Archives of Current Medical Research
Arch Curr Med Res 2026; 7(2): 321-330

Received: 10 September 2025
Revised: 21 December 2025
Accepted: 20 January 2026
DOI: 10.47482/acmr.1778992

ORIGINAL ARTICLE

Large Language Models' Responses to Patient Questions on Lateral Epicondylitis: Multi-Institutional Orthopaedic Surgeon Evaluation

Ali Geçer[1] **Emre Kaya**[2] **Alper Şükrü Kendirci**[3] **Alp Paksoy**[4] **Doruk Akgün**[4]

1 Haydarpaşa Numune Training and Research Hospital, Department of Orthopaedics and Traumatology, Istanbul, Türkiye.
2 İstanbul Kent University, Department of Orhopaedics and Traumatology, Istanbul, Türkiye.
3 Istanbul University, Faculty of Medicine, Department of Orthopaedics and Traumatology, Istanbul, Türkiye.
4 Charité University, Center for Musculoskeletal Surgery, Berlin, Germany.

Abstract

Background: Lateral epicondylitis (tennis elbow) is a common cause of elbow pain. With the increasing use of the internet and artificial intelligence (AI) for health information, large language models (LLMs) are frequently consulted by patients. This study aimed to evaluate the accuracy, reliability, content quality, and readability of responses provided by different large language models (ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot) to frequently asked patient questions about lateral epicondylitis.

Methods: The author committee reviewed patient-oriented questions on lateral epicondylitis using Google searches and selected the 12 most frequently asked questions for inclusion. These questions were presented to four LLMs: ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot. Responses were evaluated for accuracy using a five-point Likert scale, reliability using the modified DISCERN scale, quality using the Global Quality Scale (GQS), and readability using the Flesch Reading Ease Score (FRES).

Results: Perceived medical accuracy did not differ significantly among the LLMs ($p = 0.579$). Reliability differed significantly (modified DISCERN: $p < 0.001$), with Copilot and Gemini achieving higher scores than ChatGPT-4 (both $p < 0.001$) and Copilot also outperforming ChatGPT-3.5 ($p = 0.002$). Quality differed significantly (GQS: $p < 0.001$), with ChatGPT-3.5 and Gemini scoring higher than ChatGPT-4 ($p = 0.001$ and $p = 0.006$, respectively). Readability differed across models (FRES: $p = 0.049$); Gemini demonstrated higher readability than ChatGPT-3.5 ($p = 0.040$), while responses from all models were generally difficult to read. Response generation time differed significantly ($p < 0.001$), with ChatGPT-4 producing the slowest responses.

Conclusions: All evaluated LLMs provided generally accurate and moderately reliable responses to questions about tennis elbow, with differences observed across specific quality domains such as source transparency, readability, and response time. Models with citation capabilities demonstrated higher reliability in terms of source transparency, while readability remained a common limitation. LLMs show potential as supplementary patient information tools in orthopaedic; however, further refinement and improved readability are needed before widespread clinical use.

Keywords: Lateral epicondylitis, Tennis elbow, Large language models, Artificial intelligence, Patient education, Orthopaedics

Corresponding Author:
Ali Geçer, M.D.
Department of Orthopaedics and Traumatology
Haydarpaşa Numune Training and Research Hospital İstanbul, Turkey
Email: draligecer@gmail.com

INTRODUCTION

Lateral epicondylitis, more commonly known as "tennis elbow" in the community, affects 1-3% of the population (1). It represents a form of tendinosis of the extensor muscles of the forearm and is one of the most common causes of lateral elbow pain. It has been reported in the literature that a significant proportion of orthopaedic patients use internet resources to research conditions and treatment options. Burrus et al. (2) reported that 64.7% of patients used the internet for orthopaedic information. In another study, 54% of sports medicine patients reported that they had researched their condition online before consulting a physician (3). Similarly, private practice orthopaedic surgeons found that 45% of patients use the internet for medical information (4).

