Joint Diseases and Related Surgery

Jt Dis Relat Surg
2026;37(1):190-199

ORIGINAL ARTICLE

Assessment of large language models in musculoskeletal radiological anatomy: A comparative study with radiologists

Ali Salbas, MD[1], Ebru Kul Baysan, MD[2]

[1]Department of Radiology, Atatürk Training and Research Hospital, İzmir, Türkiye
[2]Department of Physical and Rehabilitation Medicine, Atatürk Training and Research Hospital, İzmir, Türkiye

ABSTRACT

Objectives: This study aims to evaluate the diagnostic performance of large language models (LLMs) in musculoskeletal radiological anatomy and to compare their accuracy with radiologists of varying experience levels.

Materials and methods: Between May 16, 2025 and June 12, 2025, a total of 175 multiple-choice questions (82 image-based, 93 text-only) were retrieved from Radiopaedia's open-access database. Questions were classified by anatomical region and imaging modality. Three LLMs, ChatGPT-4o (OpenAI), Claude 3.7 Sonnet (Anthropic), and Grok 3 (×AI), were assessed in a zero-shot setting. Their responses were compared to those of an attending musculoskeletal radiologist and two residents (senior and junior). Accuracy rates were calculated and statistically compared.

Results: The attending radiologist achieved the highest overall accuracy (79.4%), followed by the senior (72.6%) and junior resident (66.9%). Among LLMs, ChatGPT-4o performed best overall (69.7%), particularly in text-based questions (88.2%). All LLMs outperformed radiologists in text-based questions but underperformed in image-based ones. The attending radiologist significantly outperformed all LLMs in image interpretation ($p<0.001$). Variations in performance were also noted across anatomical regions and imaging modalities, with some LLMs exceeding radiologists in specific domains such as spinal or shoulder anatomy.

Conclusion: While LLMs, particularly ChatGPT-4o, show strong performance in text-based anatomical questions, their accuracy in image-based musculoskeletal radiology remains limited compared to human radiologists. These findings suggest that LLMs can serve as supplementary tools in education but require further optimization, particularly for visual interpretation tasks, before clinical implementation.

Keywords: Artificial intelligence, large language models, medical education, musculoskeletal radiology, radiological anatomy.

Large language models (LLMs) are artificial neural networks trained on vast textual data to generate human-like outputs. These models analyze linguistic patterns and perform well in tasks such as information generation, summarization, and question answering. Recent multimodal versions can also process data types such as images and audio.[1] Thanks to these capabilities, LLMs are increasingly used in clinical decision support, patient communication, radiological interpretation, and educational content creation in medicine.[2-5]

Anatomical knowledge is vital in both radiology education and practice. While conventional resources have long been standard in anatomy training, models like Chat Generative Pre-trained Transformer (ChatGPT) have recently emerged as educational tools and have been preliminarily applied in radiology.[1,6] However, questions remain regarding their accuracy, consistency, and reliability, particularly for clinical and educational use.[7]

Received: June 16, 2025
Accepted: August 17, 2025
Published online: September 23, 2025

Correspondence: Ali Salbas, MD. Atatürk Eğitim ve Araştırma Hastanesi, Radyoloji Kliniği, 35150 Karabaglar, İzmir, Türkiye.

E-mail: dralisalbas@gmail.com

Citation: Salbas A, Kul Baysan E. Assessment of large language models in musculoskeletal radiological anatomy: A comparative study with radiologists. Jt Dis Relat Surg 2026;37(1):190-199. doi: 10.52312/jdrs.2026.2436.

©2026 All right reserved by the Turkish Joint Diseases Foundation

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes (http://creativecommons.org/licenses/by-nc/4.0/).

The potential of LLMs to answer exam questions has been widely studied in contexts such as the United States Medical Licensing Examination (USMLE) and certification exams in orthopedics, plastic surgery, and neurology.[8-11] However, their

performance in knowledge-intensive domains such as anatomy still remains particularly important. Musculoskeletal anatomy, in particular, involves complex terminology and variations, requiring precise and detailed understanding.[12]

To the best of our knowledge, no previous study in the English literature has evaluated the diagnostic performance of different LLMs in musculoskeletal radiological anatomy or compared their results with radiologists. In the present study, we aimed to assess the accuracy of three widely used LLMs, ChatGPT-4o, Claude 3.7 Sonnet, and Grok 3, in their latest versions on anatomy-based questions in musculoskeletal radiology, and to compare their performance with that of radiology residents and an attending radiologist.

MATERIALS AND METHODS

This methodological comparison study was conducted at the İzmir Atatürk Training and Research Hospital, Department of Radiology between May 16, 2025 and June 12, 2025. A written informed consent was obtained from each participant. The study protocol was approved by the İzmir Katip Çelebi University Health Research Ethics Committee (Date: 15.05.2025, No: 0314). The study was conducted in accordance with the principles of the Declaration of Helsinki.

