ORIGINAL ARTICLE

European Archives of Medical Research

DOI: 10.14744/eamr.2025.24861
Eur Arch Med Res 2025;41(3):146–153

Temporal Evolution of Large Language Models For Response Evaluation Criteria in Solid Tumors-Based Response Evaluation After Locoregional Therapy

Esat Kaba,[1] Davide Giardino,[2] Arvin Naeimi,[3] Merve Solak,[1] Mehmet Beyazal,[1] Fatma Beyazal Çeliker,[1] Thomas Vogl[4]

[1]Department of Radiology, Recep Tayyip Erdogan University, Training and Research Hospital, Rize, Türkiye
[2]Department of Radiology, Università degli Studi di Udine, Udine, Italy
[3]School of Medicine, Guilan University of Medical Sciences, Rasht, Gilan, Iran
[4]Department of Diagnostic and Interventional Radiology, Frankfurt University Hospital, Frankfurt, Germany

ABSTRACT

Objective: This study aimed to evaluate the response evaluation criteria in solid tumors (RECIST) using tumor measurements from computed tomography (CT) reports of hepatocellular carcinoma (HCC) patients before and after transcatheter arterial chemoembolization with various large language models (LLMs).

Materials and Methods: Ninety-three patients were included after the exclusion criteria were applied. RECIST assessments were performed using Bard, Bing, and ChatGPT-4 in 2023, and their updated versions–ChatGPT-4, Gemini, and Copilot–in 2025. Evaluations were based on RECIST categories determined by baseline and follow-up measurements of the longest tumor diameters from contrast-enhanced CT scans. A zero-shot prompting was used for the LLM inputs. LLM-generated RECIST classifications were compared with radiologist reports. Model performance was assessed in both years, and changes over time were analyzed.

Results: ChatGPT-4 (both 2023 and 2025) and Copilot (2025) achieved perfect scores across accuracy, precision, recall, and F1 (all 1.000). Gemini improved significantly, with accuracy rising from 0.581 in 2023 (as Bard) to 0.989 in 2025. Bing's accuracy also increased from 0.839 to 1.000 after being updated to Copilot. Cohen's Kappa showed moderate agreement between ChatGPT-4 and Bing in 2023 (κ=0.612, $p<0.001$) and perfect agreement between ChatGPT-4 and Copilot in 2025 (κ=1.000). McNemar's test showed no significant change for ChatGPT-4 between 2023 and 2025 ($p=1.000$), while Gemini and Copilot improved significantly ($p<0.0001$ and $p=0.0003$).

Conclusion: LLMs demonstrate strong potential in RECIST evaluation from CT reports in HCC patients, and ongoing improvements suggest they may increasingly aid radiological assessments in the future.

Keywords: Baseline, Follow-up, Hepatocellular carcinoma, Large language model, Response evaluation criteria in solid tumors

Cite this article as: Kaba E, Giardino D, Naeimi A, Solak M, Beyazal M, Beyazal Çeliker F, et al. Temporal Evolution of Large Language Models For Response Evaluation Criteria in Solid Tumors-Based Response Evaluation After Locoregional Therapy. Eur Arch Med Res 2025;41(3):146–153.

Address for correspondence: Esat Kaba, Department of Radiology, Recep Tayyip Erdogan University, Training and Research Hospital, Rize, Türkiye
E-mail: esatkaba04@gmail.com **ORCID ID:** 0000-0001-7464-988X

Submitted: 27.04.2025 **Revised:** 13.05.2025 **Accepted:** 02.06.2025 **Available Online:** 12.09.2025

European Archives of Medical Research – Available online at www.eurarchmedres.org

INTRODUCTION

Artificial intelligence (AI) has advanced significantly in recent years and has acquired a significant role in many areas of diagnostic and interventional radiology.[1] Numerous studies on AI applications have been published, covering a broad spectrum of topics from image reconstruction to diagnostic support, report writing, and dose optimization.[2] Large language models (LLMs), a type of generative AI, have gained widespread attention in the past year, sparking discussions about their role in radiology.[3] These discussions have primarily centered on report generation, text mining in radiology reports, and report optimization.[4,5] Beyond diagnostic radiology, their use in interventional radiology is also under investigation.[6] Over the last few years, several notable LLMs have been released, including ChatGPT-3.5 and 4 by OpenAI, Bard by Google, and Bing by Microsoft. Studies have examined the knowledge levels of these models in radiology and their potential applications across various areas. For instance, a study by Bhayana et al.[7] evaluated ChatGPT's performance on the radiology board exam and found that it answered 69% of the questions correctly.

