İstanbul Med J 2026; 27(2): 143-8
DOI: 10.4274/imj.galenos.2026.95826

Evaluation of Artificial Intelligence-Generated Information About Abdominal Ultrasonography

Ali Salbas[1], Gözde Merve Tekel[2], Aslı Dilara Büyüktoka[1], Raşit Eren Büyüktoka[3], Ali Murat Koc[1], Atilla Hikmet Çilengir[2]

[1]İzmir Katip Çelebi University, Atatürk Training and Research Hospital, Department of Radiology, İzmir, Türkiye
[2]İzmir Democracy University, Buca Seyfi Demirsoy Training and Research Hospital, Department of Radiology, İzmir, Türkiye
[3]University of Health Sciences Türkiye, İzmir City Hospital, Department of Radiology, İzmir, Türkiye

ABSTRACT

Introduction: This study examined the relevance, accuracy, clarity, and completeness of ChatGPT-5 responses to frequently asked patient questions about abdominal ultrasonography and considered the potential role of large language models (LLMs) as supportive tools in patient education.

Methods: This cross-sectional study analyzed ChatGPT-5 responses to 15 frequently asked questions from patients about abdominal ultrasonography. The questions were collected from Google's "other questions" section. Each question was entered into ChatGPT-5 in a separate session, and the model's answers were recorded. Ten radiologists independently evaluated the responses using four criteria: relevance, accuracy, clarity, and completeness, with each criterion scored on a 1-to-5 scale. Interrater reliability was assessed using the intraclass correlation coefficient (ICC).

Results: ChatGPT-5 demonstrated high performance across all evaluated criteria. Mean scores were 4.97±0.18 for relevance, 4.78±0.49 for accuracy, 4.85±0.40 for clarity, and 4.68±0.53 for completeness, with an overall mean of 4.82±0.26. The minimum score assigned by the evaluators was 3. ICC values were 0.266 for relevance, 0.236 for accuracy, 0.230 for clarity, 0.582 for completeness, and 0.555 for the total score.

Conclusion: ChatGPT-5 provided generally well-rated responses to common patient questions about abdominal ultrasonography. Although interrater reliability showed variable levels of agreement, moderate agreement was observed for completeness and total scores. The model's overall performance was favorable, suggesting that LLMs may function as supportive resources for patient education. Their use should remain complementary to professional medical guidance. Further studies with broader question sets, diverse patient populations, and multiple language models are warranted.

Keywords: Artificial intelligence, health information technology, natural language processing, patient communication, ultrasonography, abdominal

Introduction

Large language models (LLMs) are artificial intelligence systems trained on extensive text datasets that can generate human-like responses and interact with users in natural language (1). Recent studies have examined LLMs in many medical specialties, exploring their use in clinical and educational settings (2). Within radiology, they have been studied for applications such as diagnostic support, report generation, and educational assessment in radiology (3,4). Additionally, LLMs are being studied as tools to help improve communication between clinicians and patients (5,6).

Ultrasonography is one of the most widely used imaging modalities in daily clinical practice because it is inexpensive, easily accessible, and free of ionizing radiation. In Türkiye, ultrasonography represents the highest imaging volume nationwide, exceeding 30 million examinations in 2020 and 35 million in 2021 according to national health statistics (7). Before undergoing imaging, many patients seek information about the procedure and often express concerns related to preparation, comfort, or diagnostic value (8).

In recent years, patients have increasingly relied on internet-based resources to obtain health-related information (9). This trend has contributed to the rising use of artificial intelligence chatbots. Although these models can provide rapid and structured answers, the accuracy and reliability of the information they offer remain important questions (10).

Address for Correspondence: Ali Salbas, MD, İzmir Katip Çelebi University, Atatürk Training and Research Hospital, Department of Radiology, İzmir, Türkiye
E-mail: dralisalbas@gmail.com **ORCID ID:** orcid.org/0000-0002-6157-6367

Cite this article as: Salbas A, Tekel GM, Büyüktoka AD, Büyüktoka RE, Koc AM, Çilengir AH. Evaluation of artificial intelligence-generated information about abdominal ultrasonography. İstanbul Med J. 2026; 27(2): 143-8

Received: 18.02.2026
Accepted: 07.04.2026
Publication Date: 12.05.2026

The potential of LLMs to support patient education has been investigated in radiology and in other medical fields (11,12). However, the quality of LLM responses to frequently asked patient questions about ultrasonography in general and abdominal ultrasonography in particular has not been systematically evaluated to date. To our knowledge, no study has evaluated how ChatGPT-5 performs in this context. The purpose of this study is to assess the relevance, accuracy, clarity, and completeness of ChatGPT-5's answers to frequently asked patient questions regarding abdominal ultrasonography, and to discuss its potential role as a supportive tool in patient education.

