TLAR 1947

Arch Rheumatol 2025;40(1):63-71
doi: 10.46497/ArchRheumatol.2025.10966

ORIGINAL ARTICLE

Artificial intelligence-generated responses to frequently asked questions on coccydynia: Evaluating the accuracy and consistency of GPT-4o's performance

Aslinur Keles, Ozge Gulsum Illeez, Berkay Erbagci, Esra Giray

Department of Physical Medicine and Rehabilitation, Health Science University, Fatih Sultan Mehmet Training and Research Hospital, İstanbul, Türkiye

Correspondence: Aslinur Keles, MD.

E-mail: aslinur.ercisli@gmail.com

Received: October 02, 2024
Accepted: January 05, 2025
Published online: March 17, 2025

Citation: Keles A, Illeez OG, Erbagci B, Giray E. Artificial intelligence-generated responses to frequently asked questions on coccydynia: Evaluating the accuracy and consistency of GPT-4o's performance. Arch Rheumatol 2025;40(1):63-71. doi: 10.46497/ArchRheumatol.2025.10966.

ABSTRACT

Objectives: This study aimed to assess whether GPT-4o's responses to patient-centered frequently asked questions about coccydynia are accurate and consistent when asked at different times and from different accounts.

Materials and methods: Questions were collected from medical websites, forums, and patient support groups and posed to GPT-4o. The responses were evaluated by two physiatrists for accuracy and consistency. Responses were categorized: *(i)* correct and comprehensive, *(ii)* correct but not inadequate, *(iii)* partially correct and partially incorrect, and *(iv)* completely incorrect. Inconsistencies in scoring were resolved by an additional reviewer as needed. Statistical analysis, including Cohen's kappa for interreviewer reliability, was performed.

Results: Of the 81 responses, 45.7% were rated as correct and comprehensive, while 49.4% were correct but incomplete. Only 4.9% of the responses contained partially incorrect information, and no responses were completely incorrect. The interreviewer agreement was substantial (kappa=0.67), but 75% of the responses differed between the two rounds. Notably, 34.9% of initially incomplete answers improved in the second round.

Conclusion: GPT-4o shows promise in providing accurate and generally reliable information about coccydynia. However, the variability observed in response consistency across repeated queries suggests that while the model is useful for patient education and general inquiries, it may not be suitable for providing specialized clinical knowledge without human oversight.

Keywords: Artificial intelligence, ChatGPT-4o, coccydynia, coccyx pain, large language models.

Coccydynia, characterized by pain in the coccyx and the surrounding tissues, can arise from multiple causes, including trauma, nontraumatic injuries, or idiopathic origins. Its diagnosis and management are often challenging due to its complex etiology and the subjective nature of the pain.[1] Accurate data on the incidence and prevalence of coccydynia are limited; however, studies show that it was the cause of pain in 2.7% of patients hospitalized for back pain and responsible for over 14,000 emergency department visits in the USA in 2014, highlighting its significant clinical burden.[2] Diagnosis of coccydynia involves a detailed evaluation, including patient history, physical examination, imaging modalities, and sometimes diagnostic injections. This diagnostic approach is crucial to identify the underlying cause and guide treatment, and given the anatomical complexity of the coccygeal region, making the correct diagnosis is crucial for effective treatment.[3] However, some physicians may assume that “it will heal on its own” or trivialize the condition and leave their patients alone with statements such as “it is only your tailbone; we do not even use it” or “it is all in your head.”[4] This attitude can further exacerbate coccydynia, discouraging patients from seeking medical help and delaying appropriate treatment. Given the precise localization of the pain and the reluctance of some patients to have a physical examination of the coccyx, many people may prefer to look for initial information about their symptoms on the internet before consulting a healthcare professional.

