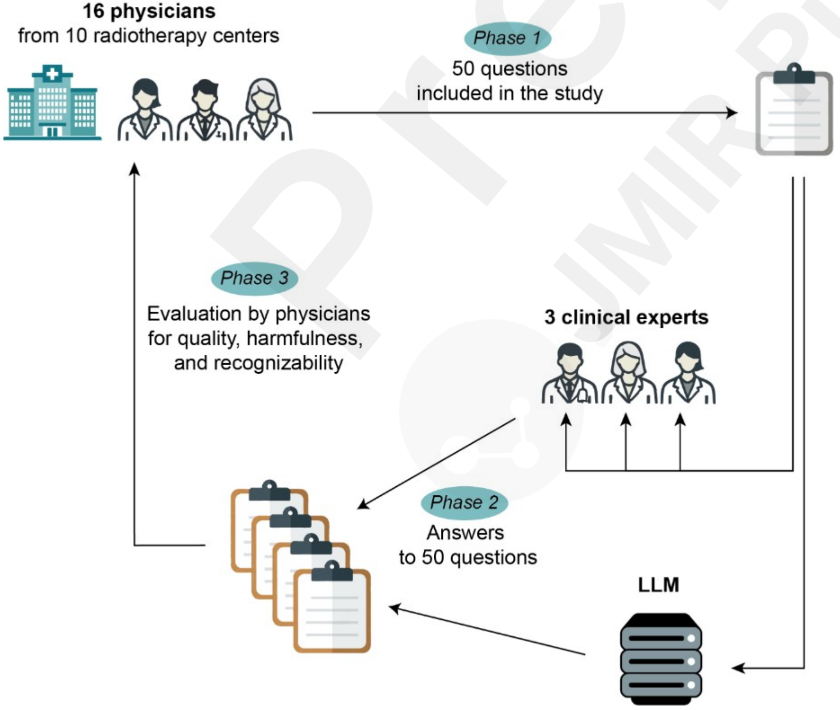
METHODS

Study design

The study had three 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 online platform *SmartOncology* [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.



Abbreviations: 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 eight weeks, from 22nd of May to 16th of June 2024, by twenty participating radiation oncologists from ten 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.

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 coming up during clinical work. Please write down any question that you would ask such an AI agent if it was 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 clinical routine of radiotherapy, valid questions included those that were not primarily related to radiotherapy (e.g., 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?"*).

For language consistency, clinicians were furthermore 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. After screening of 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" (= CNS), "lung", "palliative" and "other".

Phase 2: Answering the questions

Three radiation oncologists from different centers of the community of ISROI / DEGRO with profound knowledge in radiation oncology and at least 5 years of post-specialization work experience were selected as clinical experts to answer the questions. The clinical experts were given the following instruction: *"Please answer the given question. Imagine this question being asked to 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 online 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 (e.g., like 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, 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 like MedMCQA, MMLU Medicine and PubMedQA, while being an open-source model that can be run on a local system [13].

Details on running the LLM (e.g., prompting and hardware used) are provided in *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 (i.e., clinical expert or AI).

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

Firstly, the physicians rated the quality of each answer on a 5-point Likert score (1 – very bad, 2 – bad, 3 – acceptable, 4 – good, 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 with the evaluations are provided at *GitHub*[16].

Data and statistical analysis

The analyses were exploratory and performed using R version 4.4.2 (cran.r-project.org/). Unless stated otherwise, continuous variables were presented as median and interquartile range, while categorical variables were presented as count (i.e., % of total). There was no missing data.

We compared the quality of the answers (i.e., as a continuous characteristic) between the LLM and clinical experts (i.e., the quality) as dependent variables using Wilcoxon signed-rank test. Moreover, the quality was described for the individual thematic groups and the three 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 (i.e., clinical experts versus AI) 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 Wilcoxon signed-rank test.

Finally, we compared the categorical characteristics of the answers between the LLM and clinical experts (i.e., 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 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 McNemar test, and the false-discovery rate was applied to correct for multiple comparisons.

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

Ethical considerations

The study was deliberately designed such that no patient-/person-related medical or non-medical data were used and that no approval from an ethics committee was required. All of the data (questions, answers and evaluations) used in this work were generated by members of the research group.

RESULTS

Collected questions and length of answers

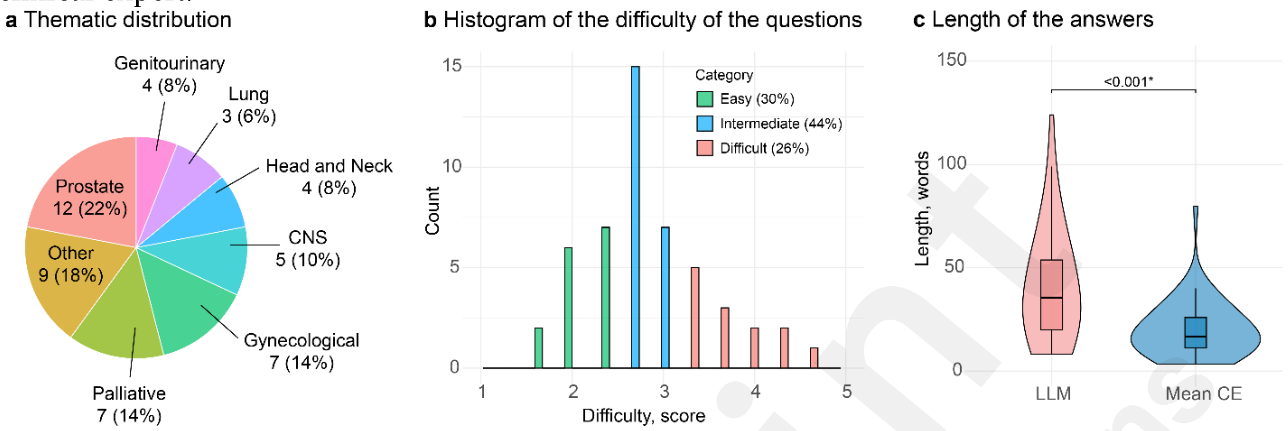
A total of 133 questions was collected by 16 of the participating physicians. Four of the initial 20 physicians did not submit any questions and did not further participate in the study. Seven questions were deemed invalid by the study coordinators due to unclarity or submission of the question in a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

