

Evaluating the Performance of ChatGPT-5 in Answering Common Patient Questions on Developmental Dysplasia of the Hip and Periacetabular Osteotomy

Cathrin Manthey

cathrin.manthey@gmail.com

Universitätsklinikum Brandenburg an der Havel

Jonathan Lettner

Universitätsklinikum Brandenburg an der Havel

Sufian S. Ahmad

Medizinische Hochschule Hannover

Mikhail Salzmann

Universitätsklinikum Brandenburg an der Havel

Nikolai Ramadanov

Universitätsklinikum Brandenburg an der Havel

Research Article

Keywords: Developmental dysplasia of the hip, Periacetabular osteotomy, ChatGPT-5, Artificial intelligence, Large language model, Patient education

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10762343/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO) require extensive patient education, yet online medical information is often inconsistent in quality and readability. Large language models such as ChatGPT have shown promising results in orthopedic patient education, although limitations regarding completeness and readability remain. The performance of ChatGPT-5 in this context has not yet been evaluated.

Methods

Using a standardized two-step prompting strategy, ChatGPT-5 generated 25 common patient questions regarding DDH and PAO and subsequently provided corresponding patient-oriented answers. Two fellowship-trained hip preservation surgeons independently evaluated all question–answer pairs across relevance, accuracy, clarity, completeness, and clinical safety using a 5-point Likert scale. Readability was additionally assessed using Flesch–Kincaid Grade Level and Flesch Reading Ease Score. Inter-rater reliability was calculated using intraclass correlation coefficients (ICC 2,1).

Results

ChatGPT-5 demonstrated consistently high performance across all evaluated domains. Mean scores were highest for relevance (5.0 ± 0.0), clarity (4.98 ± 0.10), and readability metrics (5.0 ± 0.0). Accuracy (4.76 ± 0.43) and clinical safety (4.78 ± 0.42) were also rated highly, with no responses considered unsafe or misleading. Completeness showed greater variability and lower overall scores (4.26 ± 0.50), mainly due to limited quantitative detail and lack of individualized clinical context. The overall mean score across Likert-based domains was 4.82 ± 0.28 . Inter-rater reliability ranged from moderate to excellent across domains.

Conclusion

ChatGPT-5 provides highly relevant, readable, accurate, and clinically safe responses to common patient questions regarding DDH and PAO. While completeness remains variable, ChatGPT-5 may serve as a valuable adjunctive tool for orthopedic patient education and shared decision-making.

Level of Evidence (LoE):

Level IV, Diagnostic Study

INTRODUCTION

Developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO) represent complex clinical entities within hip preservation surgery, often requiring detailed patient education regarding indications, surgical technique, expected outcomes, and potential risks. As the volume of hip preservation procedures continues to rise, patients increasingly rely on online sources to obtain medical information prior to consultation. However, the quality, accuracy, and readability of such information remain highly variable, raising concerns about misinformation and its potential impact on shared decision-making.

In this context, large language models (LLMs), particularly ChatGPT, have emerged as widely accessible tools capable of generating structured, conversational, and context-adapted medical information. Several recent studies have evaluated the performance of earlier ChatGPT versions in the domain of hip preservation. Leopold et al. [1] demonstrated that ChatGPT-4 provided generally satisfactory responses to DDH- and PAO-related patient queries, although substantial variability was observed between preoperative and postoperative content, and readability levels were considerably higher than recommended for patient education. Similarly, Gaddis et al. [2] reported high accuracy and comprehensiveness rates (with 98.3% of responses rated as excellent or satisfactory) for ChatGPT-4.0 responses to PAO-related questions, while highlighting persistent limitations in readability that could be partially mitigated through response simplification.

Beyond PAO-specific contexts, broader orthopaedic evidence further supports the overall reliability of LLM-generated patient information. Ramadanov et al. [3] conducted a systematic review and meta-analysis demonstrating high pooled accuracy (88.6%) and consistently strong performance in relevance, safety, readability, and clarity for ChatGPT-generated responses in hip arthroscopy. However, completeness remained variable, and methodological heterogeneity—including inconsistent rating systems and non-standardized question selection—limited the strength of conclusions. Notably, improved performance was consistently observed in later model generations (GPT-4 and beyond), suggesting ongoing model evolution as a key determinant of output quality.