In the medical field, large language models (LLMs) have shown significant potential in supporting clinical diagnosis and medical education.[5-7] Among these technologies, ChatGPT, a generative pretrained transducer (GPT), has become an important resource for both patients and medical professionals seeking health-related information. A recent systemic review has also highlighted that ChatGPT can become a reliable training tool if it achieves a correct response rate of over 95%, but it is unknown whether future versions will reach this level of maturity, as the training dataset was not developed with a specific focus on medical education.[8] In contrast, unlike traditional search engines that provide various search results, LLM-driven artificial intelligence (AI) chatbots offer more organized information by providing direct, practical answers to specific questions. This makes them particularly effective in providing medical information.[9]

In May 2024, GPT-4 Omni (GPT-4o), a text-based model, was introduced with enhanced capabilities, including faster response times, reduced operational costs, and improved performance in processing both English and non-English texts.[10] The 'o' in GPT-4o stands for 'omni,' representing a step toward more natural human-computer interaction. These advancements significantly expand GPT-4o's potential applications in the medical field, particularly for tasks requiring real-time data processing, multilingual communication, and personalized health management.[11] One of GPT-4o's notable advantages is its 24/7 availability, providing patients with educational support and guidance during off-hours, such as nights and weekends. This feature reduces patient anxiety, minimizes wait times, and aids in the early detection of complications, potentially improving patient satisfaction and optimizing healthcare resource utilization.[12] Additionally, GPT-4o's multilingual support enhances global patient education by eliminating language barriers and delivering high-quality medical information that is accessible to nonnative speakers. This capability fosters cross-cultural communication and engagement, which is particularly relevant in diverse healthcare settings where cultural backgrounds influence patients' understanding of diseases, treatments, and care.[12] Patients may seek information about coccydynia from GPT-4o, asking various questions regarding its causes, symptoms, and potential treatments, as they are often reluctant to consult a professional. However, the reliability of these models in providing consistent and accurate information remains a topic of interest.

Recognizing this growing trend of patients seeking AI-driven guidance, this study aimed to evaluate the accuracy and consistency of GPT-4o in providing information related to coccydynia. This study aimed to determine whether the responses provided by GPT-4o remained consistent and clinically relevant when the same set of questions was asked at different times and from various accounts. The results of this evaluation could have significant implications for how AI tools such as ChatGPT assist patients in navigating health-related concerns, particularly in conditions such as coccydynia, where accurate information and guidance are crucial for informed decision-making.

MATERIALS AND METHODS

Compilation of questions/data source

Frequently asked questions about coccydynia were derived and reformulated from publicly accessible patient-oriented content on multiple reliable sources, including medical websites, forums, and patient support groups. These sources included Pelvic Rehabilitation Medicine, Verywell Health, QI Spine Clinic, Mayo Clinic, Cleveland Clinic, Physiopedia, and Medscape. No proprietary material was reproduced, and all questions were designed to represent common queries that patients with coccydynia might ask. The purpose of this initial collection was to encompass a wide variety of information that patients commonly seek regarding coccydynia. Questions were curated, reviewed, and approved by the study authors for inclusion. Duplicate or highly similar questions from multiple sources were removed (Figure 1). In total, 81 unique questions were finalized and used to generate responses from GPT-4o. The complete list of questions can be found in Supplementary File 1. To assess the performance of AI across different aspects of

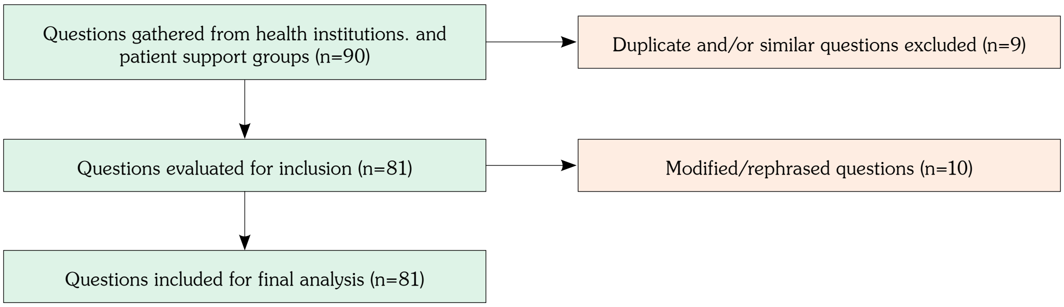
Questions gathered from health institutions. and patient support groups (n=90)
Duplicate and/or similar questions excluded (n=9)
Questions evaluated for inclusion (n=81)
Modified/rephrased questions (n=10)
Questions included for final analysis (n=81)

Figure 1. Flowchart of coccydynia-related question selection.

coccydynia, the questions were organized into several categories for analysis: *(i)* overview and general information (n=27), *(ii)* diagnosis, differential diagnosis, presentation, and workup (n=23), *(iii)* treatment options and medications (n=12), and (*iv*) quality of life (QoL) considerations (n=19). The questions were phrased in conversational, first-person language to mimic how a typical patient might input their queries into the ChatGPT interface. Ethical approval was not required because the GPT-4o used in this study is publicly accessible, and no patient data were collected or involved. The study was conducted in accordance with the principles of the Declaration of Helsinki.