The 50 randomly selected questions were mostly categorized "prostate" (22%), "gynecological" (14%), "palliative" (14%) and "other" (18%; *Figure 2a*).

The difficulty of the questions was 2.67 [2.33, 3.33] points of the 5-score Likert scale. Most questions (44%) were of intermediate difficulty, while 26% were classified as difficult and 30% as easy (*Figure 2b*).

The length of the questions was 32.0 [17.25, 47.75] words. Clinical experts had significantly shorter answers compared to those generated by the LLM (16.67 [11.25, 19.96] words vs. 35.50 [20.00, 40.08] words, p<0.001; *Figure 2c*).

Figure 2. Properties of the collected questions and 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 LLM and mean clinical expert.



Abbreviations: CE – clinical expert, LLM – large language model

Quality of the answers

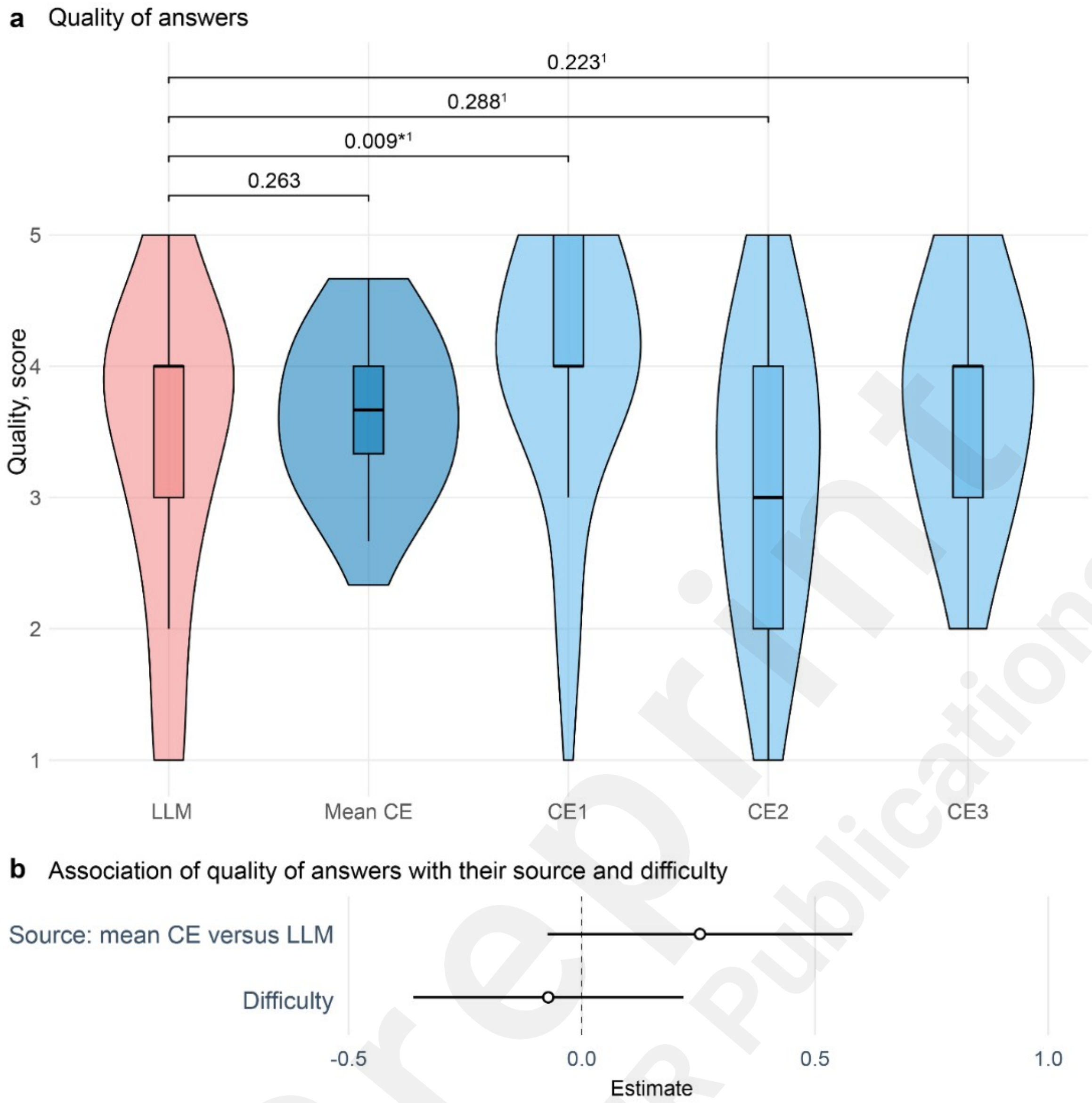
Overall, the LLM answers were deemed to be equally good as or better than the mean clinician answer in 27 cases (54 %). In 19 cases (38 %), the LLM answer was deemed at least as good as the ones of the "best clinician". In 9 cases (18 %), the LLM answer was considered worse than those of all three clinical experts and in 2 cases (4 %), the LLM answer was considered to be better than those from all three 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 of the mean quality of clinical expert answer for "CNS" (4.20 vs. 3.89), and "other" (3.78 vs. 3.63) and lower for "head and neck" (3.25 vs. 4.00), "gynecological" (3.29 vs. 3.33), "prostate" (2.82 vs. 3.37), "lung" (3.67 vs. 3.89), "palliative" (3.29 vs. 3.86) and "genitourinary" (3.25 vs. 4.00) (*Supplementary Figure 1*; see also *Supplementary Figure 2* for assessment of second reviewer). Regarding the difficulty categories, the scores of the LLM compared to mean clinical expert were 4.00 vs. 3.73 for easy, 3.00 vs. 3.65 for intermediate and 3.31 vs. 3.49 for difficult questions.

