ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/dentalsci.2025-111283

Comparison of Grok and ChatGPT in Temporomandibular Joint Magnetic Resonance Images Interpretation: Sectional Study

Grok ve Chatgpt'nin Temporomandibular Eklem Manyetik Rezonans Görüntülerinin Yorumlanmasındaki Karşılaştırılması: Kesitsel Araştırma

Melisa ÖÇBE[a], Mahmut Sabri MEDİŞOĞLU[b]

[a]Kocaeli Health and Technology University Faculty of Dentistry, Department of Oral and Maxillofacial Radiology, Kocaeli, Türkiye
[b]Kocaeli Health and Technology University Faculty of Dentistry, Department of Basic Medical Sciences, Department of Anatomy, Kocaeli, Türkiye

ABSTRACT Objective: With the rapid development of artificial intelligence, large language models (LLMs) have emerged as powerful tools capable of processing and interpreting complex medical data. These models can assist in the interpretation of imaging data, especially in conditions that require detailed anatomical analysis such as temporomandibular joint disorders. This study evaluated and compared the performance of 2 LLMs, Chat Generative Pre-trained Transformer-4 Omni (ChatGPT-4o) and Grok, in diagnosing temporomandibular joint disc displacement using magnetic resonance imaging (MRI). **Material and Methods:** A total of 129 sagittal MRI, including T1- and T2-weighted sequences, were retrospectively analyzed. The images were annotated to identify the disc and mandibular condyle, with diagnoses confirmed by oral and maxillofacial radiology experts. Both models were tasked with identifying anatomical structures and assessing the disc-condyle relationship. **Results:** Among the analyzed images, 65 showed disc displacement and 64 did not. ChatGPT-4o achieved an overall diagnostic accuracy of 67.4%, with a perfect sensitivity of 100% but lower specificity and precision. In contrast, Grok demonstrated an accuracy of 49.7% (p<0.005), but outperformed ChatGPT-4o in specificity (76.9%), precision (61.5%), and F1-score (58.1%). While ChatGPT-4o showed superior performance in identifying all pathological cases, Grok exhibited greater balance in reducing false positives. **Conclusion:** This study highlights the potential of LLMs as supplementary tools in oral and maxillofacial radiology while emphasizing the need for further advancements to improve their diagnostic capabilities.

Keywords: Temporomandibular joint;
magnetic resonance imaging;
artifical intelligence; large language models

ÖZET Amaç: Yapay zekânın hızlı gelişimiyle birlikte büyük dil modelleri [large language models (LLMs)], karmaşık tıbbi verileri işleyip yorumlayabilen güçlü araçlar olarak öne çıkmıştır. Bu modeller, özellikle temporomandibular eklem bozuklukları gibi detaylı anatomik analiz gerektiren durumlarda görüntülerin yorumlanmasına katkı sağlayabilirler. Bu çalışmada, temporomandibular eklem disk deplasmanının manyetik rezonans görüntüleme (MRG) ile tanılanmasında 2 LLMs'nin Chat Generative Pre-trained Transformer-4 Omni (ChatGPT-4o) ve Grok performansı değerlendirilmiş ve karşılaştırılmıştır. **Gereç ve Yöntemler:** Toplam 129 sagital MRG (T1 ve T2 ağırlıklı sekanslar dâhil) retrospektif olarak analiz edilmiştir. Görüntülerde disk ve mandibular kondil belirlenmiş, tanılar ağız, diş ve çene radyolojisi uzmanları tarafından doğrulanmıştır. Her iki modele de anatomik yapıları tanımlama ve disk-kondil ilişkisini değerlendirme görevi verilmiştir. **Bulgular:** Analiz edilen görüntüler arasında 65'inde disk dislokasyonu mevcutken, 64'ünde bulunmamaktaydı. ChatGPT-4o genel tanı doğruluğu açısından %67,4 başarı sağladı; duyarlılığı %100 ile mükemmel olmasına rağmen özgüllük ve kesinlik (precision) oranları daha düşüktü. Buna karşılık, Grok %49,7 doğruluk oranı elde etti (p<0,005); ancak özgüllük (%76,9), kesinlik (%61,5) ve F1 skoru (%58,1) açısından ChatGPT-4o'yu geride bıraktı. ChatGPT-4o tüm patolojik vakaları saptamada üstün bir performans sergilerken, Grok yanlış pozitifleri azaltmada daha dengeli bir performans gösterdi. **Sonuç:** Bu çalışma, LLMs'nin ağız, diş ve çene radyolojisinde yardımcı araçlar olarak potansiyelini ortaya koymakta ve tanısal yeteneklerinin geliştirilmesi gerekliliğine dikkat çekmektedir.