AI response generation

This study generated responses to the curated questions using OpenAI's GPT-4o (OpenAI Inc., San Francisco, CA, USA; released on May 13, 2024). Models from version 4.0 are not freely available; however, with the introduction of GPT-4o, users can access a limited number of queries for free before the system defaults back to version 3.5. This allows all users to test GPT-4o's functionality to a limited extent.[13] The GPT-4o model was accessed through OpenAI's public platform (https://platform.openai.com/docs/models/gpt-4). No proprietary APIs (application programming interfaces) or datasets were utilized, and all interactions adhered to OpenAI's terms of service, accessible at https://openai.com/policies/terms-of-use.

The first round of questions was conducted using a newly created account, which had no prior interactions with the model to ensure unbiased responses (access date: July 10, 2024). A new conversation was initiated for each question by selecting the "new chat" function, ensuring independence between interactions. Each session was conducted over a sufficient duration to allow GPT-4o to generate comprehensive and complete responses without interruptions. This process was repeated for every question, ensuring that all questions were addressed individually.

Following the same procedure, the same set of questions was inputted into GPT-4o using another newly created account one week later (access date: July 17, 2024). This account had no prior usage and was specifically set up for this study to eliminate any potential influence from previous interactions. This method allowed for comparing responses across different accounts and time points to evaluate consistency and reproducibility, as adapted from prior ChatGPT literature.[14]

Response evaluation

Two physiatrists independently assessed the final responses generated by GPT-4o for accuracy and consistency. The reviewers were blinded to each other's assessments. The AI-generated responses were then compared for similarity; if they were similar, only the first response was graded for accuracy. This step was crucial to determining the consistency of the AI's answers across multiple iterations. If the responses differed, both were graded individually.

The accuracy of responses was graded based on a scoring system adapted from prior ChatGPT literature:[15,16] 1=comprehensive, with no inaccurate information; a reviewer would likely have nothing significant to add; 2=correct but inadequate; while there is no

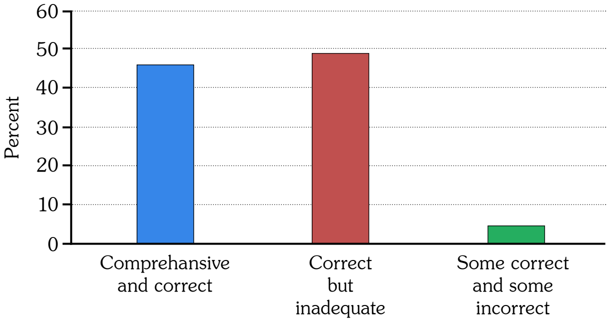

Figure 2. Distribution of the percentage of questions per score.

inaccurate information, a reviewer would almost certainly have something to add; 3=some are correct, some incorrect, contain some misinformation; 4=completely incorrect. In cases of disagreement regarding consistency or accuracy, a third reviewer, a physiatrist, was consulted to resolve discrepancies. The final evaluations were then compiled to analyze ChatGPT-4o's overall performance in answering questions regarding coccydynia.

Statistical analysis

Statistical analyses were conducted using IBM SPSS version 26.0 software (IBM Corp., Armonk, NY, USA). Descriptive statistics were utilized to calculate the proportion of responses for each score, both in aggregate and within specific question categories. To assess the performance of ChatGPT in answering questions about coccydynia, scores of 1 and 2 were labeled as "correct" and scores of 3 and 4 were labeled as "incorrect." The analysis also included the calculation of the percentages of questions that exhibited nonidentical responses when input into ChatGPT on two separate occasions. Additionally, the proportion of questions requiring additional reviewers' involvement to resolve scoring discrepancies was determined. Interrater reliability was assessed using Kappa analysis. Cohen's kappa values were interpreted as follows: 0.0 to 0.40 indicated poor agreement, 0.41 to 0.60 indicated moderate agreement, 0.61 to 0.80 indicated substantial agreement, and 0.81 to 0.99 indicated near-perfect agreement.[17] A significance level of $p<0.05$ was used to determine statistical significance.

