Joint Diseases and Related Surgery

Jt Dis Relat Surg
2026;37(2):470-476

ORIGINAL ARTICLE

Evaluating deepresearch and deepthink in total knee arthroplasty patient education: ChatGPT-4o excels in comprehensiveness, Deepseek R1 leads in clarity and readability of orthopedic information

Onur Gultekin, MD[1], Michael T. Hirschmann, MD[2,3], Halil İbrahim Arıkan, MD[4], Bekir Eray Kilinc, MD[1], Baris Yilmaz, MD[1], Süleyman Abul, MD[5], Jumpei Inoue, MD[6], Mahmut Enes Kayaalp, MD[1,7,8]

[1]Department of Orthopedics and Traumatology, University of Health Sciences, İstanbul Fatih Sultan Mehmet Training and Research Hospital, İstanbul, Türkiye
[2]Department of Orthopedic Surgery and Traumatology, Kantonsspital Baselland, Bruderholz, Switzerland
[3]Department of Clinical Research, Research Group Michael T. Hirschmann, Regenerative Medicine & Biomechanics, University of Basel, Basel, Switzerland
[4]Department of Orthopedics and Traumatology, Viranşehir State Hospital, Şanlıurfa, Türkiye
[5]Department of Orthopedics and Traumatology, İstanbul Kartal Dr. Lütfi Kırdar City Hospital, İstanbul, Türkiye
[6]Department of Orthopaedic Surgery, Nagoya Tokushukai General Hospital, Kasugai, Aichi, Japan
[7]Department of Orthopaedics and Traumatology, University Hospital Brandenburg/Havel, Brandenburg Medical School Theodor Fontane, Brandenburg/Havel, Germany
[8]Brandenburg Medical School Theodor Fontane, Faculty of Health Sciences Brandenburg, Brandenburg/Havel, Germany

In the digital age, patients increasingly rely on online sources for health information, particularly when considering procedures such as common orthopedic surgeries.[1] While digital tools improve access to medical knowledge, they also introduce variability in accuracy, readability, and currency, potentially leading to misinformation or incomplete patient education.[2,3]

Received: September 22, 2025
Accepted: November 19, 2025
Published online: March 17, 2026

Correspondence: Mahmut Enes Kayaalp, MD. SBÜ İstanbul Fatih Sultan Mehmet Eğitim ve Araştırma Hastanesi, Ortopedi ve Travmatoloji Kliniği, 34752 Ataşehir, İstanbul, Türkiye.

E-mail: mek@mek.md

Citation: Gultekin O, Hirschmann MT, Arıkan Hİ, Kilinc BE, Yilmaz B, Abul S, et al. Evaluating deepresearch and deepthink in total knee arthroplasty patient education: ChatGPT-4o excels in comprehensiveness, Deepseek R1 leads i230n clarity and readability of orthopedic information. Jt Dis Relat Surg 2026;37(2):470-476. doi: 10.52312/jdrs.2026.2645.

© 2026 Joint Diseases and Related Surgery. This is an open access article published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium, provided the original work is properly cited. http://creativecommons.org/licenses/by-nc/4.0/

ABSTRACT

Objectives: This study aims to directly compare ChatGPT and DeepSeek, both equipped with DeepResearch/DeepThink capabilities, based on their responses to frequently asked questions (FAQs) on total knee arthroplasty (TKA).

Materials and methods: Thirty frequently asked questions related to TKA were compiled from validated patient education sources, including American Academy of Orthopaedic Surgeons (AAOS) OrthoInfo, National Institute for Health and Care Excellence (NICE) guidelines, and popular patient discussion forums, and verified for clinical relevance by two independent arthroplasty surgeons. Two orthopedic surgeons, blinded to model identity, evaluated each response using a five-point Likert scale across five domains: accuracy, comprehensiveness, readability, relevance, and ethical and safety considerations. The maximum total score per response was 25. Readability was also assessed using the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES). Inter-rater and intra-rater reliability were calculated using intraclass correlation coefficients (ICCs).