Anahtar Kelimeler: Temporomandibular eklem;
manyetik rezonans görüntüleme;
yapay zekâ; büyük dil modelleri

Correspondence: Melisa ÖÇBE
Kocaeli Health and Technology University Faculty of Dentistry, Department of Oral and Maxillofacial Radiology, Kocaeli, Türkiye
E-mail: melisa.ocbe@kocaelisaglik.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Dental Sciences.

Received: 14 Apr 2025 ***Received in revised form:*** 04 Jul 2025 ***Accepted:*** 09 Jul 2025 ***Available online***: 04 Sep 2025

In recent years, the advent of advanced artificial intelligence (AI) and deep learning technologies has led to the development of large language models (LLMs) which can generate natural language, mimicking human-like text, generating images and interpretation of medical images.[1-3] LLMs have shown promise in providing rapid and accessible medical knowledge and in interpreting medical images with notable accuracy.[3,4] Chat-Generative Pre-Trained Transformer (ChatGPT, OpenAI, San Francisco, CA) is one such model that has demonstrated high accuracy in the evaluation of radiographic images.[2,4]

Recently, on November 4, 2023, Elon Musk's xAI (San Francisco, CA) introduced Grok, a novel AI model that distinguishes itself from other LLMs by synthesizing responses with "real-time knowledge" from the X platform (formerly Twitter). This capability enables users to access up-to-date information, enhancing its potential as a "powerful research assistant". Grok has been described as a tool that aids in quickly accessing relevant information, analyzing data, generating ideas, and even suggesting questions to explore further. Built on the Grok-1 architecture, which was trained using Grok-0 with 33 billion parameters, the model has reportedly outperformed ChatGPT 3.5 but falls slightly short of GPT-4 in overall capability.[5,6]

Recent advancements in AI have given rise to 2 distinct paradigms for medical image analysis: traditional image-based models and emerging LLMs.[7] Conventional models, such as convolutional neural networks (CNNs) and hybrid architectures like U-Net or ResNet, are trained directly on pixel data to identify spatial and morphological patterns within medical images.[8,9] These models typically require substantial labeled datasets, domain-specific tuning, and often operate as "black boxes", offering limited explainability. In contrast, LLMs like Chat Generative Pre-trained Transformer-4 Omni (ChatGPT-4o) and Grok function primarily through text-based reasoning but are increasingly being adapted to multimodal tasks, including visual input interpretation. Unlike CNNs, LLMs can integrate contextual textual data (e.g., prompts, anatomical landmarks, clinical questions) with visual information, enabling a more human-like diagnostic approach.[8,9] While not optimized for direct pixel-level classification, LLMs offer the potential for more generalizable and intuitive reasoning, especially in settings where explicit prompts and annotations guide their attention toward relevant features. This emerging capability positions LLMs as complementary tools to traditional AI approaches, particularly in diagnostic triage and education, where interpretability and natural language interaction are valued.[9,10]

Understanding the clinical significance of temporomandibular joint (TMJ) imaging is critical to appreciating the novelty of this study. TMJ disorders are complex conditions involving both hard and soft tissue structures, and magnetic resonance imaging (MRI) remains the gold standard for evaluating the articular disc and joint capsule.[11] MRI offers excellent soft tissue contrast without ionizing radiation, making it particularly suitable for visualizing disc displacement, joint effusion, and inflammatory changes.[12] T1-weighted sequences provide high-resolution anatomical detail, while T2-weighted sequences are more sensitive to fluid accumulation and inflammation. In TMJ assessment, sagittal oblique MRI planes are commonly used to evaluate the disc-condyle relationship in both open- and closed-mouth positions.[13] These images enable the identification of anterior disc displacement with or without reduction-key criteria for diagnosing internal derangements of the joint. Including LLMs in this diagnostic workflow introduces a novel perspective: rather than replacing radiologists, these models may serve as assistive tools that interpret annotated MRI data and provide natural-language support in cases requiring further review or explanation. Framing the study within this dual context underscores its contribution to both AI innovation and clinical imaging applications.