In recent years, there has been a notable increase in the utilization of AI in internet-based health information searches. Today, the use of AI as a source of medical information is becoming increasingly popular among healthcare professionals and patients (5). AI encompasses machines or software systems designed to carry out tasks typically requiring human intellect, such as decision-making, problem-solving, and learning from experience (6). Large Language Models (LLMs) are a subset of artificial intelligence developed using deep learning techniques. They are designed to enable machines to understand, interpret and create human-like texts, which are necessary for a variety of Natural Language Processing (NLP) tasks (7). LLMs that employ NLP algorithms represent a class of language models that are capable of autonomously understanding and responding to queries posed by human users. In studies evaluating the performance of LLMs in medical question-answering tasks, it has been stated that although LLMs are found promising, they differ in their reliability and referencing (8).

Orthopaedics conditions such as tennis elbow are frequently complex, and patients typically have a multitude of queries and concerns pertaining to their diagnoses, treatment options, and anticipated outcomes. The aim of this study was to evaluate the accuracy, reliability, content quality, and readability of responses generated by large language models to frequently asked patient questions about lateral epicondylitis. We hypothesized that LLM-generated responses would provide rapid access to information with acceptable medical accuracy, but would differ in reliability, quality, and readability across models.

MATERIALS AND METHODS

No ethical committee approval was required for this study, as no human participants or patient data were involved. An initial pool of patient-oriented questions related to lateral epicondylitis was generated by the author committee to reflect common real-world patient information needs. Frequently asked questions were identified through structured web-based searches and review of routinely encountered patient inquiries in clinical practice. Structured Google searches were performed using predefined keywords and keyword combinations, including *"lateral epicondylitis," "tennis elbow," "patient questions," "FAQ," "diagnosis," "treatment," "exercises," "surgery,"* and "return to sport." Searches were limited to the first 10 pages of results. Only English-language, patient-focused informational sources were included, such as hospital websites, professional association portals, and educational health platforms.

Content primarily intended for healthcare professionals, clinical guidelines, experimental or niche topics, case reports, advertisements, promotional material, video-only content, and social media posts or comments were excluded. A total of 100 questions were initially collected. Questions addressing the same informational goal were grouped based on question intent, defined as targeting the same underlying patient information need, and redundant questions were merged. From this pool, 12 representative questions were selected through a consensus process based on predefined criteria, including coverage of key stages of the patient journey, single-sentence and single-intent structure, clinical relevance, and clarity of wording. The final set of questions is presented in Table 1. Also, the complete pool of 100 patient-oriented questions is provided in Supplementary (Table 1).

All selected questions were posed to ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot on 15 May 2024. All queries were performed on the same day using the same device and a stable internet connection. To minimize potential contextual carryover effects, a new conversation window was opened for each question in each model. In addition, the response times were recorded using a stopwatch. To ensure impartiality, all words related to the LLM in the responses were removed and recorded on four separate forms (A, B, C, D). The responses were evaluated by 12 orthopaedic surgeons from four different centers, with three surgeons from each center, with

Table 1. Patient-oriented questions posed to ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot	
1.	What causes tennis elbow?
2.	What are the symptoms of tennis elbow?
3.	How is tennis elbow diagnosed?
4.	Can tennis elbow heal without treatment?
5.	What are the treatments of tennis elbow?
6.	Does the tennis elbow splint help?
7.	Does injection help with tennis elbow?
8.	How to prevent tennis elbow?
9.	What exercises are effective in the treatment of tennis elbow?
10.	How is tennis elbow surgery performed?
11.	Is tennis elbow recurrent?
12.	When can you return to sports after tennis elbow injury?

reference to the current literature and clinical practice. The multi-center nature of the study refers to the independent evaluation of the LLM-generated responses by orthopaedic surgeons affiliated with different institutions, rather than multi-center patient enrollment or data acquisition. Prior to the evaluation, all assessors convened for a meeting to reach a consensus about evaluation methods. Likert Scale, GQS, modified DISCERN (mDISCERN) scale and FRES were used as criteria for evaluating the LLM responses as in previous studies (9–11). All models were evaluated using their default, user-facing configurations, reflecting real-world patient use. Some models provide web-based citations or external sources by default, while others do not.

