TURKISH JOURNAL of ONCOLOGY

Turk J Oncol 2026;41(1):52–59
doi: 10.5505/tjo.2026.4733

ORIGINAL ARTICLE

The Knowledge of Large Language Models Regarding Response Evaluation Criteria in Solid Tumors: A Comparative Study with Prompt Effect

Eren ÇAMUR,[1] Turay CESUR,[2] Yasin Celal GÜNEŞ[3]

[1]Department of Radiology, Ankara 29 Mayıs State Hospital, Ankara-*Türkiye*
[2]Department of Radiology Ankara Mamak State Hospital, Ankara-*Türkiye*
[3]Department of Radiology, Kırıkkale High Specialty Hospital, Kırıkkale-*Türkiye*

OBJECTIVE

To evaluate the diagnostic performance of eight current large language models (LLMs) in applying the RECIST 1.1 guidelines for oncologic treatment response imaging and to compare their performance with that of board-certified radiologists. This study explores the potential of LLMs as supportive adjuncts in cancer follow-up imaging.

METHODS

In this observational cross-sectional study, 50 text-based and 30 case-based multiple-choice questions derived from RECIST 1.1 were administered to eight LLMs with three different prompts and two junior radiologists with seven years of experience. Responses were independently scored as correct or incorrect, and non-parametric statistical analyses were performed to compare performance across groups.

RESULTS

LLMs demonstrated promising performance in text-based interpretation about RECIST, with only minor performance variations. Claude 3.5 Sonnet had the most successful performance, achieving 83.3% accuracy on case-based and 90% on text-based questions. Other models exhibited robust performance, with no significant differences in case-based assessments between LLMs and radiologists. LLMs achieved similar results across the three different prompts with minor variations.

CONCLUSION

LLMs have great potential for response evaluation in oncological imaging and not only support radiologists but may soon redefine clinical workflows, setting a new benchmark for diagnostic excellence in radiology.

Keywords: ChatGPT; cancer; large language models; response; treatment.

INTRODUCTION

Large language models (LLMs) represent a remarkable breakthrough in natural language processing, capable of performing specific tasks in radiology without additional training.[1–4] This positions LLMs as transformative forces poised to significantly reshape radiology practice. They have the potential to usher in a new era of efficiency and excellence, both as supportive diagnostic tools and in facilitating the reporting process. Consequently, there has been a rapid increase in studies investigating the radiological knowledge of LLMs

Received: October 30, 2025
Revised: January 31, 2026
Accepted: February 01, 2026
Online: March 02, 2026

Accessible online at:
www.onkder.org

Dr. Eren ÇAMUR
Ankara 29 Mayıs Devlet Hastanesi,
Radyoloji Kliniği,
Ankara-Türkiye
E-mail: eren.camur@outlook.com

and their potential applications and contributions to radiology.[3–7] Although there are many studies evaluating the radiological knowledge of LLMs in different fields, the lack of studies evaluating their knowledge in oncology radiology is an important gap in this regard.

The radiology report is vital in guiding patient management in oncology, requiring meticulous comparison with prior studies and assessment. Response Evaluation Criteria in Solid Tumors (RECIST) guideline, revised in 2009 to RECIST 1.1, was developed to address this need. RECIST guideline comprises criteria such as defining measurable lesions (i.e., which measurement defines a measurable lymph node), identifying target lesions (i.e., which criteria the target lesion must meet), and categorizing response types (regression, stable disease, or progression). It provides a standardized approach to reporting solid tumor measurements and defines objective criteria for assessing changes in tumor size, ensuring a consistent and reliable approach to reporting.[8]

Previous studies have evaluated the proficiency and knowledge of various LLMs in different specific types of cancer.[2,9–12] Güneş et al.[13] tested the performance of current LLMs, in particular Claude 3.5 Sonnet, in interpreting BI-RADS categories via text-based questions and found that these models achieved remarkable accuracy (up to 90%), approaching the level of expertise of breast radiologists. In another study, Kaba et al.[14] demonstrated that advanced LLMs, especially ChatGPT-4, showed high accuracy (93%) in text-based questions in interpreting thyroid imaging guidelines based on the K-TIRADS classification system and emphasized the competence of LLMs in this field.

To the best of our knowledge, no study has compared the performance of LLMs in relation to RECIST 1.1, a critical guideline in the radiological reporting of follow-up imaging in cancer patients. We aimed to fill this gap by evaluating the knowledge of various LLMs in the RECIST 1.1 guideline and comparing them with that of radiologists.

MATERIALS AND METHODS

Study Design

This experimental study utilized a cross-sectional design to assess the accuracy of eight different LLMs compared with radiologists in answering text-based and case-based MCQs pertaining to RECIST 1.1. Their answers were benchmarked against responses from two board-certified radiologists (European Diploma in Radiology-EDiR): Radiologist 1 (Y.C.G.)(R1) and Radiologist 2 (T.C.)(R2), both with seven years of experience in general radiology. The text-based and case-based MCQs were designed based on the RECIST 1.1 guideline by a board-certified radiologist (EDiR), Radiologist 3 (E.Ç.)(R3), also with seven years of experience in radiology.

The MCQs did not include any authentic patient data or images; therefore, ethical committee approval was neither required nor applicable for this study. Methodological transparency and reproducibility were ensured by adhering to the Standards for Reporting Diagnostic Accuracy Studies (STARD) guideline.[15]

An overview of the flowchart is presented in Figure 1.

Data Collection for Text-based and Case-based Multiple-choice Questions

A total of 50 text-based MCQs and 30 case-based MCQs were utilized in the study. These questions comprehensively covered the all sections of RECIST 1.1 and tested the application of the information therein. Each question was carefully constructed to focus on a single, specific, and critical concept relevant to radiological practice under this guideline. Each MCQ had 5 choices and only one choice was correct. A complete list of text-based and case-based MCQs and dataset of the study are available in the Appendix.

Design of Input-output Procedures for LLMs

The three different input prompts provided to the LLMs were: Prompt 1: "Act like a professor of radiology who has 30 years of experience in oncological imaging, especially with studies on RECIST 1.1. Give just the letter of the most correct choice of multiple-choice questions that I will ask you. Each question has only one correct answer." Prompt 2: "You are a senior academic radiologist. I have some questions about RECIST 1.1. I will ask you multiple-choice questions with a single correct answer. Provide only the letter of the most accurate choice for each." Prompt 3: "I have a few questions about RECIST 1.1 criteria. Some of them are text-based, and some of them are case-based questions. I will present you with multiple-choice questions, and each has only one correct answer. Please reply with the letter corresponding to the best choice only, without any explanation." These prompts were consistently employed across eight distinct platforms with default hyperparameters by R3 in February 2025: Claude 3 Opus and 3.5 Sonnet (https://claude.ai.com), ChatGPT-o1, ChatGPT-4o (https://chat.openai.com), Gemini 1.5 Pro (https://