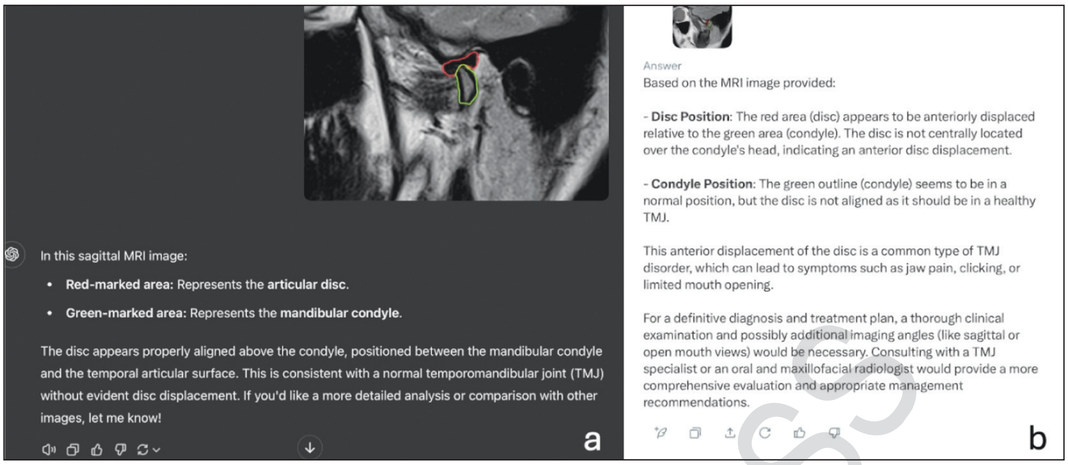
Despite the growing interest in Grok, no studies have yet evaluated or compared its performance with other LLMs in the field of oral and maxillofacial radiology. To address this gap, the present study aims to explore the potential of Grok and ChatGPT-4o in diagnosing TMJ disc displacement, providing the first comparative analysis of these models.

MATERIAL AND METHODS

This retrospective study was granted ethical approval from the Kocaeli Health and Technology University, Non-Interventional Clinical Ethics Committee (date: March 10, 2025; no: 2025/125). The study was conducted in full accordance with the ethical principles outlined in the Declaration of Helsinki, including its later amendments, and all participants provided written informed consent prior to inclusion. MRI archived between 2022-March 2024 Department of Radiology in the Kayseri City Hospital were retrospectively evaluated for this study. Only the images of patients who provided informed consent were included in the analysis.

A total of 129 sagittal MRI were examined, utilizing fast spin echo (FSE) T1- and T2-weighted sequences. All MRI were obtained from patients referred for suspected TMJ dysfunction, including complaints of joint pain, limited mandibular movement, or joint sounds. Images were obtained in either closed-mouth or open-mouth positions, and both positions were included in the dataset. The distribution of positions was recorded and considered in the results. No cases were excluded based solely on mouth position. Exclusion criteria included scans with severe motion artifacts, significant image distortion, poor contrast resolution, poor signal-to-noise ratio that impaired the ability to delineate anatomical structures, incomplete TMJ coverage that precluded reliable disc or condyle identification, patients with a documented history of TMJ surgery, neoplastic or cystic pathologies (e.g., TMJ tumors, synovial cysts), inflammatory arthritides affecting the joint, or any systemic disease known to alter joint morphology. However, minor anatomical variations, borderline cases, or mild artifacts were not excluded to preserve the clinical variability inherent in real-world data. Imaging was performed using the 3 Tesla SIGNA™ Pioneer system (GE Medical Systems, Waukesha, USA). The imaging protocol included sagittal oblique FSE sequences in both T1-weighted and T2-weighted acquisitions. For T1-weighted FSE images, the parameters were as follows: repetition time (TR)=500-800 ms, echo time (TE)=10-20 ms, slice thickness=3 mm, interslice gap=0.3 mm, field of view (FOV)=12-15 cm, matrix=256×192, and number of excitations (NEX)=2. For T2-weighted FSE sequences, parameters included: TR=3,000-5,000 ms, TE=80-100 ms, slice thickness=3 mm, interslice gap=0.3 mm, FOV=12-15 cm, matrix=256×192, and NEX=2. Imaging was performed in both closed-mouth and open-mouth positions using a bite block to standardize mouth opening during acquisition. All images were acquired in the sagittal oblique plane perpendicular to the long axis of the mandibular condyle.