Despite these encouraging findings, all available studies to date have focused on earlier ChatGPT versions (3.5, 4.0, or related iterations). No study has systematically evaluated the performance of the latest-generation model, ChatGPT-5, which incorporates substantial advancements in contextual reasoning, factual consistency, and response structuring. These theoretical improvements suggest potential gains in accuracy, completeness, and clinical applicability; however, empirical validation in orthopaedic patient education remains lacking.

Therefore, the present study aims to evaluate ChatGPT-5's performance in answering common patient questions related to DDH and PAO. Specifically, this study assesses response accuracy, completeness, clarity, and clinical suitability as judged by experienced hip preservation surgeons. By contextualizing

these findings within the existing evidence base on earlier LLM generations, this work seeks to determine whether recent advances in artificial intelligence translate into clinically meaningful improvements in patient-oriented orthopaedic information delivery.

METHODS

Study design and content generation

This descriptive observational study evaluated the educational content generated by ChatGPT-5 (OpenAI, San Francisco, CA, USA) in relation to developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO). To ensure standardized and reproducible content generation, a predefined two-step prompting strategy was applied. First, the model was instructed with the prompt: "Generate the 25 most common questions patients ask about developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO). Focus on clinically relevant, patient-oriented concerns, including symptoms, diagnosis, treatment options, surgical procedure, risks, rehabilitation, and long-term outcomes." The resulting set of questions was subsequently used as input for a second prompt: "Provide clear, accurate, and patient-friendly answers to each of these 25 questions. Use simple language suitable for a general audience, avoid technical jargon where possible, and ensure that the information is medically correct and clinically relevant." No further prompts, refinements, or follow-up instructions were provided. The model was used in a zero-shot setting without additional contextual input. This structured approach was designed to simulate real-world patient information-seeking behavior while minimizing investigator-driven bias in question formulation. The generated questions were not modified or filtered. As this study did not involve human participants, patient data, or identifiable information, formal ethical approval and informed consent were not required in accordance with institutional and national guidelines. The study is reported in accordance with the STROBE guidelines. The corresponding checklist is provided in Supplementary Table S1.

Data collection and documentation

All outputs were generated within a single session using the default ChatGPT-5 interface. Responses were recorded verbatim as produced by the model, without any manual modification, correction, or linguistic adjustment, in order to preserve the authenticity of the generated content. The complete set of question-answer pairs is provided in the Supplementary Material (Supplementary Figure S1).

Evaluation procedure

Two fellowship-trained orthopaedic surgeons with subspecialty expertise in hip preservation independently evaluated all 25 question-answer pairs. A seven-domain rating framework was applied to comprehensively assess the quality of the generated responses, including:

- (1) relevance – the extent to which the response addressed the question posed;
- (2) accuracy – the factual correctness and medical validity of the information;
- (3) clarity – readability and linguistic comprehensibility from a patient perspective;
- (4) completeness – the degree to which key aspects necessary for adequate understanding were covered;
- (5) clinical safety – the absence of misleading, incomplete, or potentially harmful medical recommendations;
- (6) linguistic complexity, assessed using the Flesch–Kincaid Grade Level (FKGL);
- (7) reading ease, assessed using the Flesch Reading Ease Score (FRES).

The first five domains were rated using a five-point Likert scale (1 = poor, 5 = excellent). Readability metrics (FKGL and FRES) were calculated objectively and not subject to reviewer scoring. Reviewers performed their assessments independently and were blinded to each other's ratings. No consensus process or post hoc adjustment of scores was undertaken. An overall quality score was calculated for each response as the mean of the five Likert-based domains.

Statistical analysis

Descriptive statistics were calculated for each evaluation domain and are presented as mean values with standard deviations (SD). The overall quality score was calculated and analyzed analogously. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC 2,1), based on a two-way random-effects model for absolute agreement, as described by McGraw and Wong [4]. The clinical safety score was included in the reliability analysis accordingly. To complement the ICC, the proportion of exact agreement between raters was additionally reported, providing a descriptive measure of concordance in the context of limited between-item variability. Readability metrics, including Flesch–Kincaid Grade Level and Flesch Reading Ease Score [5], were computed for each response using standard natural language processing libraries in Python. All statistical analyses were performed using Python version 3.12 (Python Software Foundation, Wilmington, DE, USA).