Turk J Oncol 2026;41(1):52–59
doi: 10.5505/tjo.2026.4733

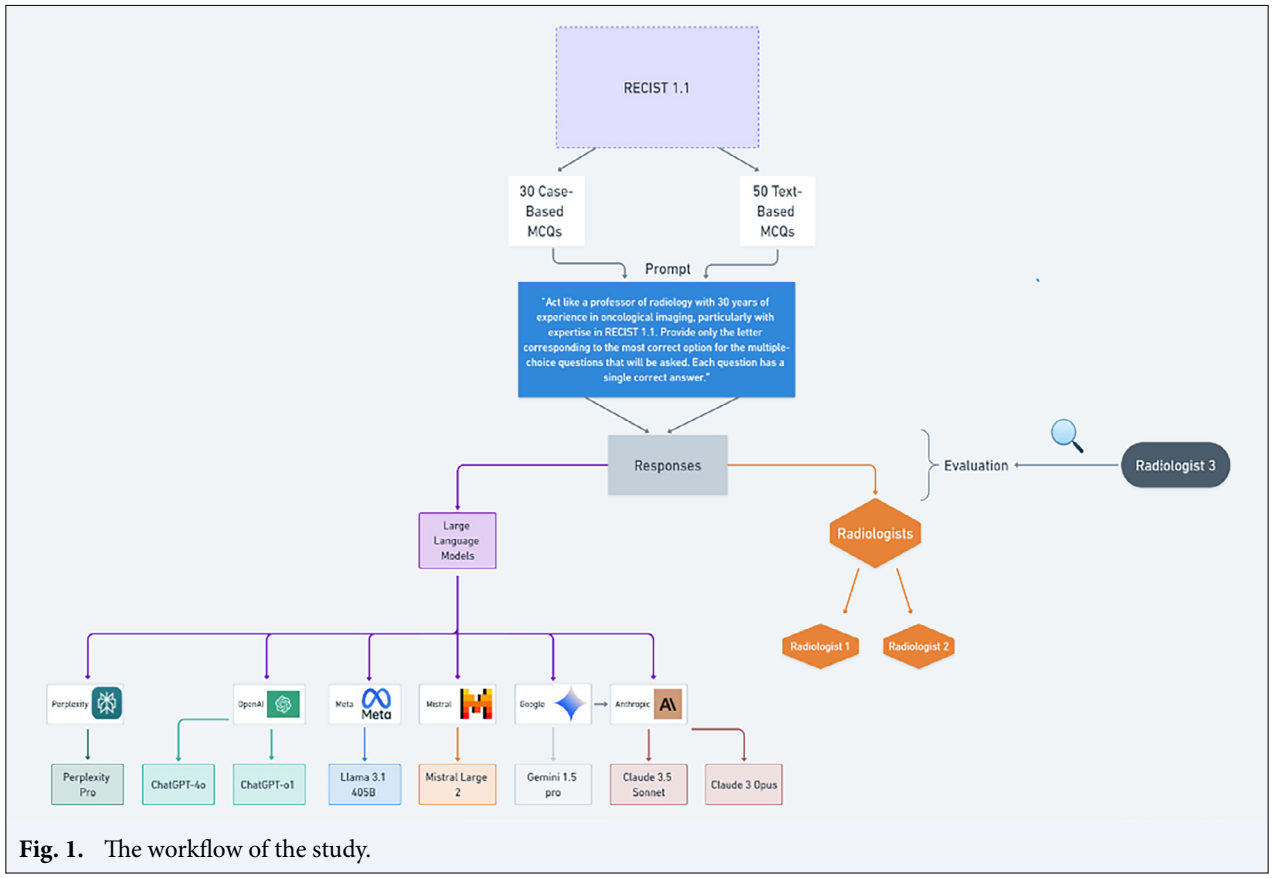

Fig. 1. The workflow of the study.

gemini.google.com), Mistral Large 2 (https://mistral.ai), Llama 3.1 405B (https://metaai.com), and Perplexity Pro (https://perplexity.ai). In order to assess the consistency in the responses of the three different prompts within each model, the responses of each prompt and the model were evaluated carefully, and the prompt with the most successful responses for all models was recorded by R3.

The MCQs were administered sequentially within a single conversation session per LLM to maintain uniformity. None of the LLMs underwent additional pre-training or fine-tuning by the study authors, and no supplementary details that could potentially affect the study results were provided (Fig. 2). R3 reviewed the LLM responses and categorized them as correct (1) or incorrect (0).

Radiologist Performance Evaluation

R1 and R2 independently answered the MCQs in a blinded manner in January 2025 using their personal computers. They completed text-based MCQs first, immediately followed by case-based MCQs without any interval. R3 separately evaluated their answers and categorized them as correct (1) or incorrect (0).

Statistical Analysis

The Kolmogorov-Smirnov test assessed data distribution. Descriptive statistics (minimum, maximum, median, interquartile range, percentages) were calculated. As the data were non-normally distributed, non-parametric tests were used. Consistency and performance across three prompts were evaluated with the Friedman and McNemar tests; the latter compared correct response rates between LLMs and radiologists. Chi-square tests assessed differences by question type. Bonferroni correction was applied for pairwise comparisons ($p \leq 0.028$ significant), while $p \leq 0.05$ indicated significance for consistency and prompt-related analyses.

RESULTS

Case-Based MCQs

Among the three different prompts (Prompt 1, 2, 3), all models achieved their highest performance on case-based MCQs with "Prompt 1". However, the variation in performance across the different prompts did not reach statistical significance, and the responses generated by all models were consistent across the three prompts ($p>0.05$) (Table 1).

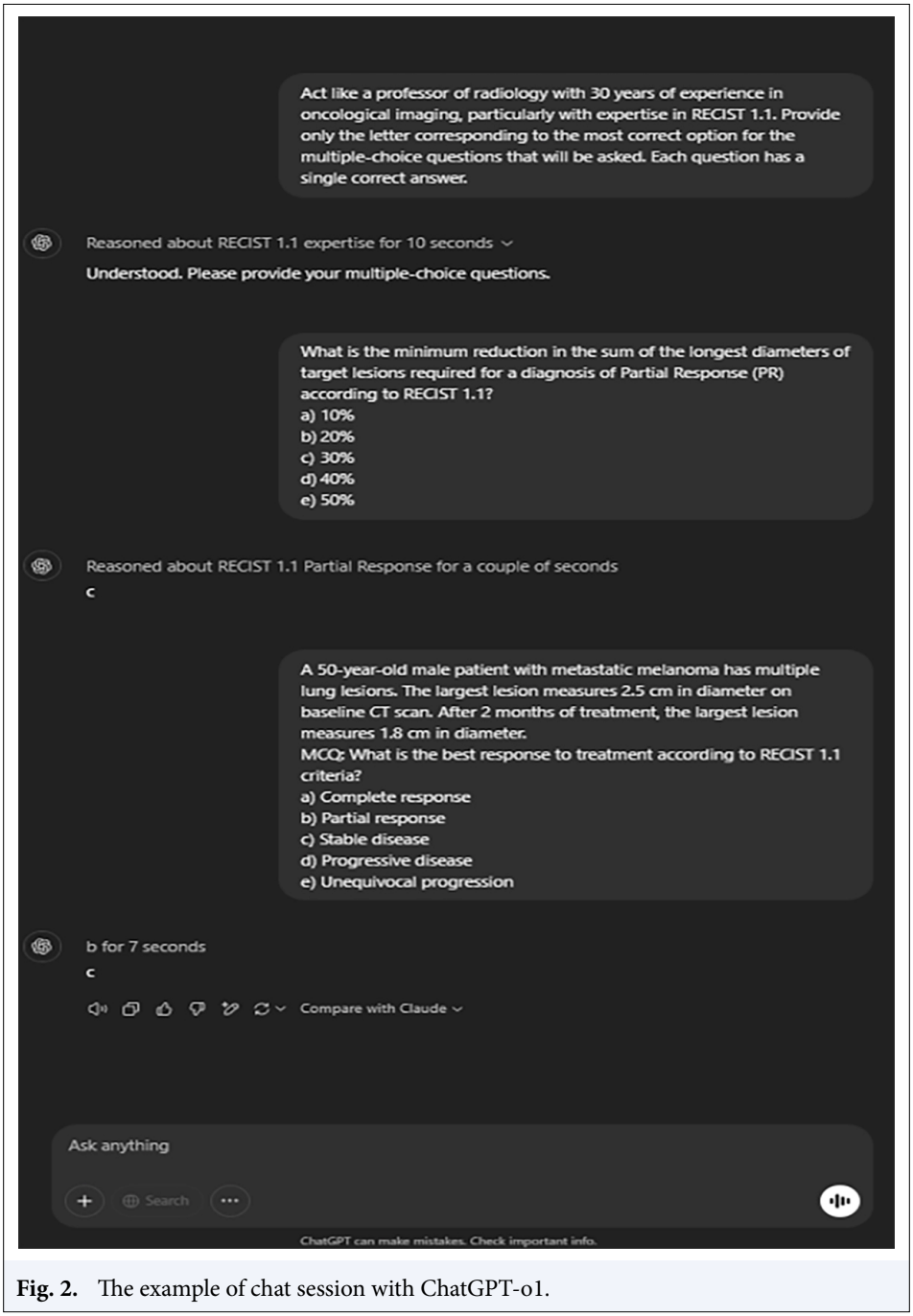

Fig. 2. The example of chat session with ChatGPT-o1.