The sagittal sections were annotated, with the disc marked in red and the mandibular condyle marked in green, to show the regions of interest. The annotation of disc and condyle and presence of disc displacement were independently confirmed by four oral and maxillofacial radiology experts an anatomy expert from three separate dentistry faculties. Disc displacement was classified into types based on the standard diagnostic criteria from the research diagnostic criteria for temporomandibular disorder, specifically distinguishing between anterior disc displacement with reduction (wADD) and without reduction (woADD).[13] Both models, ChatGPT-4o and Grok, were tasked with identifying the anatomical structures (red and green markings) and evaluating the disc-condyle relationship using the identical prompt: "Identify the green and red areas shown in this sagittal MR image and evaluate the disc-condyle relation to diagnose the disc displacement". Initially, both AI models were tasked with identifying the anatomical structures corresponding to the red and green annotations. To determine whether the models correctly identified the anatomical structures (disc and condyle), their textual responses were compared to expert-defined reference terms. An identification was considered successful only if the model explicitly used or correctly described the anatomical names "disc" and "condyle" in the appropriate context, without ambiguity or misclassification. For example, if the model identified the red marking as "cartilage" or a "joint capsule" instead of "disc", this was recorded as incorrect. Both models successfully identified the red-marked region as the disc and the green-marked region as the condyle (Figure 1). After anatomical recognition, the models' assessments of the disc-condyle relationship were evaluated. An answer was



In this sagittal MRI image:

- **Red-marked area:** Represents the **articular disc.**
- **Green-marked area:** Represents the **mandibular condyle.**

The disc appears properly aligned above the condyle, positioned between the mandibular condyle and the temporal articular surface. This is consistent with a normal temporomandibular joint (TMJ) without evident disc displacement. If you'd like a more detailed analysis or comparison with other images, let me know!

Based on the MRI image provided:

- **Disc Position**: The red area (disc) appears to be anteriorly displaced relative to the green area (condyle). The disc is not centrally located over the condyle's head, indicating an anterior disc displacement.

- **Condyle Position**: The green outline (condyle) seems to be in a normal position, but the disc is not aligned as it should be in a healthy TMJ.

This anterior displacement of the disc is a common type of TMJ disorder, which can lead to symptoms such as jaw pain, clicking, or limited mouth opening.

For a definitive diagnosis and treatment plan, a thorough clinical examination and possibly additional imaging angles (like sagittal or open mouth views) would be necessary. Consulting with a TMJ specialist or an oral and maxillofacial radiologist would provide a more comprehensive evaluation and appropriate management recommendations.

FIGURE 1: (a) ChatGPT-4 accurately identifying the absence of disc displacement. **(b)** Grok accurately identifying the presence of disc displacement. ChatGPT: Chat-Generative Pre-Trained Transformer

labeled "correct" only if the model's final interpretation of the disc-condyle relationship (e.g., indicating the presence or absence of disc displacement) matched the consensus diagnosis established by the expert panel. The images were analyzed by ChatGPT 4.0 and Grok to evaluate the disc-condyle relationship, and their responses were recorded as "correct" or "incorrect" according to the consensus of oral and maxillofacial radiology and anatomy specialists for comparison. A response was considered "incorrect" under the following conditions:

- If the model misidentified either the disc or condyle,
- If it failed to respond (refused or outputted "I don't know"),
- If it made an incorrect diagnosis of disc displacement type (e.g., stated "no displacement" when there was clear displacement),
- If the anatomical recognition was correct but the disc-condyle relationship was misinterpreted.
- An answer was considered "correct" only if ChatGPT and Grok provided the same answer with the consensus of specialists. The score was considered "incorrect" if ChatGPT or Grok refused to answer the question or provided the incorrect answer.

