

Comparison of ChatGPT-5.2 and DeepSeek-V3.2 responses to frequently asked patient questions about hallux valgus surgery

 Veysel Emre Çelik^{*1},  Ferhat Öktem¹,  Mümin Karahan²,  Berkay Kırçal¹,
 Ahmet Furkan Taşkın¹,  Hüseyin Olgun¹

¹Department of Orthopedics and Traumatology, Kocaeli City Hospital, University of Health Sciences, Kocaeli, Türkiye

²Department of Orthopedics and Traumatology, Faculty of Medicine, Kafkas University, Kars, Türkiye

Cite this article as: Çelik VE, Öktem F, Karahan M, Kırçal B, Taşkın AF, Olgun H. Comparison of ChatGPT-5.2 and DeepSeek-V3.2 responses to frequently asked patient questions about hallux valgus surgery. *J Orthop Res Rehabil.* 2026;4(4):101-105. doi:10.51271/JORR-0080

Received: 11.09.2026

Accepted: 26.10.2026

Published: 28.10.2026

ABSTRACT

Aims: To compare ChatGPT-5.2 and DeepSeek-V3.2 responses to frequently asked patient questions about hallux valgus surgery in terms of accuracy, informativeness, comprehensibility, and currency.

Methods: A Google search performed with a newly created account without previous search activity was used to identify patient-oriented questions about hallux valgus surgery. Two researchers reviewed 100 questions and consolidated overlapping items into 10 representative questions. On February 1, 2026, the same questions were submitted in English to ChatGPT-5.2 and DeepSeek-V3.2. The first response from each model was recorded without editing. Ten orthopedic specialists with at least five years of professional experience independently and blindly rated each response on a 5-point Likert scale for accuracy, informativeness, comprehensibility, and currency. For each question, the 10 ratings were averaged, and question-level mean scores (n=10 paired observations per domain) were compared using the Wilcoxon signed-rank test.

Results: Both models answered all 10 questions. ChatGPT-5.2 had numerically higher mean scores for accuracy (4.7 ± 0.3 vs 4.6 ± 0.4 ; $p=0.214$), informativeness (4.6 ± 0.4 vs 4.4 ± 0.5 ; $p=0.081$), comprehensibility (4.6 ± 0.4 vs 4.5 ± 0.4 ; $p=0.197$), and currency (4.5 ± 0.4 vs 4.3 ± 0.5 ; $p=0.064$). None of the between-model differences reached statistical significance.

Conclusion: ChatGPT-5.2 and DeepSeek-V3.2 both generated responses that orthopedic specialists rated highly across the four evaluated domains. Neither model demonstrated statistically significant superiority; however, this exploratory study was not designed to test equivalence, and the absence of significant differences should not be interpreted as evidence that the models perform comparably. These systems may support patient education as supplementary information tools, but their outputs should remain subject to clinical verification and should not replace individualized patient-physician communication.

Keywords: Hallux valgus, patient education as topic, large language models, health literacy, comprehension, generative artificial intelligence

INTRODUCTION

Hallux valgus is a common forefoot deformity whose prevalence rises with age and is consistently higher among females. A recent global meta-analysis estimated an overall prevalence of approximately 19%, emphasizing the substantial population burden of the condition.¹

For patients considering surgery, the decision extends beyond radiographic correction. Expectations regarding pain relief, footwear, recovery time, recurrence, complications, cosmetic appearance, and return to activity all influence the treatment experience. Contemporary expert work has therefore emphasized shared decision-making and patient-centered priorities, while qualitative research shows that people with painful hallux valgus may struggle to obtain information that is both trustworthy and relevant to their lived concerns.^{2,3}

Surgical management is also difficult to summarize for a lay audience because numerous procedures and technique variations are available. Systematic reviews indicate that surgery can improve pain and function, yet the certainty of evidence and comparative advantages of individual techniques vary. Open and minimally invasive approaches may provide broadly comparable outcomes in selected patients, but differences in technique, rehabilitation, and complication profiles make individualized counseling essential.^{4,5}

Patients increasingly seek this counseling outside the clinic. Online resources devoted to hallux valgus have previously shown variable accuracy, quality, and readability, creating a gap between information availability and information usability.⁶

*Corresponding Author: Veysel Emre Çelik, veyselrecelik@hotmail.com

That gap is clinically relevant because patient education in hallux valgus is not limited to naming an operation. A patient must understand why surgery is being considered, what alternatives exist, which expectations are realistic, and which postoperative restrictions are temporary or potentially persistent. Information that is technically correct but overly generic, excessively complex, or insufficiently qualified may therefore have limited value in shared decision-making. The communication task is especially demanding when the answer depends on deformity severity, procedure choice, rehabilitation protocol, activity goals, or surgeon-specific practice.

