IMAGING INFORMATICS

I J Radiol. 2025 April; 22(2): e161555 https://doi.org/10.5812/iranjradiol-161555

Published Online: 2025 April 30 **Research Article**

Could the Utilization of Large Language Models Contribute to Magnetic Resonance Imaging Acquisition Protocols?

Eren Çamur [1,*], Turay Cesur [2], Yusuf Öztürk [3], Mehmet Kutlu [3], Halil Karataş [3], Sadettin Emre Eroglu [3], Rıza Sarper Ökten [3], Semra Duran [3], Arzu Özsoy [4], Hatice Gül Hatipoglu [3]

[1] Ankara 29 Mayıs State Hospital, Ankara, Turkey
[2] Ankara Mamak State Hospital, Ankara, Turkey
[3] Ankara Bilkent City Hospital, Ankara, Turkey
[4] Department of Radiology, Faculty of Medicine, Medipol University, Istanbul, Turkey

[*] ***Corresponding Author:*** Ankara 29 Mayıs State Hospital, Ankara, Turkey. Email: eren.camur@outlook.com

Received: 24 January, 2025; **Revised:** 23 April, 2025; **Accepted:** 26 April, 2025

Abstract

Background: Large language models (LLMs) are increasingly integrated into radiology, offering potential benefits in workflow optimization and imaging study selection. The Turkish Society of Radiology published the "Turkish Society of Radiology 2018 Magnetic Resonance Imaging and Computed Tomography Acquisition Standards Guideline" (TSR-2018 MCASG) in 2018. This guideline covers sequence selection, patient positioning, scanning parameters, and specific sequence requirements. To our knowledge, no study has assessed the proficiency and knowledge of LLMs in determining magnetic resonance acquisitions and compared them to radiologists.

Objectives: This study aims to evaluate the performance of various LLMs in guiding magnetic resonance imaging (MRI) acquisition protocols based on TSR-2018 MCASG and to compare their proficiency with radiologists across different experience levels.

Materials and Methods: In this cross-sectional observational study, eight LLMs (including ChatGPT-4o models, ChatGPT-o1, Claude 3 Opus, Claude 3.5 Sonnet, Gemini 1.5 Pro, Llama 3.1 405B, and Mistral Large 2) were assessed alongside radiologists ranging from junior residents to senior radiologists (SRs). A total of 105 open-ended questions (OEQs) and 105 case-based questions (CBQs) including different sections were prepared from TSR-2018 MCASG. Statistical analyses employed non-parametric tests, including the Kruskal - Wallis test with Tamhane's T2 post hoc comparisons and McNemar's test, with a Bonferroni-adjusted significance threshold set at $P < 0.0004$.

Results: Claude 3.5 Sonnet emerged as the standout performer, achieving a mean Likert score of 3.51 ± 0.54 in OEQs and an impressive 83.8% accuracy in CBQs, outperforming other LLMs and radiology residents ($P < 0.0004$). While SRs demonstrated strong performance, Claude 3.5 Sonnet outperformed them in both OEQs and CBQs. Furthermore, LLMs have demonstrated competitive performance with junior radiologists (JRs) in both OEQs and CBQs.

Conclusion: Our findings herald a transformative era in radiology, with Claude 3.5 Sonnet leading the vanguard in MRI sequence selection and their contribution to MRI acquisitions. The LLMs can make an important contribution as supportive tools for MRI acquisition optimization.

Keywords: MR Sequences, Magnetic Resonance Imaging, ChatGPT, Large Language Models, Acquisition, Imaging Protocol

1. Background

1.1. A General Overview of Large Language Models in Radiology

Large language models (LLMs) mark a key moment in artificial intelligence. They have major effects across many fields, especially in medical sciences (1). These models use large datasets and smart algorithms. They show great language accuracy and fluency, making their output sound like human speech (2, 3). The LLMs have demonstrated particular promise in radiology for patient triage, workflow optimization, and report generation (4, 5). The LLMs can automate the selection of imaging studies. This helps rank urgent cases, streamlining workflow and improving departmental efficiency (6, 7). Given the ongoing shortage of radiologists, LLMs could play a supportive role in these areas (8). A major drawback of LLMs in clinical workflows is their risk of creating inaccurate or inappropriate information. This can result in diagnostic errors and put patient safety at risk (9).

A core aspect of radiology practice involves choosing the appropriate imaging modalities and protocols to

How to Cite: Çamur E, Cesur T, Öztürk Y, Kutlu M, Karataş H, et al. Could the Utilization of Large Language Models Contribute to Magnetic Resonance Imaging Acquisition Protocols?. I J Radiol. 2025; 22 (2): e161555. https://doi.org/10.5812/iranjradiol-161555.

ensure precise diagnoses and enhance patient outcomes (10). Previous studies underscore the utility of LLMs in recommending imaging studies across various clinical scenarios, often using established standards like the American College of Radiology Appropriateness Criteria (ACR-AC) (11, 12). Rau et al. developed accGPT, a specialized chatbot based on ChatGPT-3.5-turbo, designed to provide personalized imaging recommendations consistent with ACR-AC. In a comparison involving 50 clinical scenarios, accGPT demonstrated superior accuracy and efficiency compared to radiologists and ChatGPT-3.5 and ChatGPT-4 (11). Similarly, Zaki et al. compared Glass AI and ChatGPT across 1,075 cases from ACR panels. Glass AI significantly outperformed ChatGPT (mean scores 2.32 ± 0.67 vs. 2.08 ± 0.74, P = 0.002), notably in polytrauma, breast, and vascular imaging, though both tools showed limitations in neurologic, musculoskeletal, and cardiac imaging panels (12).

1.2. The Specific Role of Magnetic Resonance Imaging Acquisition Protocols and Turkish Society of Radiology 2018 Magnetic Resonance Imaging and Computed Tomography Acquisition Standards Guideline

The Turkish Society of Radiology published the "Turkish Society of Radiology 2018 Magnetic Resonance Imaging and Computed Tomography Acquisition Standards Guideline" (TSR-2018 MCASG) in 2018. This guideline covers sequence selection, patient positioning, scanning parameters, and specific sequence requirements (13). However, given that national guidelines often differ across countries, testing LLM performance within the context of localized standards is crucial.

1.3. The Research Gap and Study Objectives

To our knowledge, no study has assessed the proficiency and knowledge of LLMs in determining magnetic resonance acquisitions and compared them to radiologists.

2. Objectives

This study aims to address this gap by evaluating the performance of various LLMs regarding TSR-2018 MCASG and comparing them with radiologists of different experiences.

3. Materials and Methods

3.1. Study Design

This cross-sectional observational study compares the performance of various LLMs — including ChatGPT-4o with canvas, ChatGPT-4o, ChatGPT-o1, Claude 3 Opus, Claude 3.5 Sonnet, Google Gemini 1.5 Pro, Meta Llama 3.1 405B, and Mistral Large 2 — with that of two junior radiology residents (JRRs), two senior radiology residents (SRRs), two board-certified [European Diploma in Radiology (EDiR)] junior radiologists (JRs), and two senior radiologists (SRs). The comparison focused on their proficiency regarding MRI acquisition standards and their ability to select the key MRI sequence for specific conditions. To address these abilities, open-ended questions (OEQs) and case-based questions (CBQs) were utilized, which were derived from TSR-2018 MCASG.

Since all questions and cases utilized and analyzed in this study are entirely fictional, no real patient data was used in this study. Also, there were no volunteers participating in this study. No patient information and images were used to eliminate the need for ethics committee approval. Therefore, ethical approval is not applicable for this study. The study methodology adhered to the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) statement, ensuring transparency and reproducibility (14). An overview of the workflow is shown in Figure 1.