A five-point Likert scale was used to evaluate the accuracy of the responses generated by the LLMs (8,12–14). According to this scale, score 1; the chatbot's responses are completely incorrect, score 2; the chatbot's responses contain more incorrect elements than correct elements, score 3; the chatbot's responses contain an equal balance of correct and incorrect elements, score 4; the chatbot's responses contain more correct elements than incorrect elements, and score 5; the chatbot's responses are completely correct (Table 2).

To evaluate the reliability of responses generated by LLMs, a modified form of the validated DISCERN tool, which is used to assess the reliability and quality of online health information, was used (15–17). The description of the mDISCERN tool is shown in Table 2 and consists of 5 questions and for each question a 'yes' answer is scored as 1 point and a 'no' answer is scored as 0 point. The total score is the reliability score. Items within the mDISCERN scale that assess source attribution or provision of additional references were considered citation-dependent. These items were retained in the primary analysis to reflect the complete patient-facing information experience rather than isolating citation effects.

GQS was used to assess the quality of responses from the LLMs. GQS analyses the quality of written sources in the field of medicine (9,10,18,19). The description of the GQS tool is given in Table 2. According to the GQS, the lowest score is 1 and the highest score is 5. 1-2 points represent low quality, 3 points represent medium quality, and 4-5 points represent high quality.

FRES was used to assess the readability of responses from the LLMs (9,19–21). FRES is used to determine the readability of written text. The formula for Flesch Reading

Table 2. Contents of 5-point Likert Scale, mDISCERN Scale, GQS, and FRES	
Accuracy (5-point Likert Scale)	**Score**
The chatbot's responses are completely incorrect	1
The chatbot's responses are more incorrect than correct elements	2
The chatbot's responses are equal balance of correct and incorrect elements	3
The chatbot's responses are more correct than incorrect elements	4
The chatbot's responses are completely correct	5
Reliability (Modified DISCERN Scale)	**Total (5 points)**
1. Are the aims clear and achieved?	0 or 1 point
2. Are reliable sources of information used ?	
(i.e., publication cited, author or producer)	0 or 1 point
3. Is the information present both balanced and unbiased?	0 or 1 point
4. Are additional sources of information listed for patient reference?	0 or 1 point
5. Are areas of uncertainty mentioned?	
(i.e., more research is needed on this topic or there is no clear consensus)	0 or 1 point
Quality (Global Quality Scale)	**Score**
Poor quality, poor flow of the site, most information missing, not at all useful for patients	1
Generally poor quality and poor flow, some information listed but many important topics missing, of very limited use to patients	2
Moderate quality, suboptimal flow, some important information is adequately discussed but others poorly discussed, somewhat useful for patients	3
Good quality and generally good flow, most of the relevant information is listed, but some topics not covered, useful for patients	4
Excellent quality and excellent flow, very useful for patients	5
Readability (Flesch Reading Ease Score)	
206.835 - 1.015 x (total words/total sentences) - 84.6 x (total syllables/total words)	

Ease is as follows 206.835 - 1.015 x (total words/total sentences) - 84.6 x (total syllables/total words) (Table 2). This formula gives a score between 0 and 100. Scores close to 100 mean that the document is very easy to read, while scores close to 0 mean that the document is very complex and difficult to understand. The reading levels corresponding to the scoring system are shown in Table 3.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics (version 29.0; IBM Corp., Armonk, NY, USA). For each model, rater scores were aggregated at the question level (n = 12 per model). Ordinal variables (Likert score, GQS, and mDISCERN score) were summarized

Table 3. Interpretation of the FRES and Reading Level

Flesch Reading Ease Score	Reading Level
0-29	Very difficult
30-49	Difficult
50-59	Fairly difficult
60-69	Standard and/or plain
70-79	Fairly easy
80-89	Easy
90-100	Very easy
FRES: Flesch Reading Ease Score	

as median (IQR) and compared using the Kruskal–Wallis test with Dunn's post hoc test and Holm adjustment. Continuous variables (FRES and response generation time) were summarized as mean ± standard deviation and analyzed using the Kruskal–Wallis test. Effect sizes were estimated using epsilon-squared (ε^2), and $p < 0.05$ was considered statistically significant.