Methods

Ethical approval for this study was obtained from the İzmir Katip Çelebi University Health Research Ethics Committee (decision number: 0670, date: 06.11.2025). Due to the use of publicly available data and the absence of direct human participant involvement or identifiable patient information, the requirement for informed consent was waived by the ethics committee. This cross-sectional study was designed to assess the quality of responses generated by ChatGPT-5 to patients' frequently asked questions about abdominal ultrasonography. To obtain the patient questions, a new Google (Alphabet Inc., Mountain View, CA, USA) account without any prior search history was used (13). The phrase "frequently asked questions about abdominal ultrasonography" was entered into Google using its Turkish equivalent. This search yielded 150 questions listed in the "other questions" section of the results page. Two board-certified radiologists with 8 and 10 years of experience in ultrasonography independently reviewed all items. After eliminating duplicates and questions with overlapping meanings, they reached a consensus on a final set of 15 unique, patient-oriented questions through joint discussion (Table 1, Figure 1). Each question was preserved in its original form as displayed in the Google results to maintain authentic patient language.

The selected questions were entered into ChatGPT-5 (OpenAI Inc., San Francisco, CA, USA) using an account with no prior conversations. The model was accessed through its official web interface, which offers three modes of operation: Auto, Instant, and Thinking (14). In this study, the Auto mode was used because it reflects the default setting and typical real-world user interaction with the model. No model parameters, including the temperature (which is fixed in the web interface), were adjusted. No system-level prompts or hidden instructions were modified, and all questions were submitted in their original form without any additional prompting or formatting. No additional prompts or contextual information was provided. Each question was submitted separately and asked only once. To prevent any influence from previous interactions, a new session was opened for every question by clearing the chat history (15,16). All queries were entered on November 18, 2025.

The responses produced by ChatGPT-5 were evaluated by 10 radiologists from six different institutions, each with 4 to 10 years of experience in abdominal ultrasonography. Each response was assessed using four criteria: relevance, accuracy, clarity, and completeness. Scores ranged from 1 to 5 for each criterion, where 1 indicated the lowest score and 5 indicated the highest score. The scoring framework was adapted from previously published studies (5,6). Relevance referred to how directly the answer addressed the question. Accuracy reflected whether the information was medically accurate and clinically appropriate. Clarity indicated how understandable and well-organized the response was. Completeness assessed whether the answer provided sufficient information to fully address the question. The evaluators were blinded to the source of the responses and informed only that they were reviewing answers to patients' questions about abdominal ultrasonography. They were not informed that the responses were generated by a LLM. All questions were submitted in Turkish, and the radiologists evaluated the responses in Turkish.

Statistical Analysis

All data were analyzed using IBM SPSS Statistics version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were reported as means and standard deviations for each evaluation criterion. Interrater reliability

Table 1. The 15 most frequently asked patient questions about abdominal ultrasonography that were submitted to ChatGPT-5

No	Question
1	What diseases can be detected on an abdominal ultrasound?
2	What should be done before an abdominal ultrasound?
3	What is examined with an abdominal ultrasound?
4	Can intestinal problems be seen on ultrasound?
5	How long does an abdominal ultrasound take?
6	Can you drink coffee before an abdominal ultrasound?
7	How much water should be drunk before an ultrasound?
8	Do you need to remove clothing during an ultrasound?
9	What happens if I do not drink water before the ultrasound?
10	Is the uterus visible on an abdominal ultrasound?
11	Does it matter if you are fasting or not for an ultrasound?
12	Can cancer be detected on an abdominal ultrasound?
13	Can you smoke before having an ultrasound?
14	Is an abdominal ultrasound harmful?
15	Why is an abdominal ultrasound ordered?