With "Prompt 1", Claude 3.5 Sonnet demonstrated the highest accuracy at 83.3%, followed by R2 and Gemini 1.5 Pro, both of which achieved 80.0% ($p>0.028$). R1 closely followed with 76.7%, while ChatGPT-4o, Llama 3.1 405B, and Mistral Large 2 each recorded 73.3% ($p>0.028$). Claude 3 Opus and ChatGPT-o1 shared an accuracy of 66.7%. Perplexity Pro exhibited the lowest accuracy among LLMs and radiologists, with 60.0% ($p>0.028$) (Fig. 3).

There was no significant difference in accuracy on case-based MCQs among LLMs and between LLMs and radiologists ($p>0.028$) (Table 2).

Text-Based MCQs

Similar to case-based MCQ, all models reached the highest performance with 'Prompt 1' among the three different prompts. The answers given by all models to the questions with these prompts were consistent,

Turk J Oncol 2026;41(1):52–59
doi: 10.5505/tjo.2026.4733

Table 1 The consistency of LLMs responses with different prompts (Prompt 1, Prompt 2 and Prompt 3)

	Claude 3 Opus	Claude 3.5 Sonnet	Chat GPT-4o	Chat GPT-o1	Mistral Large 2	Gemini 1.5 Pro	Llama 3.1 405B	Perplexity Pro
Case-based MCQs	0.368	1	0.607	0.368	0.607	0.368	0.135	1
Text-based MCQs	0.174	0.247	0.368	0.717	0.368	0.513	0.717	0.223

Test values were obtained from Freidman test, MCQs: Multiple Choice Questions

Table 2 Comparison of the performance of LLMs and radiologist on case-based multiple-choice questions

	Claude 3 Opus	Claude 3.5 Sonnet	Chat GPT-4o	Chat GPT-o1	Mistral Large 2	Gemini 1.5 Pro	Llama 3.1 405B	Perplexity Pro	Radiologist 1	Radiologist 2
Claude 3 Opus	-	0.063	0.687	1	0.774	0.219	0.774	0.754	0.581	0.344
Claude 3.5 Sonnet	0.063	-	0.453	0.227	0.508	1	0.549	0.092	0.727	1
Chat GPT-4o	0.687	0.453	-	0.687	1	0.625	1	0.289	1	0.774
ChatGPT-o1	1	0.227	0.687	-	0.687	0.289	0.625	0.687	0.549	0.454
Mistral Large 2	0.774	0.508	1	0.687	-	0.754	1	0.344	1	0.791
Gemini 1.5 Pro	0.219	1	0.625	0.289	0.754	-	0.727	0.109	1	1
Llama 3.1 405B	0.774	0.549	1	0.625	1	0.727	-	0.344	1	0.791
Perplexity Pro	0.754	0.092	0.289	0.687	0.344	0.109	0.344	-	0.267	0.210
Radiologist 1	0.581	0.727	1	0.549	1	1	1	0.267	-	1
Radiologist 2	0.344	1	0.774	0.454	0.791	1	0.791	0.210	1	-

Test values were obtained from McNemar test, p-value ≤0.028 is considered statistically significant after Bonferroni correction

and there were no significant performance differences among models with them (p>0.05) (Table 1).

With "Prompt 1", Claude 3.5 Sonnet achieved the highest accuracy at 90.0%, followed by Claude 3 Opus and ChatGPT-o1, both scored 84.0% (p>0.028). ChatGPT-4o recorded an accuracy of 82.0%. Gemini 1.5 Pro attained 74.0%, the accuracy of Mistral Large 2 and Llama 3.1 405B at 72.0%. R2 (T.C.) had a slightly lower accuracy of 70.0% (p>0.028). Perplexity Pro demonstrated the lowest performance among all models, with an accuracy of 68.0% (Fig. 3).

Claude 3.5 Sonnet outperformed Mistral Large 2, Llama 3.1 405B, and Perplexity Pro, achieving the highest scores on text-based questions (p=0.012, p=0.022, p=0.007). It also demonstrated superior performance according to R1 and R2 (p=0.021, p=0.021).

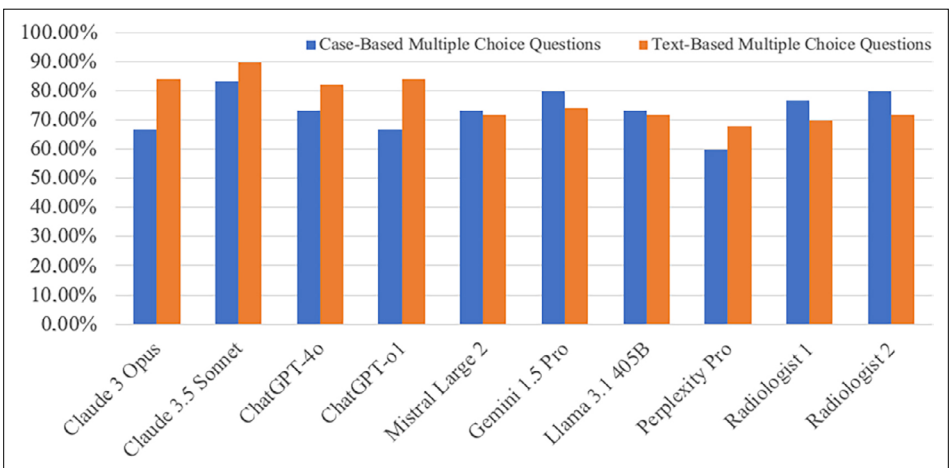

Fig. 3. The accuracy of LLMs and radiologists on multiple choice questions.

Table 3 Comparison of the performance of LLMs and radiologist on text-based multiple-choice questions

	Claude 3 Opus	Claude 3.5 Sonnet	Chat GPT-4o	Chat GPT-o1	Mistral Large 2	Gemini 1.5 Pro	Llama 3.1 405B	Perplexity Pro	Radiologist 1	Radiologist 2
Claude 3 Opus	-	0.375	1	1	0.180	0.332	0.146	0.057	0.143	0.210
Claude 3.5 Sonnet	0.375	-	0.219	0.453	**0.012**	0.077	**0.022**	**0.007**	**0.021**	**0.021**
Chat GPT-4o	1	0.219	-	1	0.125	0.388	0.267	0.118	0.238	0.359
Chat GPT-o1	1	0.453	1	-	0.109	0.332	0.210	0.115	0.167	0.238
Mistral Large 2	0.180	**0.012**	0.125	0.109	-	1	1	0.791	1	1
Gemini 1.5 Pro	0.332	0.077	0.388	0.332	1	-	1	0.607	0.804	1
Llama 3.1 405B	0.146	**0.022**	0.267	0.210	1	1	-	0.804	1	1
Perplexity Pro	0.057	**0.007**	0.118	0.115	0.791	0.607	0.804	-	1	0.824
Radiologist 1	0.143	**0.021**	0.238	0.167	1	0.804	1	1	-	1
Radiologist 2	0.210	**0.021**	0.359	0.238	1	1	1	0.824	1	-

Test values were obtained from McNemar test, p-value ≤0.028 is considered statistically significant after Bonferroni correction

When other LLMs were compared among themselves and with radiologists, there was no significant difference in performance between them (p>0.028) (Table 3).

DISCUSSION

The most striking result of our study is that LLMs included in the study demonstrated promising performance in text-based interpretation of RECIST. Our study uniquely examines the performance of several LLMs regarding the RECIST guideline, comparing their performance with that of radiologists. This approach not only identifies which LLM demonstrates a more comprehensive grasp of RECIST 1.1 but also offers insights into how radiologists' performance stacks up against that of LLMs.

Coskun et al.[16] evaluated ChatGPT using 59 prostate cancer questions from the European Urology Patient Information Society and reported suboptimal accuracy (mean: 3.62±0.49) requiring improvement. Similarly, Lombardo et al.[17] tested ChatGPT (August 2023) with 195 questions from the EAU 2023 prostate cancer guidelines; expert review showed only 26% completely correct answers, with accuracy varying by section (best in follow-up/quality of life, poorest in diagnosis/treatment. Similar to our results, in a recent study assessing LLMs in breast cancer care, three models—GPT-3.5, GPT-4, and Gemini (formerly Bard)—were evaluated using 60 MCQs covering treatment, diagnostic techniques, imaging interpretation, and pathology in breast cancer. GPT-4 achieved a 95% accuracy rate, outperforming GPT-3.5 (90%) and Gemini (80%), with statistically significant differences observed among the models (p=0.010). Furthermore, the models performed consistently across questions sourced from public databases and those formulated by radiologists.[18] Also, Cao at al.[19] evaluated LLMs in hepatocellular carcinoma diagnosis and management questions and found that ChatGPT-3.5, Gemini, and Bing answered only 45%, 60%, and 30% of basic clinical questions accurately, respectively, with even fewer responses deemed both accurate and reliable. Although there are still different results in the literature about the competence of LLMs in radiology, our results indicate that LLMs have quite a theoretical knowledge about RECIST 1.1.