3.2. Participant Radiologists

1. Junior radiology residents: The JRR1 and JRR2 both have two years of experience in radiology and one year of experience in MRI.

2. Senior radiology residents: The SRR1 and SRR2 both have four years of experience in radiology and three years of experience in MRI.

3. Junior radiologists: JR1, JR2, and JR3 are all board-certified (EDiR) with seven years of experience in radiology and six years of experience in MRI.

4. Senior radiologists: SR1, SR2, SR3, and SR4 all have twenty-three years of experience in radiology and twenty years of experience in MRI.

The background of the radiologists is provided in Table 1.

3.3. Question Development and Validation

The study utilized a total of 210 questions based on key knowledge from TSR-2018 MCASG, comprising 105 OEQs and 105 CBQs from different sections (Breast, Abdomen and Pelvis, Musculoskeletal, Brain, Cardiothoracic, Spinal, and Head and Neck). There were 15 questions for each section in both question formats

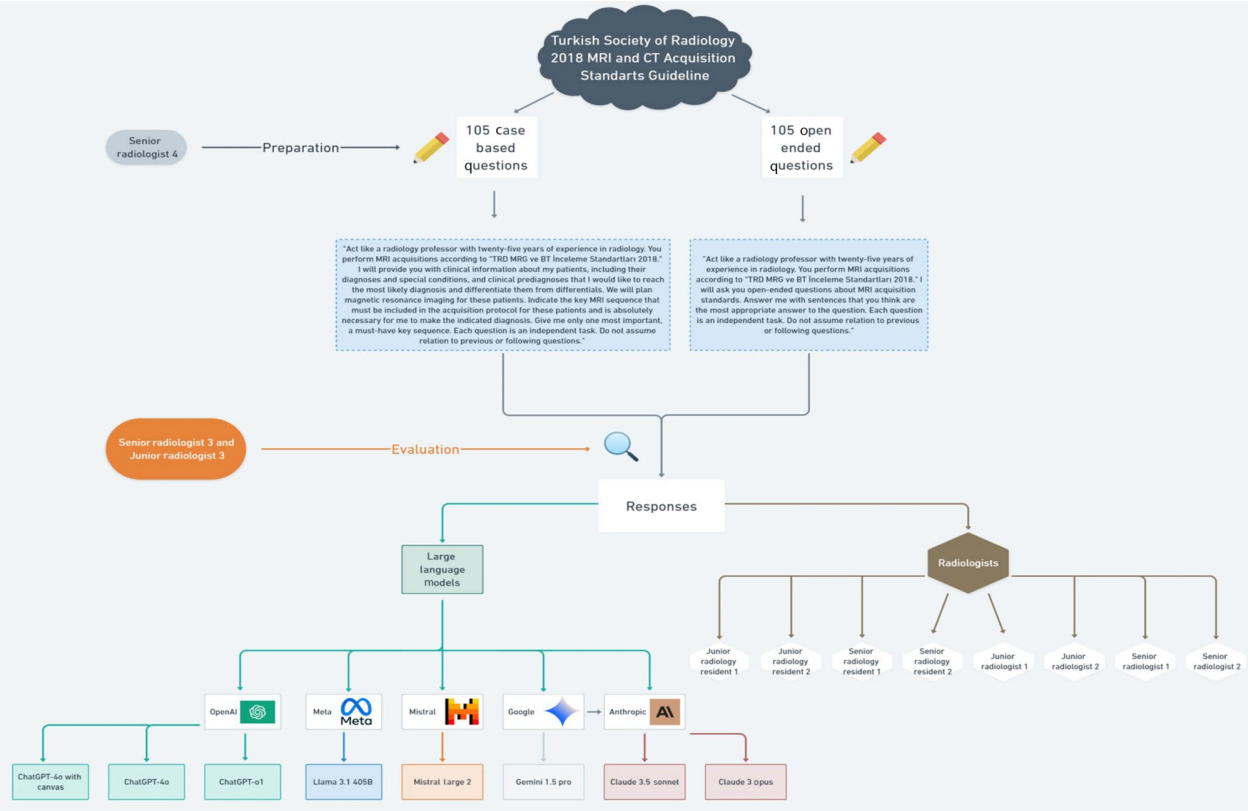

Figure 1. The workflow of the study

Table 1. The Background of Radiologists

Variables	Radiologists (name initials)	Radiology experience (y)	MRI experience (y)	Certification
JRR	JRR1 (S.E.E.); JRR2 (H.K.)	2	1	–
SRR	SRR1 (Y.Ö.); SRR2 (M.K.)	4	3	–
JR	JR1 (Y.C.G.); JR2 (T.C.); JR3 (E.Ç.)	7	6	Board-certified (EDiR)
SR	SR1 (S.D.); SR2 (R.S.Ö.); SR3 (A.Ö.); SR4 (H.G.H.Ç.)	23	20	–

Abbreviations: MRI, magnetic resonance imaging; JRR, junior radiology resident; SRR, senior radiology resident; JR, junior radiologist; EDiR, European Diploma in Radiology; SR, senior radiologist.

(Table 2). These questions were created through a three-step workflow.

A. Stage 1 (item generation): The JR3 drafted 30 OEQs and 30 CBQs for each of the seven sections (n = 420), drawing exclusively on TSR-2018 MCASG.

B. Stage 2 [expert consensus (modified Delphi)]: The SR3 and SR4 independently evaluated and rated every question for clinical relevance and clarity. Questions with a Content-Validity Index < 0.80 were revised and re-rated; after two rounds, ≥ 90% inter-rater agreement was achieved. The process yielded the final 15 OEQs and 15 CBQs per subspecialty (n = 210).

C. Stage 3 (pilot clarity testing): The consensus set was trialed with two different radiology residents (in the second and last year of their residency) who were not study participants. Feedback resulted in minor wording changes; no questions were discarded.

Table 2. The Question Set of the Study

Question type	Question number per section (n)	Sections	Total number of questions (n)	Purpose of question type
OEQs	15	7	105	Assess factual knowledge of MRI acquisition standards
CBQs	15	7	105	Identify single, key MRI sequence for a given clinical scenario

Abbreviations: OEQs, open-ended questions; MRI, magnetic resonance imaging; CBQs, case-based questions.

Because the questions are rule-based with objectively correct answers anchored in TSR-2018 MCASG, formal difficulty indexing or factor analysis was not pursued. The OEQs were clearly formulated, each addressing a single, specific concept to effectively evaluate knowledge on MRI acquisition standards such as technical parameters, protocol considerations, imaging sequences, and patient preparation and positioning. Correspondingly, the CBQs were designed to identify the most appropriate MRI sequence for particular clinical conditions, typically providing one definitive answer; however, in certain cases, two sequences were deemed equally appropriate, with either considered correct. Supplementary Materials list OEQs and CBQs with their datasets.

3.4. Prompting and Model Input Procedures

We used the following prompt for OEQs: "Act like a radiology professor with twenty-five years of experience in radiology. You perform MRI acquisitions according to 'TRD MRG ve BT İnceleme Standartları 2018.' I will ask you open-ended questions about MRI acquisition standards. Answer me with sentences that you think are the most appropriate answer to the question. Each question is an independent task. Do not assume relation to previous or following questions".

For CBQs, the following input prompt was used: "Act like a radiology professor with twenty-five years of experience in radiology. You perform MRI acquisitions according to 'TRD MRG ve BT İnceleme Standartları 2018.' I will provide you with clinical information about my patients, including their diagnoses and special conditions, and clinical pre-diagnosis that I would like to reach the most likely diagnosis and differentiate them from differentials. We will plan magnetic resonance imaging (MRI) for these patients. Indicate the key MRI sequence that must be included in the acquisition protocol for these patients and is absolutely necessary for me to make the indicated diagnosis. Give me only one most important, a must-have key sequence. Each question is an independent task. Do not assume relation to previous or following questions".