All diagnostic judgments were based on the expert panel's gold-standard classification, and the comparison was recorded in a spreadsheet as correct or incorrect for statistical analysis.

In addition to evaluating overall diagnostic accuracy across the full dataset, a qualitative error analysis was conducted. Each model's output for these cases was categorized into:

- True positive: Correct identification of disc displacement,
- False positive: Incorrect identification of disc displacement when absent,
- False negative: Failure to detect existing disc displacement,
- True negative: Correct identification of absence of displacement.

Based on these classifications, diagnostic performance metrics were calculated for each model, including accuracy, sensitivity (recall), specificity, precision (positive predictive value), and F1 score.

Statistical analysis was performed to compare the diagnostic accuracy of ChatGPT-4o and Grok in identifying the disc-condyle relationship in MRI. A chi-square test for independence was conducted to assess whether the proportion of correct diagnoses differed significantly between the 2 AI models. The observed frequencies of correct and incorrect diagnoses for each model were arranged in a 2x2 contingency table, and the chi-square statistic was computed. Diagnostic performance metrics were computed using Python (v3.11.6) and the scikit-learn library. A significance level of 0.05 was used to determine statistical significance. The test was performed using SciPy's chi-square function in Python, and the resulting p value was interpreted to assess the null hypothesis, which assumed no significant difference between the models.

RESULTS

A total of 129 sagittal section MRI (68 female and 61 male) were included in this study. Mean age of the patients were found to be 45.19 (SD 13.02). Among these, 65 images demonstrated wADD, and 64 images were woADD. Of the total images, 79 were acquired in the closed-mouth position, while 50 were acquired in the open-mouth position (Table 1).

The performance of ChatGPT-4 and Grok in identifying disc-condyle relationships in MRI was evaluated and compared. Both models were tested using the same prompt and were asked to identify the anatomical structures (red marking as the disc and green marking as the condyle) and subsequently assess the disc-condyle relationship. Their correct answer rates were recorded and analyzed.

The chi-square test revealed a statistically significant difference between ChatGPT-4o and Grok in their accuracy for diagnosing the disc-condyle relationship (p=0.0054). ChatGPT-4o correctly diagnosed 87 cases (67.4%), while Grok correctly identified 64 cases (49.7%). Given that the p value was below the standard threshold of 0.05, it was indicated that ChatGPT-4o's diagnostic performance was significantly better than Grok's (Table 2).

TABLE 1: Distribution of MRI based on disc displacement and mouth position characteristics

TMJ status	n	%
wADD	65	50.4
woADD	64	49.6
Closed mouth	79	61.2
Open mouth	50	38.8

MRI: Magnetic resonance imaging; TMJ: Temporomandibular joint; wADD: Anterior disc displacement with reduction; woADD: Anterior disc displacement without reduction

TABLE 2: Comparison of correct diagnosis rates for disc-condyle relation by ChatGPT-4o and Grok

	Correct diagnosis of disc-condyle relation		
Model	n	%	p value
ChatGPT-4o	87	67.4	0.0054*
Grok	64	49.7	

*Chi-square test; ChatGPT-4o: Chat Generative Pre-trained Transformer-4 Omni

TABLE 3: Confusion matrix classification of model responses

Model	True positive	False positive	False negative	True negative
ChatGPT-4o	65	22	0	42
Grok	62	3	3	61

ChatGPT-4o: Chat Generative Pre-trained Transformer-4 Omni

TABLE 4: Diagnostic performance metrics

Model	Accuracy	Sensitivity (recall)	Specificity	Precision (PPV)	F1 Score
ChatGPT-4o	83.7%	100.0%	65.6%	74.7%	85.5%
Grok	95.3%	95.4%	95.3%	95.4%	95.4%

ChatGPT-4o: Chat Generative Pre-trained Transformer-4 Omni; PPV:

To evaluate diagnostic robustness, a comprehensive qualitative error analysis was conducted on the full set of 129 annotated sagittal MRI. The classification of model outputs is presented in Table 3 and Table 4.