Note: Although shown in English for readability, all questions were originally submitted to ChatGPT-5 in Turkish

was assessed using the intraclass correlation coefficient (ICC), based on a two-way random-effects model with absolute agreement [ICC (2,k)], where k represents the number of raters. Ninety-five percent confidence intervals and p values were calculated for all ICC estimates. A p value <0.05 was considered statistically significant.

Results

Fifteen questions were assessed by 10 radiologists using four predefined criteria: relevance, accuracy, clarity, and completeness. All ratings were based on a 5-point scale. The lowest score assigned by the evaluators to any response was 3 out of 5 in each subcategory of the rating scale. Relevance had the highest overall mean score with a value of 4.97±0.18. The mean accuracy score was 4.78±0.49, while clarity had a mean score of 4.85±0.40. Completeness had the lowest mean among the four criteria with a value of 4.68±0.53. The total score calculated by averaging the four criteria for each evaluation was 4.82±0.26. Question-based descriptive values for all criteria are presented in Table 2. The distributions of mean scores and standard deviations for all four criteria are shown in Figure 2.

The interrater reliability analysis demonstrated varying levels of agreement among evaluators. ICC values were 0.266 for relevance, 0.236 for accuracy, 0.230 for clarity, 0.582 for completeness, and 0.555 for the total scores (Table 3). Completeness demonstrated the highest ICC value, followed by the total score; both values indicate moderate agreement, whereas relevance, accuracy, and clarity showed lower levels of agreement. Statistical significance was observed for completeness (p=0.001) and the total score (p=0.004). Relevance (p=0.144), accuracy (p=0.181), and clarity (p=0.194) did not reach statistical significance.

Discussion

This study evaluated the quality of ChatGPT-5 responses to frequently asked patient questions about abdominal ultrasonography. The analysis showed that ChatGPT-5 provided responses with favorable scores across all four evaluation criteria. Mean scores ranged from 4.68 to 4.97 on a 5-point scale, with an overall average of 4.82. Relevance received the highest scores, while completeness received the lowest scores. Interrater reliability showed variable levels of agreement, with ICC values ranging from 0.230 to 0.582. Agreement was higher for completeness and the total score, and lower for relevance, accuracy, and clarity. To our knowledge, this is the first study to evaluate LLM responses to patient questions specifically about abdominal ultrasonography. These findings suggest that while ChatGPT-5 can generate well-rated responses, variability remains in how medical experts evaluate these answers.

LLMs have increasingly been examined as tools for patient communication across different medical fields (17). A recent study comparing multiple LLMs for computed tomography (CT) and magnetic resonance imaging (MRI)-related patient questions found

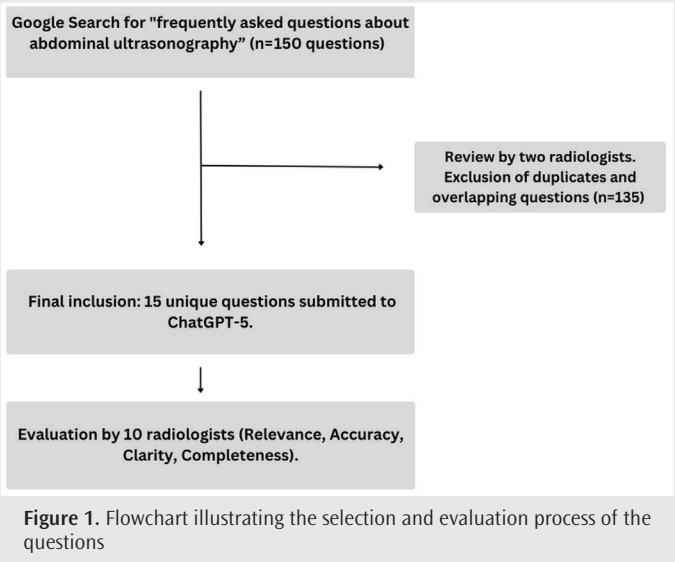

Figure 1. Flowchart illustrating the selection and evaluation process of the questions