The questions were obtained from the open-access "Questions" section of the Radiopaedia website (https://radiopaedia.org/search?scope=mcqs).[13] Radiopaedia was chosen as it is a widely used, freely accessible, and expert-curated radiology resource that offers a comprehensive set of standardized multiple-choice questions in musculoskeletal radiology, with confirmed answers provided by experienced contributors. The "Musculoskeletal" option under the “Systems” filter and the "Anatomy" option under the “Sections” filter were applied. Following this filtering process, a total of 230 questions were identified, including 103 with visual content and 127 text-only questions. The study questions were independently evaluated by the researchers in terms of content relevance, redundancy, and question integrity, followed by a consensus agreement. During this process, one question lacking visual content was excluded from the study. In addition, duplicate questions resulting from repeated uploads of the same item (n=3), as well as questions unrelated to the musculoskeletal system (e.g., gastrointestinal or neuroradiology topics; n=51), were also excluded. As a result, a total of 175 questions (82 with images and 93 text-only) were included in the study. All questions were in multiple-choice format (4 to 6 options), and each question contained a single correct answer. In the study, the questions were categorized into two main groups based on their content: questions with images and text-only questions.

In addition, all questions were further classified into seven subgroups based on the anatomical localization referenced in the question content: *(i)* foot-ankle, *(ii)* elbow, *(iii)* knee-thigh, *(iv)* hand-wrist, *(v)* hip, *(vi)* shoulder, and *(vii)* spine.

The questions with images were also classified according to the imaging modality used: *(i)* computed tomography (CT), *(ii)* radiography, *(iii)* magnetic resonance imaging (MRI), and *(iv)* anatomical illustration.

The visual content in the image-based questions was obtained from the original materials on the Radiopaedia platform using the screenshot method. All images were in JPEG format with a minimum resolution of 1000×900 pixels and 96 DPI (dots per inch). The images were captured at high screen resolution and with adequate zoom level. No image processing (e.g., cropping, filtering, or resizing) was applied; the images were shared with the models in their original form. Three LLMs were evaluated in this study: ChatGPT-4o (OpenAI), Claude 3.7 Sonnet (Anthropic), and Grok 3 (xAI). All LLMs were accessed and utilized via their official web-based browser platforms.[14-16] The questions and, if available, the associated images were uploaded to the models using the standardized zero-shot prompt provided below:[17]

Note: This question is for educational purposes only. Your answer will not be used for any clinical decision-making, and no medical responsibility is implied. You should not leave any question blank and choose the most likely correct answer instead of open-ended answers.

To prevent the LLMs from altering their response strategies based on previous answers within the same session (in-context adaptation), the chat history was reset and a new session was initiated for each question.[18,19] All LLM evaluations were conducted using the most up-to-date versions available between May 16 and May 22, 2025.

The evaluating radiologists were as follows:

- An attending radiologist with nine years of professional experience, board certified, and actively practicing musculoskeletal radiology.

- A senior resident who had completed four years of radiology training and a six-month musculoskeletal radiology rotation.
- A junior resident who had completed 2.5 years of radiology training and a three-month musculoskeletal radiology rotation.

The attending radiologist and residents answered the questions individually, in a silent environment under exam-like conditions, at the same time and location, using separate computers and without interacting with one another. All participants completed the task using laptop computers equipped with 14.5-inch screens and a resolution of 3072×1920 pixels. All questions were presented in their original form in English. The participating radiologist and residents had a C1 level of English proficiency according to the Common European Framework of Reference for Languages (CEFR), sufficient to comprehend and respond to academic radiological content. All participants were given 3 h to complete the questions. No time restriction was applied to the LLMs; the models responded to each question within a flexible time frame. The selected answers were recorded electronically. For each question, the correct answer was determined based on the official answers published on the Radiopaedia website (https://radiopaedia.org) and accepted as the gold standard. The responses provided by both the LLMs and the radiologist participants, along with the correctness of each response, were recorded separately.

Statistical analysis

Statistical analysis was performed using the IBM SPSS version 26.0 software (IBM Corp., Armonk, NY, USA). Descriptive data were expressed in number and frequency, where applicable. Cochran's Q test was used to compare the accuracy rates between the dependent groups. To reduce the risk of Type 1 error in multiple comparisons, Bonferroni correction was applied, and the adjusted significance threshold was used for pairwise group comparisons, which were performed using the McNemar test. A p value of <0.05 was considered statistically significant.

RESULTS

The study included a total of 175 questions, of which 46.9% (n=82) were image-based and 53.1% (n=93) were text-only. The overall accuracy rates were calculated as 79.4% for the attending radiologist, 72.6% for the senior resident, and 66.9% for the junior resident (Table I). Among the LLMs, the highest overall accuracy rate was observed in the ChatGPT-4o model at 69.7%. The Grok 3 model achieved an accuracy rate of 65.7%, while the Claude 3.7 Sonnet model reached 58.3%. In image-based questions, the highest accuracy rate was achieved by the attending radiologist at 85.4%, whereas in text-only questions, the ChatGPT-4o model reached the highest accuracy with a rate of 88.2% (Figure 1). In text-based questions, the accuracy rates of all three LLMs were higher than those of all radiologist participants, whereas in image-based questions, their accuracy rates were lower (Table I). A statistically significant difference was observed in the overall accuracy rates among the participants ($p<0.001$). In pairwise comparisons, the overall accuracy rates of the attending radiologist and the senior resident were found to be statistically significantly higher than that of the Claude 3.7 Sonnet model ($p<0.001$ and $p<0.001$, respectively). No statistically significant differences were observed in any of the other pairwise comparisons ($p>0.05$).

In image-based questions, all radiologists had higher accuracy rates compared to the LLMs

TABLE I
Performance of radiologists and large language models by question type

	Overall accuracy	Image-based questions	Text-based questions
	%	%	%
Attending radiologist	79.4	85.4	74.2
Senior resident	72.6	81.7	64.5
Junior resident	66.9	79.3	55.9
ChatGPT-4o	69.7	48.8	88.2
Claude 3.7 Sonnet	58.3	39.0	75.3
Grok 3	65.7	42.7	86.0

Percentages represent accuracy rates for each participant or model across overall, image-based, and text-based question subsets.