Findings demonstrated differing diagnostic strategies between the models. ChatGPT-4o achieved perfect sensitivity, correctly identifying all 65 cases of disc displacement, but at the expense of a higher false positive rate, reducing specificity and precision. In contrast, Grok exhibited higher overall accuracy and specificity, with fewer false positives, though it missed 3 cases of confirmed displacement. This trade-off between sensitivity and specificity highlights the models' complementary strengths in clinical decision support (Table 4).

DISCUSSION

The findings of this study provide novel insights into the potential applications of LLMs in the field of oral and maxillofacial radiology, specifically in evaluating TMJ disc displacement. While previous studies have examined the capabilities of LLMs in general medical applications, this study is the first to compare ChatGPT-4o and Grok in the interpretation of MR images for TMJ diagnosis.[4-8]

In this study, ChatGPT-4o demonstrated a higher correct diagnosis rate for disc-condyle relationships (67.4%) compared to Grok (49.7%). ChatGPT-4o achieved perfect sensitivity (100%), correctly identifying all 65 cases of disc displacement, but this came at the expense of lower specificity (65.6%) due to a higher rate of false positives (n=22). In contrast, Grok exhibited a more balanced performance, with high accuracy (95.3%), specificity (95.3%), and precision (95.4%), though it missed 3 confirmed cases of displacement, resulting in a slightly lower sensitivity (95.4%). ChatGPT-4o is more sensitive but less specific, favoring inclusivity and minimizing false negatives which is found to be a trait that may be valuable in screening scenarios. Conversely, Grok is more conservative and accurate, reducing false positives. The superior sensitivity of ChatGPT-4o may stem from its larger model architecture and advanced contextual reasoning capabilities, while Grok's real-time knowledge integration appeared less relevant in this static imaging task. These findings align with previous studies. For instance, Şahin et al. found that while GPT-4 exhibited more complex language structures, Grok was easier to comprehend but less accurate.[14] Similarly, Shiraishi et al. reported significantly higher accuracy rates for GPT-4 compared to Grok in answering clinical questions regarding implant-based breast reconstruction.[15]

The current study also highlights the limitations of both models in achieving a level of accuracy comparable to human experts in interpreting medical images. Although both models successfully identified the anatomical structures marked in the images (disc and condyle), their performance in assessing the disc-condyle relationship underscores the need for further advancements in AI technology to reliably interpret complex radiological findings. Qualitative error analysis revealed that both models struggled with cases where the disc-condyle relationship was obscured by imaging artifacts or atypical anatomical variations. Grok's performance, in particular, was hindered by over-reliance on positional data, leading to misclassification of borderline cases. These findings suggest that incorporating enhanced contextual understanding and training on a more diverse dataset could significantly improve diagnostic accuracy. ChatGPT-4o's superior performance can be attributed to its larger parameter size and enhanced language model architecture, which likely allowed for better pattern recognition in complex anatomical structures. In contrast, Grok's real-time knowledge synthesis appeared less relevant in this diagnostic context, which relies heavily on static image interpretation rather than dynamic updates. To date, no studies have specifically evaluated LLMs in MRI interpretation for radiology.[16-18] This gap in the literature emphasizes the novelty and importance of this research. While the use of LLMs in medical imaging remains experimental and not yet suitable for patient care, these models hold significant promise as supplementary tools for clinicians, especially as AI technologies continue to evolve and improve.