Results: The ChatGPT-4o scored significantly higher in comprehensiveness and clinical detail, whereas DeepSeek R1 produced responses with superior readability, indicated by a lower FKGL (7.5 *vs.* 10.2) and higher FRES (62.3 *vs.* 45.6) ($p < 0.05$). Both models demonstrated high accuracy and safety, with no factual errors identified. Intra-rater reliability was excellent (ICC > 0.81), and inter-rater agreement ranged from fair to substantial (ICC 0.31 to 0.63).

Conclusion: Both ChatGPT-4o and DeepSeek R1 are capable of generating accurate, ethically sound, and clinically relevant educational content for patients undergoing TKA. While ChatGPT-4o offers more comprehensive information, DeepSeek R1 provides content that is more accessible to patients with lower health literacy. Model selection should be tailored to the target population to optimize educational effectiveness in clinical practice. The ability of real-time data retrieval to incorporate the most current clinical evidence and guideline updates may further enhance the educational quality, reliability, and clinical relevance of AI-generated patient information.

Keywords: Agentic artificial intelligence, large language model, readability, retrieval augmented generation, tool augmentation, total knee arthroplasty.

Total knee arthroplasty (TKA) is a common orthopedic procedure for advanced knee osteoarthritis, significantly improving patients' quality of life.[4] However, successful outcomes depend not only on surgical expertise, but also on effective preoperative education, realistic postoperative expectations, and patient adherence to rehabilitation.[5] Large language models (LLMs) such as ChatGPT have emerged as promising tools for delivering reliable and accessible medical information while aiding in research and writing processes.[5-13] Trained on vast medical datasets, these models can distill complex clinical knowledge into digestible content for individuals with varying health literacy levels.[8,14] In addition to structured medical datasets, LLMs also integrate information from unstructured public online sources such as patient education materials and healthcare websites, further broadening their informational scope.

Nevertheless, earlier iterations faced criticism for generating incomplete, outdated, or even fabricated responses.[10] As a result, artificial intelligence (AI)-generated patient education materials, particularly in surgical fields, require rigorous evaluation for accuracy, readability, ethical safety, and overall integrity.[3,15]

Recent advancements, such as deep research and deep think features, have enhanced the credibility of AI models by enabling real-time data retrieval and reducing hallucinations. Evaluations of ChatGPT's deep research mode demonstrate its potential to synthesize valid, clinically relevant literature, signaling a transformative shift in AI-assisted healthcare communication.[15]

DeepSeek, a multimodal LLM launched in 2025, further advances this field.[13] Its retrieval-augmented generation framework-shared with ChatGPT's deep research mode-integrates current guidelines and literature, mitigating past shortcomings in accuracy and consistency and potentially limiting hallucinations. However, differences in training data and architecture may lead to divergent outputs.[13]

Total knee arthroplasty was selected as the focus of this study, as it represents one of the most common and complex orthopedic procedures which requires extensive preoperative education and postoperative rehabilitation guidance. Comparing ChatGPT-4o with DeepResearch and DeepSeek R1 with DeepThink provides insight into how real-time retrieval-augmented models perform in patient education for high-demand surgical domains.

In the present study, we aimed to directly compare ChatGPT and DeepSeek, both equipped with DeepResearch/DeepThink capabilities, based on their responses to frequently asked questions (FAQs) on TKA. By assessing critical metrics such as accuracy, clarity, completeness, and consistency, we aimed to evaluate the performance of these models in delivering evidence-based health information. Additionally, the readability of their responses would be measured using validated indices. We hypothesized that both ChatGPT and DeepSeek chatbots with DeepResearch/DeepThink features would provide with similarly accurate, clear, complete and consistent responses and readability levels.

MATERIALS AND METHODS

This observational study was designed to compare the quality and reliability of responses generated by two advanced AI chat models, ChatGPT-4o Plus (OpenAI, San Francisco, CA, USA) and DeepSeek R1 (DeepSeek Inc., Hangzhou, China), to FAQs regarding TKA. The objective was to identify which model delivered more accurate, comprehensive, and patient-friendly answers. The chatbots were instructed to answer the questions in a manner directed toward the patients.