In another study, the high text analysis capabilities of these models were leveraged to investigate the success of detecting incidental findings in radiology reports using a single-shot learning prompt technique, yielding highly satisfactory results.[8] In addition, Schmidt et al.[9] focused on the ability of LLMs to detect speech errors in radiology reports. In this study, ChatGPT-4 demonstrated high accuracy in detecting both clinically significant errors (precision, 76.9%; recall, 100%; F1 score, 86.9%) and clinically insignificant errors (precision, 93.9%; recall, 94.7%; F1 score, 94.3%). Based on such studies, the capacity of LLMs to analyze data in radiology reports appears quite noteworthy.

We designed this study to determine the level of knowledge of LLMs about response evaluation criteria in solid tumors (RECIST), which is a critical component of diagnostic and interventional radiology in the follow-up of malignancy after treatment, and to reveal the potential use and reliability of LLMs in this regard in the future. In this study, we evaluated and compared the performance of Bard, Bing, and ChatGPT-4 in the RECIST assessment. In addition, we conducted the same evaluation using the latest versions of these LLMs to demonstrate their progress in this domain over time.

MATERIALS AND METHODS

Study Cohort

A publicly available dataset comprising hepatocellular carcinoma (HCC) patients was utilized in the present study.[10,11] The dataset contained 105 HCC patients who were monitored with contrast-enhanced computed tomography (CT) both before and after the transcatheter arterial chemoembolization (TACE) procedure. The specific CT technique employed in the dataset was described in a previous study.[11] The mean baseline examination time in the dataset was 3 weeks before TACE, and the mean follow-up time was 9 weeks after the procedure. The dataset also contained reports from three radiologists, including measurements of the longest diameter of lesions and the RECIST assessment. For this study, we utilized the measurements and RECIST assessments from Radiologist 1, which included evaluations for 93 out of 105 patients. In other words, baseline and follow-up measurements for 93 HCC patients were included in our study. All measurements were taken at the longest diameter of the tumor, and the radiologist conducted a RECIST assessment based on the tumor's growth and shrinkage rates. According to the radiologist's assessment, this data set included 19 patients with complete response (CR), 40 patients with partial response (PR), 28 patients with stable disease (SD), and 6 patients with progressive disease (PD). These assessments were validated by an experienced interventional radiologist (T.J.V.) with over 25 years of experience. This study was conducted following the Helsinki Declaration. Ethics committee approval and patient informed consent were not required since a publicly available dataset was used for this retrospective study.

LLMs

This study aimed to evaluate and compare the classification performance of LLMs in assessing treatment response based on the RECIST criteria. Six LLMs were included: ChatGPT4 (2023), Bard (2023), Bing (2023), and their respective updated versions in 2025: ChatGPT4 (2025) (https://chatgpt.com/), Gemini (2025) (https://gemini.google.com/app?hl=tr), and Copilot (2025) (https://copilot.microsoft.com/). Baseline and follow-up tumor measurements were provided to LLM's chatbots by entering the following initial prompt.

Prompt

"In an interventional radiology unit where TACE for HCC is frequently performed, we need a RECIST classification to evaluate the response of lesions to treatment. Can you use the pre-treatment (baseline) and post-treatment (follow-up) values of the longest diameter of the patient's tumors to determine the percentage of tumor growth and shrinkage and perform RECIST assessment of each tumor?" This prompt provided default hyperparameters to three different models and their updated versions in November 2023 and March 2025, respectively.