The advent of advanced LLMs, such as OpenAI's GPT-4o and xAI's Grok, signifies an important step in telemedicine and virtual healthcare. In the context of this study on MR image interpretation for TMJ disorders, GPT-4o's capacity to integrate natural language processing with visual data analysis showcases its potential as a supplementary diagnostic tool. While innovations such as OpenAI's "Temporary Chat" feature aim to minimize long-term data retention by discarding conversation history, current implementations still raise unresolved concerns regarding privacy and transparency. For example, OpenAI retains interactions for up to 30 days for operational monitoring, and the specific scope of data access and storage is not always clearly disclosed. As noted in prior literature, the use of LLMs in healthcare should proceed with caution until such tools are validated within HIPAA-compliant or equivalent data protection frameworks.[12] Thus, while LLMs offer promising functionality in telemedicine and clinical decision support, their deployment must be accompanied by robust data governance and regulatory oversight to ensure patient confidentiality.[17] However, while these advancements are promising, the integration of LLMs with telemedicine systems necessitates validation and a collaborative approach that combines AI with human expertise. Also, it is obvious that LLMs are not yet ready to interpret medical images.[18,19]

Several limitations should be considered for this study. First, to evaluate the generalizability of the models, an external validation dataset comprising MRI from a different institution was planned. However, due to data-sharing restrictions, this study was limited to a single-center dataset. Future studies will aim to include external datasets to validate the findings across diverse populations and imaging protocols. In addition to that, using a single prompt to request both anatomical identification and diagnostic evaluation may have introduced bias by conflating the 2 tasks. Future studies should consider separating these instructions to independently assess model performance in each domain. Moreover, the use of pre-annotated images may have simplified the diagnostic task by directing the models' attention to key anatomical regions. In real-world applications, AI models are expected to interpret unannotated images independently, which could affect diagnostic accuracy. Future studies should evaluate model performance under more realistic clinical conditions. Finally, future research should focus on optimizing LLMs for medical imaging applications and evaluate their integration into clinical workflows. Advancements in AI training algorithms, increased dataset diversity, and enhanced contextual understanding could improve diagnostic accuracy and expand the potential applications of LLMs in healthcare.

CONCLUSION

This study presents the first comparative analysis of ChatGPT-4o and Grok in interpreting TMJ MRI for the diagnosis of disc displacement. While ChatGPT-4o achieved perfect sensitivity, and detected all positive disc displacement cases, Grok demonstrated superior performance in specificity, precision, and F1-score. This suggests that Grok was more effective at minimizing false positives and delivering a balanced prediction. Rather than a single model outperforming the other universally, each model exhibits distinct strengths. Both models exhibited limitations in achieving diagnostic accuracy comparable to that of human experts, particularly in complex or borderline cases. These findings are important to assess the role of LLMs as supplementary tools in oral and maxillofacial radiology rather than standalone diagnostic systems.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

***Idea/Concept:** Melisa Öçbe; **Design:** Melisa Öçbe; **Control/Supervision:** Mahmut Sabri Medişoğlu; **Data Collection and/or Processing:** Melisa Öçbe; **Analysis and/or Interpretation:** Mahmut Sabri Medişoğlu; **Literature Review:** Mahmut Sabri Medişoğlu; **Writing the Article:** Melisa Öçbe, Mahmut Sabri Medişoğlu; **Critical Review:** Melisa Öçbe; **References and Fundings:** Mahmut Sabri Medişoğlu.*