The quality score of the answers given by the clinical experts was 3.67 [3.33, 4.00] (mean of 3.63; range of 3.18 – 4.00) in comparison to 4.00 [3.00, 4.00] (mean of 3.38) for the answers given by the LLM based on the evaluation by the questioner reviewer. Whilst there was no statistically significant difference between the LLM and mean quality of clinical expert answer, the variability between clinical experts was large, with one clinical expert providing answers of significantly higher quality compared to 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 (*Supplementary Figure 3*).

Figure 3. Quality of the answers as assessed by the questioner reviewer. a – Box plots with violin plots for comparison of quality score between LLM and mean as well as individual clinical experts. b – Association of the quality of answers with their source and difficulty.

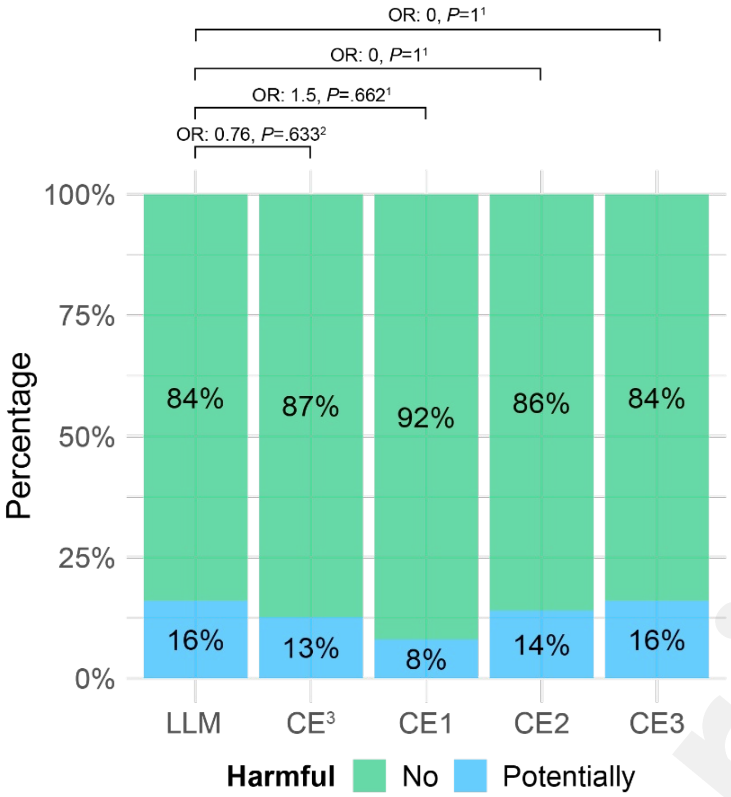


[1]Wilcoxon signed-rank test corrected for multiple comparisons with a false-discovery rate. Abbreviations: CE – clinical expert, LLM – large language model

Harmfulness of the answers

According to the questioner reviewer, 8 (16%) of the answers given by the LLM were considered "harmful" compared to 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 *Supplementary Figure 4*).

Figure 4. Percentages of answers deemed "potentially harmful" by the questioner reviewer.