Moreover, no explanation or information was provided to the models during the RECIST assessment. In other words, the zero-shot learning technique was used for prompting. RECIST assessment responses of all models were then compared with the radiologist's report as the gold standard label.

Statistical Analysis

To assess the classification performance of LLMs across different versions and RECIST categories (CR, PR, SD, PD), standard evaluation metrics, including accuracy, macro precision, macro recall, and macro F1 score, were calculated. Confusion matrices were generated to visualize the distribution of correct and incorrect predictions across classes. To evaluate the agreement between models, Cohen's Kappa coefficient was computed. Cohen's Kappa analysis is interpreted as shown in Table 1.[12] Performance differences between model versions from 2023 to 2025 were analyzed using McNemar's test. This test was applied both globally and within each RECIST category (PR, SD, PD) to determine whether the changes in classification performance were statistically significant across time. In addition, McNemar's test was used for pairwise comparisons between different models within the same RECIST class to evaluate relative classification differences. The Wilcoxon signed-rank test was also used to compare model rankings based on accuracy and F1 score. $P<0.05$ was considered indicative of statistical significance. All statistical analyses and visualizations were performed using IBM Statistical Package for the Social Sciences Statistics version 23 (IBM Corp., Armonk, NY, USA) and the Python programming language (scikit-learn, matplotlib, and seaborn libraries).

RESULTS

Of the 93 patients in the study, 59 were male and 34 were female. The mean age of patients was 68.8±10.53. The mean baseline diameter of the lesions was 79.97 mm (15.6–243.9), and the mean follow-up diameter was 45.6 mm (0–173.9 mm).

All LLMs defined it as CR if the lesion had completely disappeared, PR if there was more than 30% shrinkage, SD if there was between 20% growth and 30% shrinkage, and PD if there was more than 20% growth. As a result, all models automatically considered the RECIST version 1.1 criteria without any RECIST definition. In addition, all chatbots calculated the percentage of change according to the following formula without any information from us.

Table 1. Interpretation of Cohen's Kappa statistic

Kappa value (κ)	Level of agreement
κ<0	Poor agreement (less than chance)
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

Overall Performance

Among all models, ChatGPT 4 (2023 and 2025) and Copilot (2025) yielded optimal classification performance, achieving maximum values across all evaluated metrics (Accuracy, Precision, Recall, and F1 Score=1.00). Gemini (2025) demonstrated high overall performance (Accuracy=0.989, F1 Score=0.973), indicating a substantial improvement compared to its 2023 version (Bard), which exhibited the lowest performance across all metrics (Accuracy=0.581, F1 Score=0.437). Bing (2023) showed intermediate performance (Accuracy=0.839, F1 Score=0.776). A detailed analysis is presented in Table 2, and confusion matrices for all models are shown in Figure 1.

Model Agreement and Evolution

Cohen's Kappa analysis revealed moderate to substantial agreement among models in 2023, with the highest agreement between ChatGPT4 and Bing (κ=0.612, $p<0.001$) (Table 3). In contrast, 2025 models showed almost perfect agreement, particularly between ChatGPT4 and Copilot (κ=1.000) (Table 4).

McNemar's tests were performed to evaluate performance differences between model versions over time. No significant difference was found between ChatGPT4 (2023 vs. 2025) ($p=1.000$), suggesting temporal consistency. In contrast, both Gemini and Copilot exhibited statistically significant improvements between 2023 and 2025 ($p<0.0001$ and $p=0.0003$, respectively) (Table 5). Gemini misclassified only one patient by labeling the case as SD, although it was actually PD. Similarly, Bard and Bing also misclassified this same patient in 2023, incorrectly identifying it as a PR.