REFERENCES

1. Mira FA, Favier V, Dos Santos Sobreira Nunes H, de Castro JV, Carsuzaa F, Meccariello G, et al. Chat GPT for the management of obstructive sleep apnea: do we have a polar star? Eur Arch Otorhinolaryngol. 2024;281(4):2087-93. PMID: 37980605.
2. Mago J, Sharma M. The potential usefulness of chatgpt in oral and maxillofacial radiology. Cureus. 2023;15(7):e42133. PMID: 37476297; PMCID: PMC10355343.
3. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. 2024;66(1):73-9. PMID: 37994939.
4. Grewal H, Dhillon G, Monga V, Sharma P, Buddhavarapu VS, Sidhu G, et al. Radiology Gets Chatty: the ChatGPT Saga Unfolds. Cureus. 2023;15(6):e40135. PMID: 37425598; PMCID: PMC10329466.
5. Gupta R, Park JB, Ragsdale LB, Meggers K, Eimani A, Mailey BA. The intersection of AI Grok with aesthetic plastic surgery. Aesthet Surg J. 2024;44(6):NP437-40. PMID: 38484177.
6. xAI [Internet]. [Cited: December 14, 2024]. Introducing Grok: the conversational AI connected to X (formerly Twitter). Available from: https://x.ai/blog/grok?utm_source=chatgpt.com
7. Hasan Z, Key S, Habib AR, Aweidah L, Kumar A, Wong E, et al. Convolutional neural networks in ENT radiology: systematic review of the literature. Ann Otol Rhinol Laryngol. 2023;132(4):417-30. doi:10.1177/00034894221095899
8. Ushinsky A, Bardis M, Glavis-Bloom J, Uchio E, Chantaduly C, Nguyentat M, et al. A 3D-2D hybrid U-Net convolutional neural network approach to prostate organ segmentation of multiparametric MRI. AJR Am J Roentgenol. 2021;216(1):111-6. PMID: 32812797.
9. Cai L, Li Q, Zhang J, Zhang Z, Yang R, Zhang L. Ultrasound image segmentation based on Transformer and U-Net with joint loss. PeerJ Comput Sci. 2023;9:e1638. PMID: 38077559; PMCID: PMC10703067.
10. Al Mohamad F, Donle L, Dorfner F, Romanescu L, Drechsler K, Wattjes MP, et al. Open-source large language models can generate labels from radiology reports for training convolutional neural networks. Acad Radiol. 2025;32(5):2402-10. PMID: 39765434.
11. Al-Saleh MA, Alsufyani NA, Saltaji H, Jaremko JL, Major PW. MRI and CBCT image registration of temporomandibular joint: a systematic review. J Otolaryngol Head Neck Surg. 2016;45(1):30. PMID: 27164975; PMCID: PMC4863319.
12. Xiong X, Ye Z, Tang H, Wei Y, Nie L, Wei X, et al. MRI of temporomandibular joint disorders: recent advances and future directions. J Magn Reson Imaging. 2021;54(4):1039-52. PMID: 32869470.
13. Warzocha J, Gadomska-Krasny J, Mrowiec J. Etiologic factors of temporomandibular disorders: a systematic review of literature containing diagnostic criteria for temporomandibular disorders (DC/TMD) and research diagnostic criteria for temporomandibular disorders (RDC/TMD) from 2018 to 2022. Healthcare (Basel). 2024;12(5):575. PMID: 38470686; PMCID: PMC10931313.
14. Şahin MF, Topkaç EC, Doğan Ç, Şeramet S, Özcan R, Akgül M, et al. Still using only ChatGPT? The comparison of five different artificial intelligence Chatbots' answers to the most common questions about kidney stones. J Endourol. 2024;38(11):1172-7. PMID: 39212674.
15. Shiraishi M, Sowa Y, Tomita K, Terao Y, Satake T, Muto M, et al. Performance of artificial intelligence Chatbots in answering clinical questions on japanese practical guidelines for implant-based breast reconstruction. Aesthetic Plast Surg. 2025;49(7):1947-53. PMID: 39592492.
16. Iyengar KP, Yousef MMA, Nune A, Sharma GK, Botchu R. Perception of Chat Generative Pre-trained Transformer (Chat-GPT) AI tool amongst MSK clinicians. J Clin Orthop Trauma. 2023;44:102253. PMID: 37822477; PMCID: PMC10562840.
17. Vaishya R, Iyengar KP, Patralekh MK, Botchu R, Shirodkar K, Jain VK, et al. Effectiveness of AI-powered Chatbots in responding to orthopaedic postgraduate exam questions-an observational study. Int Orthop. 2024;48(8):1963-9. PMID: 38619565.
18. Mitsuyama Y, Tatekawa H, Takita H, Sasaki F, Tashiro A, Oue S, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. Eur Radiol. 2025;35(4):1938-47. PMID: 39198333; PMCID: PMC11913992.
19. Temsah MH, Jamal A, Alhasan K, Aljamaan F, Altamimi I, Malki KH, et al. Transforming virtual healthcare: the potentials of ChatGPT-4omni in telemedicine. cureus. 2024;16(5):e61377. PMID: 38817799; PMCID: PMC11139454.