These prompts followed a structured, zero-shot format without any iterative refinement during the study. They employed role-based contextualization to emulate the reasoning process of a SR, with the intent to enhance clinical relevance and promote detailed differential generation. To avoid potential bias from variable prompt construction across whole sessions, a single prompt format was used. To eliminate carry-over context, each model was tested in a fresh session per format with no prior conversation history or memory activated. Context-resetting measures were taken where applicable.

All models were used with default hyperparameter settings as provided in their publicly available web interfaces as of January 2025. No fine-tuning or API-level parameter manipulation was applied. We used the web-based front ends to ensure evaluation under standard user conditions. These standardization procedures were applied across all eight LLMs to control for prompt variability and ensure that observed differences in performance were attributable to model behavior rather than prompt structure or system configuration. These prompts were administered in January 2025 across eight different models: Anthropic's Claude 3 Opus and Claude 3.5 Sonnet (https://claude.ai.com), OpenAI's ChatGPT-4o with canvas, ChatGPT-4o, and ChatGPT-o1 (https://chat.openai.com), Google Gemini 1.5 Pro (https://aistudio.google.com), Mistral Large 2 (https://mistral.ai), and Meta Llama 3.1 405B (https://metaai.com).

The LLMs were not subjected to any additional pre-training or fine-tuning by the authors before the study, and no specific information or criteria that might influence the research objectives or outcomes were provided (Figures 2A and B).

3.5. Performance Evaluation

The responses to OEQs were independently evaluated by JR3 and SR3 using a 4-point Likert scale:

A. 1 point: Completely incorrect.

B. 2 points: Mostly incorrect.

C. 3 points: Mostly correct.

D. 4 points: Completely correct.

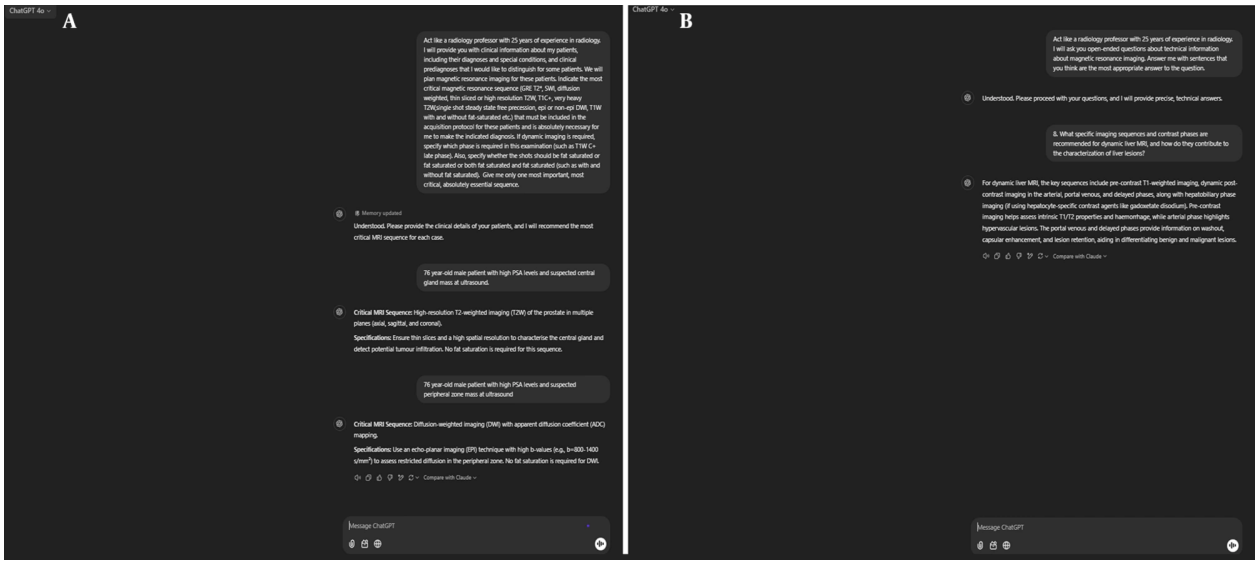

Figure 2. The examples of chat-session with ChatGPT-4o

To reduce potential bias, all responses for each question were exported as plain-text files and anonymized using alphanumeric codes by an independent radiology resident not involved in scoring and not participating in the study. This ensured that evaluators were blinded to the identity of the respondent (LLMs and radiologists) and scored the responses without knowledge of the source.

For the evaluation of OEQs, SR4 authored concise model answers encapsulating the core technical or protocol concept mandated by the guideline. These reference answers, plus a 4-point Likert rubric, were supplied to the two evaluators (JR3 and SR3). Ratings were assigned independently using the reference answer key as the benchmark. The JR3 and SR3 reviewed the responses independently. After that, the discrepancies were resolved via in-person discussion. The pre-consensus interobserver agreement was substantial, with a weighted Cohen's κ of 0.864 (95% CI, 0.821 - 0.902), indicating high scoring consistency.

For the evaluation of CBQs, SR4 derived a single key (indispensable) MRI sequence for each question strictly from TSR-2018 MCASG. Where the guideline allowed two equally critical sequences, both were entered into the answer key. This answer key constituted the gold standard. CBQ responses were independently scored by JR3 and SR3 using binary scoring (correct = 1, incorrect = 0).

3.6. Statistical Analysis

Descriptive statistics were represented using percentages. Subsequently, Tamhane's T2 procedure was employed for post hoc multiple comparisons to delineate specific intergroup differences. McNemar's test was used to compare the proportion of correct responses between different questions. The Wilcoxon test was used to compare Likert scores. For paired comparisons of Likert scores and accuracy between radiologists and LLMs, a Bonferroni correction was applied to adjust for multiple comparisons. Specifically, with 120 pairwise comparisons across eight LLMs and eight radiologists, statistical significance was defined as $P < 0.0004$. All statistical analyses were performed using SPSS version 26.0 (IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0. Armonk, NY: IBM Corp).

4. Results

4.1. Open-ended Question Performance

Claude 3.5 Sonnet achieved the highest performance of 3.51 ± 0.54 (median: 4), followed by ChatGPT-4o with canvas at 3.45 ± 0.73 (median: 3) and ChatGPT-4o at 3.39 ± 0.71 (median: 3). ChatGPT-o1 and Claude 3 Opus demonstrated comparable performance, with scores of 3.25 ± 0.74 (median: 3) and 3.24 ± 0.70 (median: 3),

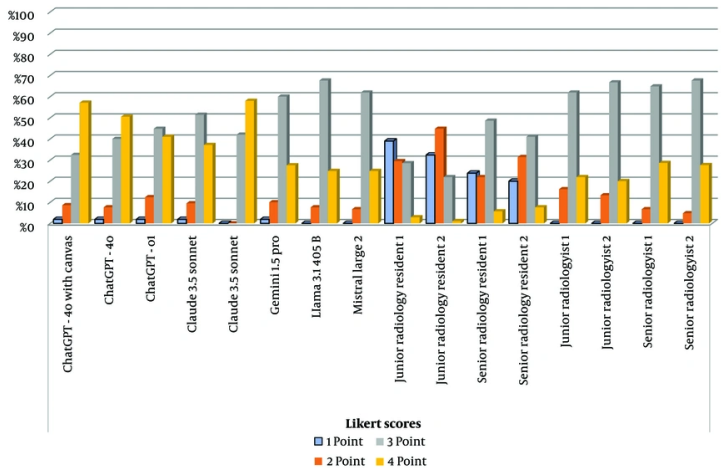

Figure 3. The performance of large language models (LLMs) and radiologists at open-ended questions (OEQs)