Large language models (LLMs) offer a different form of access because users can ask questions in natural language and receive immediate, conversational responses. Unlike static webpages, an LLM can adapt the wording and scope of an answer to the user's question, which may make it attractive for preoperative information seeking. At the same time, the same flexibility can produce variation in depth, emphasis, and certainty across prompts. Hallux valgus-specific work has already shown that LLM answers can be detailed but may exceed ideal reading levels.⁷

More recent studies have extended this question to newer models. ChatGPT-5.2 and Gemini produced similarly high-quality responses about minimally invasive bunion surgery but remained limited in actionability and reading accessibility, while simulated hallux valgus cases showed progressively improved surgical recommendation consistency across newer ChatGPT versions.^{8,9}

Outside foot and ankle surgery, studies have likewise found that generative AI can produce clinically useful patient-facing information, sometimes approaching physician-rated quality, although limitations in reliability and contextual judgment remain.^{10,11}

In orthopedics, direct comparisons between ChatGPT and DeepSeek have expanded rapidly. Studies involving hand fractures, anterior cruciate ligament surgery, meniscus injuries, rotator cuff disorders, spinal surgery, and knee osteoarthritis show that relative performance varies by clinical topic: one model may be more comprehensive while another is clearer or easier to read.¹²⁻¹⁷ Complementary work on orthopaedic spine and oncology patient-education materials further suggests that LLMs can improve readability without necessarily sacrificing accuracy, although the balance between simplification, completeness, and comprehension remains model- and task-dependent.^{18,19}

The ability of LLMs to answer specialist-level orthopedic questions has also been demonstrated in examination settings, but good performance on knowledge tasks does not necessarily establish suitability for patient counseling.²⁰

Against this background, a direct comparison of ChatGPT-5.2 and DeepSeek-V3.2 using the same frequently asked hallux valgus surgery questions may clarify whether either model offers a meaningful advantage for patient-facing information. The aim of this exploratory study was to compare the two models in accuracy, informativeness, comprehensibility, and currency as judged by experienced orthopedic specialists. The study hypothesis was that the two models would differ significantly in at least one evaluation domain.

METHODS

Ethics

No patients, patient records, biological samples, identifiable personal data, or personal health information were used. The orthopedic specialists participated solely as voluntary professional evaluators of anonymized AI-generated responses, and no personal, sensitive, or clinical data were collected from them. No institutional ethics committee approval or formal exemption was obtained for this study. The study was conducted in accordance with the relevant ethical principles and publication ethics standards.

Study Design and Question Selection

This cross-sectional comparative study evaluated responses generated by two LLMs to frequently asked questions about hallux valgus surgery. The study was designed to assess the quality of model-generated information rather than to test diagnosis, treatment decisions, or outcomes in individual patients.

To identify questions that reflect common public information needs, a newly created Google account with no previous search history was used. The search was performed on February 1, 2026 using the single keyword "hallux valgus" in the Google search engine (Google LLC, Mountain View, CA, USA) from Kocaeli, Türkiye, with the interface language set to English. Questions appearing in the "People also ask" section were progressively expanded and collected until a total of 100 questions had been recorded.

Two researchers experienced in hallux valgus surgery independently reviewed the initial pool of 100 questions obtained from Google's "People also ask" section. No exclusion criteria were applied during the initial collection of questions. Questions addressing identical or substantially overlapping topics were grouped according to their content. Disagreements regarding question selection were resolved through discussion until consensus was reached between the two researchers. Through this process, 10 representative questions covering the main aspects of hallux valgus surgery, including surgical indications, operative techniques, postoperative recovery, pain, recurrence, footwear, cosmetic expectations, complications, return to physical activity, and postoperative joint mobility, were selected (**Table 1**). This approach was intended to evaluate responses to commonly encountered online questions rather than to represent all possible clinical scenarios associated with hallux valgus surgery.