Another important result of our study is that LLMs responded as well as radiologists on text-based and case-based MCQs that require analysis of the findings and data obtained. This result suggests that LLMs are successful in analyzing and reasoning texts such as radiology reports and providing the status of the disease (progression, stable, or regression) according to the RECIST guideline, which is the most critical for clinicians. To our best knowledge, there are no studies evaluating the performance of LLMs on case-based questions about cancer. Previous studies have evaluated LLMs' knowledge of cancer and cancer-related guidelines, which were only text-based. Çıtır reported that ChatGPT-3.5 gave largely correct answers to questions about oral cancer, 51.25% gave "very good" and 46.25% gave "good" answers, and the overall reliability was 97.5%.[20] Similarly, Yurtcu et al.[21] demonstrated that ChatGPT has strong accuracy in answering frequently asked questions about cervical cancer. Beyond these studies, our study uniquely demonstrates that

Turk J Oncol 2026;41(1):52–59
doi: 10.5505/tjo.2026.4733

LLMs perform quite adequately on case-based MCQs in line with the correct analysis. With this finding, we believe that our study may be a leading point for further multicenter studies that evaluate the performance of LLMs in real-patient scenarios.

In our study, all LLMs with three different prompts showed great consistency with minor differences in responses. Due to the nature of LLMs, it is a surprising result that these models, which largely determine their answers according to the given prompt, perform similarly with different prompts to the questions about RECIST 1.1.[21,22] In contrast to our results, Russe et al.[22] demonstrated the prompt effect, transforming a generic request into a precision prompt increased ChatGPT-4's factual correctness and decreased hallucinations, while a zero-shot chain-of-thought format further improved explain ability and user trust. Nguyen et al.[23] tested ChatGPT mixed text-and-image multiple-choice questions from the 2022 ACR in-training examination; an "encouraging" prompt boosted overall accuracy to 61%, whereas a "threatening" prompt reduced accuracy to 48% and tripled the non-response rate. These studies suggest that when a task is open-ended or cognitively complex, the model's probabilistic reasoning is highly sensitive to contextual cues embedded in the prompt. RECIST 1.1 evaluation, however, is a narrowly defined, rule-based exercise, so the knowledge retrieved is limited, and the model's decision space is tightly constrained; once the required rules are invoked, rewording the prompt affords little additional leverage. Therefore, the prompting effect could diminish as the task approaches deterministic guideline application rather than broad clinical reasoning.

The impressive performance of Claude 3.5 Sonnet—achieving 83.3% accuracy on case-based MCQs and 90% on text-based MCQs—indicates great potential of its model in this field. The observed minor variations in accuracy among LLMs can be largely attributed to differences in their underlying architectures. Models with real-time web access capability, such as Gemini 1.5 Pro and Perplexity Pro, frequently derive their responses from non-scientific sources, which may account for the comparatively lower performance of web-enabled LLMs relative to those without internet access. In contrast, some of the ChatGPT and Claude models are trained on closed datasets, potentially contributing to their enhanced reliability.

Limitations of the Study

Our study has a few limitations. First, the number of questions was limited, and the assessment relied solely on MCQs. The performance of LLMs on open-ended questions was not evaluated in the study, which may have led to exaggerated LLM performances.

Second, we compared the accuracy of LLMs against two general radiologists with seven years of experience. It is likely that more experienced senior radiologists, particularly those with more specialized knowledge about oncological imaging, would achieve higher performance. Senior radiologists who are specialized and/or subspecialized in this field may perform even better than LLMs, but since follow-up images of cancer patients are often evaluated by general radiologists in daily practice for many different reasons, the radiologists included in this study are general radiologists to better reflect real-life practice.

Lastly, this study assessed the performance of LLMs about RECIST 1.1 textually, while visual evaluation remains an integral component of radiological assessment. As such, the results of our study may not fully reflect the real-world applicability of LLMs in this field. It is important to emphasize that while LLMs performed well on structured MCQs, this could not directly reflect their ability in actual imaging interpretation for RECIST.

CONCLUSION

Radiologists can benefit from understanding how LLMs interpret RECIST 1.1, as these models may soon assist in standardized follow-up reporting. Incorporating LLM-based educational modules and decision-support tools can enhance consistency and reduce interpretive variability. Ongoing evaluation of model accuracy and bias is essential before clinical deployment.

Informed Consent: The authors declare that this study was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

Funding: No funding was received for this study.

Use of AI for Writing Assistance: No AI technologies utilized.

Author Contributions: Concept – E.Ç.; Design – E.Ç., T.C.; Supervision – Y.C.G.; Materials – E.Ç.; Data collection and/or processing – E.Ç.; Data analysis and/or interpretation – E.Ç., T.C.; Literature search – E.Ç., T.C.; Writing – E.Ç.; Critical review – E.Ç., T.C., Y.C.G.

Peer-review: Externally peer-reviewed.

REFERENCES

1. Nakaura T, Ito R, Ueda D, Nozaki T, Fushimi Y, Matsui Y, et al. The impact of large language models on radiology: A guide for radiologists on the latest innovations in AI. Jpn J Radiol 2024;42(3):1–12
2. Chung EM, Zhang SC, Nguyen AT, Atkins KM, Sandler HM, Kamrava M. Feasibility and acceptability of ChatGPT-generated radiology report summaries for cancer patients. Digit Health 2023;9:1–7
3. Keshavarz P, Bagherieh S, Nabipoorashrafi SA, Chalian H, Rahsepar AA, Kim GHJ, et al. ChatGPT in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagn Interv Imaging 2024;105(7–8):251–65
4. Bhayana R. Chatbots and large language models in radiology: A practical primer for clinical and research applications. Radiology 2024;310(1):1–8
5. Bera K, O'Connor G, Jiang S, Tirumani SH, Ramaiya N. Analysis of ChatGPT publications in radiology: Literature so far. Curr Probl Diagn Radiol 2024;53(2):215–25
6. 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(2):80–90
7. Srivastav S, Chandrakar R, Gupta S, Babhulkar V, Agrawal S, Jaiswal A, et al. ChatGPT in radiology: The advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus 2023;15(7):e41435
8. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer 2009;45(2):228–47
9. Liu X, Duan C, Kim MK, Zhang L, Jee E, Maharjan B, et al. Claude 3 Opus and ChatGPT with GPT-4 in dermoscopic image analysis for melanoma diagnosis: Comparative performance analysis. JMIR Med Inform 2024;12:e59273
10. Chiarelli G, Stephens A, Finati M, Cirulli GO, Beatrici E, Filipas DK, et al. Adequacy of prostate cancer prevention and screening recommendations provided by an artificial intelligence-powered large language model. Int Urol Nephrol 2024;56(4):1–7
11. Aghamaliyev U, Karimbayli J, Giessen-Jung C, Matthias I, Unger K, Andrade D, et al. ChatGPT's gastrointestinal tumor board tango: A limping dance partner? Eur J Cancer 2024;205:114100
12. Sorin V, Barash Y, Konen E, Klang E. Large language models for oncological applications. J Cancer Res Clin Oncol 2023;149(11):9505–8
13. Güneş YC, Cesur T, Çamur E, Günbey Karabekmez L. Evaluating text and visual diagnostic capabilities of large language models on questions related to the Breast Imaging Reporting and Data System Atlas 5th edition. Diagn Interv Radiol 2025;31(2):1–8
14. Kaba E, Hürsoy N, Solak M, Çeliker FB. Accuracy of large language models in thyroid nodule-related questions based on the Korean Thyroid Imaging Reporting and Data System (K-TIRADS). Korean J Radiol 2024;25(5):499–500
15. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. Radiology 2015;277(3):826–32
16. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? Urology 2023;180:35–58
17. Lombardo R, Gallo G, Stira J, Turchi B, Santoro G, Riolo S, et al. Quality of information and appropriateness of OpenAI outputs for prostate cancer. Prostate Cancer Prostatic Dis 2024;27(2):1–7
18. Irmici G, Cozzi A, Della Pepa G, Berardinis CD, D'Ascoli E, Cellina M, et al. How do large language models answer breast cancer quiz questions? A comparative study of GPT-3.5, GPT-4 and Google Gemini. Radiol Med 2024;129(10):1–8
19. Cao JJ, Kwon DH, Ghaziani TT, Kwo P, Tse G, Kesslman A, et al. Large language models' responses to liver cancer surveillance, diagnosis, and management questions: Accuracy, reliability, readability. Abdom Radiol 2024;49(12):4286–94
20. Çi Ti RM. ChatGPT and oral cancer: A study on informational reliability. BMC Oral Health 2025;25(1):86
21. Yurtcu E, Ozvural S, Keyif B. Analyzing the performance of ChatGPT in answering inquiries about cervical cancer. Int J Gynaecol Obstet 2025;168(2):502–7
22. Russe MF, Reisert M, Bamberg F, Rau A. Improving the use of LLMs in radiology through prompt engineering: From precision prompts to zero-shot learning. Rofo 2023;196(11):1–8
23. Nguyen D, MacKenzie A, Kim YH. Encouragement vs liability: How prompt engineering influences ChatGPT-4's radiology exam performance. Clin Imaging 2024;115:110276