Table 3. Descriptive Statistics for Open-ended Questions

Variables	Mean Likert score (point)	95% CI lower (point)	95% CI upper (point)
Claude 3.5 Sonnet	3.51	3.39	3.63
ChatGPT-4o with canvas	3.45	3.28	3.62
ChatGPT-4o	3.39	3.22	3.56
ChatGPT-o1	3.25	3.08	3.42
Claude 3 Opus	3.24	3.08	3.4
Mistral Large 2	3.25	3.12	3.38
Llama 3.1 405B	3.17	3.04	3.3
Gemini 1.5 Pro	3.13	2.99	3.27
JRR1	1.95	1.73	2.17
JRR2	1.91	1.65	2.17
SRR1	2.36	2.13	2.59
SRR2	2.36	2.13	2.59
JR1	3.06	2.92	3.2
JR2	3.07	2.93	3.21
SR1	3.22	3.09	3.35
SR2	3.23	3.1	3.36

Abbreviations: CI, confidence interval; JRR, junior radiology resident; SRR, senior radiology resident; JR, junior radiologist; SR, senior radiologist.

respectively. Gemini 1.5 Pro recorded a mean score of 3.13 ± 0.67 (median: 3), while Llama 3.1 405B and Mistral Large 2 achieved mean scores of 3.17 ± 0.55 (median: 3) and 3.25 ± 0.57 (median: 3), respectively. The JRR1 and JRR2 recorded means of 1.95 ± 0.89 (median: 2) and 1.91 ± 0.76 (median: 2), respectively. The SRR1 and SRR2 demonstrated slightly higher performances of 2.36 ± 0.90 (median: 3) and 2.36 ± 0.89 (median: 3), respectively. The JR1 and JR2 achieved a mean of 3.06 ± 0.62 (median: 3) and 3.07 ± 0.58 (median: 3), respectively. The SR1 recorded a mean of 3.22 ± 0.55 (median: 3), while SR2 achieved the highest performance among radiologists with 3.23 ± 0.52 (median: 3) (Figure 3 and Table 3).

There was no significant difference in LLM performance across the sections (e.g., brain, abdomen, MSK) at OEQs. The overall effect size for OEQ performance was large ($\eta^2 = 0.48$), indicating

Table 4. Comparison of the Performance of Large Language Models and Radiologists at Open-ended Questions [a]

Variables	Claude 3 Opus	Claude 3.5 Sonnet	ChatGPT-4o	Mistral Large 2	ChatGPT-4o with canvas	Gemini 1.5 Pro	ChatGPT-o1	Llama 3.1 405B	JRR-1	JRR-2	SRR-1	SRR-2	JR-1	JR-2	SR-1	SR-2
Claude 3 Opus	-	0.0020	0.0640	0.9070	0.0100	0.2470	0.9110	0.3400	0.0002	0.0002	0.0002	0.0002	0.0420	0.0400	0.8390	0.7590
Claude 3.5 Sonnet	0.0020	-	0.1580	0.0002	0.4550	0.0001	0.0040	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002	0.0002
ChatGPT-4o	0.0640	0.1580	-	0.0910	0.3170	0.0050	0.0800	0.3200	0.0001	0.0001	0.0003	0.0003	0.0010	0.0010	0.4900	0.5200
Mistral Large 2	0.9070	0.0002	0.0910	-	0.0230	0.1350	0.9590	0.2580	0.0001	0.0001	0.0002	0.0003	0.0340	0.0380	0.7230	0.7020
ChatGPT-4o with canvas	0.0100	0.4550	0.3170	0.0230	-	0.0010	0.0050	0.0010	0.0001	0.0001	0.0002	0.0002	0.0010	0.0010	0.0120	0.0110
Gemini 1.5 Pro	0.2470	0.0001	0.0050	0.1350	0.0010	-	0.3390	0.6150	0.0003	0.0002	0.0002	0.0001	0.3800	0.3110	0.2700	0.2340
ChatGPT-o1	0.9110	0.0040	0.0800	0.9590	0.0050	0.3390	-	0.3790	0.0001	0.0001	0.0002	0.0003	0.0350	0.0310	0.7920	0.7000
Llama 3.1 405B	0.3400	0.0001	0.3200	0.2580	0.0010	0.6150	0.3790	-	0.0001	0.0001	0.0001	0.0001	0.1700	0.1230	0.5400	0.5080
JRR-1	0.0002	0.0001	0.0001	0.0001	0.0001	0.0003	0.0001	0.0001	-	0.7510	0.0002	0.0003	0.0002	0.0002	0.0002	0.0002
JRR-2	0.0002	0.0001	0.0001	0.0001	0.0001	0.0002	0.0001	0.0001	0.7510	-	0.0002	0.0002	0.0001	0.0002	0.0001	0.0001
SRR-1	0.0002	0.0001	0.0003	0.0002	0.0002	0.0002	0.0003	0.0001	0.0003	0.0002	-	0.8210	0.0001	0.0002	0.0002	0.0002
SRR-2	0.0002	0.0001	0.0003	0.0003	0.0002	0.0001	0.0003	0.0001	0.0003	0.0002	0.8210	-	0.0002	0.0002	0.0001	0.0001
JR-1	0.0420	0.0002	0.0010	0.0340	0.0010	0.3800	0.0350	0.1700	0.0002	0.0001	0.0001	0.0002	-	0.8600	0.0002	0.0002
JR-2	0.0400	0.0002	0.0010	0.0380	0.0010	0.3110	0.0310	0.1230	0.0002	0.0002	0.0002	0.0002	0.8600	-	0.0003	0.0002
SR-1	0.8390	0.0002	0.4900	0.7230	0.0120	0.2700	0.7920	0.5400	0.0002	0.0001	0.0002	0.0001	0.0002	0.0003	-	0.6400
SR-2	0.7590	0.0002	0.5200	0.7020	0.0110	0.2340	0.7000	0.5080	0.0002	0.0001	0.0002	0.0001	0.0002	0.0002	0.6400	-

Abbreviations: JRR, junior radiology resident; SRR, senior radiology resident; JR, junior radiologist; SR, senior radiologist.

[a] P-values are obtained from Wilcoxon test.

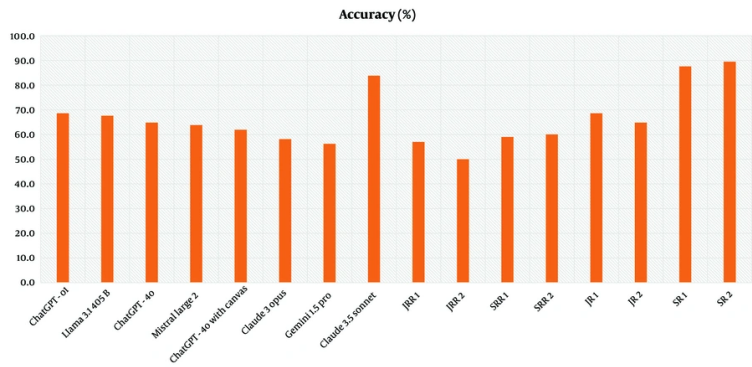

Figure 4. The accuracies of large language models (LLMs) and radiologists at case-based questions (CBQs) (abbreviations: JRR, junior radiology resident; SRR, senior radiology resident; JR, junior radiologist; SR, senior radiologist).

substantial variability between groups. The pairwise effect size between Claude 3.5 Sonnet and Gemini 1.5 Pro was large (Cohen's d = 0.79, 95% CI: 0.52 - 1.06). Similarly, the effect size between Claude 3.5 Sonnet and Mistral Large 2 was Cohen's d = 0.57 (95% CI: 0.30 - 0.83). Claude 3.5 Sonnet outperformed Gemini 1.5 Pro (P = 0.0001), Mistral Large 2 (P = 0.0002), Llama 3.1 405B (P = 0.0001), and radiologists (JRR1, JRR2, SRR1, SRR2, JR1, JR2, SR1, and

Table 5. Descriptive Statistics for Case-based Questions

Variables	CBQ accuracy (%)	95% CI lower (%)	95% CI upper (%)
Claude 3.5 Sonnet	84	78	90
SR2	90	84	96
SR1	88	82	94
ChatGPT-o1	69	61	77
Llama 3.1 405B	68	60	76
JR1	69	61	77
ChatGPT-4o	65	57	73
JR2	65	57	73
Mistral Large 2	64	56	72
ChatGPT-4o with canvas	62	54	70
SRR2	60	52	68
SRR1	59	51	67
Claude 3 Opus	58	50	66
JRR1	57	49	65
Gemini 1.5 Pro	56	48	64
JRR2	50	42	58

Abbreviations: CBQ, case-based question; CI, confidence interval; SR, senior radiologist; JR, junior radiologist; SRR, senior radiology resident; JRR, junior radiology resident.