Artificial Intelligence Models and Query Procedure

The models evaluated were ChatGPT-5.2 (OpenAI, San Francisco, CA, USA) and DeepSeek-V3.2 (Hangzhou DeepSeek Artificial Intelligence Co., Ltd., Hangzhou, China). Separate newly created user accounts without prior interaction history were used for each platform. ChatGPT-5.2 was accessed through the web interface using a ChatGPT Plus subscription in Thinking mode. DeepSeek-V3.2 was accessed through its web interface with the DeepThink feature enabled. Web search was disabled on both platforms. Each question was submitted in a separate, newly initiated chat session to minimize potential contextual influences from previous questions. On February 1, 2026, all 10 questions were

Table 1. Ten representative patient questions about hallux valgus surgery selected from Google's "People also ask" section.

Question no	Question
1	Is hallux valgus surgery necessary? When should I have surgery?
2	How is hallux valgus surgery performed? What surgical methods are there?
3	How long does the recovery process after surgery take?
4	Is there pain after surgery?
5	Is there a risk of recurrence after hallux valgus surgery?
6	When can I start wearing shoes after surgery?
7	What are the aesthetic (cosmetic) results? Will the foot appearance improve?
8	What are the risks and complications of hallux valgus surgery?
9	Can I do sports or exercise after surgery?
10	Is there a permanent limitation or loss of movement after hallux valgus surgery?

entered in English using identical wording for both models. Questions were submitted at different times throughout the day rather than in immediate succession.

The models were not instructed to cite sources, use a particular format, or restrict response length. The first response generated for each question was saved. Regeneration was not used, and no wording was added to, removed from, or otherwise edited within the model outputs before evaluation.

Expert Evaluation of Responses

Ten orthopedic and traumatology specialists with clinical expertise in foot surgery and at least five years of experience after completion of specialty training independently reviewed the 20 model responses. The evaluators were recruited from the same hospital through direct personal invitations from the researchers. None of the evaluators was a co-author of the study, and all participated voluntarily after providing verbal consent.

Before evaluation, model names and other model-identifying information were removed from the responses. All responses were converted into plain text, with headings, bullet points, bold formatting, and other distinctive formatting elements removed to minimize potential identification of the models. The anonymized responses were presented to the evaluators in Microsoft Word documents, and the presentation order of the two models' responses was varied across questions to reduce potential order-related bias. All evaluators assessed the responses independently, without any consensus meetings or joint scoring sessions.

Each response was rated on a 5-point Likert scale, where 1 indicated very poor performance and 5 indicated very good performance, across four predefined domains: accuracy (medical correctness), informativeness (adequacy of scope and detail), comprehensibility (clarity for a reader without professional medical training), and currency (alignment with contemporary clinical knowledge). No external reference standard was specified for the assessment of information currency. The rating scale was developed specifically for this study and had not undergone formal validation. Each specialist therefore scored responses from both models to all 10 questions in each domain.

Statistical Analysis

The data analyses were performed with IBM SPSS Statistics version 28.0 (IBM Corp., Armonk, NY, USA). For each question, the scores assigned by the 10 evaluators were averaged separately for each model and domain, and the question was used as the unit of analysis ($n=10$ paired observations per domain). Likert scores were summarized as mean and standard deviation to facilitate descriptive comparison with prior LLM studies. Because the same evaluators assessed responses from both models to corresponding questions, model comparisons were treated as paired. Given the ordinal scale of the ratings, the Wilcoxon signed-rank test was used. No a priori sample-size calculation was performed, and no equivalence margin was prespecified; the analysis was therefore exploratory. Tests were two-sided, and $p<0.05$ was considered statistically significant.

RESULTS

All 10 selected questions were answered by both LLMs, yielding 20 responses for expert assessment. The questions are shown in **Table 1**. On the investigators' qualitative reading, ChatGPT-5.2 generally produced longer and more elaborated answers, whereas DeepSeek-V3.2 tended to provide shorter responses; response length was not quantified and was not used as an outcome measure.

Both models received high scores in all four domains (**Table 2**). Mean accuracy was 4.7 ± 0.3 for ChatGPT-5.2 and 4.6 ± 0.4 for DeepSeek-V3.2, with no significant difference ($p=0.214$). Mean informativeness scores were 4.6 ± 0.4 and 4.4 ± 0.5 , respectively ($p=0.081$).

Table 2. Expert evaluation scores for ChatGPT-5.2 and DeepSeek-V3.2 responses.