Inter-rater reliability was assessed using weighted Fleiss' kappa for ordinal outcomes and a two-way random effects intraclass correlation coefficient [ICC (2,k)] for continuous variables, in accordance with established methodological guidelines (22,23).

RESULTS

The 12 questions primarily addressed diagnostic information, treatment options, rehabilitation, recurrence, and return-to-activity concerns, reflecting key stages of the patient information journey. Median Likert scores did not differ significantly among the four large language models ($p = 0.579$), indicating comparable perceived medical accuracy across models (Table 4).

A statistically significant difference was observed in modified DISCERN (mDISCERN) scores ($p < 0.001$). Copilot and Gemini demonstrated higher median mDISCERN scores than ChatGPT-4 (both $p < 0.001$), while Copilot also achieved higher scores than ChatGPT-3.5 ($p = 0.002$) (Table 5).

Global Quality Scale (GQS) scores differed significantly between models ($p < 0.001$). ChatGPT-3.5 and Gemini achieved higher GQS scores than ChatGPT-4 ($p = 0.001$ and $p = 0.006$, respectively), whereas no statistically significant difference was observed between ChatGPT-3.5 and Gemini. According to GQS classification, 33% of Gemini responses and 25% of ChatGPT-3.5 responses were rated as high quality, while all ChatGPT-4 responses were classified as moderate quality (Table 4).

Readability differed significantly across models ($p = 0.049$). Gemini responses demonstrated higher Flesch Reading Ease Scores than ChatGPT-3.5 ($p = 0.040$), whereas no significant differences were observed among Gemini, Copilot, and ChatGPT-4 (Tables 4 and 5). Overall, responses generated by all models were classified as difficult to read according to established readability thresholds.

Response generation time differed significantly between models ($p < 0.001$). ChatGPT-4 had the longest response generation time and was significantly slower than all other models. Copilot responses were generated more slowly than Gemini responses ($p = 0.015$) (Tables 4 and 5).

Inter-rater reliability analysis demonstrated substantial agreement among evaluators for ordinal outcome measures. Weighted Fleiss' kappa values indicated substantial agreement for Likert scores ($\kappa = 0.72$), modified DISCERN scores ($\kappa = 0.68$), and Global Quality Scale scores ($\kappa = 0.65$). Excellent agreement was observed for continu-

Table 4. Comparison of Evaluation Scores, Readability and Generation Time Across Large Language Models

Outcome measure	ChatGPT-3.5	ChatGPT-4	Gemini	Copilot	p*
Likert score	4.0 [3.0–5.0]	4.0 [3.0–4.0]	4.0 [3.0–5.0]	4.0 [3.0–5.0]	0.579
mDISCERN score	3.0 [3.0–4.0]	3.0 [2.0–3.0]	4.0 [3.0–4.0]	4.0 [3.0–4.0]	<0.001†
GQS score	4.0 [3.0–4.0]	3.0 [3.0–4.0]	4.0 [3.0–4.0]	4.0 [3.0–4.0]	<0.001†
FRES	29.73 ± 16.03	37.72 ± 15.73	47.71 ± 19.78	43.17 ± 10.91	0.049†
Generation time (s)	8.69 ± 2.72	23.13 ± 7.99	7.13 ± 1.31	13.07 ± 3.44	<0.001†

