Comparison of the Knowledge of Large Language Models and General Radiologist on RECIST

Eren Çamur, Turay Cesur, Yasin Celal Güneş

Table of Contents

Comparison of the Knowledge of Large Language Models and General Radiologist on RECIST

Eren Çamur[1]; Turay Cesur[2*]; Yasin Celal Güne?[3*]

[1]Ministry of Health Ankara 29 Mayis State Hospital Department of Radiology Ankara TR
[2]Ankara Mamak State Hospital Ankara TR
[3]TC Saglik Bakanligi Kirikkale Yuksek Ihtisas Hastanesi K?r?kkale TR
[*]these authors contributed equally

Corresponding Author:
Eren Çamur
Ministry of Health Ankara 29 Mayis State Hospital
Department of Radiology
Ayd?nlar, Dikmen Cd No:312
Çankaya
Ankara
TR

Abstract

Background: Large language models (LLMs) represent a remarkable breakthrough in natural language processing. What sets the current generation of LLMs apart is their remarkable ability to perform very specific tasks in radiology, as in many other fields, without the need for additional training. LLMs have the potential to usher in a new era of efficiency and excellence in radiology practice, both in their potential as supportive diagnostic tool and in their ability to facilitate the reporting process. This great potential of LLMs is of great importance in both oncology and oncological radiology, and researchers have been conducting many new studies in these fields in order to demonstrate the position of LLMs.

Objective: We aimed to provide a new perspective on their potential to facilitate reporting and improve reporting efficiency in oncological imaging by comparatively assessing LLMs' knowledge of RECIST 1.1 both among themselves and with general radiologist (GR).

Methods: Radiologist (E.Ç.) prepared the 25 multiple-choice questions in this study utilizing the information in RECIST 1.1, thus eliminating the need for ethics committee approval. We initiated the input prompt as follows: ''Act like a professor of radiology who has 30 years of experience in oncological radiology.. Give just letter of the most correct choice of multiple choice questions. Each question have only one correct answer.''This prompt was tested in June 2024 on seven different LLMs using the default settings. The testing included models from various developers: Claude 3 Opus, ChatGPT 4 and ChatGPT 4o, Gemini 1.5 pro, Mistral Large, Meta Llama 3 70B, Perplexity pro. Also GR (T.C.) board certified by EDiR and with 6 years of experience each, answered the same questions

Results: The results revealed that Claude 3 Opus achieved the highest accuracy of 100% (25/25 questions), followed by newest model of Open AI's ChatGPT 4o with 96% accuracy (24/25 questions). ChatGPT 4 and Mistral Large 92% (23/25 questions), Meta Llama 3 70 B, Perplexity pro and Gemini 1.5 pro had accuracy of 88% (21/25 questions). GR (T.C.) has accuracy of 92% (23/25 questions).

Conclusions: The outstanding success of Claude 3 Opus by knowing all the questions raises the question of whether oncology can be a new star among LLMs in radiology. Our study reveals that the majority of LLM models exhibit a commendable level of proficiency and comperable to GR in answering RECIST 1.1 related questions. Our findings show that current LLM models have more than sufficient text-based information about RECIST 1.1. Additionally, our findings underscore the high potential of LLMs as tools to assist radiologists in oncology reporting. However, to take full advantage of LLMs' abilities in oncology reporting, it is of great importance that their visual abilities are also evaluated. Visual evaluation forms the basis of radiology. Therefore, future studies should focus on evaluating the visual information of multimodal LLMs with visual evaluation ability.
In conclusion, LLMs have great potential in oncological radiology as in every field of radiology, and new studies that will reveal this potential will allow LLMs to be more easily integrated into radiological practice. Clinical Trial: There is no need to trial registration.

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

Intrıduction

Large language models (LLMs) represent a remarkable breakthrough in natural language processing. What sets the current generation of LLMs apart is their remarkable ability to perform very specific tasks in radiology, as in many other fields, without the need for additional training. This positions LLMs as a transformative force, poised to significantly reshape the way radiology practice is conducted [1]. LLMs have the potential to usher in a new era of efficiency and excellence in radiology practice, both in their potential as supportive diagnostic tool and in their ability to facilitate the reporting process. It is therefore not surprising that there has been a rapid increase in studies investigating the radiological knowledge of LLMs and their potential applications and contributions to radiology [2,3]. This great potential of LLMs is of great importance in both oncology and oncological radiology, and researchers have been conducting many new studies in these fields in order to demonstrate the position of LLMs [4,5,6].

The radiology report serves as a vital tool in guiding patient management decisions in oncology, but the process of evaluating these patients' images requires meticulous comparison with prior studies. This demands a highly precise and time-consuming assessment from the radiologist. Even today, this challenge persists as a significant hurdle that radiology clinics must overcome. To address this need Response Evaluation Criteria in Solid Tumors (RECIST) guideline was developed. In 2009, RECIST underwent a substantial revision, resulting in the publication of RECIST 1.1. This updated guideline provides a standardized approach to reporting by solid tumor measurement and defines objective criteria for assessing changes in tumor size. Radiologists can ensure a more consistent and reliable approach to oncological reporting by using RECIST 1.1guideline [7].