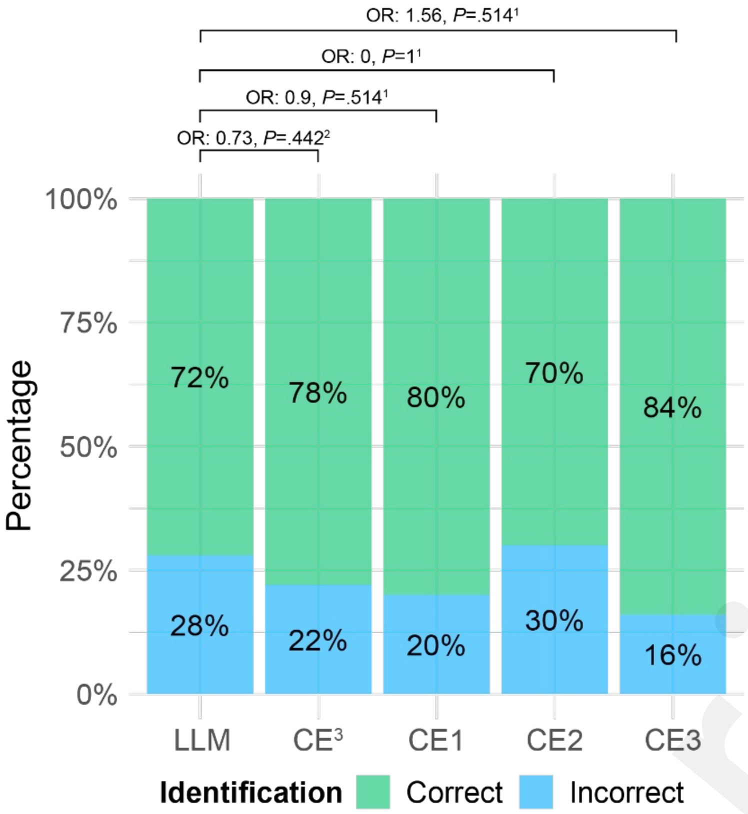


[1]McNemar test corrected for multiple comparisons with a false-discovery rate. [2]Fisher Exact test. [3]Cumulative value for clinical experts. Abbreviations: CE – clinical expert, LLM – large language model, OR – Odds ratio

Identification of LLM vs. clinical expert

The physicians correctly identified the source of an answer in most cases with 72% for answers given by the LLM and 78% for answers given by a clinician (*Figure 5*; Results for the second reviewer are presented in *Supplementary Figure 5*). When the interaction between participant and length was considered, the interaction term between length and clinical experts versus LLM was significant (OR -0.07 95% confidence interval [-0.13, -0.003], *P*=.04), indicating that the likelihood of correct identification of a clinical expert decrease with the increase of the length of the answers.

Figure 5. Percentages of correct and incorrect identifications of the source (LLM or clinical expert) by the questioner reviewer.



[1]McNemar test corrected for multiple comparisons with a false-discovery rate. [2]Fisher Exact test. [3]Cumulative value for clinical experts. Abbreviations: CE – clinical expert, LLM – large language model, OR – Odds ratio

DISCUSSION

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. 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 [17]. 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 [18]. Several approaches have been proposed, including classical exams, Elo systems [19] or logical benchmarks [1 17].

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 those of experts in this clinical domain.

We believe that this approach is an essential component of a comprehensive evaluation. Firstly, many questions arising in real-life are not exam-style, but open-ended, on topics with only limited data available. Secondly, the "quality of an answer" needs to be evaluated without a clearly defined ground truth existing. Comparing 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 to clinical experts.

Are LLMs ready to be used as "AI assistants" in clinical practice?

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 design and different data sets, 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 questioner reviewer in 8 cases (16 %).

Although determining the threshold at which LLMs are ready for clinical implementation is challenging, we would currently discourage their use in clinical practice. Nonetheless, it is interesting that, even within the study setting, a similar proportion of responses provided by the average clinical expert were rated as “bad”, “very bad”, or “potentially harmful”.

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 a general, better-known models provided by private companies (e.g., ChatGPT, Claude or Gemini). A model like OpenBioLLM-70B can be run in a local environment with all data staying within the hospital and avoidance of transmission of sensitive healthcare 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 – the system used in our study ran on hardware that can be bought for 3000-5000€. Of course, many legal and regulatory issues have to 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 to most clinical experts in such benchmarking studies. Additional performance gains may be obtained adding context-specific information in the form of guidelines in full via retrieval augmented generation systems [20]. Beyond that, newer systems will not only process text data, but include multimodal medical data [21].

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.

Limitations

Our study has several limitations. While the aim was to investigate the performance of LLMs on questions from clinical practice of radiation therapy, no real patient-/person-related healthcare data were used. This is clearly a considerable limitation as many questions in daily clinical life stem from patient-specific information. Most importantly, 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. Beyond that, we only involved three clinical experts, which may limit the generalizability of the results.

Finally, evaluators were able to identify which answers were given by LLM and which by a clinical expert in most of the cases. We hypothesize that this is due to the different language style used by the LLM, generating considerably longer answers. It is likely that 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 quality of 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. Despite seemingly satisfactory LLM performance, LLMs should be used with caution for answering medical question. Yet, these systems have shown rapid advancements in recent years and are expected to continue improving. Future studies, investigating whether their application leads to an improvement in outcome, are warranted.

Acknowledgements
We thank the researchers from Meta AI 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.

Conflicts of Interest
Dr. Cihoric is a technical lead for the *SmartOncology* project and medical advisor for Wemedoo AG, Steinhausen AG, Switzerland.
The authors declare no other conflicts of interest.

Abbreviations
AI – Artificial Intelligence
API – Application Programming Interface
CE – Clinical expert
CNS – Central Nervous System
DEGRO – German Society for Radiation Oncology
ISROI – International Society for Radiation Oncology Informatics
LLM – Large Language Model
USMLE – United States Medical Licensing Exam

Data availability statement
The questions and the answers of the LLM are provided at *GitHub*[16]. The Llama3-OpenBioLLM-70B model is publicly available on the internet (see also *Appendix 1* containing further information for running the model).