*Kruskal–Wallis test, †p<0.05

Table 5. Dunn Post Hoc Pairwise Comparisons with Holm Correction

Comparison	MDISCERN (p)	GQS (p)	FRES (p)	Generation time (p)
ChatGPT-3.5 vs ChatGPT-4	0.042†	0.001†	0.613	0.004†
ChatGPT-3.5 vs Gemini	0.123	1.000	0.040†	0.841
ChatGPT-3.5 vs Copilot	0.002†	0.205	0.180	0.106
ChatGPT-4 vs Gemini	<0.001†	0.006†	0.425	<0.001†
ChatGPT-4 vs Copilot	<0.001†	0.184	0.836	<0.001†
Gemini vs Copilot	0.123	0.468	0.897	0.015†

† Holm-adjusted p < 0.05

ous variables, with ICC (2,k) values of 0.89 for the Flesch Reading Ease Score and 0.94 for response generation time.

DISCUSSION

LLMs are increasingly used by patients to obtain information about orthopaedic conditions, raising concerns regarding the accuracy, reliability, and clinical safety of AI-generated content (24). Focusing on frequently asked questions about lateral epicondylitis, the present study evaluates the strengths and limitations of commonly used LLMs in providing patient-oriented medical information.

Various indices have been used in the literature to assess the accuracy of LLM responses; in this study, a five-point Likert scale was used for evaluation (19,25,26). Youssef et al. (26) stated that ChatGPT-3.5 provided accurate answers to patients' frequently asked questions about glenohumeral osteoarthritis. Zhang et al. (27) reported high accuracy of ChatGPT-3.5 responses to frequently asked

questions about total knee arthroplasty. In the present study, all evaluated LLMs demonstrated comparable perceived medical accuracy when addressing frequently asked questions about tennis elbow, with no statistically significant differences observed between models. Although the responses given by LLMs in our study were generally accurate, Giuffrè et al. (28) evaluated LLMs in the context of digestive diseases and concluded that despite their potential, their current accuracy and reliability are insufficient for clinical use. This finding raises important considerations about the variability in LLM performance depending on the complexity of the condition being addressed and the specificity of the questions asked. Furthermore, although LLMs such as ChatGPT, Gemini and Copilot show strong potential for patient education, the small differences in their accuracy scores highlight the need for continuous improvement and rigorous validation. LLMs are typically trained on large text datasets, which include publicly accessible sources such as web pages, blogs, online encyclopedias, news websites, and forums (19). Therefore, websites created by experts and supported by evidence-based scientific data may contribute to improved response quality of language models.

Evaluation of information sources, citations, and bias is important in assessing the reliability of health information, and the mDISCERN scale was used in our study (19). Yeramosu et al. (29) stated that ChatGPT was moderately reliable compared to DISCERN scale in frequently asked questions about total shoulder arthroplasty. Similarly, in our study, responses to questions about tennis elbow were moderately reliable in all LLMs. Higher mDISCERN scores were observed for Copilot and Gemini compared with ChatGPT-3.5 and ChatGPT-4. This finding aligns with the high scores achieved by Copilot and Gemini on the DISCERN tool, particularly in questions evaluating the inclusion of information sources used in content creation. In contrast, a notable drawback of chatbots like ChatGPT-3.5 and ChatGPT-4 in addressing health-related inquiries is their lack of capacity to cite or reference the sources of the information they provide (30). The ability of Copilot and Gemini to add references or citations may be related to their real-time access to up-to-date information and the integration of internet search engines (31). Gilmore et al (32) valuated the effectiveness of ChatGPT-3.5 in answering common patient questions about knee osteoarthritis and found that responses were of moderate quality, lacked citations, and that appropriate patient education by orthopaedic surgeons was still needed. Gupta et al. (33) stated that most references in both Copilot and Gemini were provided by journals, followed by academic sources, and Copilot provided a higher number of references compared to Gemini for responses about anterior cruciate ligament injury and repair. Similar to our study, Reyhan et al. (34) evaluated the answers of different chatbots to questions about keratoconus and found that Copilot's reliability score measured by mDISCERN was higher than ChatGPT-3.5. In this study, Gemini and Copilot are thought to have higher mDISCERN scores, reflecting more comprehensive and source-transparent information, which may be related to differences in training data, fine-tuning strategies, or underlying algorithms (34). Differences in mDISCERN scores may partly reflect variations in default citation and web-retrieval capabilities; however, this was considered an integral component of real-world patient information quality rather than a confounding factor. This divergence may also be explained by model-specific safety guardrails and instruction-tuning policies rather than differences in foundational medical knowledge, as previously demonstrated in orthopedic LLM evaluations (35).