Class-level Analysis

The accuracies of the models in all groups are shown in Figure 2. Class-based comparisons showed that Gemini (2025) demonstrated statistically significant improvements in both PR and SD categories ($p<0.001$), while the difference in PD did not reach statistical significance ($p=0.0736$). Similarly, Copilot

Table 2. Performance metrics of the large language models in 2023 and 2025

Model	Accuracy	Precision	Recall	F1 score
ChatGPT4 (2023)	1.0	1.0	1.0	1.0
Bard (2023)	0.581	0.429	0.544	0.437
Bing (2023)	0.839	0.921	0.738	0.776
ChatGPT4 (2025)	1.0	1.0	1.0	1.0
Gemini (2025)	0.989	0.991	0.958	0.973
Copilot (2025)	1.0	1.0	1.0	1.0

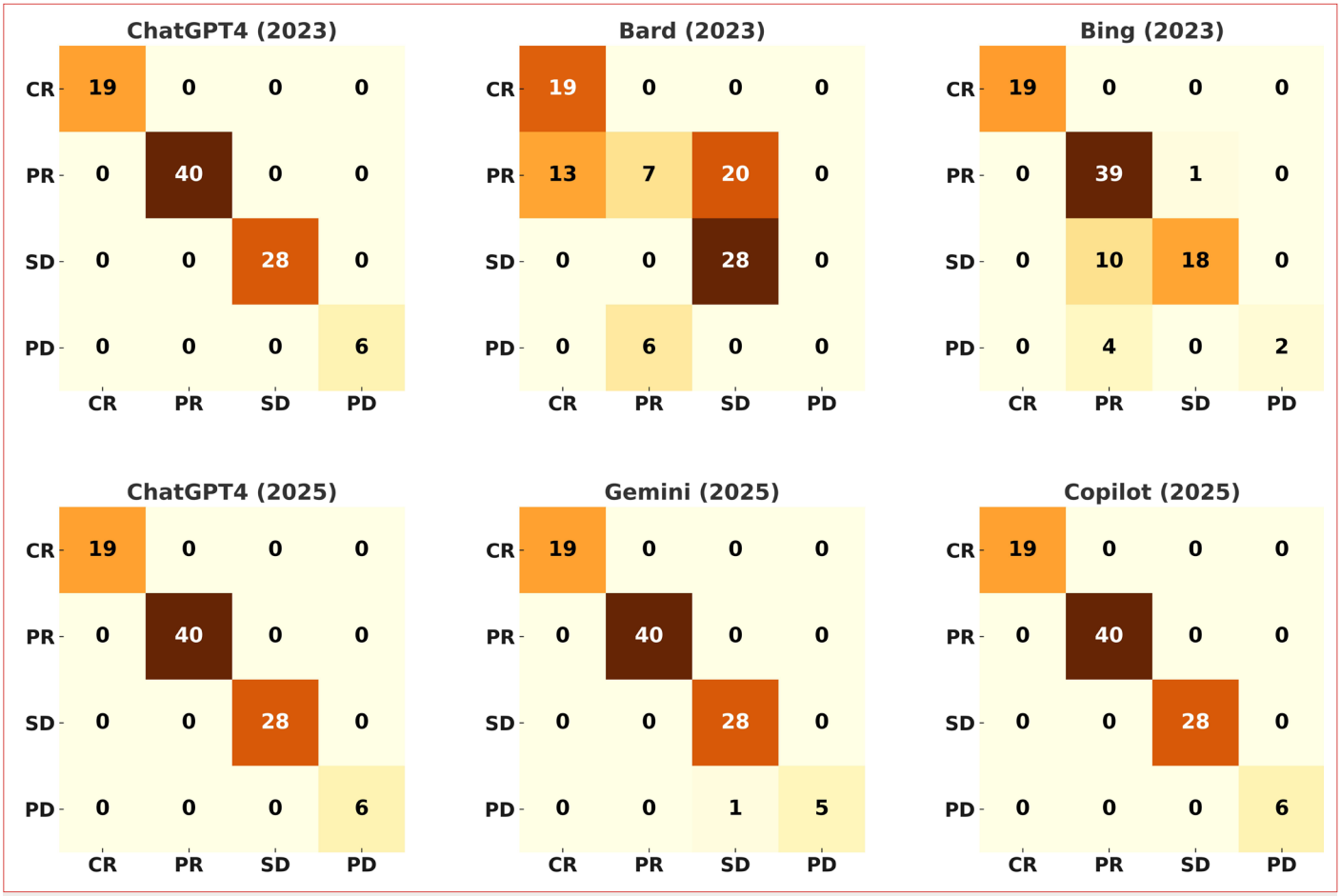

Figure 1. Confusion matrices for all models.