A total of 30 FAQs were selected from public patient forums such as Reddit (r/KneeReplacement) and HealthUnlocked, as well as from official patient education materials published by the American Academy of Orthopaedic Surgeons (AAOS) and the United Kingdom National Institute for Health and Care Excellence (NICE).[16,17] Selection was further refined based on input from two experienced arthroplasty surgeons to ensure clinical relevance. Real patients were not involved in the question selection process to ensure methodological standardization and reproducibility.

These questions spanned key clinical aspects such as preoperative decision-making, surgical techniques, recovery phases, complications, and long-term outcomes.

AI model configurations

The ChatGPT-4o model (OpenAI, San Francisco, CA, USA) was utilized with its DeepResearch mode activated, enabling retrieval from verified external medical resources beyond its static pretraining database. Through this functionality, the model could incorporate contemporary clinical literature, updated treatment guidelines, and trustworthy educational materials into its generated

responses.[18] Each answer was designed to maintain clinical accuracy while using simplified terminology suitable for patients, and no bibliographic references were displayed to preserve the conversational format.

The DeepSeek R1 model (DeepSeek Inc., Hangzhou, China) was evaluated in an equivalent setup with the DeepThink feature activated. This configuration supports reasoning and synthesis based on current medical data obtained through external database connections. DeepSeek was similarly instructed to provide comprehensive yet patient-focused explanations. Each question was processed independently in isolated sessions to avoid any form of contextual or output contamination between models.

Evaluation criteria

Responses generated by both models were systematically compiled per question. To ensure objectivity, all responses were anonymized and blinded to model identity. Two experienced orthopedic surgeons independently assessed each response using a five-point Likert scale (1 = very poor, 5 = excellent), based on five validated criteria derived from established measures of health information quality:[2,19]

1. Accuracy
2. Comprehensiveness
3. Readability
4. Relevance
5. Ethical and safety considerations

Total scores were calculated by summing the ratings across all five criteria, yielding a maximum score of 25 per response.

Readability

The readability of responses produced by both models was analyzed using a validated online tool (https://www.readabilityformulas.com) which calculates text complexity through established linguistic algorithms.[20] The assessment focused on two standard indices: the Flesch-Kincaid Grade Level (FKGL) and the Flesch Reading Ease Score (FRES), which estimate the literacy level and ease of comprehension required to understand a text.[21,22]

These measures are based on sentence length and word structure, offering a reproducible metric for language accessibility. In this context, lower FKGL values correspond to simpler and more patient-accessible language, whereas higher FRES values indicate greater readability. Each model's compiled set of answers was analyzed separately, and average scores for both indices were recorded to allow direct comparison of linguistic simplicity. This standardized approach ensured that readability evaluation was objective and consistent across both AI systems.

Reliability and bias reduction

To ensure consistency in scoring and confirm that both evaluators applied the assessment criteria uniformly, inter-rater reliability was analyzed using the intraclass correlation coefficient (ICC) within a two-way mixed-effects model. Interpretation followed conventional thresholds, where ICC values below 0.50 indicate poor, 0.50–0.75 moderate, 0.75–0.90 good, and above 0.90 excellent agreement.[23]

Intra-rater reliability was evaluated by having each reviewer re-assess all responses after a two-week interval. Both inter- and intra-rater ICCs were calculated and reported separately for each criterion.

To minimize bias, all responses were anonymized, randomized in order, and coded without identifiers to conceal the model's identity. Evaluators did not participate in content generation and did not communicate during scoring. Any observed differences in scores were included in the reliability analysis. The complete set of 30 standardized patient questions used in the study is available in Supplementary File 1.