The analysis of GQS scores highlighted differences in the quality of medical information generated by LLMs. Gemini responses were 33% high quality and 67% medium quality. 25% of ChatGPT-3.5 responses were high quality and 75% were medium quality. Copilot responses were 8% high quality and 92% medium quality. All ChatGPT-4 responses were medium quality. In general, LLMs produced moderate to high quality answers to frequently asked questions about lateral epicondylitis. Reyhan et al. (34) evaluated the responses of six popular chatbots (ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity) to questions about keratoconus. The lowest GQS score was observed in the ChatGPT-3.5 model, and the highest score was observed in the Gemini model. Gemini achieved relatively higher GQS scores, indicating stronger overall quality in this metric. Onder et al. (9) evaluated the reliability and readability of ChatGPT-4 responses to questions about hypothyroidism in pregnancy and stated that the majority were of high quality (84.2%), followed by medium quality (10.5%) according to GQS. In our study, the

highest scores were observed in ChatGPT-3.5 and Gemini, while the lowest score was observed in ChatGPT-4. Differences in GQS scores may reflect variations in training datasets, response structuring, and access to external information sources. In addition, the complexity of the topic, the capacity of the models to understand user questions and the wording of the answers also play an important role in quality. These factors may contribute to variations between models depending on their technical infrastructure and intended use.

In literature, readability is recognized as a crucial component of health literacy, ensuring that written materials can be easily understood. Readability is assessed using various indices that consider factors such as sentence length and the presence of complex words. The FRES ranges from 0 to 100, with higher scores indicating greater readability; a score of 60-70 represents standard English, and 65 is generally considered an acceptable target (19). Gemini demonstrated higher readability compared with ChatGPT-3.5, indicating that its responses were easier to read. Despite this, all models produced responses that were generally classified as difficult to read, which may limit their accessibility for individuals with lower health literacy levels. Similarly, Tepe et al. (9) evaluated the responses of LLMs (ChatGPT-4, Gemini and Copilot) to frequently asked questions in breast imaging and stated that Gemini and Copilot had higher readability scores and were easier to understand compared to ChatGPT-4. Since each LLM is trained on different datasets, the structure and complexity of the language used may differ. Gemini and Copilot may have been trained on simpler or more conversational language patterns, which may partly explain their higher readability scores. Similar to our findings, Goktas et al.(36) performed an analysis between different LLMs for melanoma-specific information and found no significant difference between Gemini and Bing (now Copilot).

ChatGPT-3.5 demonstrated faster response times than ChatGPT-4, consistent with previous reports attributing this difference to architectural and processing characteristics (37,38). To date, no studies have directly compared the response generation time of Copilot with Gemini or ChatGPT models.

This study has several limitations. First, evaluating only English-language queries may limit generalizability to other languages. Second, despite standardized conditions, LLM outputs may evolve over time due to model updates. Third, the exploratory design and limited sample size precluded a formal priori power analysis; thus, effect size estimates should be interpreted with caution. Finally, although manual querying reflects real-world patient use, API-based access could offer higher reproducibility in future research. Subsequent studies should consider two-phase frameworks combining expert assessment with patient-level validation, as proposed by Wang et al. (39).