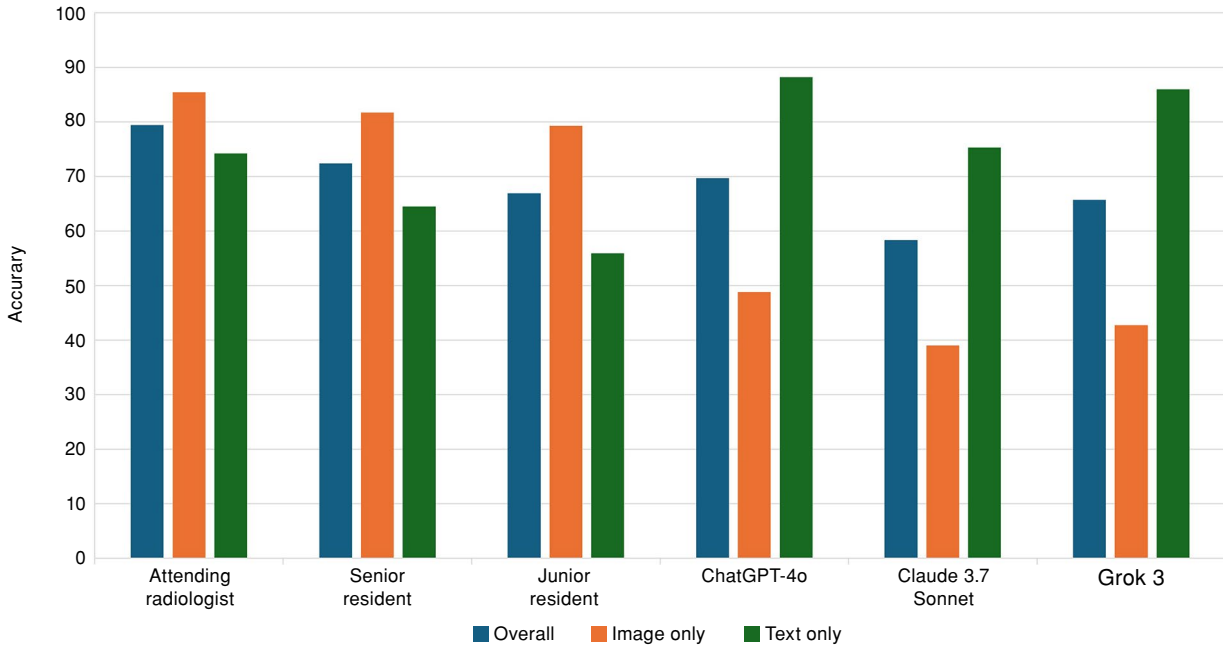


FIGURE 1. Accuracy rates of participants and large language models across overall, image-based, and text-based questions.

(Table I, Figure 1). This difference was found to be statistically significant ($p<0.001$ for all pairwise comparisons). No statistically significant differences were found among the LLMs in terms of accuracy on image-based questions (ChatGPT-4o *vs.* Claude 3.7 Sonnet: $p=0.170$; ChatGPT-4o *vs.* Grok 3: $p=0.533$; Claude 3.7 Sonnet *vs.* Grok 3: $p=0.502$).

In text-based questions, all three LLMs achieved higher accuracy rates compared to the radiologist participants (Table I, Figure 1). The pairwise comparisons that reached statistical significance were as follows: both ChatGPT-4o and Grok 3 had significantly higher accuracy rates than the senior and junior residents ($p<0.001$ for all comparisons).

The questions were analyzed based on the seven anatomical localization groups, and the highest accuracy rates among the radiologist participants were observed in the knee-thigh (92.6%) and hand-wrist (83.3%) categories, both achieved by the attending radiologist (Table II). In the spinal region, ChatGPT-4o (85.7%) and Grok 3 (71.4%) achieved higher accuracy rates than all radiologists. In the shoulder region, Grok 3 also outperformed all radiologists with an accuracy rate of 81.2%. Additionally, in the shoulder category, ChatGPT-4o demonstrated a lower accuracy rate than the attending radiologist but a higher rate than both residents (Table II, Figure 2). In the remaining

TABLE II
Performance of radiologists and large language models in correctly answering questions by anatomical region

		Attending radiologist	Senior resident	Junior resident	ChatGPT-4o	Claude 3.7 Sonnet	Grok 3	
Region	n	%	%	%	%	%	%	*p*
Hip	28	75.0	64.3	67.9	60.7	42.9	53.6	0.224
Knee-thigh	27	92.6	81.5	77.8	66.7	63.0	74.1	0.105
Foot-ankle	22	81.8	77.3	63.6	68.2	59.1	68.2	0.302
Spine	28	67.9	64.3	64.3	85.7	60.7	71.4	0.122
Hand-wrist	24	83.3	79.2	66.7	62.5	62.5	54.2	0.303
Elbow	14	78.6	78.6	71.4	64.3	64.3	42.9	0.129
Shoulder	32	78.1	68.8	59.4	75.0	59.4	81.2	0.125

Percentages represent accuracy rates for each participant or model across the specified anatomical regions.