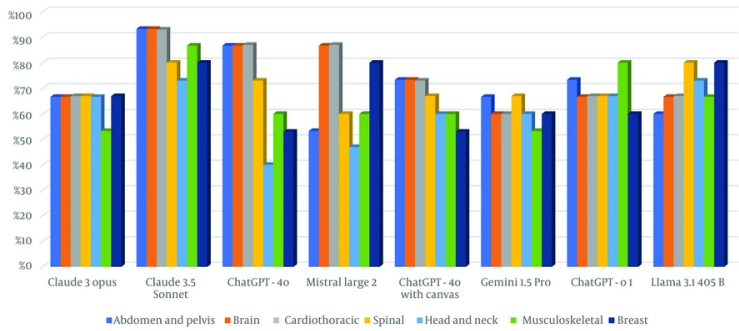

Figure 5. The performance of large language models (LLMs) across the sections at case-based questions (CBQs)

SR2) ($P < 0.0004$). There was no significant difference between the performance of other LLMs ($P > 0.0004$). Radiology residents underperformed significantly when compared to LLMs and other radiologists ($P < 0.0004$). The comparison of the performance of LLMs and radiologists at OEQs is shown in Table 4.

4.2. Case-based Question Accuracy

The LLMs revealed the following accuracy rates (the proportion of correctly selected key MRI sequences by the models and radiologists in response to clinical case-based scenarios) at CBQs: ChatGPT-4o with canvas (61.9%), ChatGPT-4o (64.8%), ChatGPT-o1 (68.6%), Claude 3 Opus (58.1%), Claude 3.5 Sonnet (83.8%), Gemini 1.5 Pro (56.2%), Llama 3.1 405B (67.6%), and Mistral Large 2 (63.8%). The JRRs achieved 57% and 50%, while SRRs scored 59% and 60%. JR1 and JR2 achieved accuracy rates of 68.6% and 64.8%, respectively, while SR1 recorded 87.6%, and SR2 demonstrated accuracy at 89.5% (Figure 4 and Table 5).

Claude 3.5 Sonnet demonstrated superior performance among LLMs across all sections at CBQs. It consistently achieved ≥ 73% accuracy across all subspecialties and excelled in the "Abdomen and Pelvis", "Brain", "Cardiothoracic", and "Spinal" sections with 93.3% accuracy (Figures 5 and 6). The performance of

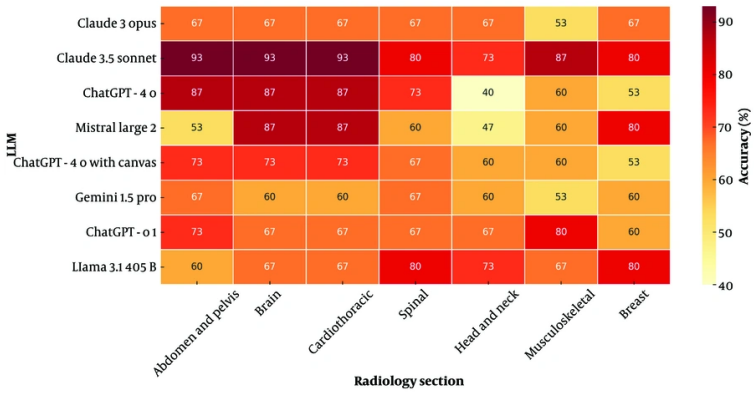

Figure 6. Heat-map of model accuracy in selecting key magnetic resonance imaging (MRI) sequences by radiology section

LLMs at CBQs and OEQs according to sections is shown in Table 6 ($P > 0.05$).

The overall effect size for CBQ accuracy was large ($\eta^2 = 0.52$), suggesting substantial between-group variability. The pairwise effect size between Claude 3.5 Sonnet and Gemini 1.5 Pro was large (Cohen's d = 1.41, 95% CI: 1.07 - 1.75). Between Claude 3.5 Sonnet and Mistral Large 2, Cohen's d was 1.34 (95% CI: 1.01 - 1.67). Claude 3.5 Sonnet outperformed Claude 3 Opus ($P = 0.0001$), ChatGPT-4o with canvas ($P = 0.0001$), Gemini 1.5 Pro ($P = 0.0002$), Mistral Large 2 ($P = 0.0003$), Llama 3.1 405B ($P = 0.0001$), and all radiologists ($P < 0.0004$).

There was no significant difference between the performance of other LLMs ($P > 0.0004$). Radiology residents underperformed significantly when compared to SRs ($P < 0.0004$). The comparison of the performance of all LLMs and radiologists at CBQs is shown in Table 7.

5. Discussion

5.1. Summary of Key Findings

The most noteworthy finding of this study is the superior performance of the Claude 3.5 Sonnet model in adhering to MRI acquisition standards across nearly all sections of radiology. This model outperformed SRs in selecting the most important sequence that must be included in the MRI acquisition protocol. Claude 3.5 Sonnet shows impressive performance with 83.8% accuracy on CBQs and scored high with an average of 3.51 (± 0.54) and a median of 4 on OEQs. This suggests that it could be a pioneer model in this field.

Another important result of our study is that all LLMs had comparable proficiency in both OEQs and CBQs to JRs. This shows that recent LLMs show strong potential as helpful tools in MRI acquisition and sequence selection. In addition to this, our study uniquely compares the performance of different LLMs on MRI acquisition standards knowledge and key sequence selection, and it shows the valuable potential of these models by comparing them to radiologists across different experience levels.

The performance variations among LLMs can be primarily attributed to their distinct architectural frameworks and training methodologies. A notable observation was that internet-enabled models like Gemini 1.5 Pro demonstrated lower accuracy compared to their offline counterparts, likely due to their tendency to retrieve information from non-peer-reviewed sources. In contrast, ChatGPT models and Claude models operate on proprietary, curated datasets specifically designed for medical and scientific applications.

When the performances of LLMs included were compared according to sections, there was no significant difference in both OEQs and CBQs according to the sections ($P > 0.05$). This result shows that current LLMs have a similar level of competence and knowledge not only in certain sections but also in almost all sections that we frequently encounter in radiology practice.