We aimed to provide a new perspective on their potential to facilitate reporting and improve reporting efficiency in oncological imaging by comparatively assessing LLMs' knowledge of RECIST 1.1 both among themselves and with general radiologist (GR).

METHODS

Radiologist (E.Ç.) prepared the 25 multiple-choice questions in this study utilizing the information in RECIST 1.1, thus eliminating the need for ethics committee approval. To ensure transparency and reproducibility, all the questions used in this letter and datas are included in Supplementary Material 1. We initiated the input prompt as follows: ''Act like a professor of radiology who has 30 years of experience in oncological radiology.. Give just letter of the most correct choice of multiple choice questions. Each question have only one correct answer.''This prompt was tested in June 2024 on seven different LLMs using the default settings. The testing included models from various developers: Claude 3 Opus, ChatGPT 4 and ChatGPT 4o, Gemini 1.5 pro, Mistral Large, Meta

Llama 3 70B, Perplexity pro. Also GR (T.C.) board certified by EDiR and with 6 years of experience each, answered the same questions. Radiologist (E.Ç.) assessed the responses provided by LLMs and GR (T.C.), categorizing them as either correct (1) or incorrect (0).

Results

The results revealed that Claude 3 Opus achieved the highest accuracy of 100% (25/25 questions), followed by newest model of Open AI's ChatGPT 4o with 96% accuracy (24/25 questions). ChatGPT 4 and Mistral Large 92% (23/25 questions), Meta Llama 3 70 B, Perplexity pro and Gemini 1.5 pro had accuracy of 88% (21/25 questions). GR (T.C.) has accuracy of 92% (23/25 questions).

Discussion

The outstanding success of Claude 3 Opus by knowing all the questions raises the question of whether oncology can be a new star among LLMs in radiology. Our study reveals that the majority of LLM models exhibit a commendable level of proficiency and comperable to GR in answering RECIST 1.1 related questions. However, it is important to note that there are some variations in the performance of different models. These disparities can be ascribed to the distinct architectural design and training methodologies employed by each LLM. Our findings show that current LLM models have more than sufficient text-based information about RECIST 1.1. Additionally, our findings underscore the high potential of LLMs as tools to assist radiologists in oncology reporting. However, to take full advantage of LLMs' abilities in oncology reporting, it is of great importance that their visual abilities are also evaluated. Visual evaluation forms the basis of radiology. Therefore, future studies should focus on evaluating the visual information of multimodal LLMs with visual evaluation ability.

In conclusion, LLMs have great potential in oncological radiology as in every field of radiology, and new studies that will reveal this potential will allow LLMs to be more easily integrated into radiological practice.

REFERENCES

[1]   S. Kim, C.-K. Lee, and S.-S. Kim, "Overview of Large Language Models (LLMs) Large Language Models: A Guide for Radiologists," 2024, doi: 10.3348/kjr.2023.0997.

[2]   E. Çamur, T. Cesur, and Y. C. Güneş, "Accuracies of large language models in answering radiation protection questions," Journal of Radiological Protection, vol. 44, no. 2, p. 024501, May 2024, doi: 10.1088/1361-6498/AD4B29.

[3]   E. Çamur, T. Cesur, and Y. C. Güneş, "Comparative Evaluation of the Accuracies of Large Language Models in Answering VI-RADS-Related Questions," Korean J Radiol, vol. 25, 2024, doi:

10.3348/KJR.2024.0438.

[4]   A. Caglayan, W. Slusarczyk, R. D. Rabbani, A. Ghose, V. Papadopoulos, and S. Boussios, "Large
        Language Models in Oncology: Revolution or Cause for Concern?," Current Oncology 2024, Vol.
        31,   Pages   1817-1830,   vol.   31,   no.   4,   pp.   1817–1830,   Mar.   2024,   doi:
        10.3390/CURRONCOL31040137.

[5]   R. Odabashian et al., "Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study
        with ASCO-SEP Benchmarks.," JMIR AI, vol. 3, no. 1, p. e50442, Jan. 2024, doi: 10.2196/50442.

[6]   V. Sorin, Y. Barash, E. Konen, and E. Klang, "Large language models for oncological applications,"
        J Cancer Res Clin Oncol, vol. 149, no. 11, pp. 9505–9508, Sep. 2023, doi: 10.1007/S00432-023-
        04824-W.

[7]   E. A. Eisenhauer et al., "New response evaluation criteria in solid tumours: Revised RECIST
        guideline (version 1.1)," Eur J Cancer, vol. 45, pp. 228–247, doi: 10.1016/j.ejca.2008.10.026.

Supplementary Files