RESULTS

A total of 25 question-answer pairs (Table 1) generated by ChatGPT-5 were evaluated across seven predefined domains by two independent reviewers. Individual ratings are presented in Table 2, while mean scores per item and overall summary statistics are shown in Tables 3 and 4.

Table 1

Patient questions and corresponding answers on developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO) generated using a standardized two-step prompting approach with ChatGPT-5 and used for subsequent evaluation.

Prompt provided to ChatGPT-5: "Generate the 25 most common questions patients ask about developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO). Focus on clinically relevant, patient-oriented concerns, including symptoms, diagnosis, treatment options, surgical procedure, risks, rehabilitation, and long-term outcomes."	
1. What is developmental dysplasia of the hip (DDH)?	DDH means that the hip socket does not fully cover the ball of the hip joint. This can make the joint less stable and lead to increased wear over time.
2. What symptoms does DDH cause?	Common symptoms include groin pain, pain during activity, and sometimes a feeling of instability or catching in the hip.
3. Can DDH be treated without surgery?	In mild cases, physiotherapy, muscle strengthening, and pain management may help. Surgery is considered if symptoms persist or worsen.
4. What is a periacetabular osteotomy (PAO)?	PAO is a surgical procedure where the hip socket is repositioned to better cover the ball of the joint, improving stability.
5. Who is a good candidate for PAO?	Typically younger patients with symptomatic DDH and little or no arthritis.
6. What is the goal of PAO?	To reduce pain, improve hip function, and delay or prevent the development of arthritis.
7. How is the surgery performed?	The surgeon cuts the bone around the hip socket, repositions it, and fixes it in place with screws.
8. How long does the surgery take?	Usually around 2–3 hours, depending on complexity.
9. Is PAO a risky procedure?	Like all surgeries, it carries risks, but it is considered safe when performed in experienced centers.
10. What are the risks of PAO?	Possible risks include infection, blood loss, nerve irritation, and delayed bone healing.
11. How long is the hospital stay?	Typically between 5 and 10 days, depending on recovery.
12. When can I start walking again?	You can usually begin walking early with crutches, but only partial weight-bearing at first.
13. How long does recovery take?	Recovery can take several months, often 3–6 months or longer.
14. When can I return to sports?	Light activity may be possible after a few months, while high-impact sports often require 6–12 months.
15. Will the screws stay in my body?	Yes, they are usually left in place but can be removed later if needed.
16. How successful is PAO?	Many patients experience significant pain relief and improved function, especially when treated early.
17. Will I be completely pain-free after surgery?	Some patients become pain-free, but the main goal is significant improvement rather than guaranteed elimination of pain.
18. Can PAO prevent a hip replacement?	It can delay or sometimes avoid the need for a hip replacement, especially if done early.
19. What happens if DDH is not treated?	It can lead to increasing joint damage and early arthritis.
20. Are there alternatives to PAO?	Depending on the case, options include physiotherapy, hip arthroscopy, or later hip replacement.

Prompt provided to ChatGPT-5: "Generate the 25 most common questions patients ask about developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO). Focus on clinically relevant, patient-oriented concerns, including symptoms, diagnosis, treatment options, surgical procedure, risks, rehabilitation, and long-term outcomes."	
21. How painful is recovery after PAO?	
	Pain is expected but is usually well controlled with medication.
22. When can I return to work?	
	This depends on your job. Office work may be possible after a few weeks, while physical jobs take longer.
23. Can DDH come back after surgery?	
	The anatomical correction is permanent, but long-term outcomes depend on various factors.
24. Will I have a scar after surgery?	
	Yes, there will be a scar near the hip.
25. How do I find an experienced surgeon?	
	Look for specialized centers and surgeons with experience in hip preservation procedures.

Table 2

Individual ratings of all 25 question–answer pairs across the seven evaluation domains by two independent raters. Each domain was assessed using a five-point scale (1 = poor, 2 = fair, 3 = good, 4 = very good, 5 = excellent), except for linguistic complexity and reading ease, which were calculated objectively.