5.2. Comparison with Existing Literature

Prior studies have largely addressed the radiological knowledge of LLMs and their support for post-imaging tasks such as lesion detection, image segmentation, and

Table 6. The Performances of Large Language Models Across Sections [a]

Variables	Abdomen and pelvis	Brain	Cardiothoracic	Spinal	Head and neck	Musculoskeletal	Breast	P
Claude 3 Opus (CBQ)								0.239 X^2
False	5 (33.3)	5 (33.3)	5 (33.3)	5 (33.3)	5 (33.3)	7 (46.7)	5 (33.3)	
True	10 (66.7)	10 (66.7)	10 (66.7)	10 (66.7)	10 (66.7)	8 (53.3)	10 (66.7)	
Claude 3 Opus Likert Score (OEQ)								0.592 K
Mean ± SD	3.07± 0.594	3.47± 0.516	3.07 ± 0.799	3.33 ± 0.617	3.40 ± 0.632	3.13 ± 0.990	3.20 ± 0.676	
Median	3	3	3	3	3	3	3	
Claude 3.5 Sonnet (CBQ)								0.710 X^2
False	1 (6.7)	1 (6.7)	1 (6.7)	3 (20.0)	4 (26.7)	2 (13.3)	3 (20.0)	
True	14 (93.3)	14 (93.3)	14 (93.3)	12 (80.0)	11 (73.3)	13 (86.7)	12 (80.0)	
Claude-3.5 Sonnet Likert Score (OEQ)								0.670 K
Mean ± SD	3.60 ± 0.507	3.60± 0.507	3.60 ± 0.507	3.67 ± 0.488	3.67 ± 0.488	3.47 ± 0.516	3.47 ± 0.516	
Median	4	4	4	4	4	3	3	
ChatGPT-4o (CBQ)								0.050 X^2
False	2 (13.3)	2 (13.3)	2 (13.3)	4 (26.7)	9 (60.0)	6 (40.0)	7 (46.7)	
True	13 (86.7)	13 (86.7)	13 (86.7)	11 (73.3)	6 (40.0)	9 (60.0)	8 (53.3)	
ChatGPT-4o Likert Score (OEQ)								0.717 K
Mean ± SD	3.47 ± 0.516	3.47± 0.516	3.53 ± 0.640	3.33 ± 0.724	3.47 ± 0.640	3.20 ± 1.014	3.13 ± 0.834	
Median	3	3	4	4	4	3	3	
Mistral Large 2 (CBQ)								0.210 X^2
False	7 (46.7)	2 (13.3)	2 (13.3)	6 (40.0)	8 (53.3)	6 (40.0)	3 (20.0)	
True	8 (53.3)	13 (86.7)	13 (86.7)	9 (60.0)	7 (46.7)	9 (60.0)	12 (80.0)	
Mistral Large 2 Likert Score (OEQ)								0.423 K
Mean ± SD	3.20 ± 0.561	3.20± 0.561	3.33 ± 0.617	3.13 ± 0.516	3.33 ± 0.617	3.47 ± 0.640	3.07 ± 0.458	
Median	3	3	3	3	3	4	3	
ChatGPT-4o with canvas (CBQ)								0.709 X^2
False	4 (26.7)	4 (26.7)	4 (26.7)	5 (33.3)	6 (40.0)	6 (40.0)	7 (46.7)	
True	11 (73.3)	11 (73.3)	11 (73.3)	10 (66.7)	9 (60.0)	9 (60.0)	8 (53.3)	
ChatGPT-4o with canvas Likert Score (OEQ)								0.128 K
Mean ± SD	3.20 ± 0.775	3.53± 0.640	3.80 ± 0.414	3.60 ± 0.737	3.53 ± 0.640	3.27 ± 1.033	3.20 ± 0.676	
Median	3	3	4	3	4	4	3	
Gemini 1.5 Pro (CBQ)								0.648 X^2
False	5 (33.3)	6 (40.0)	6 (40.0)	5 (33.3)	6 (40.0)	7 (46.7)	6 (40.0)	
True	10 (66.7)	9 (60.0)	9 (60.0)	10 (66.7)	9 (60.0)	8 (53.3)	9 (60.0)	
Gemini 1.5 Pro Likert Score (OEQ)								0.761 K
Mean ± SD	3.20 ± 0.561	2.93± 0.799	3.07 ± 0.594	3.27 ± 0.594	3.33 ± 0.617	3.07 ± 0.799	3.07 ± 0.704	
Median	3	3	3	3	3	3	3	
ChatGPT-o1 (CBQ)								0.948 X^2
False	4 (26.7)	5 (33.3)	5 (33.3)	5 (33.3)	5 (33.3)	3 (20.0)	6 (40.0)	
True	11 (73.3)	10 (66.7)	10 (66.7)	10 (66.7)	10 (66.7)	12 (80.0)	9 (60.0)	
ChatGPT-o1 Likert Score (OEQ)								0.912 K
Mean ± SD	3.27 ± 0.799	3.47± 0.516	3.20 ± 0.775	3.27 ± 0.594	3.27 ± 0.704	3.20 ± 1. 014	3.07 ± 0.799	
Median	3	3	3	3	3	3	3	
Llama 3.1 405B (CBQ)								0.459 X^2
False	6 (40.0)	5 (33.3)	5 (33.3)	3 (20.0)	4 (26.7)	5 (33.3)	3 (20.0)	
True	9 (60.0)	10 (66.7)	10 (66.7)	12 (80.0)	11 (73.3)	10 (66.7)	12 (80.0)	
Llama 3.1 405B Likert Score (OEQ)								0.752 K
Mean ± SD	3.20 ± 0.561	3.27± 0.458	3.13 ± 0.516	3.07 ± 0.594	3.00 ± 0.535	3.27 ± 0.594	3.27 ± 0.594	
Median	3	3	3	3	3	3	3	

Abbreviations: X^2,chi-square test; CBQ, case based questions; OEQ, open-ended questions; K, Kruskal-wallis test; SD, standart deviation.

[a] Values are expressed as No. (%), Mean ± SD, or Median.

automated reporting (5, 15-18). Bhayana et al. evaluated ChatGPT's performance on 150 MCQs designed to match the style and rigor of Canadian Royal College and American Board of Radiology examinations, demonstrating that the model correctly answered nearly 70% of all questions. Notably, ChatGPT excelled in lower-order cognitive tasks — those requiring recall or understanding — achieving an 84% success rate, and also performed strongly (89%) on higher-order clinical management questions (16).

In another study, Ariyaratne et al. assessed ChatGPT's suitability for radiology board-style assessments; GPT-4

was tested on question banks mirroring parts 1 and 2A of the Fellowship of the Royal College of Radiologists (FRCR) examination. Although GPT-4 answered nearly 75% of part 1 true/false questions correctly — a score marginally below the established passing mark — its performance on 2A single best answer questions was notably stronger, achieving a 74.2% accuracy rate and comfortably surpassing the passing threshold of 63.3% (19).

Beyond text-based knowledge, Horiuchi et al. found that ChatGPT 4 performed comparably well to a radiology assistant, although not as well as a board-certified radiologist, with an accuracy of 43% on 106 "Test Yourself" cases from Skeletal Radiology (20).

5.3. Implications for Radiology Practice

The contribution that LLMs can make to radiology goes beyond radiological image assessment. Mese et al. emphasized that by generating customized assessments, translating complex materials, and summarizing large volumes of data, ChatGPT can serve as a flexible tutor around the clock. Using these capabilities of ChatGPT in radiology education can provide a more holistic, student-centered environment by advancing critical thinking and professional skill development without eliminating the essential role of traditional teaching methods (21).

Lyu et al. translated both computed tomography and MRI reports into different languages with ChatGPT, and radiologists evaluated the reports translated by ChatGPT on a 5-point system for accuracy and adequacy. In this study, ChatGPT-4 achieved a score of 4.27, and it was stated that it has great potential in this regard (22).

In another study, Sievert et al. performed risk stratification of thyroid nodules according to the Thyroid Imaging Reporting and Data System (TI-RADS) with ChatGPT on anonymized radiology reports and stated that it offers an important future to guide clinicians in this regard (23). Zaki et al. demonstrated that LLMs are highly effective in determining the most appropriate imaging modality (12).