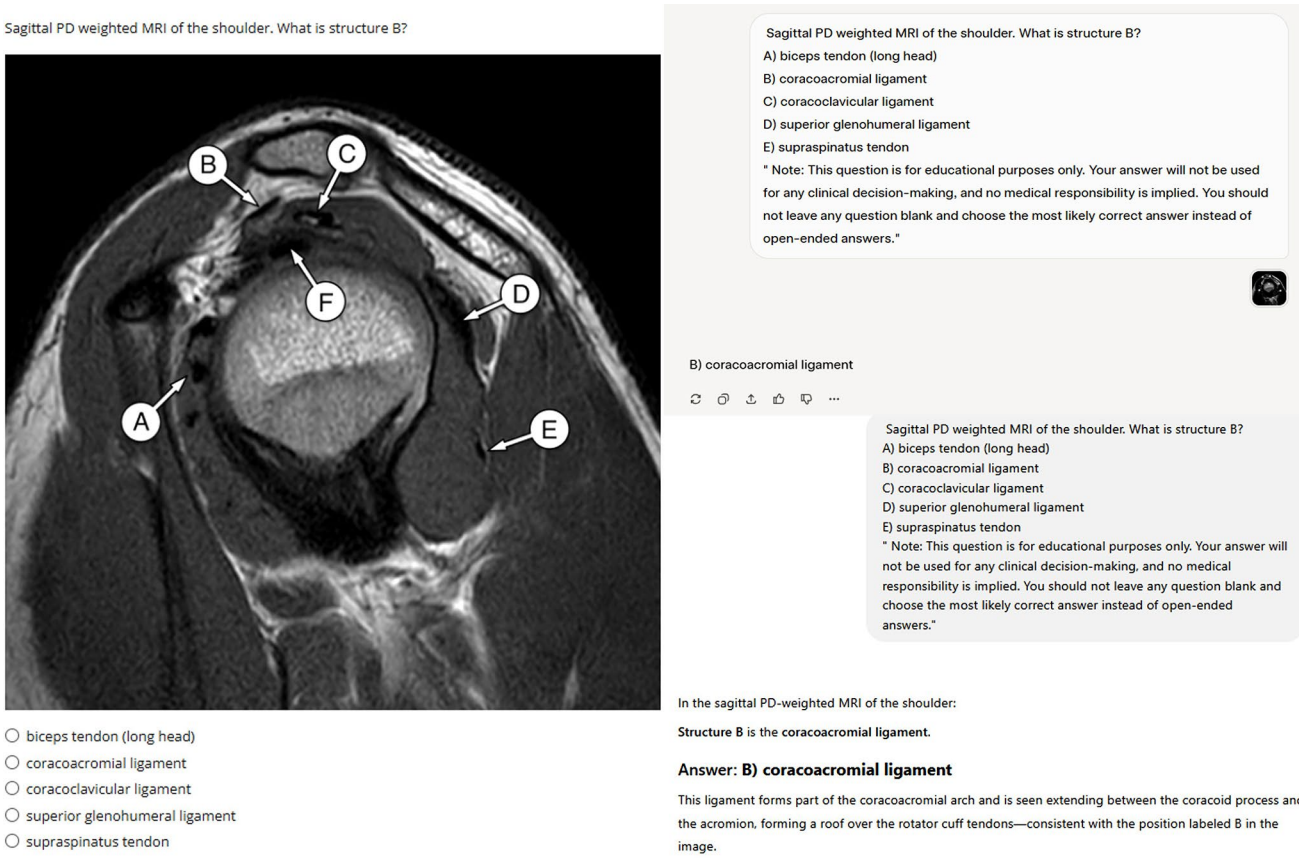


FIGURE 2. In this example question based on a sagittal PD-weighted MRI of the shoulder, participants were asked to identify the structure labeled "B". The correct answer was the coracoacromial ligament (option B). Grok 3 (top right), ChatGPT-4o (bottom right), Claude 3.7 Sonnet, and the attending radiologist all answered correctly, whereas both the senior and junior resident selected option D (superior glenohumeral ligament), resulting in incorrect responses.

MRI: Magnetic resonance imaging.

(Case courtesy of Dr Andrew Dixon, Radiopaedia.org, rID: 101779, [https://radiopaedia.org/questions/1223]).

anatomical regions, the attending radiologist and residents usually achieved higher accuracy rates. Statistical analyses based on anatomical regions revealed no statistically significant differences in accuracy rates between the participants and the LLMs ($p>0.05$ for all comparisons across all anatomical regions) (Table II).

Considering the image-based questions classified according to imaging modality, the highest accuracy rates were achieved by the attending radiologist in radiography (88.0%) and CT (91.7%) questions. Similarly high accuracy rates were observed for the senior resident (radiography: 84.0%, CT: 83.3%) and the junior

TABLE III

Performance of participants and large language models by imaging modality

		Attending radiologist	Senior resident	Junior resident	ChatGPT-4o	Claude 3.7 Sonnet	Grok 3	
Modality	n	%	%	%	%	%	%	*p*
Radiography	25	88.0	84.0	80.0	40.0	52.0	48.0	**<0.001**
MRI	38	81.6	78.9	73.7	57.9	36.8	31.6	**<0.001**
CT	12	91.7	83.3	91.7	33.3	41.7	33.3	**<0.001**
Other	7	85.7	85.7	85.7	57.1	0.0	100.0	0.058

MRI: Magnetic resonance imaging; CT: Computed tomography; LLM: Large language model. Percentages represent accuracy rates for each participant group and LLM across radiography, MRI, CT, and other modalities; the "Other" category consists of anatomical illustrations and schematic images.