MULTIPLE CHOICE QUESTIONS

1. Which of the following is NOT a RECIST criterion for evaluating tumor response?
    a) Complete Response (CR)
    b) Partial Response (PR)
    c) Stable Disease (SD)
    d) Progressive Disease (PD)
    e) Minor Response (MR)

2. What is the minimum reduction in the sum of the longest diameters of target lesions required for a diagnosis of Partial Response (PR) according to RECIST 1.1?
    a) 10%
    b) 20%
    c) 30%
    d) 40%
    e) 50%

3. Which of the following statements is FALSE regarding the assessment of non-target lesions in RECIST 1.1?
    a) Non-target lesions must be recorded at baseline and followed throughout the study.
    b) New non-target lesions appearing during treatment are considered evidence of Progressive Disease (PD).
    c) Non-target lesions that decrease in size by 30% or more are considered a Partial Response (PR).
    d) Non-target lesions that increase in size by 20% or more are considered Progressive Disease (PD).
    e) Non-target lesions that remain stable are not considered in the overall response assessment.

4. What is the maximum number of target lesions that can be used for RECIST assessment in a given patient?
    a) 2
    b) 5
    c) 10
    d) 15

e) 20

5. Which of the following imaging modalities is NOT commonly used for RECIST assessment?
    a) Computed Tomography (CT)
    b) Magnetic Resonance Imaging (MRI)
    c) Positron Emission Tomography (PET)
    d) Ultrasound
    e) X-ray

6. What is the term used to describe the situation when a patient has both a reduction in the size of target lesions and the appearance of new non-target lesions?
    a) Mixed Response
    b) Discordant Response
    c) Indeterminate Response
    d) Pseudo-Progression
    e) Hyper-Progression

7. According to RECIST 1.1, what is the minimum follow-up interval required between consecutive tumor assessments?
    a) 1 week
    b) 2 weeks
    c) 4 weeks
    d) 6 weeks
    e) 8 weeks

8. Which of the following is NOT a RECIST criterion for evaluating lymph node response?
    a) Complete Response (CR)
    b) Partial Response (PR)
    c) Stable Disease (SD)
    d) Progressive Disease (PD)
    e) Minor Response (MR)

9. What is the minimum reduction in the short axis diameter of lymph nodes required for a diagnosis of Partial Response (PR) according to RECIST 1.1?

a) 10%
b) 20%
c) 30%
d) 40%
e) 50%

10. Which of the following statements is FALSE regarding the assessment of bone lesions in RECIST 1.1?
    a) Bone lesions must be recorded at baseline and followed throughout the study.
    b) New bone lesions appearing during treatment are considered evidence of Progressive Disease (PD).
    c) Bone lesions that decrease in size by 30% or more are considered a Partial Response (PR).
    d) Bone lesions that increase in size by 20% or more are considered Progressive Disease (PD).
    e) Bone lesions that remain stable are not considered in the overall response assessment.

11. What is the maximum number of target bone lesions that can be used for RECIST assessment in a given patient?
    a) 2
    b) 5
    c) 10
    d) 15
    e) 20

12. Which of the following is NOT a RECIST criterion for evaluating soft tissue lesions?
    a) Complete Response (CR)
    b) Partial Response (PR)
    c) Stable Disease (SD)
    d) Progressive Disease (PD)
    e) Minor Response (MR)

13. What is the minimum reduction in the sum of the longest diameters of soft tissue lesions required for a diagnosis of Partial Response (PR) according to RECIST 1.1?

a) 10%
b) 20%
c) 30%
d) 40%
e) 50%

14. Which of the following statements is FALSE regarding the assessment of non-target soft tissue lesions in RECIST 1.1?
    a) Non-target soft tissue lesions must be recorded at baseline and followed throughout the study.
    b) New non-target soft tissue lesions appearing during treatment are considered evidence of Progressive Disease (PD).
    c) Non-target soft tissue lesions that decrease in size by 30% or more are considered a Partial Response (PR).
    d) Non-target soft tissue lesions that increase in size by 20% or more are considered Progressive Disease (PD).
    e) Non-target soft tissue lesions that remain stable are not considered in the overall response assessment.

15. What is the maximum number of target soft tissue lesions that can be used for RECIST assessment in a given patient?
    a) 2
    b) 5
    c) 10
    d) 15
    e) 20

16. Which of the following imaging modalities is NOT commonly used for RECIST assessment of soft tissue lesions?
    a) Computed Tomography (CT)
    b) Magnetic Resonance Imaging (MRI)
    c) Positron Emission Tomography (PET)
    d) Ultrasound
    e) X-ray

17. What is the term used to describe the situation when a patient has both a reduction in the size of target lesions and the appearance of new non-target soft tissue lesions?
    a) Mixed Response
    b) Discordant Response
    c) Indeterminate Response
    d) Pseudo-Progression
    e) Hyper-Progression

18. According to RECIST 1.1, what is the minimum follow-up interval required between consecutive tumor assessments for soft tissue lesions?
    a) 1 week
    b) 2 weeks
    c) 4 weeks
    d) 6 weeks
    e) 8 weeks

19. Which of the following is NOT a RECIST criterion for evaluating ascites or pleural effusion?
    a) Complete Response (CR)
    b) Partial Response (PR)
    c) Stable Disease (SD)
    d) Progressive Disease (PD)
    e) Minor Response (MR)

20. What is the definition of Complete Response (CR) for ascites or pleural effusion according to RECIST 1.1?
    a) Complete disappearance of all ascites or pleural effusion
    b) Reduction in the volume of ascites or pleural effusion by 50% or more
    c) Stabilization of the volume of ascites or pleural effusion
    d) Increase in the volume of ascites or pleural effusion by less than 20%
    e) Appearance of new ascites or pleural effusion

21. Which of the following statements is FALSE regarding the assessment of ascites or pleural effusion in RECIST 1.1?

a) Ascites or pleural effusion must be recorded at baseline and followed throughout the study.
b) New ascites or pleural effusion appearing during treatment are considered evidence of Progressive Disease (PD).
c) Ascites or pleural effusion that decreases in volume by 30% or more are considered a Partial Response (PR).
d) Ascites or pleural effusion that increases in volume by 20% or more are considered Progressive Disease (PD).
e) Ascites or pleural effusion that remain stable are not considered in the overall response assessment.