Table 3. In 2023, the agreement analysis between the large language models

Models (2023)	Cohen's kappa	p*
ChatGPT4 versus bard	0.418	<0.001
ChatGPT4 versus bing	0.612	<0.001
Bard versus bing	0.349	<0.001

*Wilcoxon signed-rank test.

Table 4. In 2025, the agreement analysis between the LLMs. The Wilcoxon signed-rank test could not be computed for the comparison between ChatGPT4 and Copilot, as all paired predictions were identical, resulting in zero differences across the answers

Models (2025)	Cohen's kappa	p*
ChatGPT4 versus Gemini	0.978	0.317
ChatGPT4 versus Copilot	1.000	-
Gemini versus Copilot	0.978	0.317

*Wilcoxon signed-rank test. LLMs: Large language models.

(2025) significantly improved in the SD category (p=0.004) but not in the PR or PD classes (Table 6). All models demonstrated 100% accuracy in identifying cases within the CR group.

DISCUSSION

This study investigated the accuracy of Bard, Bing, and ChatGPT-4 in 2023 and their updated versions in 2025 in the assessment of response to locoregional therapy in patients with HCC. We provided the models with the baseline and follow-up longest tumor diameters as input. Afterward, the models were prompted to calculate tumor growth/shrinkage percentages and then perform a RECIST evaluation. No information other than baseline and follow-up values was provided during the LLM evaluation. The results were obtained without manipulating the models in any direction (zero-shot learning). As a result,