Statistical analysis

Statistical analysis was performed using the IBM SPSS version 28.0 software (IBM Corp., Armonk, NY, USA). For each of the five evaluation criteria, accuracy, completeness, readability, relevance, and ethical and safety considerations, scores were averaged across two raters for each AI model's response to 30 frequently asked patient questions. Composite total scores were calculated by summing all five domain scores for each item, with a maximum of 25 points per response. The normality of data was confirmed with Shapiro-Wilk tests. Paired sample t-tests were used to compare mean scores between ChatGPT and DeepSeek. Descriptive data were expressed in mean ± standard deviation (SD) or number and frequency, where applicable. A p value of < 0.05 was considered statistically significant.

Artificial intelligence tools

The ChatGPT-4o with the DeepResearch feature and DeepSeek R1 (DeepSeek Inc., Hangzhou,

China) with the DeepThink feature were used to generate patient-directed responses to standardized questions. These outputs constituted the study material for subsequent blinded expert evaluation. No AI tool was used in data interpretation, statistical analysis, or manuscript writing.

RESULTS

Model performance comparison

The ChatGPT received notably high scores for accuracy, comprehensiveness, and ethical and safety considerations. However, it scored lower in readability and relevance compared to DeepSeek. The results are represented in Table I.

Flesch-Kincaid readability analysis

Table II represents readability metrics results. The ChatGPT responses had a FKGL of 10.2, corresponding to a reading level appropriate for a college. In contrast, the DeepSeek responses yielded a lower FKGL of 7.5, indicating a readability level consistent with middle to early high school education. The FRES was 45.6 for ChatGPT, suggesting a more challenging text, whereas DeepSeek achieved a FRES of 62.3, reflecting a moderately readable output. These results are consistent with evaluator scores, suggesting that DeepSeek tends to produce simpler and more understandable texts.

Reliability analysis

Intra-rater ICCs:

All intra-rater ICCs exceeded 0.81, with most values ranging between 0.86 and 1.00. These findings indicate a high degree of scoring consistency for both ChatGPT and DeepSeek across repeated evaluations.

Inter-rater ICCs:

Inter-rater reliability ranged from poor to moderate across domains (ICC: 0.31–0.63), reflecting usually acceptable agreement between raters. Highest consistency was observed in Ethical and Safety ratings for DeepSeek (ICC: 0.63), whereas Completeness and Relevance demonstrated more variation.

DISCUSSION

In the present study, we compared ChatGPT and DeepSeek, both equipped with DeepResearch/DeepThink capabilities, based on their responses to FAQs on TKA. The main findings of this study were the differences in produced text by two major LLMs on providing orthopedic information on TKA for patients. The ChatGPT-4o Plus with DeepResearch feature produced significantly more comprehensive responses, but at the expense of reduced readability. Whereas

TABLE I

Statistical comparison of ChatGPT and DeepSeek performance across evaluation criteria

	ChatGPT	DeepSeek	
Criterion	Mean ± SD	Mean ± SD	*p*
Accuracy	4.8 ± 0.2	3.8 ± 0.4	< 0.001
Comprehensiveness	4.7 ± 0.2	3.9 ± 0.3	< 0.001
Readability	3.2 ± 0.4	4.3 ± 0.4	< 0.001
Relevance	4 ± 0.5	4.9 ± 0.3	< 0.001
Ethical and safety considerations	4.8 ± 0.4	4.4 ± 0.5	0.004
Total score	21.4 ± 1.1	21.3 ± 1	NS

SD, standard deviation; NS, not significant.

TABLE II

Readability metrics

	ChatGPT	DeepSeek	
Metric	Mean ± SD	Mean ± SD	*p*
FKGL	10.2 ± 1.3	7.5 ± 1.1	< 0.001
FRES	45.6 ± 5.8	62.3 ± 4.9	< 0.001

SD, standard deviation; FKGL, Flesch-Kincaid Grade Level; FRES, Flesch Reading Ease Score.

DeepSeek with DeeptThink feature prioritized accessibility, aligning with recommended health literacy standards. These observations emphasize the inherent challenge of maintaining an optimal balance between informational depth and linguistic simplicity when artificial intelligence systems are used to communicate with patients.