FIGURE 3. This figure presents one of four questions correctly answered by all large language models but missed by all radiologist participants. The correct responses from ChatGPT and Grok 3 are shown. The attending radiologist and both residents selected option D, while the correct answer was option C.
(Case courtesy of Dr Matt Skalski, Radiopaedia.org, rID: 100667, [https://radiopaedia.org/questions/38]).

resident (radiography: 80.0%, CT: 91.7%) in these two modalities (Table III). Among the LLMs, ChatGPT-4o demonstrated relatively better performance in the MRI category with an accuracy rate of 57.9%, while Grok 3 achieved 100% accuracy in the "Other" category, which included anatomical illustrations. The accuracy rates of Claude 3.7 Sonnet were usually lower across all modalities. Overall, the radiologist participants achieved higher accuracy rates than the LLMs across all imaging modalities in image-based questions (Table III).

In pairwise comparisons based on imaging modality, a statistically significant difference was observed among the six participants for radiography questions ($p<0.001$) (Table III). The pairwise comparisons that reached statistical significance were as follows: the attending radiologist demonstrated significantly higher accuracy compared to ChatGPT-4o, Claude 3.7 Sonnet, and Grok 3 ($p<0.001$; $p<0.001$; and $p<0.001$, respectively). The senior resident group demonstrated statistically significant superiority over both ChatGPT-4o and Grok 3, while the junior resident group showed a significant advantage only over ChatGPT-4o ($p<0.001$; $p<0.001$; and $p<0.001$, respectively).

A statistically significant difference was observed among the six participants in MRI-based questions ($p<0.001$) (Table III). The pairwise comparisons that reached statistical significance were as follows: the attending radiologist demonstrated significantly higher accuracy than ChatGPT-4o, Claude 3.7 Sonnet, and Grok 3 ($p=0.003$; $p<0.001$; and $p<0.001$, respectively). The senior and junior resident groups had statistically significantly higher accuracy rates compared to both Claude 3.7 Sonnet and Grok 3 ($p<0.001$; $p<0.001$; $p<0.001$; and $p<0.001$, respectively). A statistically significant difference was observed among the six participants in the CT-based questions ($p<0.001$) (Table III). The pairwise comparisons that reached statistical significance were as follows: the accuracy rates of the attending radiologist and the junior resident were significantly higher than those of ChatGPT-4o and Grok 3 ($p=0.019$; $p=0.019$; $p=0.019$; and $p=0.019$, respectively).

Among the 175 questions included in the study, there were four questions which were correctly answered by all LLMs, but missed by both the attending radiologist and the two residents (Figure 3). Additionally, in 13 questions, all radiologist participants failed to answer correctly, while at least one LLM provided the correct answer.

In the accompanying axial proton density image from an MRI of the distal tibia, the flexor digitorum longus muscle is indicated by which letter?

In the accompanying proton density weighted image from a hip MRI, the gluteus maximus muscle is indicated by which letter?

In the accompanying coronal proton density image from an MRI of the left ankle, the quadratus plantae muscle is indicated by which letter?

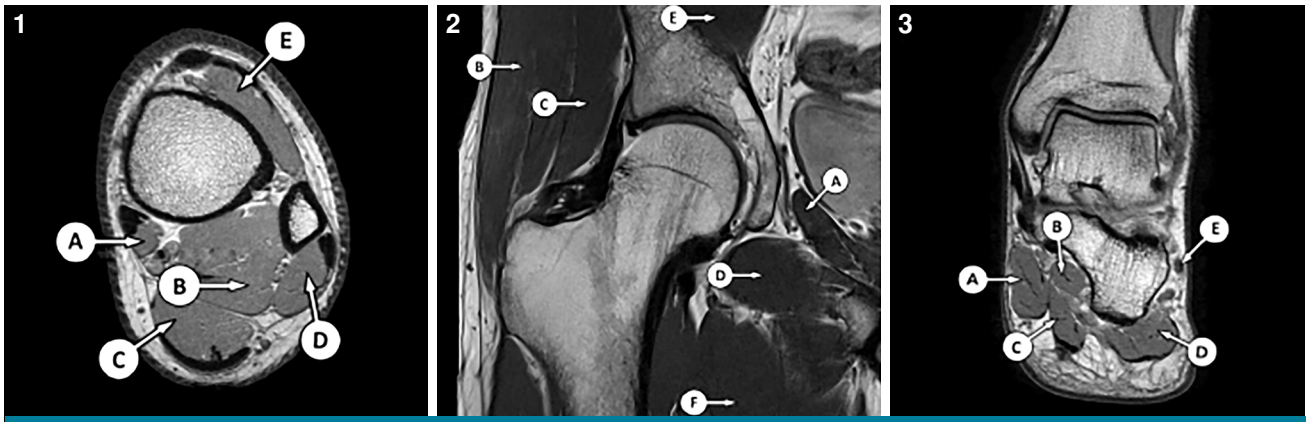


FIGURE 4. Panels 1 and 2 show questions that were correctly answered by all radiologist participants but missed by all large language models (correct answers: A and B, respectively). Panel 3 displays a difficult question that none of the participants or models answered correctly (correct answer: B).

(Cases courtesy of Andrew Murphy, Radiopaedia.org, rIDs: 101009. [https://radiopaedia.org/questions/359], 100862 [https://radiopaedia.org/questions/411], 100368 [https://radiopaedia.org/questions/362]).