RESULTS

Of the 81 questions, the responses were categorized as follows: 37 (45.7%) questions received correct and comprehensive answers (score of 1), 40 (49.4%) questions received correct but not comprehensive answers (score of 2), four (4.9%) questions received partially incorrect answers (score of 3), and none received completely incorrect answers (score of 4; Figure 2). The breakdown of scores across the question categories is demonstrated in Figure 3. The percentage of responses rated

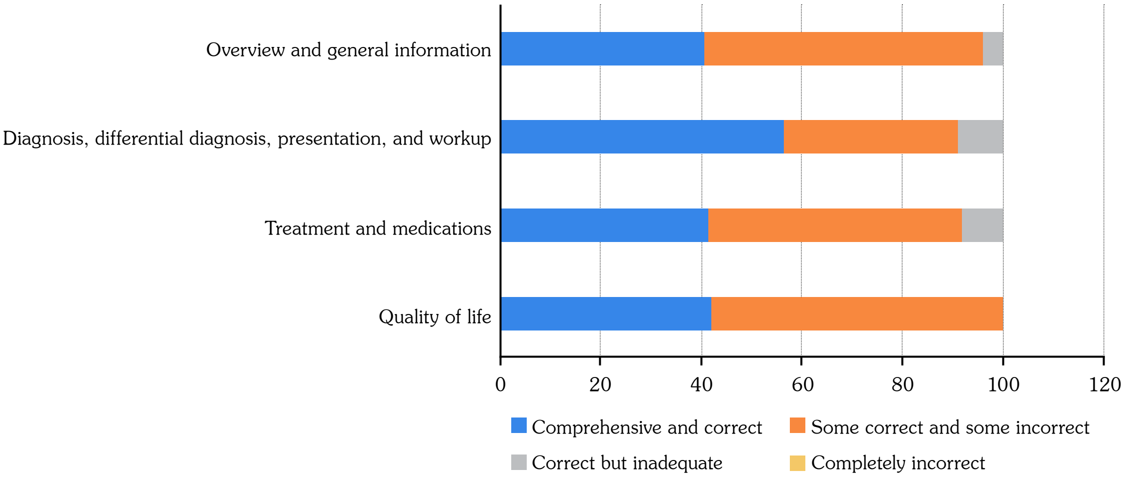

Figure 3. Distribution of scores across question categories.

Since there are four evaluation categories, the fourth category also needs to be indicated in yellow. However, it was not included in the figure as it was not present in the evaluation results.

Table 1. Agreement between two reviewers across subcategories

	Kappa	95% CI	*p*
Overview and general information	0.64	0.36-0.91	<0.0001
Diagnosis, differential diagnosis, presentation, and work-up	0.76	0.50-1.01	<0.0001
Treatment and medications	0.66	0.22-1.09	0.02
Quality of Life	0.58	0.22-0.93	0.009

CI: Confidence interval.

Table 2. The proportions of grades assigned by reviewer 3 for response 1 and response 2.

	Reviewer 3's grade for response 1		Reviewer 3's grade for response 2	
	n	%	n	%
Comprehensive and correct	37	45.7	37	45.7
Correct, but inadequate	40	49.4	43	53.1
Some correct, some incorrect	4	4.9	1	1.2
Completely incorrect	0	0	0	0

as correct versus incorrect was 96.3% and 3.7% for overview and general information, 91.3% and 8.7% for diagnosis, differential diagnosis, presentation, and work-up, 91.7% and 8.3% for treatment and medications, and 100% and 0% for QoL.

The proportion of questions that required additional reviewers to resolve scoring discrepancies was 14%. The kappa value for agreement between the first and second reviewers across all questions was 0.67 (95% confidence interval 0.51-0.82, p<0.001), indicating a substantial level of agreement. Moderate to substantial agreement between reviewers was observed in most response subcategories (Table 1).

When the same question was entered into ChatGPT a second time, 75% of the responses were inconsistent between the first and second interactions. Table 2 presents the grading of responses from both the first and second queries, as evaluated by reviewer 3. Notably, the percentage of responses graded as 3 in the second query decreased from 4.9% to 1.2%. Furthermore, using the "Select Cases" feature in SPSS, it was found that 34.9% of the questions that initially received scores of 3 or 2 in the first query were upgraded to a grade of 1 after the second query, demonstrating a clear improvement in GPT's performance.