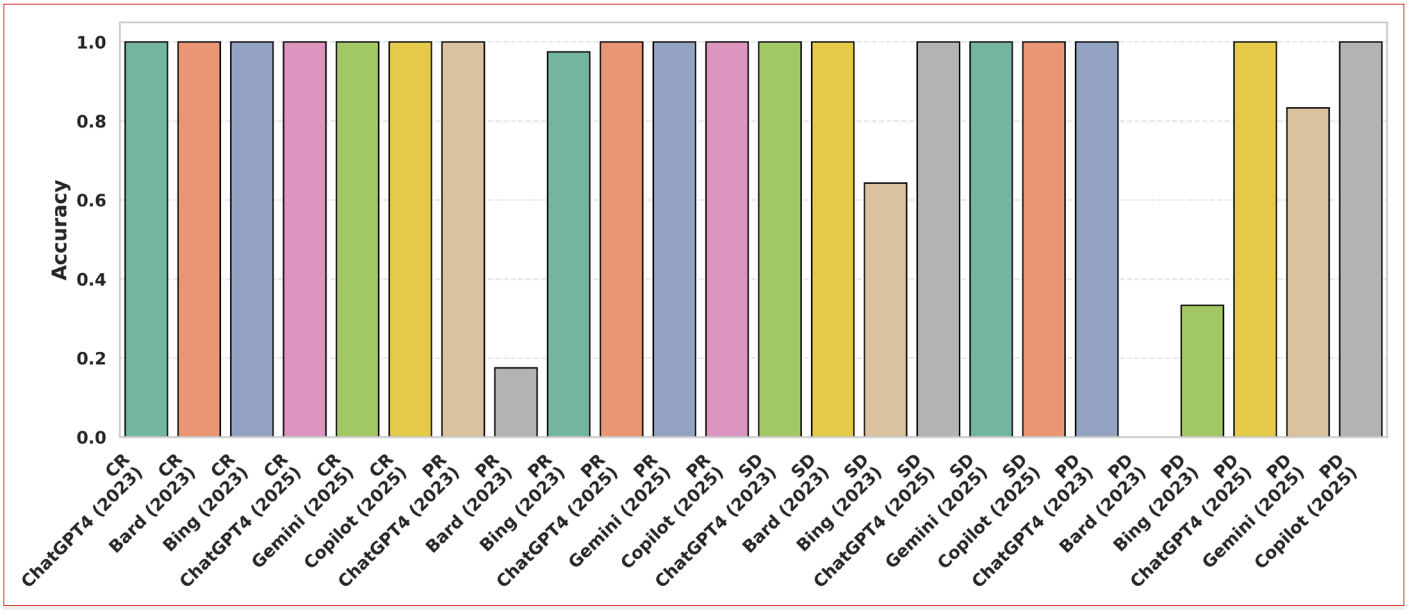

Figure 2. The accuracies of the large language models in all response evaluation criteria in solid tumors groups.

Table 5. Accuracy analysis of answers across years. The number of patients for which the models answered correctly in 2023 and incorrectly in 2025, and the number of patients for which the models answered incorrectly in 2023 and correctly in 2025 are indicated, and statistical analysis is presented

Model evolution	2023 correct/ 2025 incorrect	2023 incorrect/ 2025 correct	McNemar p-value
ChatGPT4 (2023 → 2025)	0	0	0.000
Bard → Gemini	0	38	<0.0001
Bing → Copilot	0	15	0.0003

while minor variations existed between the models, all LLMs demonstrated high accuracy in calculating the percentage of tumor diameter change. In the RECIST assessment based on these calculations, both ChatGPT-4 versions achieved 100% accuracy. Bard and Bing, on the other hand, significantly improved their accuracy from their 2023 version to the 2025 version.

LLMs have rapidly influenced millions worldwide. In medicine, researchers are actively exploring their potential to transform the field, particularly through their capacity to process vast medical texts using natural language processing. This ability holds promise for analyzing reports, identifying patient data patterns, and supporting clinical decisions. Their applications in medicine fall into three domains: Research, education, and

Table 6. Accuracy analysis of class-based answers across years. The number of patients for which the models answered correctly in 2023 and incorrectly in 2025, and the number of patients for which the models answered incorrectly in 2023 and correctly in 2025 are indicated, and statistical analysis is presented

Class	Models	2023 correct / 2025 incorrect	2023 incorrect / 2025 correct	McNemar p-value
PR	Bing (2023) → Copilot (2025)	0	1	1.000
SD	Bing (2023) → Copilot (2025)	0	10	0.004
PD	Bing (2023) → Copilot (2025)	0	4	0.134
PR	Bard (2023) → Gemini (2025)	0	33	<0.001
SD	Bard (2023) → Gemini (2025)	0	32	<0.001
PD	Bard (2023) → Gemini (2025)	0	5	0.0736

PR: Partial response; SD: Stable disease; PD: Progressive disease.

patient care.[13] In research, they assist with data analysis, literature review, and coding. In education, both medical and patient education are being explored, with studies assessing LLMs' success in medical examinations and their role in producing patient education materials.[14,15] For instance, Gilson et al.[14] evaluated ChatGPT's performance on the United States Medical Licensing Examination, finding it achieved over 60% accuracy, comparable to a 3rd-year medical student's basic competency. In patient care, LLMs may summarize histories, analyze images, and interpret clinical data, underscoring their versatility and broad applicability in clinical settings.[16]

Potential applications of LLM in radiology also cover a wide range of topics.[17] Given the advanced text analysis capabilities of LLMs, the simplification of radiology reports and the extraction of some analytics from them are one of the main topics of research. In this regard, Fink et al.[18] compared GPT-4 and ChatGPT to extract lesion parameters from CT reports and identify metastatic status and label oncological progression. The authors report that GPT-4 achieved 98.6% accuracy in extracting lesion parameters, 98.1% accuracy in identification of metastatic status, and an F1 score of 0.96 in labeling oncological progression. In all tasks, GPT-4 showed superior performance in their study.