Another key outcome of this study is the high accuracy and consistency shown by both ChatGPT-4o and DeepSeek R1 in generating educational content about TKA. The integration of real-time medical data through the deep research feature allowed both models to deliver up-to-date, evidence-based clinical information as assessed by experts. This represents a major advancement in the reliability of AI-supported patient education tools similar or beyond the level of previously reported results.[8,12]

The combination of research capability with real-time access to medical knowledge and transparent rationale represents one of the most recent innovations introduced in AI chatbots in early 2025.[13] Preliminary evaluations of OpenAI's Deep Research tool suggest that it delivers well-cited summaries that are accurate, but still has limitations in prioritizing recent findings, distinguishing conceptual nuance, and integrating broader scientific context.[18] In this study, real-time data access clearly improved the accuracy and consistency of AI responses. This direct linkage between real-time retrieval and patient education quality arises from the models' capacity to integrate the latest clinical data and updated guidelines into their responses. By reducing the reliance on outdated or static training information, these systems deliver more evidence-based and trustworthy educational content to patients. It should also be acknowledged that LLMs frequently draw upon publicly available web-based educational materials and health promotion content. However, such sources may vary widely in their scientific validity and evidence base. This variability highlights the importance of continuous human oversight and clinical verification of AI-generated information before its use in patient education.

Previous versions of ChatGPT were reported to provide only about 65% accuracy in orthopedic or surgical contexts and often lacked alignment with current treatment guidelines.[8] The current versions of ChatGPT-4 and DeepSeek, both equipped with real-time data retrieval, have largely overcome earlier limitations, particularly those related to outdated or hallucinated content.[6,10] More importantly, none of the answers generated via deep research by either ChatGPT or DeepSeek were deemed factually incorrect by expert reviewers. This aligns with recent findings in other medical fields such as diabetes education, where enhanced data access led to improved accuracy.[24]

Both models also demonstrated high alignment with expert clinical perspectives, suggesting that access to credible sources not only supports internal consistency, but also facilitates alignment with evidence-based clinical guidelines. These outcomes contrast with prior reports on ChatGPT-3.5 and 4, which often produced responses deviating from expert consensus.[15] Despite these strengths in accuracy and consistency, the models differed significantly in terms of clarity and comprehensiveness. The ChatGPT consistently produced broader content, offering detailed information on post-TKA rehabilitation, surgical techniques, prosthesis types, and complication management. However, this comprehensiveness came at the expense of readability, as reflected by a FKGL of 10.2 and a FRES of 45.6, levels that may be difficult for the average patient to understand.[3,21] These results support previous studies showing that ChatGPT-generated educational materials often exceed health literacy thresholds, with readability levels significantly above the average American reading grade.[3] In contrast, DeepSeek's responses were closer to the 6th to 8th grade level recommended by the American Medical Association (AMA) and the Centers for Disease Control and Prevention.[25] The simplified response structure of DeepSeek aligns with previous research showing that easier-to-read content is more effective at improving patient comprehension.[26] This contrast also reflects a common clinical dilemma: while too much detail may overwhelm patients, excessive simplification risks omitting important information.[27]

Considering that most TKA patients belong to older age groups, content which minimizes cognitive load and enhances comprehension is particularly crucial.[2] The ChatGPT's higher reading level may hinder access to information for patients with limited education or cognitive capacity, potentially increasing anxiety during the information process. In contrast, DeepSeek's patient-friendly language may offer an advantage for elderly populations. Meanwhile, ChatGPT's more comprehensive outputs may be more suitable for clinician-focused materials or multidisciplinary educational content.[19]

Nonetheless, this study is limited by its evaluation of only 30 patient questions, and the readability assessments relied on expert review. A control condition without real-time data retrieval was not included, limiting the ability to directly quantify the contribution of retrieval augmentation to performance. Patient comprehension testing was not performed, which limits the generalizability of the findings to real-world educational outcomes. Future studies incorporating direct patient surveys across varied age groups could provide a more holistic evaluation of AI-generated content's impact on real patient experiences. However, the novel features of ChatGPT and DeepSeek as evaluated in the current study were not available previously. Future AI systems may incorporate a dual-prompt approach in which the model first generates a comprehensive, evidence-based answer and subsequently simplifies it to an appropriate readability level. Furthermore, adaptive AI models could autonomously assess the user's health literacy, determine the optimal reasoning pathway, and tailor the response's complexity accordingly. Such evolution toward dynamic, patient-centric communication may overcome the trade-offs identified in the present study.