DISCUSSION

Advanced LLMs, some of them also equipped with image analysis capabilities, have extended their use to the diagnosis and interpretation of medical images such as skin conditions or X-rays.[5] This study focuses on the text-based capabilities of GPT-4o and evaluates its performance in delivering accurate and consistent responses to patient inquiries, specifically related to coccydynia. The results demonstrate that GPT-4o is generally effective in delivering correct information, suggesting its potential to automate healthcare tasks such as generating patient education materials and offering real-time responses to medical queries. This capability holds promise for enhancing patient access to reliable health information while also reducing the administrative burden on healthcare professionals. In addition to enhancing the efficiency and quality of medical services, GPT models can significantly improve patient engagement and empowerment by offering educational resources and addressing medical inquiries through interactive AI-driven communication.[5]

Different medical specialties have assessed ChatGPT's accuracy in responding to frequently asked questions related to their respective areas of expertise.[18-21] While these studies consistently demonstrate that ChatGPT delivers a high percentage of correct answers, the depth of its responses has often been questioned. Similarly, our study revealed variability in the comprehensiveness of GPT-4o's responses, with some answers lacking the nuanced detail necessary for more complex medical queries. This highlights that, although GPT-4o is effective for addressing general medical queries, its capacity to provide specialized clinical knowledge remains limited. Although the responses were evaluated based on patient-centered questions, the clinical evaluation revealed that the answers often lacked the depth expected from a clinical standpoint. For example, when asked about specific signs on an X-ray for coccydynia, the initial response lacked details on dynamic imaging, a detail even many general practitioners may not find critical. This demonstrates that while GPT-4o's responses may be adequate for patient understanding, the reviewers' advanced expertise in the field led them to anticipate a more comprehensive and nuanced explanation. However, in the second round, GPT-4o provided a more detailed response that included dynamic imaging, which is important for specialists managing coccydynia.[22,23] This also suggests that despite being asked at different times and from different accounts, GPT-4o can evolve by continuously updating its knowledge. This process allows the model to refine its accuracy and provide more comprehensive responses over time. Although individual interactions do not immediately impact the model’s learning, aggregated data from various users contributes to its ongoing improvement during periodic updates using larger datasets, enabling it to deliver increasingly precise information and enhanced training methodologies.

Additionally, the model exhibited significant variability in consistency between repeated queries. Notably, 34.9% of initially inadequate responses improved during the second query, suggesting the model's potential to deliver more accurate and detailed information over time. This improvement could be attributed to enhanced contextual understanding or the evolving nature of its training data. Natural language processing techniques, such as those employed by GPT-4o, enable AI to process vast amounts of information efficiently.[24] During the COVID-19 (coronavirus 2019) pandemic and natural language processing tools were instrumental in summarizing scientific literature and providing accurate information to the public.[25] Despite these advancements, the variability observed in GPT-4o emphasizes the need for continuous refinement. Real-time learning mechanisms could improve consistency by allowing the model to adapt to new data, particularly in complex or underrepresented cases.

Regarding accuracy, GPT-4o performed best in the QoL, where all responses were correct. It showed the lowest accuracy in the diagnosis, differential diagnosis, presentation, and work-up categories, with 8.7% of responses rated as incorrect. This suggests that the complexity of the medical category may influence the model’s performance. Additionally, the inherent unpredictability in the output of generative models, with the potential to confidently present incorrect information, raises concerns in clinical applications, particularly when evaluated across diverse queries or larger data sets. While the accuracy and consistency of ChatGPT's responses to questions about coccydynia are impressive for an LLM, they are imperfect; by clinical standards, the frequency of false statements we observed precludes ChatGPT from being used without careful human oversight or review.