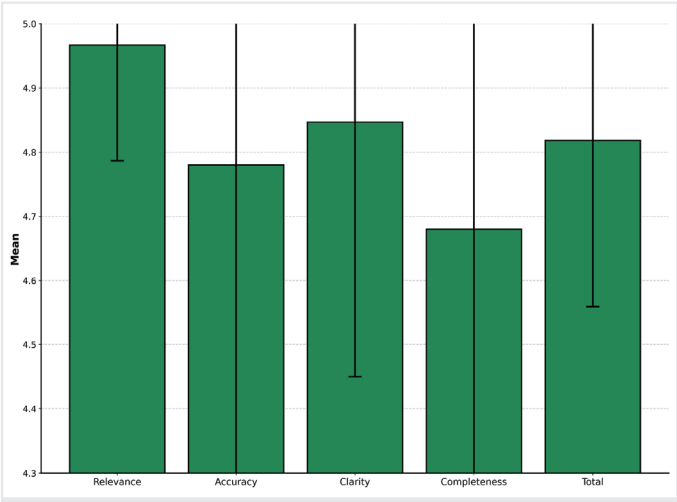

Figure 2. Mean scores and standard deviations for relevance, accuracy, clarity, completeness, and total scores

Table 2. Descriptive evaluation scores for each question across four assessment criteria

No	Relevance	Accuracy	Clarity	Completeness
1	4.9±0.32	4.7±0.48	4.8±0.42	4.3±0.67
2	5.0±0.00	4.8±0.42	4.9±0.32	4.9±0.32
3	5.0±0.00	5.0±0.00	4.9±0.32	4.3±0.67
4	4.8±0.42	4.4±0.70	4.5±0.85	4.3±0.82
5	5.0±0.00	4.6±0.84	4.9±0.32	4.7±0.48
6	5.0±0.00	4.6±0.70	5.0±0.00	4.7±0.48
7	4.9±0.32	4.8±0.42	4.9±0.32	4.8±0.42
8	5.0±0.00	4.8±0.42	4.9±0.32	4.5±0.53
9	5.0±0.00	4.9±0.32	5.0±0.00	4.9±0.32
10	4.9±0.32	4.9±0.32	4.6±0.52	4.9±0.32
11	5.0±0.00	5.0±0.00	4.9±0.32	4.8±0.42
12	5.0±0.00	4.7±0.67	4.9±0.32	4.5±0.71
13	5.0±0.00	4.9±0.32	4.9±0.32	4.9±0.32
14	5.0±0.00	4.9±0.32	4.9±0.32	4.9±0.32
15	4.9±0.32	4.9±0.32	4.9±0.32	4.8±0.42
Total	4.97±0.18	4.78±0.49	4.85±0.40	4.68±0.53

Values are presented as mean ± standard deviation. Each value represents the average score of 10 radiologists who evaluated the answers to 15 patient questions across four criteria: relevance, accuracy, clarity, and completeness

Table 3. ICCs, confidence intervals, and p values for all evaluation criteria

	ICC	Lower bound (95% CI)	Upper bound (95% CI)	p
Relevance	0.266	-0.220	0.730	0.144
Accuracy	0.236	-0.350	0.680	0.181
Clarity	0.230	-0.390	0.680	0.194
Completeness	0.582	0.240	0.830	0.001
Total	0.555	0.220	0.822	0.004

CI: Confidence interval, ICC: Intraclass correlation coefficient. ICC (2,k) model with absolute agreement was used, where k=10 raters.

that ChatGPT-4o achieved the highest scores for CT questions (mean 4.52) and shared the top position with Claude 3.5 Sonnet for MRI questions (both 4.79) (18). In cardiac imaging questions, ChatGPT-4 demonstrated 78.3% accuracy, with 86.7% of responses rated as clear and 81.7% as comprehensive (19). Similarly, Gordon et al. (20) found that 83% of ChatGPT responses to patient questions about radiological imaging were accurate, with 99% being at least partially relevant. However, performance varies with procedural complexity. While ChatGPT-3.5 adequately conveyed basic information for interventional radiology patient brochures, errors were infrequent for simpler procedures such as breast biopsy, whereas significant issues were noted for more complex topics including lung ablation and transarterial radioembolization (21). On questions about obstetric ultrasonography, both ChatGPT-3.5 and ChatGPT-4.0 answered 95% of questions correctly, significantly outperforming Microsoft Copilot (22). When evaluating responses to thyroid nodule questions, Campbell et al. (23) found 69.2% accuracy, though hallucination issues emerged when the model generated references.