In conclusion, both ChatGPT-4o (DeepResearch) and DeepSeek R1 (DeepThink) effectively generated accurate, safe, and clinically relevant educational content on TKA. The ChatGPT provided more comprehensive and technically detailed explanations, whereas DeepSeek produced language that was easier for patients to understand. The incorporation of real-time information retrieval represents a key methodological advance, contributing to the reliability and educational value of AI-generated outputs. Applying these models complementarily-or selecting one according to patient literacy level-may enhance the clarity and accessibility of orthopedic patient education while reinforcing the supportive, rather than substitutive, role of artificial intelligence in clinical care.

Data Sharing Statement: The datasets generated and analyzed during the current study (AI-generated responses and blinded evaluator scoring sheets) are available from the corresponding author.

Author Contributions: M.E.K., M.T.H., J.I.: Idea/concept, analysis; O.G., H.İ.A., B.E.K., B.Y.: Design; H.İ.A., S.A.: Data collection; O.G., M.E.K.: Literature review; O.G., M.E.K., H.İ.A., S.A.: Writing the article; B.Y., M.E.K., M.T.H., J.I., B.E.K. Critical review.

Conflict of Interest: The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding: The authors received no financial support for the research and/or authorship of this article.

AI Disclosure: The authors declare that artificial intelligence (AI) tools were not used, or were used solely for language editing, and had no role in data analysis, interpretation, or the formulation of conclusions. All scientific content, data interpretation, and conclusions are the sole responsibility of the authors. The authors further confirm that AI tools were not used to generate, fabricate, or 'hallucinate' references, and that all references have been carefully verified for accuracy.