Code availability statement
The source code to run the LLM used in this study is provided at *GitHub*[16] (see also *Appendix 1* containing further information for running the model). The *SmartOncology* platform used in this study for data collection is an OpenSource software, publicly available on the internet [14] [15].

Author contributions
Conceptualization – FD, JH, JP, FP, NC
Methodology, technical implementation – RG, LC
Methodology, data collection – MS, ER, LM, KB, NB, PHM, MA, DS, DR Z, SJ R, OR, MM, EG, HH,
JCP, PMP, MG, DH, SMC
Formal Analysis – FD, NC
Statistical Analysis – FD, IF, NC
Writing, original draft preparation – FD, NC
Writing, illustrations – FD, IF
Writing, review and editing – All authors
Project administration – NC, FD, DA

References

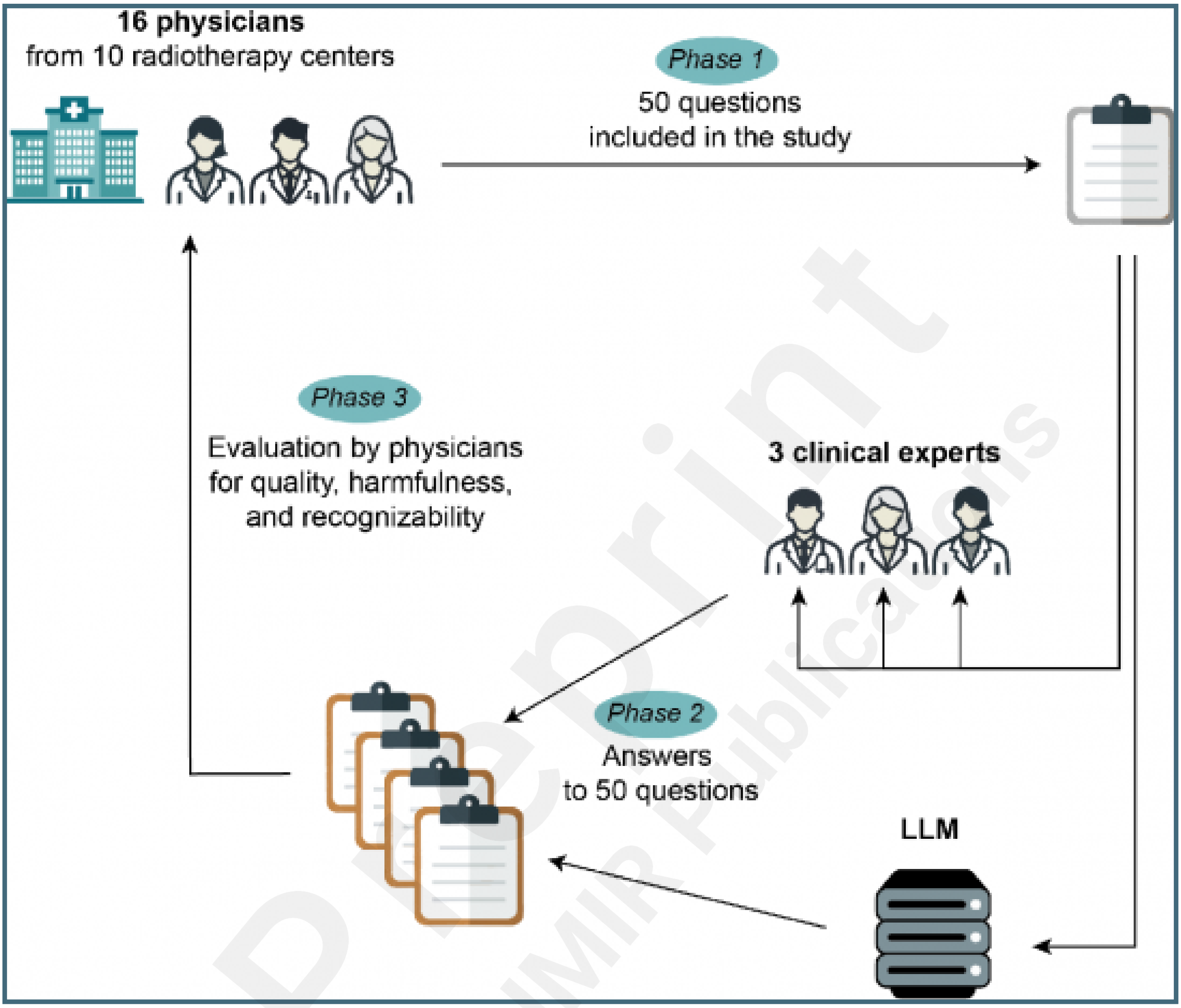
1. Clusmann, J. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023).
2. Singhal, K. *et al.* Towards Expert-Level Medical Question Answering with Large Language Models. Preprint at http://arxiv.org/abs/2305.09617 (2023).
3. Meng, X. *et al.* The application of large language models in medicine: A scoping review. *iScience* **27**, 109713 (2024).
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* **15**, 2050 (2024).
5. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* **30**, 2613–2622 (2024).
6. Dennstädt, F. *et al.* Exploring Capabilities of Large Language Models such as ChatGPT in Radiation Oncology. *Advances in Radiation Oncology* **9**, 101400 (2024).
7. Huang, Y. *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.* **13**, 1265024 (2023).
8. Putz, F. *et al.* Exploring the Capabilities and Limitations of Large Language Models for Radiation Oncology Decision Support. *International Journal of Radiation Oncology*Biology*Physics* **118**, 900–904 (2024).
9. Gilson, A. *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* **9**, e45312 (2023).
10. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