Beyond radiology, studies in other specialties have also evaluated LLM performance. Physicians from 17 specialties rated ChatGPT responses with a median accuracy of 5.5 on a 6-point scale (24). In inflammatory bowel disease, ChatGPT responses showed no significant difference from physician responses in overall quality and accuracy, with superior completeness (25). For questions about anterior cruciate ligament surgery, both ChatGPT-4o and DeepSeek R1 demonstrated high accuracy with mean scores of 3.9 out of 4 (26). In contrast, breast reconstruction materials generated by ChatGPT-3.5 showed only 50% accuracy compared to expert-written content (27). In our study, ChatGPT-5 achieved an overall mean score of 4.82 across all evaluation criteria for abdominal ultrasonography questions, and scores for individual criteria ranged from 4.68 to 4.97. These findings are broadly consistent with previous research showing that LLMs generally perform well in answering patient-focused imaging questions. However, their reliability may vary depending on the clinical domain and the complexity of the information being evaluated.

The responses generated by ChatGPT-5 for patient questions about abdominal ultrasonography received generally high scores in this study. However, interrater agreement varied across the evaluation criteria and was higher for completeness and the total score but lower for relevance, accuracy, and clarity. This pattern has also been observed in studies evaluating LLMs across different medical fields. Hofmann et al. (28), who examined informed consent materials in interventional radiology, reported weak interrater agreement despite the high quality scores achieved by ChatGPT-4. Similarly, a study evaluating ChatGPT responses about celiac disease reported low reliability between two specialists (29). Studies in the field of orthopedics have also documented low agreement among evaluators, further supporting this observation (5,6). These findings suggest that variability in experts' evaluations may still occur when assessing text generated by LLMs, even when overall response quality is high.

Although the evaluators used the same scoring scales, aspects such as the perceived level of detail, tone of the responses, or subtle nuances in the content may have been interpreted differently. This should not be viewed as an error on the part of the evaluators. The ICC quantifies the proportion of total variance attributable to shared variance. For the ICC to yield a meaningful result, there must be sufficient variability in the scores (30). In this study, variability was limited for some of the evaluation criteria. For example, the mean relevance score of 4.97 indicates that the radiologists assigned a score of 5 to most of the questions. This limited variability may have influenced the ICC estimates, particularly for relevance, accuracy, and clarity, for which agreement remained lower despite generally high scores.

The high scores achieved by ChatGPT-5 in this study suggest that LLMs may serve as supportive tools for developing patient-education materials. Even so, the general limitations of these systems should be taken into account. Their training data may not fully reflect the most recent developments in medical practice, and they may occasionally generate information that is plausible yet incorrect (31). For this reason, model-generated responses cannot substitute for professional medical advice or individualized clinical judgment. Such tools are best used under expert oversight, particularly in settings where patients may rely heavily on the accuracy and clarity of the information provided. Given the widespread use of abdominal ultrasonography in routine clinical practice, a reliable and understandable supplementary resource may support patients' health literacy.

Study Limitations

This study has several limitations. The most important limitation is the lack of a patient perspective, as all evaluations were conducted solely by radiologists. In addition, no standardized readability or patient-oriented assessment tools were used, which may limit assessment of clarity from a patient perspective. A second limitation is that only a single LLM was examined, and the findings may differ when other models are included. Third, all patient questions were gathered from a single source, and broader sampling from platforms such as social media or patient forums might have provided greater diversity. In addition, the study included only Turkish-language questions, relied on responses collected at a single time point, and used a limited set of fifteen questions. Furthermore, each question was submitted only once, and potential variability in responses across repeated queries was not assessed. Given the stochastic nature of LLMs, repeated submissions of the same question may yield different responses, which could influence evaluation scores. Therefore, the results of this study should be interpreted as reflecting a single instance of model output rather than as a comprehensive assessment of response variability. Moreover, some of the patient questions used in this study were derived

from publicly available online sources, some of which may have been included in the model's training data. This potential overlap could have influenced the responses generated by the model. In addition, the scoring system was adapted from previously published studies and was not formally validated for the assessment of radiology-specific patient information. Furthermore, no predefined reference standard was used as the accuracy criterion, and evaluations were based on the radiologists' clinical judgment, which may introduce subjectivity. The results are specific to abdominal ultrasonography and may not be generalizable to other imaging modalities or types of ultrasonography. Despite these limitations, this study provides useful preliminary insights into the performance of LLMs in addressing common patient questions about abdominal ultrasonography.