In our study, we tested and compared the accuracy of four different LLMs in calculating the growth and shrinkage rate of lesions and subsequent RECIST assessment, which are very important in diagnostic and interventional radiology. The findings of our study reveal that LLMs, especially updated versions, correctly classify patients, providing the most important hope for the future. Several studies in the literature have evaluated the performance of LLMs at different time points, highlighting their potential for continued improvement over time. A study evaluated the performance of LLMs (GPT-4, GPT-3.5, Claude, and Google Bard) on radiology exam questions over 3 months, focusing on accuracy, subspecialty performance, and internal consistency.[19] Among the models, GPT-4 achieved the highest overall accuracy, although a slight downward trend was observed over time, while Claude demonstrated gradual improvement. The observed reduction in intra-model discordance across all models suggests increasing internal consistency, yet persistent variability across subspecialties and difficulty with fact-based questions highlight the current limitations of LLMs in domain-specific medical reasoning. Our study also found that while ChatGPT-4 was the most accurate model among the 2023 versions, both Bard and Bing showed a marked improvement in their 2025 versions.

Considering that LLMs can potentially be integrated into hospital image archiving and communication systems in the future, treatment response assessment applications can be created quickly and effectively by accessing patients' reports and using baseline and follow-up measurement results.[20] Evaluating our findings with other studies in the literature, we conclude that in the future, LLMs, particularly advanced models like ChatGPT-4, will not only support physicians in their demanding daily routines but will also enable them to perform certain analyses of patients' radiology reports. These models can also support the patient-physician interview by offering high accuracy in assessing treatment responses based on radiology reports. This capability is crucial in oncology, where accurate assessment of treatment responses, such as in HCC, directly influences clinical decision-making. In addition, LLMs could enhance patient understanding by converting complex medical information into clear, accessible language, aiding patients and their families in comprehending diagnoses, treatment options, and expected outcomes. This improved communication can lead to better patient engagement and adherence to treatment plans. Ultimately, integrating LLMs into clinical workflows can promote a more efficient and patient-centered approach in healthcare, leading to improved outcomes for both physicians and patients.

The rapid development and strong performance of LLMs offer significant promise for healthcare innovation. However, their deployment brings several limitations and ethical concerns. A major issue is the inheritance of biases from training data, whose sources and quality are often uncertain, making bias correction challenging.[20,21] In addition, LLMs cannot verify the accuracy of their outputs, leading to potential misinformation or "hallucinations," which may cause serious harm in medical contexts. Complex architectures also limit interpretability and hinder transparency and accountability.[22] While techniques such as the Chain of thought may help address some issues, robust solutions remain necessary. Patient privacy is another critical consideration; although anonymized data can be used, complete anonymization cannot always be guaranteed and increases ethical risks. Therefore, LLMs must be strictly secured against unauthorized access in healthcare settings.[20,22,23]

The current study has some limitations. First, the number of patients was limited. Studies investigating the RECIST assessment of LLMs using much larger datasets are needed. Second, the present study only reported the baseline and follow-up longest diameters instead of providing the complete radiology reports. In addition, the RECIST assessment is not only based on the longest diameter but also includes many other criteria. Hence, more studies with complete reports and larger sample sizes that include all the criteria used in the RECIST assessment are recommended. However, in our study, we have demonstrated the capabilities of LLM in the field of radiology for exploratory purposes. As more comprehensive reports become available, a comprehensive assessment of RECIST with LLMs can be conducted in the future.

CONCLUSION

This study has demonstrated the evolution of LLMs over the years in terms of some medical knowledge and as an effective tool in RECIST assessment and revealed their potential advantages in various medical and radiological contexts. LLMs can efficiently assess treatment responses in radiology reports and offer valuable support in managing the increasing workload of radiologists. Integrating LLMs into clinical workflows can support radiologists in current challenges and improve the overall quality of care, ultimately contributing to better outcomes for both physicians and patients.

DECLARATIONS

Ethics Committee Approval: Not required.

Informed Consent: Not required.

Conflict of Interest: The authors declare that there is no conflict of interest.

Funding: The authors received no financial support for the research and/or authorship of this article.

Use of AI for Writing Assistance: The authors used ChatGPT to correct grammar and English translations. The content of the publication is entirely the responsibility of the authors, and the authors reviewed it and made the necessary corrections.