22. What is the maximum number of target ascites or pleural effusion sites that can be used for RECIST assessment in a given patient?
    a) 2
    b) 5
    c) 10
    d) 15
    e) 20

23. Which of the following imaging modalities is NOT commonly used for RECIST assessment of ascites or pleural effusion?
    a) Computed Tomography (CT)
    b) Magnetic Resonance Imaging (MRI)
    c) Positron Emission Tomography (PET)
    d) Ultrasound
    e) X-ray

24. What is the term used to describe the situation when a patient has both a reduction in the volume of ascites or pleural effusion and the appearance of new target lesions?
    a) Mixed Response
    b) Discordant Response
    c) Indeterminate Response
    d) Pseudo-Progression
    e) Hyper-Progression

25. According to RECIST 1.1, what is the minimum follow-up interval required between consecutive tumor assessments for ascites or pleural effusion?
    a) 1 week
    b) 2 weeks
    c) 4 weeks
    d) 6 weeks
    e) 8 weeks

26. According to RECIST 1.1, what is the maximum number of lesions that can be assessed for response determination?
    a) 2
    b) 5
    c) 10
    d) 15
    e) 20

27. What is the minimum short axis measurement for a lymph node to be considered measurable and assessable as a target lesion?
    a) 5 mm
    b) 10 mm
    c) 15 mm
    d) 20 mm
    e) 25 mm

28. When a target lesion becomes too small to measure on a CT scan, what should be recorded on the case report form?
    a) 0 mm
    b) 5 mm
    c) 10 mm
    d) 15 mm
    e) 20 mm

29. What is the minimum size for a measurable lesion on a CT scan with a slice thickness of 5 mm or less?
    a) 5 mm

b) 10 mm

c) 15 mm

d) 20 mm

e) 25 mm

30. In which of the following scenarios is confirmation of response required?

a) Trials with response as the primary endpoint

b) Randomized studies

c) Phase II trials

d) Phase III trials

e) All of the above

31. How are non-target lesions recorded on the case report form?

a) By measuring their size

b) By indicating their presence or absence

c) By describing their appearance

d) By noting their location

e) By all of the above

32. For a target lesion to meet the criteria for Complete Response (CR), what must happen to any pathological lymph nodes?

a) They must disappear completely.

b) They must reduce in size by at least 30%.

c) They must reduce in short axis to <10 mm.

d) They must increase in size by at least 20%.

e) None of the above

33. What is the minimum percentage increase in the sum of diameters of target lesions required to qualify for Progressive Disease (PD)?

a) 10%

b) 15%

c) 20%

d) 25%

e) 30%

34. In addition to the relative increase for PD, what is the minimum absolute increase required in the sum of diameters?

a) 2 mm

b) 5 mm

c) 10 mm

d) 15 mm

e) 20 mm

35. Which of the following is not considered a measurable lesion according to RECIST 1.1 criteria?

a) Skin nodule with a diameter of 12 mm

b) Lung lesion measuring 18 mm on chest X-ray

c) Blastic bone lesion

d) Lymph node with a short axis of 20 mm

e) All of the above are measurable lesions

36. Which of the following is not a criterion for assessing tumor response in RECIST 1.1?

a) Change in tumor size

b) Change in tumor markers

c) Presence of new lesions

d) Duration of response

e) All of the above are criteria for assessing tumor response

37. What is the minimum follow-up duration required after the end of treatment to assess for complete response?

a) 1 month

b) 3 months

c) 6 months

d) 12 months

e) 24 months

38. Which of the following is not considered unequivocal progression of non-measurable/non-target disease?

a) Appearance of new lesions

b) Increase in the size of non-target lesions

c) Worsening of symptoms

d) Development of malignant ascites

e) All of the above are considered unequivocal progression

39. What is the term used to describe a situation where a patient's disease initially responds to treatment but later progresses?

a) Partial response

b) Stable disease

c) Progressive disease

d) Relapse

e) Remission

40. Which of the following statements is true about RECIST 1.1 criteria?

a) RECIST 1.1 criteria are used to assess response to treatment in patients with solid tumors.

b) RECIST 1.1 criteria are used to assess response to treatment in patients with leukemia.

c) RECIST 1.1 criteria are used to assess response to treatment in patients with brain tumors.

d) RECIST 1.1 criteria are used to assess response to treatment in patients with lymphoma.

e) RECIST 1.1 criteria are used to assess response to treatment in patients with all types of cancer.

41. How is the short axis of a lymph node determined?

a) By measuring the longest diameter of the node

b) By measuring the smallest diameter of the node

c) By measuring the diameter of the node in the axial plane

d) By measuring the diameter of the node in the sagittal plane

e) By measuring the diameter of the node in the coronal plane

42. What is the recommended course of action when unequivocal progression is observed in non-measurable disease?

a) Continue treatment without any changes

b) Consider the patient to have had overall PD

c) Repeat imaging studies to confirm progression

d) Consult with a specialist for further evaluation

e) Discontinue treatment and enroll in a new trial

43. Which of the following is not an example of unequivocal progression in non-measurable disease?

a) Increase in pleural effusion from 'trace' to 'large'

b) Increase in lymphangitic localized to widespread

c) Increase in tumor burden representing an additional 73% increase in 'volume'

d) Increase in skin nodules from 2 to 5

e) Increase in liver metastases from 3 to 7

44. What is the term used to describe a situation where a patient's disease shows no significant change in tumor size after treatment?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Relapse

45. What is the term used to describe a situation where a patient's disease worsens or develops new lesions after an initial response to treatment?

a) Partial response

b) Stable disease

c) Progressive disease

d) Relapse

e) Remission

46. Which of the following is not a criterion for assessing tumor response in RECIST 1.1?

a) Change in tumor size

b) Change in tumor markers

c) Presence of new lesions

d) Duration of response

e) Overall survival

47. What is the recommended method for measuring the size of lung lesions?

a) Chest X-ray

b) CT scan

c) MRI

d) Ultrasound

e) None of the above

48.Which of the following is not considered a measurable lesion according to RECIST 1.1 criteria?

a) Skin nodule with a diameter of 12 mm

b) Lung lesion measuring 18 mm on chest X-ray

c) Blastic bone lesion

d) Lymph node with a short axis of 20 mm

e) All of the above are considered measurable lesions

49. What is the minimum follow-up duration required after the end of treatment to assess for complete response?

a) 1 month

b) 3 months

c) 6 months

d) 12 months

e) 24 months

50. Which of the following is not a type of response defined in RECIST 1.1?

a) Complete response

b) Partial response

c) Stable disease

d) Minimal response

e) All of the above are types of response defined in RECIST 1.1

ANSWERS

1. E
2. C

3. C
4. C
5. D
6. A
7. C
8. E
9. B
10. C
11. C
12. E
13. C
14. C
15. C
16. D
17. A
18. C
19. E
20. A
21. C
22. C
23. C
24. A
25. C
26. B
27. C
28. B
29. B
30. A
31. B
32. C
33. C
34. B
35. C
36. B
37. C

38. C
39. D
40. A
41. B
42. B
43. D
44. C
45. C
46. E
47. B
48. E
49. C
50. D

CASE-BASED QUESTIONS

1. A 50-year-old male patient presents with acute myeloid leukemia (AML). Bone marrow biopsy reveals 80% blasts. The patient is started on induction chemotherapy.

MCQ: Can RECIST 1.1 criteria be used to assess the response to treatment in this patient?

a) Yes

b) No

c) Only for lymph node involvement

d) Only for bone involvement

e) None of the above

2. A 60-year-old female patient presents with a 5 cm GIST in the stomach. The patient undergoes surgical resection of the tumor.

MCQ: Can RECIST 1.1 criteria be used to assess the response to treatment in this patient?

a) Yes

b) No

c) Only for lymph node involvement

d) Only for distant metastasis

e) None of the above

3. A 40-year-old male patient presents with a 2 cm glioblastoma in the right frontal lobe. The patient undergoes surgical resection of the tumor followed by radiation therapy.

MCQ: Can RECIST 1.1 criteria be used to assess the response to treatment in this patient?

a) Yes

b) No

c) Only for main lesion

d) Only for distant metastasis

e) None of the above

4. A 50-year-old male patient with metastatic melanoma has multiple lung lesions. The largest lesion measures 2.5 cm in diameter on baseline CT scan. After 2 months of treatment, the largest lesion measures 1.8 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

5. A 45-year-old female patient with breast cancer has a 3 cm lymph node in the axilla. After 4 cycles of chemotherapy, the lymph node measures 1.5 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

6. A 60-year-old male patient with prostate cancer has multiple bone lesions. The largest lesion measures 1.5 cm in diameter on baseline bone scan. After 6 months of hormone therapy, the largest lesion measures 1.2 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

7. A 30-year-old female patient with Hodgkin lymphoma has multiple lymph nodes in the neck. The largest lymph node measures 3 cm in diameter on baseline CT scan. After 4 cycles of chemotherapy, the largest lymph node measures 2.7 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

8. A 70-year-old male patient with lung cancer has a 4 cm lung lesion. After 3 cycles of chemotherapy, the lesion has completely disappeared on CT scan.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

9. A 50-year-old male patient with metastatic melanoma has multiple lung lesions. The largest lesion measures 2.5 cm in diameter on baseline CT scan. After 2 months of treatment, the largest lesion measures 3 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

10. A 45-year-old female patient with breast cancer has a 3 cm lymph node in the axilla. After 4 cycles of chemotherapy, the lymph node measures 4 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

11. A 50-year-old male patient with metastatic melanoma has multiple lung lesions. The largest lesion measures 2.5 cm in diameter and the second largest lesion measures 2 cm in diameter.