Evaluation criterion	Model	Mean	SD*	p-value
Accuracy	ChatGPT-5.2	4.7	0.3	0.214
	DeepSeek-V3.2	4.6	0.4	
Informativeness	ChatGPT-5.2	4.6	0.4	0.081
	DeepSeek-V3.2	4.4	0.5	
Comprehensibility	ChatGPT-5.2	4.6	0.4	0.197
	DeepSeek-V3.2	4.5	0.4	
Currency	ChatGPT-5.2	4.5	0.4	0.064
	DeepSeek-V3.2	4.3	0.5	

*SD: Standard deviation. Ratings used a 1–5 Likert scale (1=very poor; 5=very good). Means and SDs were calculated from question-level mean scores ($n=10$ questions, each rated by 10 evaluators). p-values were calculated with the Wilcoxon signed-rank test for paired model scores.

Comprehensibility was rated 4.6 ± 0.4 for ChatGPT-5.2 and 4.5 ± 0.4 for DeepSeek-V3.2 ($p=0.197$). Currency scores were 4.5 ± 0.4 and 4.3 ± 0.5 , respectively ($p=0.064$). Although ChatGPT-5.2 had a numerically higher mean score in each domain, none of the paired comparisons reached the predefined threshold for statistical significance. All mean scores lay between 4.3 and 4.7 on the 5-point scale.

DISCUSSION

The principal finding was that both ChatGPT-5.2 and DeepSeek-V3.2 generated hallux valgus surgery answers that experienced orthopedic specialists rated highly. ChatGPT-5.2 was numerically higher in all four domains, but the differences were statistically non-significant. The

data therefore do not support a claim of superiority for either model. Equally, they do not demonstrate equivalence: the study was not powered for small between-model differences, and no equivalence margin was prespecified.

These findings should be interpreted in the context of rapidly evolving foot and ankle AI literature. Hlavinka et al.⁷ found that LLMs could address common hallux valgus concerns but that readability remained a limitation. A more recent study of minimally invasive bunion surgery compared ChatGPT-5.2 with Gemini and found no significant difference in overall information quality, while actionability and reading complexity remained problematic.⁸ Our study extends this literature by evaluating DeepSeek-V3.2 against ChatGPT-5.2 across broader hallux valgus surgery questions and by using blinded ratings from 10 orthopedic specialists.

Hallux valgus counseling is particularly suitable for this type of evaluation because patients commonly ask questions that involve uncertainty rather than a single factual answer. Recovery duration, return to shoes or exercise, recurrence, residual stiffness, and complication risk vary with deformity severity, procedure selection, surgeon preference, and patient-specific factors. Evidence comparing operative strategies confirms that multiple procedures can be effective and that no single approach can be presented as universally optimal.^{4,5} Consequently, a useful patient-education response should not only be factually correct; it should also communicate conditionality and avoid presenting population-level information as individualized advice.

The broader orthopedic literature also suggests that model strengths are domain-dependent. ChatGPT and DeepSeek have shown different profiles of comprehensiveness, clarity, and readability in hand fracture, anterior cruciate ligament, meniscus, and rotator cuff education,¹²⁻¹⁵ and studies of spinal surgery and knee osteoarthritis indicate that readability and information quality do not necessarily move together.^{16,17} LLM-assisted simplification of orthopaedic materials can improve readability while largely preserving accuracy, but performance varies with the model, prompt, and target reading level.^{18,19}

The consistently high ratings in our study are also compatible with the broader ability of LLMs to generate medically coherent information.^{10,20} However, knowledge performance should not be equated with clinical decision-making, and persuasive language can mask error: LLMs can produce unsupported claims and fabricated references, and reviews emphasize variable reliability and the need for human verification.^{21,22} In the present study the models were deliberately not asked to provide references, so citation accuracy and evidence hierarchy were not tested. A high expert score therefore indicates that a response appeared clinically appropriate within the study framework; it does not establish that every statement was traceable to a high-quality source.

The high scores also need to be read alongside the rating instrument. With means of 4.3–4.7 on a 5-point scale, ratings clustered near the upper limit, and a ceiling effect may have compressed real differences between the models. Together with the unvalidated, endpoint-anchored scale, this further limits the ability of the study to discriminate between two

high-performing systems. The near-significant differences in informativeness ($p=0.081$) and currency ($p=0.064$) may reflect a tendency for ChatGPT-5.2 to provide longer answers, but response length could not be quantified, so this explanation remains speculative. A longer answer can increase perceived completeness while creating cognitive burden; a shorter response may be easier to follow yet omit important caveats. Future comparisons should therefore combine expert ratings with validated instruments such as DISCERN and PEMAT-P,^{23,24} objective length and readability measures, inter-rater reliability statistics, and patient-based assessment.