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. Preprint at https://doi.org/10.48550/ARXIV.2404.14883 (2024).
12. Naveed, H. *et al.* A Comprehensive Overview of Large Language Models. Preprint at https://doi.org/10.48550/ARXIV.2307.06435 (2023).
13. Huggingface: Llama3-OpenBioLLM-70B. https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B.
14. GitHub: SmartOncology. https://github.com/wemedoo/smartoncology.
15. Wemedoo: SmartOncology. https://wemedoo.com/smart-oncology/.
16. GitHub: LLM-evaluation-in-RO. https://github.com/med-data-tools/LLM-evaluation-in-RO.
17. Park, Y.-J. *et al.* Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak* **24**, 72 (2024).
18. Kanithi, P. K. *et al.* MEDIC: Towards a Comprehensive Framework for Evaluating LLMs in Clinical Applications. Preprint at https://doi.org/10.48550/ARXIV.2409.07314 (2024).
19. Elo, A. E. *The Rating of Chessplayers, Past and Present*. (Arco Pub, New York, 1978).
20. Kresevic, S. *et al.* Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *npj Digit. Med.* **7**, 102 (2024).
21. Kaczmarczyk, R., Wilhelm, T. I., Martin, R. & Roos, J. Evaluating multimodal AI in medical diagnostics. *npj Digit. Med.* **7**, 205 (2024).