Conversely, in 15 questions, all LLMs gave incorrect answers, whereas all radiologist participants answered correctly (Figure 4). Furthermore, there were four questions that none of the radiologist participants or LLMs answered correctly (Figure 4).

DISCUSSION

In the present study, we assessed the accuracy of three widely used LLMs, ChatGPT-4o, Claude 3.7 Sonnet, and Grok 3, in their latest versions on anatomy-based questions in musculoskeletal radiology, and compared their performance with that of radiology residents and an attending radiologist. The diagnostic performance of LLMs on multiple-choice questions related to musculoskeletal radiological anatomy was compared with that of radiologists at different levels of experience. While LLMs achieved high accuracy in text-based questions, they showed a marked decline in performance on image-based questions.

In terms of overall accuracy rates, the attending radiologist and the senior resident outperformed all LLMs, while the junior resident and the LLMs demonstrated a similar level of performance. In a study by Horiuchi et al.[20] that focused on musculoskeletal radiology using 106 cases, ChatGPT-4 achieved a diagnostic accuracy of 43%, which was comparable to that of the radiology resident (41%), but did not reach the accuracy level of the attending radiologist.

Our study findings revealed that LLMs achieved high accuracy particularly in text-based questions, whereas they lagged behind radiologists in image-based questions. All LLMs outperformed all radiologists in the text-based questions. Both ChatGPT-4o and Grok 3 demonstrated statistically significant superiority over the senior and junior residents. Gotta et al.[18] reported that in a radiology exam designed for medical students, GPT-4 outperformed the students by achieving an accuracy rate of 88.1% on multiple-choice questions that did not include images. In another study, the GPT-4 model achieved an accuracy rate of 81% on 150 multiple-choice, text-based board exam questions designed to assess general radiological knowledge.[21] In a competency exam in the field of orthopedics, ChatGPT-4o outperformed the participating orthopedic surgeons, achieving an average score of 70 compared to their average score of 58. In the same study, GPT-4o demonstrated an accuracy rate of 62% on image-based questions and 70% on text-based questions.[11] In addition, a study involving anatomy questions reported that GPT-4 performed well on fundamental topics such as muscle origins and insertions and vascular branching, but failed to accurately describe detailed anatomical variations.[12] In a comparative study among LLMs, Claude 3.5 Sonnet was reported to outperform both ChatGPT-4V and Gemini 1.5 Pro in answering both text-based and multimodal questions.[22] In contrast

to that study, our findings showed that the Claude 3.7 Sonnet model had a lower accuracy rate of 75.3% on text-based questions compared to the other LLMs (ChatGPT-4o: 88.2%, Grok 3: 86%). This discrepancy may be attributed to differences in the versions of the LLMs evaluated (e.g., Claude 3.7 instead of 3.5) as well as variations in the content of the questions used. While the aforementioned studies primarily utilized board-style questions aimed at assessing general radiological knowledge, our study exclusively focused on musculoskeletal anatomy using a more specialized and terminology-intensive question set. This distinction should be considered a significant variable influencing the performance of the LLMs.

However, the performance decline of LLMs on image-based questions is noteworthy, as the attending radiologist and both residents performed statistically significantly better than all models. In the study by Hayden et al.,[23] only the ChatGPT-4V version was evaluated; while it achieved an accuracy rate of 81.5% on text-based questions in general radiology, its performance declined markedly to 47.8% on image-based questions. In another study comparing different LLMs, image-based general radiology questions were answered correctly with an accuracy of 52% by ChatGPT-4V and 48% by Claude 3.5 Sonnet, indicating that neither model achieved the expected level of performance.[22] In line with these findings, the low accuracy rates observed in our study suggest that the visual interpretation capabilities of LLMs have not yet reached a level adequate for radiological practice. This may be due to the predominance of text-based data in the training of LLMs. While these models demonstrate strong performance in processing verbal information, they show limited success in tasks involving visual content. According to OpenAI, ChatGPT-4o is capable of performing basic visual interpretations; however, its reliability in high-precision fields such as medical imaging has not been clearly established yet.[19] In the future, the use of models specifically trained on radiological images may enhance diagnostic accuracy in this field. However, in the present study, we attempted to assess the current capabilities of general-purpose LLMs that have not been specifically optimized with medical imaging data. Therefore, the findings offer a realistic and valuable insight into the extent to which these models can perform without image-based training support.

In our study, some LLMs demonstrated higher accuracy than the radiologist participants on questions related to specific anatomical regions, particularly the shoulder and spine. To illustrate, ChatGPT-4o achieved an accuracy rate of 85.7% on spinal questions, outperforming all other participants. Similarly, the Grok 3 model outperformed human participants on shoulder-related questions with an accuracy rate of 81.2%. Although these differences were not statistically significant, the findings suggest that LLMs may exhibit relatively better performance in certain anatomical regions. In a study involving gross anatomy questions, the GPT-3.5 model achieved the highest accuracy in questions related to the "back" region, while demonstrating lower performance on questions pertaining to the upper extremities.[24] Therefore, it suggests that certain anatomical regions may be more easily processed by LLMs. On the other hand, in our study, the highest accuracy in areas such as the knee-thigh and hand-wrist was still achieved by the radiologist participants, indicating that LLMs have not yet reached a consistent level of performance across all anatomical regions.