MCQ: Which lesion should be selected as the target lesion?

a) The largest lesion

b) The second largest lesion

c) Both lesions

d) Neither lesion

e) None of the above

12. A 45-year-old female patient with breast cancer has a 3 cm lymph node in the axilla and a 2 cm lymph node in the supraclavicular fossa.

MCQ: Which lymph node should be selected as the target lesion?

a) The lymph node in the axilla

b) The lymph node in the supraclavicular fossa

c) Both lymph nodes

d) Neither lymph node

e) None of the above

13. A 60-year-old male patient with prostate cancer has multiple bone lesions. The largest lesion measures 1.5 cm in diameter and the second largest lesion measures 1.2 cm in diameter.

MCQ: Which lesion should be selected as the target lesion?

a) The largest lesion

b) The second largest lesion

c) Both lesions

d) Neither lesion

e) None of the above

14. A 50-year-old male patient with metastatic melanoma has multiple lung lesions. The largest lesion measures 2.5 cm in diameter, the second largest lesion measures 2 cm in diameter, and the third largest lesion measures 1.5 cm in diameter.

MCQ: How many target lesions should be selected?

a) One

b) Two

c) Three

d) Four

e) Five

15. A 45-year-old female patient with breast cancer has a 3 cm lymph node in the axilla, a 2 cm lymph node in the supraclavicular fossa, and a 1 cm lymph node in the internal mammary chain.

MCQ: How many target lesions should be selected?

a) One

b) Two

c) Three

d) Four

e) Five

16. A 60-year-old male patient with prostate cancer has multiple bone lesions. The largest lesion measures 1.5 cm in diameter, the second largest lesion measures 1.2 cm in diameter, and the third largest lesion measures 1 cm in diameter.

MCQ: How many target lesions should be selected?

a) One

b) Two

c) Three

d) Four

e) Five

17. A 50-year-old male patient with metastatic melanoma has multiple lymph nodes in the neck. The largest lymph node measures 2.5 cm in diameter on baseline CT scan. After 2 months of treatment, the largest lymph node measures 1.8 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

18. A 45-year-old female patient with breast cancer has a 3 cm lymph node in the axilla. After 4 cycles of chemotherapy, the lymph node measures 1.5 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

19. A 60-year-old male patient with prostate cancer has multiple lymph nodes in the mediastinum. The largest lymph node measures 1.5 cm in diameter on baseline CT scan. After 6 months of hormone therapy, the largest lymph node measures 1.2 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

20. A 30-year-old female patient with Hodgkin lymphoma has multiple lymph nodes in the neck. The largest lymph node measures 3 cm in diameter on baseline CT scan. After 4 cycles of chemotherapy, the largest lymph node measures 2.7 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

21. A 70-year-old male patient with lung cancer has a 4 cm lymph node in the mediastinum. After 3 cycles of chemotherapy, the lymph node has completely disappeared on CT scan.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

22. A 50-year-old male patient with metastatic melanoma has multiple lung lesions. The largest lesion measures 2.5 cm in diameter on baseline CT scan. After 2 months of treatment, the largest lesion measures 1.8 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

23. A 45-year-old female patient with breast cancer has a 3 cm lymph node in the axilla. After 4 cycles of chemotherapy, the lymph node measures 1.5 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

24. A 60-year-old male patient with prostate cancer has multiple bone lesions. The largest lesion measures 1.5 cm in diameter on baseline bone scan. After 6 months of hormone therapy, the largest lesion measures 1.2 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

25. A 30-year-old female patient with Hodgkin lymphoma has multiple lymph nodes in the neck. The largest lymph node measures 3 cm in diameter on baseline CT scan. After 4 cycles of chemotherapy, the largest lymph node measures 2.7 cm in diameter.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

26. A 70-year-old male patient with lung cancer has a 4 cm lung lesion. After 3 cycles of chemotherapy, the lesion has completely disappeared on CT scan.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

27. A 50-year-old male patient with metastatic melanoma has a new lesion in the liver that was not present on baseline imaging. The new lesion measures 1 cm in diameter.

MCQ: How should the new lesion be classified according to RECIST 1.1 criteria?

a) Target lesion

b) Non-target lesion

c) Unequivocal progression

d) Indeterminate

e) None of the above

28. A 45-year-old female patient with breast cancer has a new lymph node in the axilla that was not present on baseline imaging. The new lymph node measures 1.5 cm in diameter.

MCQ: How should the new lymph node be classified according to RECIST 1.1 criteria?

a) Target lesion

b) Non-target lesion

c) Unequivocal progression

d) Indeterminate

e) None of the above

29. A 60-year-old male patient with prostate cancer has a new bone lesion that was not present on baseline imaging. The new bone lesion measures 2 cm in diameter.

MCQ: How should the new bone lesion be classified according to RECIST 1.1 criteria?

a) Target lesion

b) Non-target lesion

c) Unequivocal progression

d) Indeterminate

e) None of the above

30. A 50-year-old male patient with metastatic melanoma has multiple lymph nodes in the neck. The largest lymph node measures 2.5 cm in diameter on baseline CT scan. After 2 months of treatment, the largest lymph node measures 1.8 cm in diameter, but a new lymph node measuring 1.5 cm in diameter has appeared in the supraclavicular fossa.

MCQ: What is the best response to treatment according to RECIST 1.1 criteria?

a) Complete response

b) Partial response

c) Stable disease

d) Progressive disease

e) Unequivocal progression

ANSWERS

1. B
2. B
3. B
4. B

5. B
6. C
7. D
8. A
9. D
10. D
11. A
12. A
13. A
14. B
15. C
16. B
17. B
18. B
19. C
20. D
21. A
22. B
23. B
24. C
25. D
26. A
27. C
28. C
29. C
30. D

Number of Question	ChatGPT o1 (Prompt 1)	ChatGPT o1 (Prompt 2)
1.	1	1
2.	1	1
3.	1	1
4.	1	1
5.	1	1
6.	1	1
7.	1	1

8.	1	1
9.	1	1
10.	1	1
11.	1	1
12.	1	1
13.	1	1
14.	1	1
15.	1	1
16.	1	1
17.	0	1
18.	1	1
19.	1	1
20.	1	0
21.	0	0
22.	1	1
23.	1	1
24.	1	1
25.	1	1
26.	1	1
27.	1	1
28.	1	1
29.	0	0
30.	1	1
31.	1	1
32.	1	1
33.	1	1
34.	0	1
35.	1	0
36.	1	1
37.	0	0
38.	1	1
39.	1	0
40.	1	1
41.	0	0
42.	1	1
43.	0	0
44.	1	1
45.	1	1
46.	1	1

47.	1	1
48.	1	1
49.	0	0
50.	1	1
True: 1, False:0		

answer." Prompt 2: "You are a senior academic radiologist. I have some questio

Number of Question	ChatGPT o1 (Prompt 1)	ChatGPT o1 (Prompt 2)
1.	1	1
2.	1	1
3.	1	0
4.	0	0
5.	1	1
6.	1	1
7.	0	0
8.	1	1
9.	1	1
10.	1	1
11.	0	0
12.	0	0
13.	0	0
14.	1	1
15.	0	0
16.	1	1
17.	0	0
18.	1	1
19.	1	1
20.	0	0
21.	1	1
22.	0	0
23.	1	1
24.	1	1
25.	0	0
26.	1	1
27.	1	1
28.	1	1
29.	1	1
30.	1	1
True: 1, False:0		

answer." Prompt 2: "You are a senior academic radiologist. I have some question

ChatGPT o1 (Prompt 3)	ChatGPT 4o (Prompt 1)	ChatGPT 4o (Prompt 2)
1	1	1
1	1	1
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	0	0
1	1	1
0	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	0	0
1	1	1
1	0	0
1	1	1