All evaluated large language models provided generally accurate and moderately reliable responses to frequently asked questions about tennis elbow, with comparable perceived medical accuracy across models. Differences were observed in specific quality dimensions, including information structure, source transparency, readability, and response generation time. Models with built-in citation or web-retrieval capabilities demonstrated higher reliability in terms of source transparency, whereas readability remained a common limitation, as most responses were classified as difficult to read. LLMs show promise as supplementary patient information tools in orthopaedics; however, improvements in readability, source transparency, and evidence-based risk communication are necessary before widespread clinical adoption.

REFERENCES

1. Finestone HM, Rabinovitch DL. Tennis elbow no more: practical eccentric and concentric exercises to heal the pain. Can Fam Physician. 2008;54(8):1115-6.
2. Tyrrell Burrus M, Werner BC, Starman JS, Kurkis GM, Pierre JM, Diduch DR, et al. Patient perceptions and current trends in internet use by orthopedic outpatients. HSS J. 2017;13(3):271-5.
3. Koenig S, Nadarajah V, Smuda MP, Meredith S, Packer JD, Henn RF. Patients' use and perception of internet-based orthopaedic sports medicine resources. Orthop J Sports Med. 2018;6(9):232596711879646.
4. Krempec J, Hall J, Biermann JS. Internet use by patients in orthopaedic surgery. Iowa Orthop J. 2003;23:80-2.
5. Abu Arqub S, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod. 2024;94(3):263-72.
6. Nagendraswamy C, Amogh S. A review article on artificial intelligence. Ann Biomed Sci Eng. 2021;5(1):13-4.
7. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell. 2023;6:1253929.
8. 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 ChatGPT model. Res Sq. 2023. doi:10.21203/rs.3.rs-2566942/v1.
9. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep. 2024;14(1):243.
10. 10. Giorgino R, Alessandri-Bonetti M, Del Re M, Verdoni F, Peretti GM, Mangiavini L. Google Bard and ChatGPT in orthopedics: which is the better doctor in sports medicine and pediatric orthopedics? The role of AI in patient education. Diagnostics (Basel). 2024;14(12):1253.
11. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr Google to Dr ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. Surg Endosc. 2024;38(5):2887-93.
12. Yalamanchili A, Sengupta B, Song J, Lim S, Thomas TO, Mittal BB, et al. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open. 2024;7(4):e244630.
13. Cardona G, Argiles M, Pérez-Mañá L. Accuracy of a large language model as a new tool for optometry education. Clin Exp Optom. 2023.
14. Sullivan GM, Artino AR. Analyzing and interpreting data from Likert-type scales. J Grad Med Educ. 2013;5(4):541-2.
15. Griffiths KM, Christensen H. Website quality indicators for consumers. J Med Internet Res. 2005;7(5):e55.
16. Sanger S. DISCERN in practice. Health Expect. 1998;1(2):135-6.
17. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105-11.
18. Cakir H, Caglar U, Sekkeli S, Zerdali E, Sarilar O, Yildiz O, et al. Evaluating ChatGPT ability to answer urinary tract infection-related questions. Infect Dis Now. 2024;54(4):104884.
19. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024;24(1):211.
20. Fahy S, Niemann M, Böhm P, Winkler T, Oehme S. Assessment of the quality and readability of information provided by ChatGPT in relation to the use of platelet-rich plasma therapy for osteoarthritis. J Pers Med. 2024;14(5):495.
21. Temel MH, Erden Y, Bağcıer F. Information quality and readability: ChatGPT's responses to the most common questions about spinal cord injury. World Neurosurg. 2024;181:e1138-44.
22. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. NPJ Digit Med. 2024;7:312.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63.
24. White CA, Masturov YA, Haunschild E, Michaelson E, Shukla DR, Cagle PJ. Can ChatGPT reliably answer the most common patient questions regarding total shoulder arthroplasty? J Shoulder Elbow Surg. 2025;34(5):e254-64.
25. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofacial Orthop. 2024;165(6):652-62.
26. Youssef Y, Youssef S, Melcher P, Henkelmann R, Osterhoff G, Theopold J. How accurately can ChatGPT 3.5 answer frequently asked questions by patients on glenohumeral osteoarthritis? Obere Extrem. 2025;20:205-10.
27. Zhang S, Liau ZQG, Tan KLM, Chua WL. Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement. Knee Surg Relat Res. 2024;36(1):15.
28. Giuffrè M, Kresevic S, You K, Dupont J, Huebner J, Grimshaw AA, et al. Systematic review: The use of large language models as medical chatbots in digestive diseases. Aliment Pharmacol Ther. 2024;60(2):144–66.
29. Yeramosu T, Johns WL, Onor G, Menendez ME, Namdari S, Hammoud S. ChatGPT is capable of providing satisfactory responses to frequently asked questions regarding total shoulder arthroplasty. Shoulder Elbow. 2024;16(4):407–12.
30. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J. 2024;57(3):305–14.
31. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod. 2025;48(1):cjae017.