Beyond previous studies, this study broadens the scope of LLMs' role to include strategic decision-making before an acquisition is even performed by demonstrating the great performance of LLMs in patient-specific key sequence selection (24-27). The MRI acquisition optimization and sequence selection have critical importance as fundamental aspects for radiologists. Our study contributes valuable insights and a different approach by evaluating their proficiency in MRI acquisition standards and key sequence selection.

5.4. Study Limitations

This study has several limitations. First, we used one standardized prompt for each question. We did not test how different prompts might affect LLM responses, as this was beyond our study's aim. Since the input prompt significantly influences LLM performance, using optimized prompts could potentially yield better results. Further studies are needed to understand how different prompts affect LLM responses in this subject.

Second, both question formats employed in this study were intentionally simplified for controlled benchmarking. The OEQs required concise, single-point answers rather than the stepwise narrative reasoning that characterizes day-to-day consultation with clinical teams. Likewise, the CBQs focused on selecting a single, highest-yield MRI sequence per vignette, thereby omitting the diagnostic uncertainty and multi-sequence protocols often necessary in routine practice. Consequently, our design does not fully capture the complexity, ambiguity, and iterative decision-making present in real-world radiology. Future work should therefore validate LLM performance within authentic clinical workflows — preferably through prospective studies or large retrospective imaging datasets that incorporate the full spectrum of patient presentations and imaging requirements.

Also, while the question set was developed using expert consensus and piloted for clarity, formal psychometric validation — such as item difficulty analysis, discrimination indices, or test - retest reliability — was not performed. This may limit the generalizability of item-level findings. Future studies should incorporate comprehensive psychometric evaluation across broader learner cohorts to further establish the reliability and validity of the assessment tool.

Third, MRI acquisition standards in different centers and countries could show differences with minor variations. In this study, we performed our evaluations according to TSR-2018 MCASG, but the performance of LLMs may vary with different MRI acquisition guidelines. Further studies including different MRI acquisition guidelines are needed to show that the performance of LLMs in our study is generalizable to other guidelines. Finally, LLM development is an ongoing process, with models continuously improving through new knowledge and reinforcement learning. Therefore, our results reflect the models' capabilities

Table 7. Comparison of the Performance of Large Language Models and Radiologists at Case Based Questions [a]

Variables	Claude 3 Opus	Claude 3.5 Sonnet	ChatGPT-4o	Mistral Large 2	ChatGPT-4o with canvas	Gemini 1.5 Pro	ChatGPT-o1	Llama 3.1405B	JRR-1	JRR-2	SRR-1	SRR-2	JR-1	JR-2	SR-1	SR-2
Claude 3 Opus	-	0.0001	0.0907	0.1650	0.1700	0.7280	0.0300	0.0270	0.5510	0.6710	0.2810	0.4100	0.0300	0.2810	0.0002	0.0002
Claude 3.5 Sonnet	0.0001	-	0.0020	0.0003	0.0001	0.0002	0.0150	0.0120	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0002
ChatGPT-4o	0.0970	0.0020	-	1	0.5130	0.1760	0.5960	0.6350	0.2650	0.0500	0.7600	0.5960	0.6430	0.7750	0.0001	0.0001
Mistral Large 2	0.1650	0.0003	1	-	0.8620	0.2300	0.4860	0.4050	0.4610	0.0610	1	0.6880	0.4880	0.8880	0.0001	0.0001
ChatGPT-4o with canvas	0.1700	0.0001	0.5130	0.8620	-	0.3550	0.1940	0.2620	0.1510	0.0700	1	0.7850	0.2500	1	0.0002	0.0002
Gemini 1.5 Pro	0.7280	0.0002	0.1760	0.2300	0.3550	-	0.0310	0.0430	0.8770	0.3710	0.4960	0.6880	0.0990	0.4700	0.0001	0.0001
ChatGPT-o1	0.0300	0.0150	0.5960	0.4860	0.1940	0.0310	-	1	0.1090	0.0070	0.3600	0.2220	1	0.3600	0.0030	0.0010
Llama 3.1 405B	0.0270	0.0120	0.6350	0.4050	0.2620	0.0430	1	-	0.1530	0.0030	0.3710	0.2720	1	0.4010	0.0010	0.0001
JRR-1	0.5510	0.0001	0.2650	0.4610	0.1510	0.8770	0.1090	0.1530	-	0.3130	0.6260	0.8900	0.1090	0.6580	0.0001	0.0002
JRR-2	0.6710	0.0001	0.0500	0.0610	0.0700	0.3710	0.0070	0.0030	0.3130	-	0.0800	0.1360	0.0140	0.1060	0.0002	0.0002
SRR-1	0.2810	0.0001	0.7600	1	1	0.4960	0.3600	0.3710	0.6260	0.0800	-	0.8900	0.3370	1	0.0002	0.0002
SRR-2	0.4100	0.0001	0.5960	0.6880	0.7850	0.6880	0.2220	0.2720	0.8900	0.1360	0.8900	-	0.2720	0.8740	0.0002	0.0002
JR-1	0.0300	0.0001	0.6430	0.4880	0.2500	0.0990	1	1	0.1090	0.0140	0.3370	0.2720	-	0.4420	0.0002	0.0010
JR-2	0.2810	0.0001	0.7750	0.8880	1	0.4700	0.3600	0.4010	0.6580	0.1060	1	0.8740	0.4420	-	0.0800	0.1360
SR-1	0.0002	0.5410	0.0001	0.0001	0.0002	0.0001	0.0030	0.0010	0.0001	0.0002	0.0002	0.0002	0.0002	0.0800	-	0.8900
SR-2	0.0002	0.0002	0.0001	0.0001	0.0002	0.0001	0.0010	0.0001	0.0002	0.0002	0.0002	0.0002	0.0010	0.1360	0.8900	-

Abbreviations: JRR, junior radiology resident; SRR, senior radiology resident; JR, junior radiologist; SR, senior radiologist.

[a] P-values are obtained from McNemar test.

only at the time of this study. Future versions of these models may show improved performance in this field.

5.5. Directions for Future Research

Further multicenter studies are warranted to corroborate our findings under authentic clinical conditions, wherein diagnostic uncertainty, patient heterogeneity, and complex multi-sequence protocols more closely mirror routine radiologic workflows. Such studies should incorporate large, longitudinal imaging datasets governed by diverse acquisition guidelines — extending beyond TSR-2018 MCASG — to determine the external validity of current LLMs. Rigorous experimentation with advanced prompt-engineering strategies, adaptive conversational frameworks, and iterative human-in-the-loop feedback will be essential to optimize model performance and mitigate context-specific biases. Finally, longitudinal evaluations charting successive model iterations, together with implementation studies that assess integration into radiologist reporting, protocol planning, and trainee education, will be critical for delineating the true clinical utility and safety profile of LLM-enabled decision support in MRI practice.

In conclusion, particularly Claude 3.5 Sonnet, performed robustly in our simulated benchmark, accurately selecting key MRI sequences and adhering to guideline-based acquisition standards across multiple subspecialties. These results highlight the potential of advanced LLMs as decision-support aids during protocol planning. Future studies incorporating real clinical scenarios and authentic patient data are critical for realizing this great potential of LLMs and evaluating their prospective transformative role in radiologic practice.

Supplementary Material

Supplementary material(s) is available here [To read supplementary materials, please refer to the journal website and open PDF/HTML].

Footnotes

Authors' Contribution: Study conception and design: E. C.; Material preparation and data collection: E. C. and H. G. H. C.; Statistical analysis: E. C. and A. O.; Drafting of the manuscript: E. C. All authors did critical revision of the manuscript for important intellectual content.

Conflict of Interests Statement: The authors declare no conflict of interest.