1	1	1
1	0	0
0	0	0
1	1	1

s about RECIST 1.1. I will ask you multiple choice questions with a single correct a

ChatGPT o1 (Prompt 3)	ChatGPT 4o (Prompt 1)	ChatGPT 4o (Prompt 2)
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	0
0	1	1
0	0	0
0	0	0
1	1	1
0	0	0
1	0	0
0	1	1
1	1	1
1	1	1
0	0	0
1	1	1
0	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
1	0	0
1	1	1

s about RECIST 1.1. I will ask you multiple choice questions with a single correct an

ChatGPT 4o (Prompt 3)	Gemini 1.5 Pro (Prompt 1)	Gemini 1.5 Pro (Prompt 2)
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	0	0
1	0	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	0
1	1	1
1	1	1
0	1	1
1	1	1
1	0	0
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	0	0
1	1	1
1	0	0
1	1	1
0	0	0
1	0	0
1	0	0
0	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
0	1	0
1	1	1
0	0	0
1	1	1

1	1	1
0	0	0
0	1	1
1	0	0

swer. Provide only the letter of the most accurate choice for each.", Prompt 3: "I ha

ChatGPT 4o (Prompt 3)	Gemini 1.5 Pro (Prompt 1)	Gemini 1.5 Pro (Prompt 2)
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
0	1	1
0	1	1
1	0	0
0	0	0
0	0	0
1	1	1
1	1	0
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
0	1	1
1	1	1

swer. Provide only the letter of the most accurate choice for each.", Prompt 3: "I ha

Gemini 1.5 Pro (Prompt 3)	Perplexity Pro (Prompt 1)
1	1
1	1
1	1
1	1
1	0
1	1
1	0
1	1
1	1
0	1
1	1
1	1
1	1
1	1
0	1
1	1
1	1
1	1
1	1
1	0
1	1
0	1
0	1
1	1
0	1
1	0
0	0
1	1
1	1
0	0
1	0
0	0
1	1
0	1
0	0
0	1
1	1
1	1
1	0
1	0
0	1
1	1
1	0
1	1
0	0
1	1

1	1
0	0
1	0
0	0
ve a few questions about RECIST 1	

Gemini 1.5 Pro (Prompt 3)	Perplexity Pro (Prompt 1)
1	1
1	1
1	0
1	0
1	1
1	1
0	0
1	1
1	1
1	0
1	0
1	1
1	0
0	1
0	0
0	1
1	0
1	1
1	1
0	0
1	1
1	1
1	1
1	1
0	0
1	1
1	1
1	1
1	0
1	0

ve a few questions about RECIST 1	

TEXT-BASED MULTIPLE CHOICE QUESTIONS		
Perplexity Pro (Prompt 2)	Perplexity Pro (Prompt 3)	istral Large 2 (Prompt 1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	1
1	1	1
0	0	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	1
1	1	1
0	0	1
1	1	0
1	1	1
1	1	1
0	0	1
0	0	0
1	1	1
0	1	0
0	0	0
0	0	1
0	0	1
1	1	1
1	1	0
0	0	1
1	1	1
1	1	0
1	1	1
0	0	0
0	0	1
1	1	0
1	1	1
0	0	0
1	1	1
0	0	0
1	1	1

1	1	1
0	0	0
0	0	0
0	0	0

CASE-BASED MULTIPLE CHOICE QUESTIONS		
Perplexity Pro (Prompt 2)	Perplexity Pro (Prompt 3)I	istral Large 2 (Prompt 1
1	1	1
1	1	1
0	0	1
0	0	0
1	1	1
1	1	1
0	0	1
1	1	1
1	1	0
0	0	1
0	0	1
1	1	0
0	0	0
1	1	1
0	0	0
1	1	1
0	0	0
1	1	1
1	1	1
0	0	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	0
1	1	1
0	0	1
0	0	1

Mistral Large 2 (Prompt 2	Mistral Large 2 (Prompt 3	lama 3.1 405B (Prompt 1	lama 3.1 405B (Prompt 2
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	0	0
1	1	1	1
1	1	0	0
0	0	1	1
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	0	1	1
1	1	1	1
1	1	1	1
1	1	0	0
0	0	1	1
1	1	1	1
1	1	1	1
1	1	0	0
0	0	1	1
1	1	1	1
0	0	1	1
0	0	1	1
1	1	0	0
1	1	1	1
1	1	1	1
0	0	0	0
1	1	1	1
1	1	0	0
0	0	1	1
1	1	0	0
0	0	1	1
1	1	1	1
0	0	1	1
1	1	1	1
0	0	0	0
1	1	1	1
0	0	0	0
1	1	1	1

1	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Mistral Large 2 (Prompt 2	Mistral Large 2 (Prompt 3	lama 3.1 405B (Prompt 1	lama 3.1 405B (Prompt 2
1	1	1	1
1	1	0	0
1	1	1	1
0	0	1	1
1	1	1	1
1	1	1	1
1	0	0	0
1	1	1	1
0	0	1	0
1	1	1	1
0	1	0	0
0	0	0	0
0	0	1	1
1	1	1	1
0	0	0	0
1	1	1	1
0	0	0	0
1	1	1	0
1	1	1	1
1	1	0	0
1	1	1	1
1	1	0	0
1	1	1	1
1	1	1	1
0	0	1	1
1	1	1	1
0	0	1	1
1	1	1	1
1	1	1	1
1	1	1	1

lama 3.1 405B (Prompt	ude 3.5 Sonnet (Prompt	ude 3.5 Sonnet (Prompt	ude 3.5 Sonnet (Prompt
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
1	1	0	1
0	1	1	1
1	1	1	1
1	1	1	1
0	1	1	0
1	1	1	1
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	0
1	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
1	1	0	1
1	1	1	1
0	1	1	1
1	1	1	1
0	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
1	1	1	1
0	1	1	1
1	0	0	0
0	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
0	0	0	0
1	1	1	1
0	0	0	0
1	0	0	0

0	1	1	1
0	1	1	1
0	0	0	0
0	1	1	1

lama 3.1 405B (Prompt	ude 3.5 Sonnet (Prompt	ude 3.5 Sonnet (Prompt	ude 3.5 Sonnet (Prompt
1	1	1	1
0	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	0	0	0
0	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
0	1	1	1
1	0	0	0
1	1	1	1
0	0	0	0
1	0	0	0
0	1	1	1
1	1	1	1
1	1	1	1
0	1	1	1
1	1	1	1
0	1	1	1
1	1	1	1
1	0	0	0
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

laude 3 Opus (Prompt 1	laude 3 Opus (Prompt	Claude 3 Opus (Prompt 3)
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	0
1	0	1
1	1	1
1	1	1
1	1	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
1	0	0
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
0	0	0
0	0	0

1	1	1
0	0	0
0	0	0
1	1	1

laude 3 Opus (Prompt 1	laude 3 Opus (Prompt 2	Claude 3 Opus (Prompt 3)
1	1	1
1	1	1
1	1	1
1	1	1
1	1	0
0	0	0
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0
1	1	1
0	0	0
0	0	0
1	1	1
1	1	1
0	0	0
0	0	0
1	1	1
1	1	1
1	1	1
0	0	0
0	0	0
1	1	1
1	1	1
1	1	1
1	1	1
0	0	0

Radiologist 1(Y.C.G.)	Radiologist 2(T.C.)
1	1
1	1
1	1
0	1
0	0
1	1
1	1
1	1
1	1
1	1
0	0
1	1
1	1
0	0
1	1
1	1
1	0
1	1
1	1
0	0
1	1
0	0
1	1
1	1
1	0
1	1
0	1
1	1
0	0
0	1
1	1
0	1
1	0
1	1
1	1
0	0
1	1
1	1
0	0
1	1
1	1
1	0
1	1
1	1
0	0
1	1

0	1
1	0
0	1
1	0

Radiologist 1(Y.C.G.)	Radiologist 2(T.C.)
1	1
1	1
1	1
1	1
0	1
1	0
1	1
1	1
1	1
1	1
0	1
1	1
0	1
1	0
0	1
1	0
1	1
0	0
1	1
1	1
1	0
1	1
0	1
1	1
1	1
1	1
1	1
0	1
1	1
1	0