Future studies could compare multiple LLMs to determine whether performance differs across platforms. Incorporating patient perspectives and standardized readability assessments will be important for better evaluation of response quality from the end-user perspective. Expanding the question set through broader sources such as patient forums, social media platforms, or direct clinical encounters may improve representativeness. Evaluating model performance across different languages, including English and Turkish, could also provide insight into the effect of language on response quality. In addition, larger question sets and longitudinal designs that assess consistency over time and across model updates may strengthen the evidence base. Future research may also benefit from repeated submissions of the same questions to assess response variability and improve the robustness of performance evaluation. The use of validated assessment tools, such as DISCERN or PEMAT, and the development of consensus-based reference standards for accuracy evaluation may further enhance methodological rigor. Applying similar methods to other imaging modalities, such as MRI, CT, or alternative types of ultrasonography, could help determine the generalizability of these findings across radiology subspecialties.

Conclusion

ChatGPT-5 provided responses to frequently asked patient questions about abdominal ultrasonography, which were rated favorably with consistently positive scores for relevance, accuracy, clarity, and completeness. Although the model performed well, interrater reliability showed variable levels of agreement, with moderate agreement observed for completeness and total scores. The findings indicate that LLMs may be valuable as tools to support patient education. At the same time, information generated by these systems should be viewed as complementary rather than a substitute for professional medical guidance. Integrating such tools into clinical communication may offer benefits but their use should be approached with caution. Further studies involving diverse patient groups, multiple models and broader imaging contexts will help clarify the potential role of LLMs in medical settings.

Ethics

Ethics Committee Approval: Ethical approval for this study was obtained from the İzmir Katip Çelebi University Health Research Ethics Committee (decision number: 0670, date: 06.11.2025).

Informed Consent: Due to the use of publicly available data and the absence of direct human participant involvement or identifiable patient information, the requirement for informed consent was waived by the ethics committee.

Footnotes

Authorship Contributions: Concept - A.S., G.M.T., A.D.B., R.E.B., A.H.Ç.; Design - A.S., R.E.B., A.M.K., A.H.Ç.; Data Collection or Processing - G.M.T., A.D.B., R.E.B.; Analysis or Interpretation - A.S., A.D.B., A.M.K.; Literature Search - A.S., G.M.T., A.M.K., A.H.Ç.; Writing - A.S., A.M.K., A.H.Ç.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, et al. A comprehensive overview of large language models. ACM Trans Intell Syst Technol. 2025; 16: 106:1-72.
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023; 29: 1930-40.
3. Artsi Y, Klang E, Collins JD, Glicksberg BS, Nadkarni GN, Korfiatis P, et al. Large language models in radiology reporting: a systematic review of performance, limitations, and clinical implications. Intell Based Med. 2025; 12: 100287.
4. Horiuchi D, Tatekawa H, Oura T, Oue S, Walston SL, Takita H, et al. Comparing the diagnostic performance of GPT-4-based ChatGPT, GPT-4V-based ChatGPT, and radiologists in challenging neuroradiology cases. Clin Neuroradiol. 2024; 34: 779-87.
5. Ayık G, Ercan N, Demirtaş Y, Yıldırım T, Çakmak G. Evaluation of ChatGPT-4o's answers to questions about hip arthroscopy from the patient perspective. Jt Dis Relat Surg. 2025; 36: 193-9.
6. Magruder ML, Rodriguez AN, Wong JCJ, Erez O, Piuzzi NS, Scuderi GR, et al. Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. J Arthroplasty. 2024; 39: 2022-7.
7. Republic of Türkiye Ministry of Health GD of HIS. Health Statistics Yearbook. Available at: https://www.saglik.gov.tr/TR-84930/saglik-istatistikleri-yilliklari.html.
8. Forshaw KL, Boyes AW, Carey ML, Hall AE, Symonds M, Brown S, et al. Raised anxiety levels among outpatients preparing to undergo a medical imaging procedure: prevalence and correlates. J Am Coll Radiol. 2018; 15: 630-8.
9. Fu Y, Han P, Wang J, Shahzad F. Digital pathways to healthcare: a systematic review for unveiling the trends and insights in online health information-seeking behavior. Front Public Health. 2025; 13: 1497025.
10. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025; 25: 117.
11. Einspänner E, Schwab R, Hupfeld S, Thormann M, Fuchs E, Gawlitza M, et al. Evaluating the role of large language models in supporting patient education during the informed consent process for routine radiology procedures. Br J Radiol. 2025; 98: 2184-90.
12. Khaldi A, Machayekhi S, Salvagno M, Maniaci A, Vaira LA, La Via L, et al. Accuracy of ChatGPT responses on tracheotomy for patient education. Eur Arch Otorhinolaryngol. 2024; 281: 6167-72.
13. Google Search Engine [Internet]. Mountain View (CA): Alphabet Inc.; [cited 2025 Nov 18]. Available from: https://www.google.com
14. OpenAI. ChatGPT [Internet]. San Francisco (CA): OpenAI. Available from: https://openai.com/chatgpt/