32. Gilmore N, Kushner JN, Redden A, Hansen AW, Yerke Hansen P, Martinez L. Assessing ChatGPT Responses to Common Patient Questions on Knee Osteoarthritis. Journal of Orthopaedic Experience & Innovation. 2024 Nov 1.

33. Gupta S, Tarapore R, Haislup B, Fillar A. Microsoft Copilot Provides More Accurate and Reliable Information About Anterior Cruciate Ligament Injury and Repair Than ChatGPT and Google Gemini; However, No Resource Was Overall the Best. Arthrosc Sports Med Rehabil. 2024;7(2):101043.

34. Reyhan AH, Mutaf Ç, Uzun İ, Yüksekyayla F. A Performance Evaluation of Large Language Models in Keratoconus: A Comparative Study of ChatGPT-3.5, ChatGPT-4.0, Gemini, Copilot, Chatsonic, and Perplexity. J Clin Med. 2024;13(21):6512.

35. Chundi G, Dawar A, Sarwar S, Prasad S, Vosbikian M, Ahmed I. Comparative evaluation of LLMs in orthopedic surgery. Journal of Orthopaedic Reports. 2026;5(2):100728

36. Goktas P, Grzybowski A. Assessing the Impact of ChatGPT in Dermatology: A Comprehensive Rapid Review. J Clin Med. 2024;13(19):5909.

37. Rana N, Katoch N. AI for Biophysical Phenomena: A Comparative Study of ChatGPT and Gemini in Explaining Liquid–Liquid Phase Separation. Applied Sciences. 2024;14(12):5065.

38. Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Forte AJ. Large Language Models for Intraoperative Decision Support in Plastic Surgery: A Comparison between ChatGPT-4 and Gemini. Medicina (B Aires). 2024;60(6):957.

39. Wang YL, Tian LC, Meng JY, Zhang JC, Nie ZX, Wei WR, et al. Evaluation of large language models in patient education and clinical decision support for rotator cuff injury: a two-phase benchmarking study. BMC Medical Informatics and Decision Making. 2025;25(1):289.

Abbreviations

AI: Artificial Intelligence
GQS: Global Quality Score
FRES: Flesch Reading Ease Score
LLMs: Large Language Models
NLP: Natural Language Processing

Ethics Approval and Consent to Participate

This study did not involve human participants, animal experiments, or patient data. The analyses were conducted using publicly available information and outputs generated by large language models. Therefore, ethics committee approval was not required.

Consent for Publication

Not applicable.

Availability of data and Materials

No datasets were generated or analyzed during the current study.

Competing Interests

The authors declare that they have no competing interests.

Funding

The authors did not receive any financial support for the submitted work.

Authors Contributions

Idea/Concept: A.G., Design: A.G., Control/Supervision: A.G., D.A., Data Collection and/or Processing: A.G., E.K., A.Ş.K., A.P. Analysis and/or Interpretation: A.G., Literature Review: A.G., A.P., Writing the Article: A.G., A.P., Critical Review: E.K., A.Ş.K, D.A., References and Fundings: A.G., Materials: Not applicable, Other: Not applicable, All authors have read and approved the final manuscript.

Acknowledgements

Not applicable.