The interreviewer agreement was notably strong in the diagnosis-related queries despite the lower accuracy in this category, which underscores the robustness of the evaluation process. This suggests that while the model may face challenges with more complex queries, the reviewers' consistent evaluations highlight the reliability of the assessment framework. Nevertheless, some responses required additional reviewer input to resolve ambiguities, highlighting areas where further model refinement may be necessary. An illustrative example of this variability in evaluation can be observed in the question, “Can diagnostic injections help confirm coccydynia?” The first reviewer, who frequently performs injection therapies for coccydynia, rated the initial response as correct but deemed the second response as “correct but inadequate” due to the omission of a critical detail in the usage of imaging guidance, which is crucial for increasing diagnostic accuracy and therapeutic outcomes. This discrepancy suggests that the first reviewer’s deeper expertise in injection techniques for coccydynia may have influenced their expectations, particularly given the complexity of differentiating pain sources, such as joint, ligamentous, or muscular structures. Fluoroscopic or ultrasound-guided injections allow for more specific, differential diagnosis-focused treatment, which can enhance therapeutic outcomes.[26] For patients, mentioning imaging techniques also provides sufficient information and encourages preparedness and greater involvement in the treatment process. This detail improves understanding, strengthens trust, and supports the patient-physician relationship.

To the best of our knowledge, no reports of any LLMs being used in the context of coccydynia exist. However, this study had several limitations. The fast-paced advancements in AI technology make it challenging to consistently evaluate GPT-4o’s performance, as the responses may shift with model updates. This dynamic nature raises concerns about the validity and reliability of AI-generated information, particularly when evaluated over time. Additionally, while physicians specializing in coccydynia evaluated the responses, the absence of patient perspectives limits insights into how patients interpret or utilize the provided information. Variations in question phrasing and context are also expected to influence the model’s output, as prompts can significantly shape responses. Furthermore, the study was conducted solely in English, which is the primary language of GPT-4o. The chatbot's accuracy and reliability when responding in other languages remain unknown and warrant further investigation. Finally, while this study focused on text-based outputs, evaluating visual content generated by multimodal AI models (such as GPT-4o) remains an important area for future research, particularly in medical education and diagnostics. These limitations underscore the need for cautious interpretation of the findings, particularly regarding their generalizability to diverse patient populations. While GPT-4o shows promise in providing accurate and generally reliable responses, its variability in consistency and depth highlights its role as a supplementary tool rather than a standalone source of medical information. Future studies incorporating diverse languages, patient-centered evaluations, and multimodal AI capabilities are critical for enhancing the reliability and applicability of AI-generated responses in clinical practice.

In conclusion, these findings highlight the need for further refinement of GPT-4o to improve its ability to deliver more comprehensive and contextually detailed responses, specifically in areas requiring a higher level of clinical precision. Enhancing the model's depth of knowledge will be essential to address complex medical queries better and support healthcare professionals in their decision-making processes. However, GPT-4o shows promise in patient education, where its ability to provide general information and create educational materials can improve patient engagement and understanding. GPT-4o's ability to provide accurate general information can assist patients in understanding their conditions, but this tool must be used as a supplement to professional medical advice rather than as a replacement. Healthcare providers and students must approach the model’s outputs cautiously, viewing them as a starting point for further research and validation rather than definitive answers.

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

Author Contributions: All authors contributed equally to this article.

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. White WD, Avery M, Jonely H, Mansfield JT, Sayal PK, Desai MJ. The interdisciplinary management of coccydynia: A narrative review. PM R 2022;14:1143-54. doi: 10.1002/pmrj.12683.