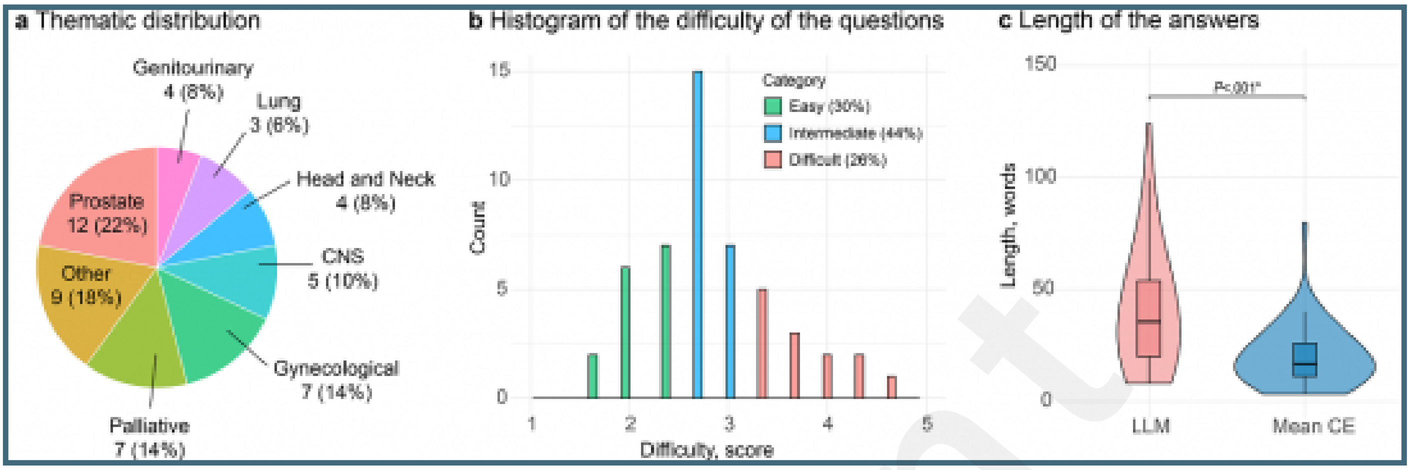
Supplementary Files

Figures

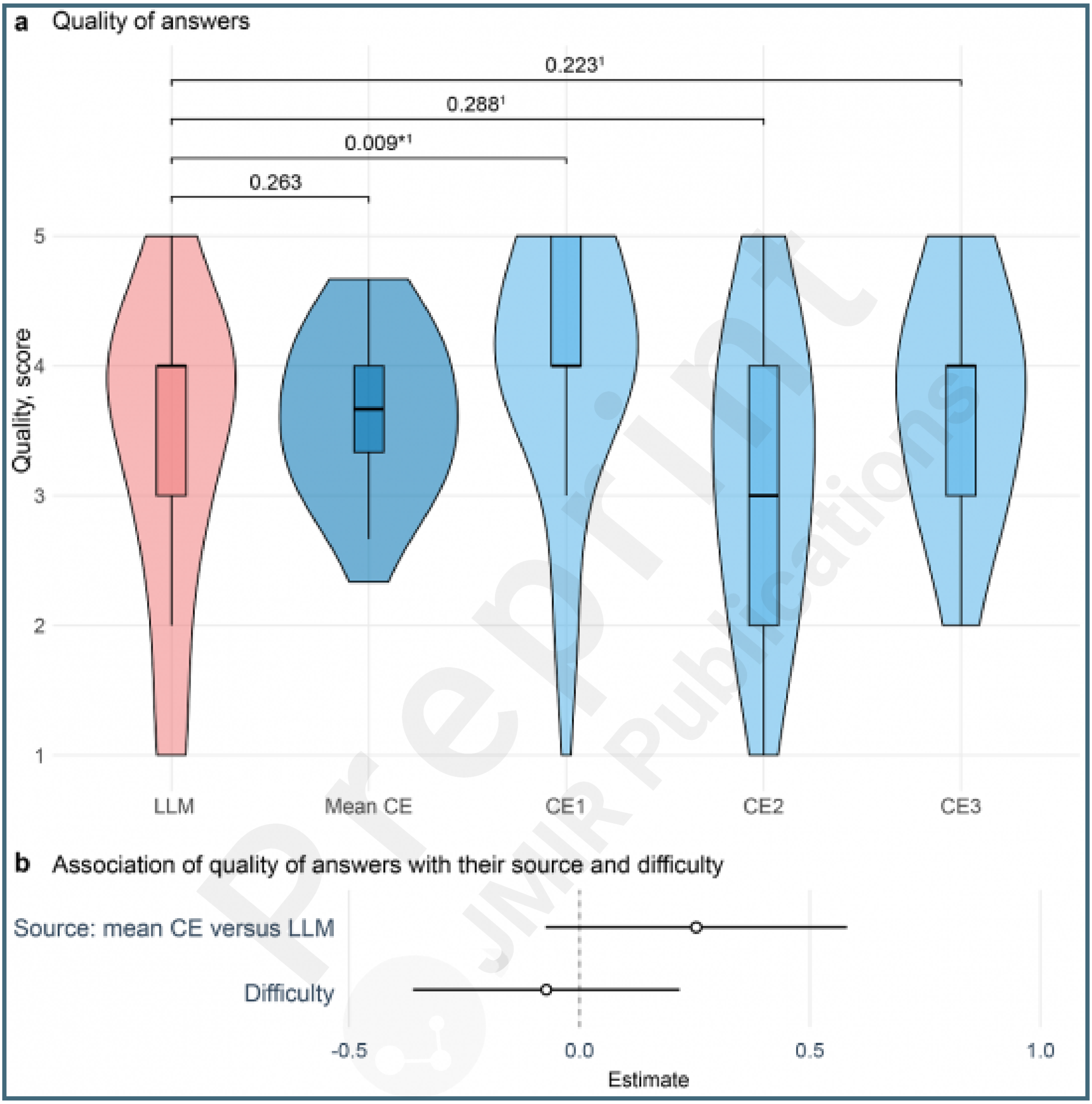
Schematic illustration of the study design.



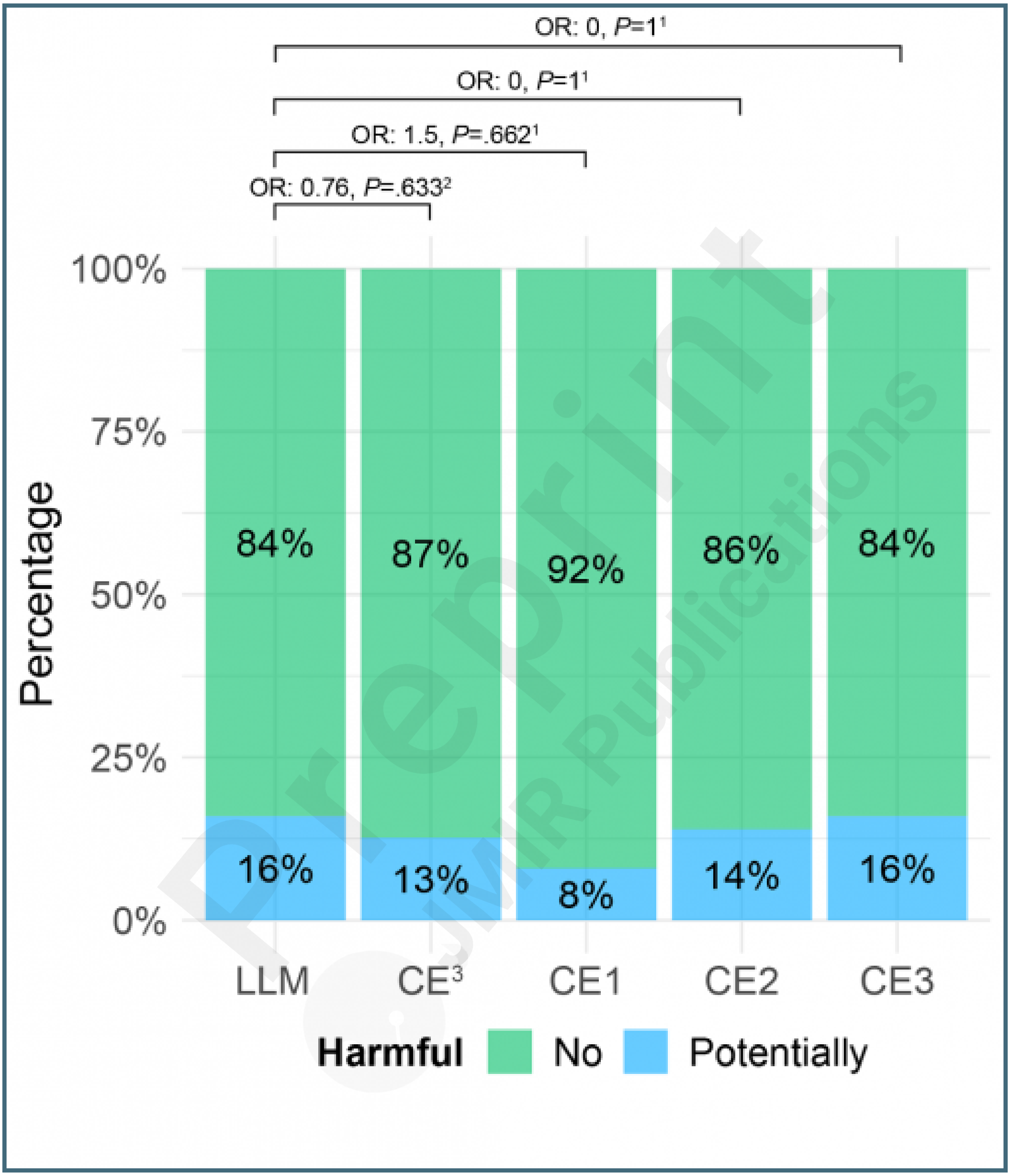
Properties of the collected questions and length of answers.