Data Availability: All data supporting the findings of this study are available within the paper and its Supplementary Materials.

Ethical Approval: Since all questions and cases utilized and analyzed in this study are entirely fictional, no real human data used in this study. No patient information and images are used to eliminate the need for ethics committee approval. Therefore, ethical approval is not applicable for this study.

Funding/Support: The present study received no funding/support.

References

1. Sarker IH. LLM potentiality and awareness: a position paper from the perspective of trustworthy and responsible AI modeling. *Discover Artificial Intelligence*. 2024;**4**(1). https://doi.org/10.1007/s44163-024-00129-0.
2. Biswas SS. Role of Chat GPT in Public Health. *Ann Biomed Eng*. 2023;**51**(5):868-9. [PubMed ID: 36920578]. https://doi.org/10.1007/s10439-023-03172-7.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;**29**(8):1930-40. [PubMed ID: 37460753]. https://doi.org/10.1038/s41591-023-02448-8.
4. 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. [PubMed ID: 37789676]. [PubMed Central ID: PMC10916534]. https://doi.org/10.4274/dir.2023.232417.
5. Kim K, Hong GS, Kim N. [Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging]. *J Korean Soc Radiol*. 2024;**85**(5):848-60. [PubMed ID: 39416320]. [PubMed Central ID: PMC11473984]. https://doi.org/10.3348/jksr.2024.0066.
6. Infante A, Gaudino S, Orsini F, Del Ciello A, Gulli C, Merlino B, et al. Large language models (LLMs) in the evaluation of emergency radiology reports: performance of ChatGPT-4, Perplexity, and Bard. *Clin Radiol*. 2024;**79**(2):102-6. [PubMed ID: 38087683]. https://doi.org/10.1016/j.crad.2023.11.011.
7. Preiksaitis C, Ashenburg N, Bunney G, Chu A, Kabeer R, Riley F, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. *JMIR Med Inform*. 2024;**12**. e53787. [PubMed ID: 38728687]. [PubMed Central ID: PMC11127144]. https://doi.org/10.2196/53787.
8. 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. [PubMed ID: 37891083]. https://doi.org/10.1067/j.cpradiol.2023.10.013.
9. Sblendorio E, Dentamaro V, Lo Cascio A, Germini F, Piredda M, Cicolini G. Integrating human expertise & automated methods for a dynamic and multi-parametric evaluation of large language models' feasibility in clinical decision-making. *Int J Med Inform*. 2024;**188**:105501. [PubMed ID: 38810498]. https://doi.org/10.1016/j.ijmedinf.2024.105501.
10. Cascade PN. The American College of Radiology. ACR Appropriateness Criteria project. *Radiology*. 2000;**214 Suppl**:3-46. [PubMed ID: 10646480]. https://doi.org/10.1148/radiology.214.1.r00ja493.
11. Rau A, Rau S, Zoeller D, Fink A, Tran H, Wilpert C, et al. A Context-based Chatbot Surpasses Trained Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. *Radiology*. 2023;**308**(1). e230970. [PubMed ID: 37489981]. https://doi.org/10.1148/radiol.230970.
12. Zaki HA, Aoun A, Munshi S, Abdel-Megid H, Nazario-Johnson L, Ahn SH. The Application of Large Language Models for Radiologic Decision Making. *J Am Coll Radiol*. 2024;**21**(7):1072-8. [PubMed ID: 38224925]. https://doi.org/10.1016/j.jacr.2024.01.007.
13. Turkish Society of Radiology Executive Committee Members. *TRD MRG ve BT İnceleme Standartları*. 2018. Available from: https://www.turkrad.org.tr/dernekten-haberler/trd-mrg-ve-bt-inceleme-standartlari-2018/.
14. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiol Artif Intell*. 2024;**6**(4). e240300. [PubMed ID: 38809149]. [PubMed Central ID: PMC11304031]. https://doi.org/10.1148/ryai.240300.
15. 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**(7):685-96. [PubMed ID: 38551772]. [PubMed Central ID: PMC11217134]. https://doi.org/10.1007/s11604-024-01552-0.
16. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology*. 2023;**307**(5). e230582. [PubMed ID: 37191485]. https://doi.org/10.1148/radiol.230582.
17. Sarangi PK, Lumbani A, Swarup M, Panda S, Sahoo SS, Hui P, et al. Assessing ChatGPT's Proficiency in Simplifying Radiological Reports for Healthcare Professionals and Patients. *Cureus*. 2023. https://doi.org/10.7759/cureus.50881.
18. 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. [PubMed ID: 38679540]. https://doi.org/10.1016/j.diii.2024.04.003.
19. Ariyaratne S, Jenko N, Mark Davies A, Iyengar KP, Botchu R. Could ChatGPT Pass the UK Radiology Fellowship Examinations? *Acad Radiol*. 2024;**31**(5):2178-82. [PubMed ID: 38160089]. https://doi.org/10.1016/j.acra.2023.11.026.
20. Horiuchi D, Tatekawa H, Oura T, Shimono T, Walston SL, Takita H, et al. ChatGPT's diagnostic performance based on textual vs. visual information compared to radiologists' diagnostic performance in musculoskeletal radiology. *Eur Radiol*. 2025;**35**(1):506-16. [PubMed ID: 38995378]. [PubMed Central ID: PMC11632015]. https://doi.org/10.1007/s00330-024-10902-5.
21. Mese I, Altintas Taslicay C, Kuzan BN, Kuzan TY, Sivrioglu AK. Educating the next generation of radiologists: a comparative report of ChatGPT and e-learning resources. *Diagn Interv Radiol*. 2024;**30**(3):163-74. [PubMed ID: 38145370]. [PubMed Central ID: PMC11095068]. https://doi.org/10.4274/dir.2023.232496.
22. Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis*

Comput Ind Biomed Art. 2023;**6**(1):9. [PubMed ID: 37198498]. [PubMed Central ID: PMC10192466]. https://doi.org/10.1186/s42492-023-00136-5.

23. Sievert M, Conrad O, Mueller SK, Rupp R, Balk M, Richter D, et al. Risk stratification of thyroid nodules: Assessing the suitability of ChatGPT for text-based analysis. *Am J Otolaryngol*. 2024;**45**(2):104144. [PubMed ID: 38113774]. https://doi.org/10.1016/j.amjoto.2023.104144.
24. Gordon EB, Maxfield CM, French R, Fish LJ, Romm J, Barre E, et al. Large Language Model Use in Radiology Residency Applications: Unwelcomed but Inevitable. *J Am Coll Radiol*. 2025;**22**(1):33-40. [PubMed ID: 39299618]. https://doi.org/10.1016/j.jacr.2024.08.027.
25. Chizhikova M, Lopez-Ubeda P, Martin-Noguerol T, Diaz-Galiano MC, Urena-Lopez LA, Luna A, et al. Automatic TNM staging of colorectal cancer radiology reports using pre-trained language models. *Comput Methods Programs Biomed*. 2025;**259**:108515. [PubMed ID: 39602989]. https://doi.org/10.1016/j.cmpb.2024.108515.
26. Lee S, Youn J, Kim H, Kim M, Yoon SH. CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images. *Eur Radiol*. 2025;**35**(7):4374-86. [PubMed ID: 39812665]. [PubMed Central ID: PMC12166004]. https://doi.org/10.1007/s00330-024-11339-6.
27. Liu M, Okuhara T, Dai Z, Huang W, Gu L, Okada H, et al. Evaluating the Effectiveness of advanced large language models in medical Knowledge: A Comparative study using Japanese national medical examination. *Int J Med Inform*. 2025;**193**:105673. [PubMed ID: 39471700]. https://doi.org/10.1016/j.ijmedinf.2024.105673.