From a clinical perspective, the most defensible role for these tools is supplementary. An LLM may help patients formulate questions, review general concepts before or after a consultation, and revisit information in accessible language. It cannot examine the foot, interpret weight-bearing radiographs, determine deformity severity, incorporate comorbidities or patient priorities reliably, or select a procedure with the accountability expected of a treating surgeon. The model response should therefore be treated as an informational aid that remains subordinate to individualized clinical assessment and shared decision-making.^{2,11}

Limitations

Several limitations should be considered. First, the evaluation was restricted to 10 frequently asked questions and may not represent uncommon, technically complex, or patient-specific scenarios; the small number of questions also limited statistical power, and no sample-size calculation was performed. Second, the question set was derived from a Google “People also ask” search; search-engine results are dynamic and can vary with time, region, language, and platform configuration, and the reduction from 100 to 10 questions may have introduced selection bias. Third, the responses were generated once for each question, so within-model variability and reproducibility across repeated prompts were not examined. Fourth, the expert scoring system, although blinded and clinically intuitive, was a study-specific, unvalidated scale with anchors defined only at its endpoints and without an operational reference standard for currency; ratings clustered near the ceiling of the scale. Fifth, comprehensibility was judged by orthopedic specialists rather than by patients, and objective readability metrics and response length were not calculated; because the archived responses are no longer available, these measures cannot be derived retrospectively, and question-level scores and the content of the lowest-rated responses cannot be reported. Sixth, the models were not asked to provide citations, preventing assessment of source quality, reference accuracy, or evidence hierarchy. Finally, LLMs are updated frequently; the findings apply specifically to the ChatGPT-5.2 and DeepSeek-V3.2 versions accessed on February 1, 2026, under the settings described above, and may not generalize to later model iterations.

CONCLUSION

ChatGPT-5.2 and DeepSeek-V3.2 both produced responses to common hallux valgus surgery questions that received high expert ratings for accuracy, informativeness,

comprehensibility, and currency. ChatGPT-5.2 scored numerically higher in every domain, but no comparison reached statistical significance. Because this exploratory study was small, used an unvalidated scale with a ceiling effect, and was not designed to test equivalence, the results neither establish superiority nor show that the models perform comparably. These findings support the potential use of current LLMs as adjunctive patient-education tools while reinforcing the need for clinical verification, appropriate health-literacy assessment, and preservation of patient-physician communication as the basis of individualized surgical counseling.

ETHICAL DECLARATIONS

Ethics Committee Approval

The study did not involve patients, patient records, identifiable personal data, biological materials, or any clinical intervention. The orthopedic specialists served solely as voluntary professional evaluators of AI-generated responses. No institutional ethics committee approval or formal exemption was obtained.

Informed Consent

Patient informed consent was not applicable because no patients or identifiable patient data were included in the study. The orthopedic specialists participated voluntarily as professional evaluators and provided verbal consent to participate.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: VEÇ, FÖ, MK; Design: VEÇ, FÖ; Data Collection and/or Processing: VEÇ, FÖ, BK; Analysis and/or Interpretation: VEÇ, BK, AFT, HO; Literature Review: VEÇ, FÖ, MK; Writing the Article: VEÇ, FÖ, AFT; Critical Review: VEÇ, FÖ, MK, BK, AFT, HO.