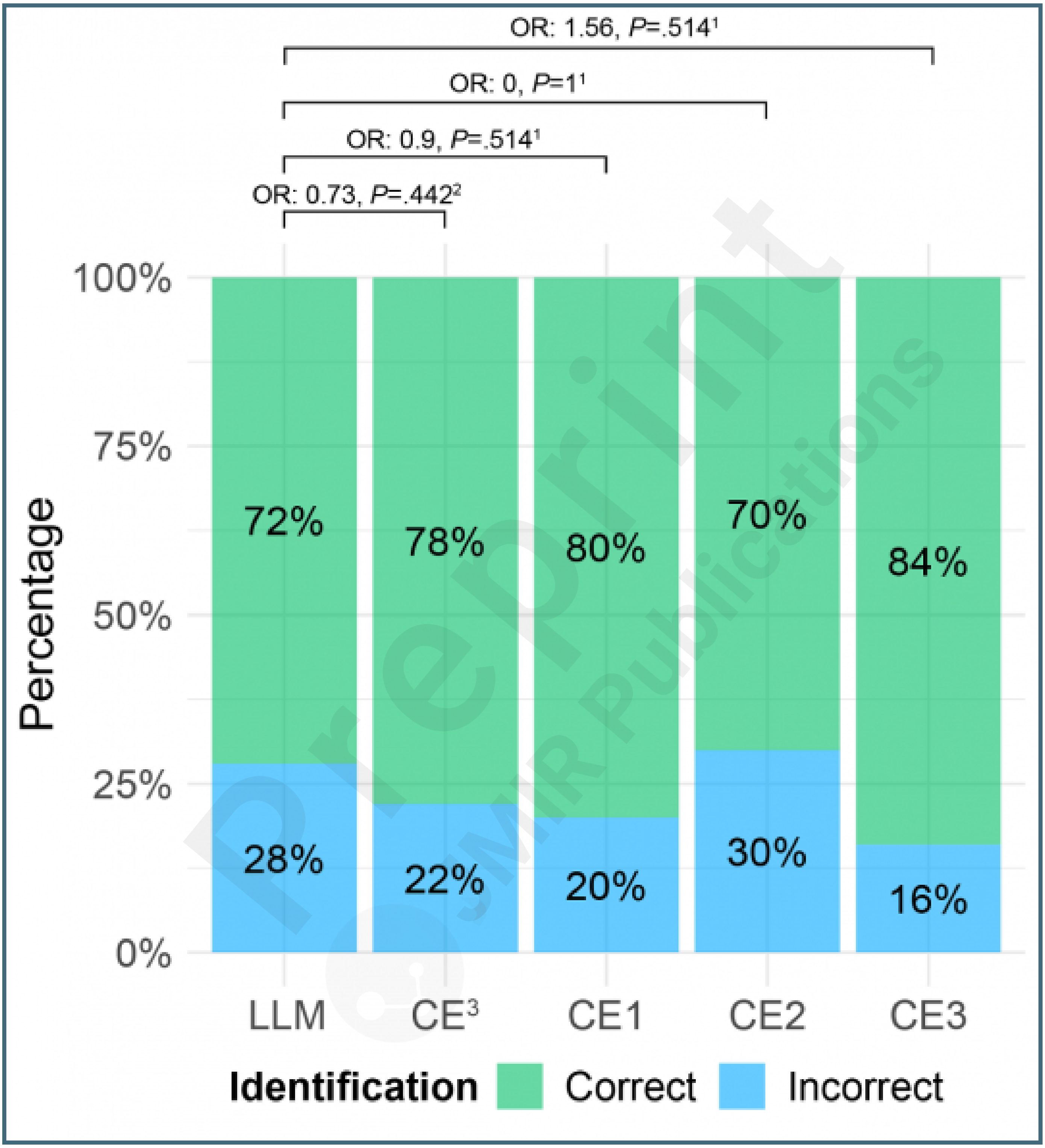
Quality of the answers as assessed by the questioner reviewer.



Percentages of answers deemed “potentially harmful” by the questioner reviewer.



Percentages of correct and incorrect identifications of the source (LLM or clinical expert) by the questioner reviewer.



Multimedia Appendixes

Technical details for running the LLM.
URL: http://asset.jmir.pub/assets/46a075e9261758297ae84f411d344f39.docx

Supplementary Figures.
URL: http://asset.jmir.pub/assets/4067483d076908c3f922a80ed7d26b2a.docx