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

Eren Çamur, Turay Cesur, Yasin Celal Güneş

Submitted to: JMIR AI
on: July 26, 2024

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 Ankara TR
[2]Ankara Mamak State Hospital Ankara TR
[3]TC Saglik Bakanligi Kirikkale Yuksek Ihtisas Hastanesi K?r?kkale TR

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

Abstract

This study aims to assess the potential of large language models (LLMs) to enhance reporting efficiency and accuracy in oncological imaging, specifically evaluating their knowledge of RECIST 1.1 guidelines. While the capabilities of LLMs have been explored across various domains, their specific applications in radiology are of significant interest due to the intricate and time-consuming nature of image evaluation in oncology. We conducted a comparative analysis involving seven different LLMs and a general radiologist (GR) to determine their proficiency in responding to RECIST 1.1-based multiple-choice questions.

Our methodology involved the creation of 25 multiple-choice questions by a board-certified radiologist, ensuring alignment with RECIST 1.1 guidelines. These questions were presented to seven LLMs—Claude 3 Opus, ChatGPT 4, ChatGPT 4o, Gemini 1.5 Pro, Mistral Large, Meta Llama 3 70B, and Perplexity Pro—as well as to a GR with six years of experience. The LLMs were prompted to answer as an experienced radiologist, and their responses were compared to those of the GR.

The results demonstrated that Claude 3 Opus achieved a perfect accuracy of 100% (25/25), followed closely by ChatGPT 4o with 96% (24/25). ChatGPT 4 and Mistral Large both scored 92% (23/25), while Meta Llama 3 70B, Perplexity Pro, and Gemini 1.5 each scored 88% (21/25). The GR also achieved a score of 92% (23/25). These findings highlight the impressive proficiency of current LLMs in understanding and applying RECIST 1.1 guidelines, suggesting their potential as valuable tools in radiology.

The outstanding performance of Claude 3 Opus raises the prospect of LLMs becoming integral to oncology practices, potentially enhancing the accuracy and efficiency of radiology reporting. However, the variations in performance among different models underscore the need for further refinement and evaluation. Additionally, while this study focused on text-based responses, the visual assessment capabilities of multimodal LLMs remain unexplored. Given the visual nature of radiology, future research should investigate the integration of visual analysis in LLMs to fully harness their potential in clinical settings.

In conclusion, our study underscores the high potential of LLMs to assist radiologists in oncological reporting, providing a consistent and reliable approach to interpreting RECIST 1.1 guidelines. These findings advocate for the continued development and integration of LLMs in radiology to enhance diagnostic accuracy and reporting efficiency.

(JMIR Preprints 26/07/2024:64805)
DOI: https://doi.org/10.2196/preprints.64805

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

Dear Editor,

We read with great interest studies by Odabashian et al., published in January issue of JMIR Bioinformatics and Biotechnology [1]. This paper reveals the potential of ChatGPT which is one of the best known large language models (LLMs) to serve as a supportive tool for clinical decisions in oncology practice, to provide instant information for healthcare providers, and to answer the question of whether it can be a new and effective educational aid in oncology . 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. 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. Given the potential of LLMs, researchers are keen to harness the full power of these innovative tools. 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].

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).

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 1.1 guideline [4].

Radiologist (E.Ç.) who obtained board certified (EDiR) prepared the 25 multiple-choice questions in this letter 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 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 oncology radiology, especially has studies on RECIST 1.1. Give just letter of the most correct choice of multiple choice questions which i’m going to ask you. 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 in radiology, answered the same questions. Radiologist (E.Ç.) assessed the answers provided by LLMs and GR (T.C.), categorizing them as either correct (1) or incorrect (0).

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 Google Gemini 1.5 had an accuracy of 88% (21/25 questions). GR (T.C.) has accuracy of 92% (23/25 questions).

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 comparable 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.

References

[1] 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.

[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] K. Nassiri and M. A. Akhloufi, "Recent Advances in Large Language Models for Healthcare," *BioMedInformatics 2024, Vol. 4, Pages 1097-1143*, vol. 4, no. 2, pp. 1097–1143, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020062.

[4] 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.