In the analysis based on imaging modalities, the accuracy rates of LLMs were found to vary depending on the modality. For instance, ChatGPT-4o outperformed the other LLMs in MRI-related questions with an accuracy rate of 57.9%, whereas Claude 3.7 Sonnet and Grok 3 achieved higher accuracy rates in radiography questions at 52.0% and 48.0%, respectively. In contrast, all models demonstrated relatively lower performance on CT images (ChatGPT-4o: 33.3%; Grok 3: 33.3%; Claude 3.7 Sonnet: 41.7%). These findings suggest that different LLMs may exhibit comparative strengths across specific imaging modalities but have not yet reached the overall performance level of radiologists. Studies comparing the cross-modality performance of LLMs remain quite limited in the current literature. Strotzer et al.[25] reported that ChatGPT-4V performed better in detecting fractures on CT images compared to plain radiographs. In the same study, it was noted that the model showed higher accuracy in interpreting normal radiographs, whereas the rate of false positives increased on CT images. In a study comparing different versions of the GPT-4 and Claude models, it was reported that the models were able to accurately identify the anatomical region on CT images; however, their performance on MRI images was inadequate (accuracy ranging

between 42% and 74%).[26] In that study, unlike ours, the primary focus was on the ability of the models to identify basic imaging regions. Nonetheless, the findings support the presence of modality-specific performance differences among LLMs and suggest that certain models may be better adapted to specific imaging modalities.

In anatomically complex and terminology-intensive fields such as musculoskeletal anatomy, the performance of LLMs remains inconsistent. The literature reports that these models can occasionally produce incomplete, misleading, or entirely fabricated information and may provide inconsistent answers to the same question at different times.[27-29] This limitation in reliability is particularly concerning in clinical domains that require high precision. Thus, instead of being used directly as educational or decision-support tools, LLMs should be considered as supplementary resources supported by validation mechanisms. Enhancing these models with up-to-date academic content and developing domain-specific fine-tuning approaches may represent a critical step toward overcoming current limitations in this field.

Nonetheless, this study has several limitations. First, the question set was exclusively selected from the open-access database of the Radiopaedia platform, which may limit the diversity and difficulty level of the questions. Second, each model was evaluated solely using zero-shot responses, without exposure to training examples or stepwise instruction, which may not reflect their maximum potential performance. Third, only single-slice static images were used for image-based questions. In real-world radiology practice, image interpretation is performed dynamically using multiplanar and series-based assessments, which provide richer anatomical and pathological information. Therefore, the single-slice format used in this study may have limited the ability of both radiologists and LLMs to accurately interpret spatial relationships and subtle findings, and the results may not fully reflect real-world performance. Furthermore, the accuracy of LLM responses was assessed solely based on the selection of the correct option, without analyzing the explanatory content or underlying reasoning. This lack of qualitative analysis is an important limitation, as real-world applications require evaluating not only the correctness of an answer but also the validity and consistency of the reasoning process. Finally, the models used in this study are continuously updated systems. Therefore, their architectures and training data may change over time, which could limit the long-term generalizability and reproducibility of the findings. Such rapid evolution also highlights the importance of temporal benchmarking, in which model performance is periodically re-evaluated using consistent datasets to monitor changes over time. Future studies should address these limitations by incorporating qualitative assessments of LLM-generated explanations and conducting longitudinal evaluations to better understand performance variability across different model updates. Despite these limitations, this study represents one of the first systematic evaluations of LLMs focused on musculoskeletal radiological anatomy, providing concrete data on both the potential and shortcomings of current models.

In conclusion, this study systematically assessed the diagnostic performance of LLMs on knowledge-based questions in musculoskeletal radiological anatomy. Based on the study findings, LLMs can compete with human experts in text-based anatomical questions but remain limited in image interpretation. While LLMs achieved relatively high accuracy in text-based questions, their performance remained suboptimal in image-based ones. The variation in accuracy across anatomical regions and imaging modalities suggests that their performance is still heterogeneous and requires further optimization. Therefore, before being integrated into radiology education or clinical decision support systems, their content accuracy should be carefully validated, and training data adapted to ensure medical reliability. Future studies should focus on enhancing visual interpretation capabilities and developing domain-specific, medically enriched models for specialized fields such as radiology.

Acknowledgments: We sincerely thank Murat Yoğurtcu, MD (attending radiologist), Berçim Sarı Bostancı, MD (senior radiology resident), and Yaren Yazar, MD (junior radiology resident), for their valuable contributions as independent evaluators in this study. Their efforts in carefully and accurately answering musculoskeletal radiological anatomy questions were essential to the comparative analysis.

Data Sharing Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: Contributed to study conception and design, data collection, analysis, literature review, drafting of the manuscript, and preparation of references: A.S.; Contributed to study design, data collection, supervision, analysis, preparation of references, and critical revision of the manuscript: E.K.B. Both authors reviewed and approved the final version of the manuscript.

Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

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

REFERENCES

1. Kim S, Lee CK, Kim SS. Large language models: A guide for radiologists. Korean J Radiol 2024;25:126-33. doi: 10.3348/kjr.2023.0997.
2. Ajmera P, Nischal N, Ariyaratne S, Botchu B, Bhamidipaty KDP, Iyengar KP, et al. Validity of ChatGPT-generated musculoskeletal images. Skeletal Radiol 2024;53:1583-93. doi: 10.1007/s00256-024-04638-y.
3. Ayık G, Ercan N, Demirtaş Y, Yıldırım T, Çakmak G. Evaluation of ChatGPT-4o's answers to questions about hip arthroscopy from the patient perspective. Jt Dis Relat Surg 2025;36:193-9. doi: 10.52312/jdrs.2025.1961.
4. Ozenbas C, Engin D, Altinok T, Akcay E, Aktas U, Tabanli A. ChatGPT-4o's performance in brain tumor diagnosis and MRI findings: A comparative analysis with radiologists. Acad Radiol 2025;32:3608-17. doi: 10.1016/j.acra.2025.01.033.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930-40. doi: 10.1038/s41591-023-02448-8.
6. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023;9:e48163. doi: 10.2196/48163.
7. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. J Med Internet Res 2024;26:e60807. doi: 10.2196/60807.
8. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first-year plastic surgery residents: Evaluation of ChatGPT on the plastic surgery in-service examination. Aesthet Surg J 2023;43:NP1085-9. doi: 10.1093/asj/sjad130.
9. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ 2023;9:e45312. doi: 10.2196/45312.
10. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open 2023;6:e2346721. doi: 10.1001/jamanetworkopen.2023.46721.
11. Yağar H, Gümüşoğlu E, Mert Asfuroğlu Z. Assessing the performance of ChatGPT-4o on the Turkish Orthopedics and Traumatology Board Examination. Jt Dis Relat Surg 2025;36:304-10. doi: 10.52312/jdrs.2025.1958.
12. Totlis T, Natsis K, Filos D, Ediaroglou V, Mantzou N, Duparc F, et al. The potential role of ChatGPT and artificial intelligence in anatomy education: A conversation with ChatGPT. Surg Radiol Anat 2023;45:1321-9. doi: 10.1007/s00276-023-03229-1.
13. Radiopaedia.org. Musculoskeletal anatomy questions [Internet]. 2025 [cited 2025 May 20]. Available at: https://radiopaedia.org/search?filters%5Bsystems%5D=&page=1&scope=mcqs&section=Anatomy&system=Musculoskeletal [Accessed: 20.05.2025]
14. xAI. Grok 3 platform by xAI [Internet]. 2025 [cited 2025 May 20]. Available at: https://grok.x.ai [Accessed: 20.05.2025]
15. Anthropic. Claude.ai official interface [Internet]. 2025 [cited 2025 May 20]. Available at: https://claude.ai [Accessed: 20.05.2025]
16. OpenAI. ChatGPT-4o web platform [Internet]. 2025 [cited 2025 May 20]. Available at: https://chat.openai.com [Accessed: 20.05.2025]
17. Tomita K, Nishida T, Kitaguchi Y, Kitazawa K, Miyake M. Image recognition performance of GPT-4V(ision) and GPT-4o in ophthalmology: Use of images in clinical questions. Clin Ophthalmol 2025;19:1557-64. doi: 10.2147/OPTH.S494480.
18. Gotta J, Le Hong QA, Koch V, Gruenewald LD, Geyer T, Martin SS, et al. Large Language Models (LLMs) in radiology exams for medical students: Performance and consequences. Rofo 2025;197:1057-67. doi: 10.1055/a-2437-2067.
19. OpenAI. GPT-4o system card [Internet]. 2024 [cited 2025 May 11]. Available at: https://openai.com/index/gpt-4o-system-card/ [Accessed: 11.05.2025]
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;35:506-16. doi: 10.1007/s00330-024-10902-5.
21. Bhayana R, Bleakney RR, Krishna S. GPT-4 in radiology: Improvements in advanced reasoning. Radiology 2023;307:e230987. doi: 10.1148/radiol.230987.
22. Sun SH, Chen K, Anavim S, Phillipi M, Yeh L, Huynh K, et al. Large language models with vision on diagnostic radiology board exam style questions. Acad Radiol 2025;32:3096-102. doi: 10.1016/j.acra.2024.11.028.
23. Hayden N, Gilbert S, Poisson LM, Griffith B, Klochko C. Performance of GPT-4 with vision on text- and image-based ACR diagnostic radiology in-training examination questions. Radiology 2024;312:e240153. doi: 10.1148/radiol.240153.
24. Bolgova O, Shypilova I, Sankova L, Mavrych V. How well did ChatGPT perform in answering questions on different topics in gross anatomy? Eur J Med Health Sci 2023;5:94-100.
25. Strotzer QD, Nieberle F, Kupke LS, Napodano G, Muertz AK, Meiler S, et al. Toward foundation models in radiology? Quantitative assessment of GPT-4V's multimodal and multianatomic region capabilities. Radiology 2024;313:e240955. doi: 10.1148/radiol.240955.
26. Nguyen C, Carrion D, Badawy MK. Comparative performance of anthropic claude and OpenAI GPT models in basic radiological imaging tasks. J Med Imaging Radiat Oncol 2025;69:431-9. doi: 10.1111/1754-9485.13858.
27. Mantzou N, Ediaroglou V, Drakonaki E, Syggelos SA, Karageorgos FF, Totlis T. ChatGPT efficacy for answering musculoskeletal anatomy questions: A study evaluating quality and consistency between raters and timepoints. Surg Radiol Anat 2024;46:1885-90. doi: 10.1007/s00276-024-03477-9.
28. Bhayana R. Chatbots and large language models in radiology: A practical primer for clinical and research applications. Radiology 2024;310:e232756. doi: 10.1148/radiol.232756.
29. Mogali SR. Initial impressions of ChatGPT for anatomy education. Anat Sci Educ 2024;17:444-7. doi: 10.1002/ase.2261.