REFERENCES

- Cai Y, Song Y, He M, et al. Global prevalence and incidence of hallux valgus: a systematic review and meta-analysis. *J Foot Ankle Res.* 2023; 16(1):63. doi:10.1186/s13047-023-00661-9
- Lewis TL, See A, Thomas M, et al. The inaugural UK National Hallux Valgus Think Tank: identification of key issues and strategies to improve clinical care for patient benefit. *Bone Jt Open.* 2025;6(4):432-439. doi:10.1302/2633-1462.64.BJO-2024-0212.R1
- Cotchett M, Bramston C, Bergin S, Menz HB, Jessup R. Lived experience of people with painful hallux valgus: a descriptive qualitative study. *Musculoskeletal Care.* 2023;21(4):1421-1428. doi:10.1002/msc.1822
- Dias CGP, Godoy-Santos AL, Ferrari J, Ferretti M, Lenza M. Surgical interventions for treating hallux valgus and bunions. *Cochrane Database Syst Rev.* 2024;7(7):CD013726. doi:10.1002/14651858.CD013726.pub2
- Ji L, Wang K, Ding S, Sun C, Sun S, Zhang M. Minimally invasive vs open surgery for hallux valgus: a meta-analysis. *Front Surg.* 2022;9: 843410. doi:10.3389/fsurg.2022.843410
- Tartaglione JP, Rosenbaum AJ, Abousayed M, Hushmendi SF, DiPreta JA. Evaluating the quality, accuracy, and readability of online resources pertaining to hallux valgus. *Foot Ankle Spec.* 2016;9(1):17-23. doi:10.1177/1938640015592840
- Hlavinka WJ, Sontam TR, Gupta A, Croen BJ, Abdullah MS, Humbyrd CJ. Are large language models a useful resource to address common patient concerns on hallux valgus? A readability analysis. *Foot Ankle Surg.* 2025;31(1):15-19. doi:10.1016/j.fas.2024.08.002
- Yüksel B, Uçar SB, Özdemir E, Ayık G, Kaymakoglu M, Huri G. Can artificial intelligence provide reliable patient education in minimally invasive bunion surgery? A comparative study of ChatGPT 5.2 and Gemini. *Foot Ankle Surg.* Published online June 16, 2026. doi:10.1016/j.fas.2026.06.006
- Karadamar ÖL, Aydılek A. Evaluation of surgical recommendation algorithms in hallux valgus cases using ChatGPT models: an artificial intelligence approach based on 50 simulated scenarios. *Acta Orthop Traumatol Turc.* 2026;60:e25580. doi:10.5152/j.aott.2026.25580
- Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med.* 2023;388(13):1233-1239. doi:10.1056/NEJMSr2214184
- Marcaccini G, Seth I, Xie Y, et al. Breaking bones, breaking barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. *J Clin Med.* 2025;14(6):1983. doi:10.3390/jcm14061983
- Gültekin O, Inoue J, Yilmaz B, 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(8):3025-3031. doi:10.1002/ksa.12711
- Bozgeyik B, Ögümsöğütü E. Can artificial intelligence educate patients? Comparative analysis of ChatGPT and DeepSeek models in meniscus injuries. *Healthcare (Basel).* 2025;13(22):2980. doi:10.3390/healthcare13222980
- Bozgeyik B, Ögümsöğütü E, Huri G. artificial intelligence-based patient information in rotator cuff injuries: a cross-sectional comparative study of ChatGPT and DeepSeek models. *Medicine (Baltimore).* 2026;105(27):e49658. doi:10.1097/MD.00000000000049658
- Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. *Int J Med Inform.* 2025;198:105871. doi:10.1016/j.ijmedinf.2025.105871
- Öner SK, Demirkiran ND, Canlı EA, Bilir A. Gonarthrosis Advisor vs ChatGPT-5: quality and readability of artificial intelligence-generated patient education for knee osteoarthritis. *Acta Orthop Traumatol Turc.* 2026;60(2):e25618. doi:10.5152/j.aott.2026.25618
- Romoff M, Brunette M, Peterson MK, Hashmi SZ, Kim MS. The role of large language models in improving the readability of orthopaedic spine patient educational material. *J Orthop Surg Res.* 2025;20(1):531. doi:10.1186/s13018-025-05955-1
- Nian PP, Williams CJ, Baidya J, Marsh IG, Senthilnathan IS, Maheshwari AV. Evaluation of commonly utilized artificial intelligence large language models in optimizing readability, accuracy, and comprehension of orthopaedic oncology patient educational materials. *J Am Acad Orthop Surg.* Published online May 26, 2026. doi:10.5435/JAAOS-D-25-00883
- Pamuk Ç, Uyanık AF, Kuyucu E, Uğurlar M. Can ChatGPT pass the Turkish Orthopedics and Traumatology Board Examination? Turkish orthopedic surgeons versus artificial intelligence. *Ulus Travma Acil Cerrahi Derg.* 2025;31(3):310-315. doi:10.14744/tjtes.2025.07724
- Gravel J, D'Amours-Gravel M, Osmanliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin Proc Digit Health.* 2023;1(3):226-234. doi:10.1016/j.mcpdig.2023.05.004
- Keshavarz P, Bagherieh S, Nabipoorashrafi SA, et al. ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives. *Diagn Interv Imaging.* 2024;105(7-8):251-265. doi:10.1016/j.diii.2024.04.003
- 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):105-111. doi:10.1136/jech.53.2.105
- Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient Educ Couns.* 2014;96(3):395-403. doi:10.1016/j.pec.2014.05.027