Authorship Contributions: Concept – None; Design – EK, DG, AN, MS, MB, FBÇ, TV; Supervision – None; Fundings – None; Materials – EK, DG, AN, MS, MB, FBÇ, TV; Data collection &/or processing – EK, DG, AN, MS, MB, FBÇ, TV; Analysis and/or interpretation – EK, DG, AN, MS, MB, FBÇ, TV; Literature search – EK; Writing – EK, DG, AN, MS, MB, FBÇ, TV; Critical review – EK, DG, AN, MS, MB, FBÇ, TV.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Boeken T, Feydy J, Lecler A, Soyer P, Feydy A, Barat M, et al. Artificial intelligence in diagnostic and interventional radiology: Where are we now? Diagn Interv Imaging 2023;104:1–5.
2. Barreiro-Ares A, Morales-Santiago A, Sendra-Portero F, Souto-Bayarri M. Impact of the rise of artificial intelligence in radiology: What do students think? Int J Environ Res Public Health 2023;20:1589.
3. Bajaj S, Gandhi D, Nayar D. Potential applications and impact of ChatGPT in radiology. Acad Radiol 2024;31:125–61.
4. Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. AJR Am J Roentgenol 2023;221:373–6.
5. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023;309:e232561.
6. Scheschenja M, Viniol S, Bastian MB, Wessendorf J, König AM, Mahnken AH. Feasibility of GPT-3 and GPT-4 for in-depth patient education prior to interventional radiological procedures: A comparative analysis. Cardiovasc Intervent Radiol 2024;47:245–50.
7. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. Radiology 2023;307:e230582.
8. Bhayana R, Elias G, Datta D, Bhambra N, Deng Y, Krishna S. Use of GPT-4 with single-shot learning to identify incidental findings in radiology reports. AJR Am J Roentgenol 2024;222:e2330651.
9. Schmidt RA, Seah JCY, Cao K, Lim L, Lim W, Yeung J. Generative large language models for detection of speech recognition errors in radiology reports. Radiol Artif Intell 2024;6:e230205.
10. Moawad AW, Fuentes D, Morshid A, Khalaf AM, Elmohr MM, Abusaif A, et al. Multimodality annotated HCC cases with and without advanced imaging segmentation [dataset]. Cancer Imaging Arch; 2021.
11. Morshid A, Elsayes KM, Khalaf AM, Elmohr MM, Yu J, Kaseb AO, et al. A machine learning model to predict hepatocellular carcinoma response to transcatheter arterial chemoembolization. Radiol Artif Intell 2019;1:e180021.
12. McHugh ML. Interrater reliability: The kappa statistic. Biochem Med (Zagreb) 2012;22:276–82.
13. 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;3:141.
14. 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. Erratum in: JMIR Med Educ 2024;10:e57594.
15. Kuckelman IJ, Yi PH, Bui M, Onuh I, Anderson JA, Ross AB. Assessing AI-powered patient education: A case study in radiology. Acad Radiol 2024;31:338–42
16. Qiu J, Yuan W, Lam K. The application of multimodal large language models in medicine. Lancet Reg Health West Pac 2024;45:101048.
17. Kim K, Cho K, Jang R, Kyung S, Lee S, Ham S, et al. Updated primer on generative artificial intelligence and large language models in medical imaging for medical professionals. Korean J Radiol 2024;25:224–42.
18. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text

CT reports on lung cancer. Radiology 2023;308:e231362.

19. Gupta M, Virostko J, Kaufmann C. Large language models in radiology: Fluctuating performance and decreasing discordance over time. Eur J Radiol 2025;182:111842.
20. Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: Fundamentals, applications, ethical considerations, risks, and future directions. Diagn Interv Radiol 2024;30:80–90.
21. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023;6:120.
22. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023;5:e333–5.
23. Stahl BC, Eke D. The ethics of ChatGPT – exploring the ethical issues of an emerging technology. Int J Inf Manag 2024;74:102700.