2. Skalski MR, Matcuk GR, Patel DB, Tomasian A, White EA, Gross JS. Imaging coccygeal trauma and coccydynia. Radiographics 2020;40:1504. doi: 10.1148/rg.2020209006.
3. Lee SH, Yang M, Won HS, Kim YD. Coccydynia: Anatomic origin and considerations regarding the effectiveness of injections for pain management. Korean J Pain 2023;36:272-80. doi: 10.3344/kjp.23175.
4. Foye PM. Stigma against patients with coccyx pain. Pain Med 2010;11:1872. doi: 10.1111/j.1526-4637.2010.00999.x.
5. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. JMIR Med Educ 2023;9:e46885. doi: 10.2196/46885.
6. Gencer G, Gencer K. A comparative analysis of ChatGPT and medical faculty graduates in medical specialization exams: Uncovering the potential of artificial intelligence in medical education. Cureus 2024;16:e66517. doi: 10.7759/cureus.66517.
7. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023;2:e0000198. doi: 10.1371/journal.pdig.0000198.
8. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: A systematic review and a meta-analysis. BJOG 2024;131:378-80. doi: 10.1111/1471-0528.17641.
9. Liu M, Okuhara T, Dai Z, Huang W, Okada H, Furukawa E, et al. Performance of advanced large language models (GPT-4o, GPT-4, Gemini 1.5 Pro, Claude 3 Opus) on Japanese medical licensing examination: A comparative study. medRxiv 2024. doi: 10.1101/2024.07.09.24310129
10. Shahriar S, Lund BD, Mannuru NR, Arshad MA, Hayawi K, Bevara RVK, et al. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. Applied Sciences 2024;14:7782.
11. Zhang N, Sun Z, Xie Y, Wu H, Li C. The latest version ChatGPT powered by GPT-4o: What will it bring to the medical field? Int J Surg 2024;110:6018-9. doi: 10.1097/JS9.0000000000001754.
12. Zhao B, Zhang W, Zhou Q, Zhang Q, Du J, Jin Y, et al. Revolutionizing patient education with GPT-4o: A new approach to preventing surgical site infections in total hip arthroplasty. Int J Surg 2024;111:1571–5. doi: 10.1097/JS9.0000000000002023.
13. Jaworski A, Jasiński D, Sławińska B, Błecha Z, Jaworski W, Kruplewicz M, et al. GPT-4o vs. Human candidates: Performance analysis in the Polish Final Dentistry Examination. Cureus 2024;16:e68813. doi: 10.7759/cureus.68813.
14. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the accuracy and reliability of AI-generated medical responses: An evaluation of the Chat-GPT model. Res Sq [Preprint] 2023:rs.3.rs-2566942. doi: 10.21203/rs.3.rs-2566942/v1.
15. Hermann CE, Patel JM, Boyd L, Growdon WB, Aviki E, Stasenko M. Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions. Gynecol Oncol 2023;179:164-8. doi: 10.1016/j.ygyno.2023.11.008.
16. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33:1790-6. doi: 10.1007/s11695-023-06603-5.
17. Giray E, Kenis-Coskun O, Karadag-Saygi E, Ozyemisci-Taskiran O. Interrater reliability, acceptability, and practicality of real-time video pediatric gait, arms, legs, and spine for musculoskeletal assessment of children during telemedicine visits. J Clin Rheumatol 2022;28:235-9. doi: 10.1097/RHU.0000000000001840.
18. Caglar U, Yildiz O, Ozervarli MF, Aydin R, Sarilar O, Ozgor F, et al. Assessing the performance of chat generative pretrained transformer (ChatGPT) in answering andrology-related questions. Urol Res Pract 2023;49:365-9. doi: 10.5152/tud.2023.23171.
19. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29:721-32. doi: 10.3350/cmh.2023.0089.
20. Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with Chat-GPT in radiology: Evaluating the efficacy and readability of answers to common imaging-related questions. J Am Coll Radiol 2024;21:353-9. doi: 10.1016/j.jacr.2023.09.011.
21. Villarreal-Espinosa JB, Berreta RS, Allende F, Garcia JR, Ayala S, Familiari F, et al. Accuracy assessment of ChatGPT responses to frequently asked questions regarding anterior cruciate ligament surgery. Knee 2024;51:84-92. doi: 10.1016/j.knee.2024.08.014.
22. Garg B, Ahuja K. Coccydynia-A comprehensive review on etiology, radiological features and management options. J Clin Orthop Trauma 2021;12:123-9. doi: 10.1016/j.jcot.2020.09.025.
23. Skalski MR, Matcuk GR, Patel DB, Tomasian A, White EA, Gross JS. Imaging coccygeal trauma and coccydynia. Radiographics 2020;40:1090-106. doi: 10.1148/rg.2020190132.
24. Zhang P, Kamel Boulos MN. Generative AI in Medicine and Healthcare: Promises, Opportunities and Challenges. Future Internet 2023;15:286.

25. Bekbolatova M, Mayer J, Ong CW, Toma M. Transformative potential of AI in healthcare: Definitions, applications, and navigating the ethical landscape and public perspectives. Healthcare (Basel) 2024;12:125. doi: 10.3390/healthcare12020125.

26. Lee SH, Yang M, Won HS, Kim YD. Coccydynia: Anatomic origin and considerations regarding the effectiveness of injections for pain management. Korean J Pain 2023;36:272-80. doi: 10.3344/kjp.23175.