15. OpenAI. GPT-5 system card [Internet]. Available from: https://cdn.openai.com/gpt-5-system-card.pdf

16. Salbas A, Baysan EK. Assessment of large language models in musculoskeletal radiological anatomy: a comparative study with radiologists. Jt Dis Relat Surg. 2026; 37: 190-9.

17. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024; 11: 1477898.

18. Eminovic S, Levita B, Dell'Orco A, Leppig JA, Nawabi J, Penzkofer T. Comparison of multiple state-of-the-art large language models for patient education prior to CT and MRI examinations. J Pers Med. 2025; 15: 235.

19. Marey A, Saad AM, Tanas Y, Ghorab H, Niemierko J, Backer H, et al. Evaluating the accuracy and reliability of AI chatbots in patient education on cardiovascular imaging: a comparative study of ChatGPT, Gemini, and Copilot. Egypt J Radiol Nucl Med. 2025; 56: 37.

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.

21. Kooraki S, Hosseiny M, Jalili MH, Rahsepar AA, Imanzadeh A, Kim GH, et al. Evaluation of ChatGPT-generated educational patient pamphlets for common interventional radiology procedures. Acad Radiol. 2024; 31: 4548-53.

22. Du Y, Ji C, Xu J, Wei M, Ren Y, Xia S, et al. Performance of ChatGPT and Microsoft Copilot in Bing in answering obstetric ultrasound questions and analyzing obstetric ultrasound reports. Sci Rep. 2025; 15: 14627.

23. Campbell DJ, Estephan LE, Sina EM, Mastrolonardo EV, Alapati R, Amin DR, et al. Evaluating ChatGPT responses on thyroid nodules for patient education. Thyroid. 2024; 34: 371-7.

24. 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.

25. Yan Z, Liu J, Fan Y, Lu S, Xu D, Yang Y, et al. Ability of ChatGPT to replace doctors in patient education: cross-sectional comparative analysis of inflammatory bowel disease. J Med Internet Res. 2025; 27:e62857.

26. Gültekin O, Inoue J, Yilmaz B, Cerci MH, Kilinc BE, Yilmaz H, et al. Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. Knee Surg Sports Traumatol Arthrosc. 2025; 33: 3025-31.

27. Hung YC, Chaker SC, Sigel M, Saad M, Slater ED. Comparison of patient education materials generated by chat generative pre-trained transformer versus experts: an innovative way to increase readability of patient education materials. Ann Plast Surg. 2023; 91: 409-12.

28. Hofmann HL, Vairavamurthy J. Large language model doctor: assessing the ability of ChatGPT-4 to deliver interventional radiology procedural information to patients during the consent process. CVIR Endovasc. 2024; 7: 83.

29. Mahmoudi Ghehsareh M, Asri N, Azizmohammad Looha M, Sadeghi A, Ciacci C, Rostami-Nejad M. Expert evaluation of ChatGPT accuracy and reliability for basic celiac disease frequently asked questions. Sci Rep. 2025; 15: 29871.

30. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15: 155-63. Erratum in: J Chiropr Med. 2017; 16: 346.

31. Salbas A, Buyuktoka RE. Performance of large language models in recognizing brain MRI sequences: a comparative analysis of ChatGPT-4o, Claude 4 Opus, and Gemini 2.5 Pro. Diagnostics (Basel). 2025; 15: 1919.