REFERENCES

1. Bujnowska-Fedak MM, Węgierek P. The ımpact of online health information on patient health behaviours and making decisions concerning health. Int J Environ Res Public Health 2020;17:880. doi: 10.3390/ijerph17030880.
2. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: Current role in orthopaedics. Clin Orthop Relat Res. 2010;468:2572-80. doi: 10.1007/s11999-010-1380-y.
3. Fahy S, Oehme S, Milinkovic D, Jung T, Bartek B. Assessment of quality and readability of information provided by ChatGPT in relation to anterior cruciate ligament ınjury. J Pers Med 2024;14:104. doi: 10.3390/jpm14010104.
4. Sadoghi P, Koutp A, Prieto DP, Clauss M, Kayaalp ME, Hirschmann MT. The projected economic burden and complications of revision hip and knee arthroplasties: Insights from national registry studies. Knee Surg Sports Traumatol Arthrosc 2025;33:3211-7. doi: 10.1002/ksa.12678.
5. Cieremans DA, Arraut J, Marwin S, Slover J, Schwarzkopf R, Rozell JC. Patellar component design does not impact clinical outcomes in primary total knee arthroplasty. J Arthroplasty 2023;38:1493-8. doi: 10.1016/j.arth.2023.01.061.
6. Dahmen J, Kayaalp ME, Ollivier M, Pareek A, Hirschmann MT, Karlsson J, et al. Artificial intelligence bot ChatGPT in medical research: The potential game changer as a double-edged sword. Knee Surg Sports Traumatol Arthrosc 2023;31:1187-9. doi: 10.1007/s00167-023-07355-6.
7. Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. J Exp Orthop 2023;10:74. doi: 10.1186/s40634-023-00642-8.
8. Kaarre J, Feldt R, Keeling LE, Dadoo S, Zsidai B, Hughes JD, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee Surg Sports Traumatol Arthrosc 2023;31:5190-8. doi: 10.1007/s00167-023-07529-2.
9. Ollivier M, Pareek A, Dahmen J, Kayaalp ME, Winkler PW, Hirschmann MT, et al. A deeper dive into ChatGPT: History, use and future perspectives for orthopaedic research. Knee Surg Sports Traumatol Arthrosc 2023;31:1190-2. doi: 10.1007/s00167-023-07372-5.
10. Kayaalp ME, Ollivier M, Winkler PW, Dahmen J, Musahl V, Hirschmann MT, et al. Embrace responsible ChatGPT usage to overcome language barriers in academic writing. Knee Surg Sports Traumatol Arthrosc 2024;32:5-9. doi: 10.1002/ksa.12014.
11. 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. doi: 10.52312/jdrs.2025.1961.
12. 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. doi: 10.1002/ksa.12711.
13. Kayaalp ME, Prill R, Sezgin EA, Cong T, Królikowska A, Hirschmann MT. DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation-Redefining innovation in research and practice. Knee Surg Sports Traumatol Arthrosc 2025;33:1553-6. doi: 10.1002/ksa.12628.
14. Yapar D, Demir Avcı Y, Tokur Sonuvar E, Eğerci ÖF, Yapar A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. Jt Dis Relat Surg 2024;35:169-76. doi: 10.52312/jdrs.2023.1402.
15. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: An analysis of ChatGPT 3.5, ChatGPT 4, and Bard. J Orthop 2023;50:1-7. doi: 10.1016/j.jor.2023.11.063.
16. American Academy of Orthopaedic Surgeons. Total knee replacement. OrthoInfo. 2025 [cited 22.09.2025]. Available from: https://orthoinfo.aaos.org.
17. National Institute for Health and Care Excellence. Total knee replacement guidelines. NICE. 2025 [cited 22.09.2025]. Available from: https://www.nice.org.uk/guidance.
18. OpenAI. Deep Research system card. 2025 [cited 22.09.2025]. Available from: https://cdn.openai.com/deep-research-system-card.pdf.
19. Sun Y, Zhang Y, Gwizdka J, Trace CB. Consumer evaluation of the quality of online health information: Systematic literature review of relevant criteria and indicators. J Med Internet Res 2019;21:e12522. doi: 10.2196/12522.
20. Readability Formulas. Free readability tools. 2025 [cited 22.09.2025]. Available from: https://www.readabilityformulas.com.
21. Flesch R. A new readability yardstick. J Appl Psychol 1948;32:221-33. doi: 10.1037/h0057532.
22. Kincaid JP. Derivation of new readability formulas: automated readability index, fog count and Flesch reading ease formula for navy enlisted personnel. Memphis: Chief of Naval Technical Training, Naval Air Station; 1975.
23. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155-63. doi: 10.1016/j.jcm.2016.02.012.
24. Wang D, Liang J, Ye J, Li J, Li J, Zhang Q, et al. Enhancement of the performance of large language models in diabetes education through retrieval-augmented generation: Comparative study. J Med Internet Res 2024;26:e58041. doi: 10.2196/58041.
25. Eltorai AE, Ghanian S, Adams CA Jr, Born CT, Daniels AH. Readability of patient education materials on the american association for surgery of trauma website. Arch Trauma Res 2014;3:e18161. doi: 10.5812/atr.18161.
26. Dubin JA, Bains SS, Chen Z, Hameed D, Nace J, Mont MA, et al. Using a Google web search analysis to assess the utility of ChatGPT in total joint arthroplasty. J Arthroplasty 2023;38:1195-202. doi: 10.1016/j.arth.2023.04.007.
27. Oviedo-Trespalacios O, Peden AE, Cole-Hunter T, Costantini A, Haghani M, Javier E. Rodriguez, JE, et al. The risks of using ChatGPT to obtain common safety-related information and advice. Saf Sci 2023;167:106024.