Question	Relevance (Rater 1)	Relevance (Rater 2)	Accuracy (Rater 1)	Accuracy (Rater 2)	Clarity (Rater 1)	Clarity (Rater 2)	Completeness (Rater 1)	Completeness (Rater 2)	Clinical safety (Rater 1)	Clinical safety (Rater 2)	Linguistic complexity (Rater 1)
Question 1	5	5	5	5	5	5	4	5	5	5	5
Question 2	5	5	4	5	5	5	3	5	4	5	5
Question 3	5	5	4	5	5	5	4	5	4	5	5
Question 4	5	5	5	5	5	5	3	5	4	5	5
Question 5	5	5	5	5	5	5	4	5	5	5	5
Question 6	5	5	5	5	5	5	5	5	5	5	5
Question 7	5	5	5	5	5	5	4	5	5	5	5
Question 8	5	5	5	5	5	5	5	4	5	5	5
Question 9	5	5	4	5	5	5	4	4	4	5	5
Question 10	5	5	5	5	5	5	4	5	4	5	5
Question 11	5	5	3	5	4	5	3	5	5	5	5
Question 12	5	5	5	5	5	5	4	5	4	5	5
Question 13	5	5	5	5	5	5	4	5	5	5	5
Question 14	5	5	5	5	5	5	4	5	4	5	5
Question 15	5	5	5	5	5	5	5	5	5	5	5
Question 16	5	5	5	5	5	5	4	5	5	5	5
Question 17	5	5	5	5	5	5	4	5	5	5	5
Question 18	5	5	5	5	5	5	4	5	5	5	5
Question 19	5	5	5	5	5	5	4	5	5	5	5
Question 20	5	5	3	5	5	5	4	4	4	5	5
Question 21	5	5	4	5	5	5	4	5	4	5	5
Question 22	5	5	5	5	5	5	4	4	5	5	5
Question 23	5	5	5	5	5	5	4	4	5	5	5
Question 24	5	5	5	5	5	5	4	4	5	5	5
Question 25	5	5	4	5	5	5	4	4	4	5	5

Table 3

Mean scores for each question–answer pair across all evaluation domains, including overall mean scores. Values represent the average of ratings from both reviewers, with readability metrics reported as calculated values.

Question	Mean Relevance	Mean Accuracy	Mean Clarity	Mean Completeness	Clinical safety	Linguistic complexity	Reading ease	Overall Mean
1	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
2	5.0	4.5	5.0	4.0	4.5	5.0	5.0	4.60
3	5.0	4.5	5.0	4.5	4.5	5.0	5.0	4.70
4	5.0	5.0	5.0	4.0	4.5	5.0	5.0	4.70
5	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
6	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.00
7	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
8	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
9	5.0	4.5	5.0	4.0	4.5	5.0	5.0	4.60
10	5.0	5.0	5.0	4.5	4.5	5.0	5.0	4.80
11	5.0	4.0	4.5	4.0	5.0	5.0	5.0	4.50
12	5.0	5.0	5.0	4.5	4.5	5.0	5.0	4.80
13	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
14	5.0	5.0	5.0	4.5	4.5	5.0	5.0	4.80
15	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.00
16	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
17	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
18	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
19	5.0	5.0	5.0	4.5	5.0	5.0	5.0	4.90
20	5.0	4.0	5.0	4.0	4.5	5.0	5.0	4.50
21	5.0	4.5	5.0	4.5	4.5	5.0	5.0	4.70
22	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.80
23	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.80
24	5.0	5.0	5.0	4.0	5.0	5.0	5.0	4.80
25	5.0	4.5	5.0	4.0	4.5	5.0	5.0	4.60

Table 4

Summary of evaluation results across all domains, including mean values with standard deviations, inter-rater reliability (ICC 2,1), and exact agreement between raters. The overall mean score represents the average across all Likert-based domains.

Domain	Mean $\pm$ SD (items)	ICC (2,1)	Exact agreement (%)
Relevance	5.00 $\pm$ 0.00	1.00	100
Accuracy	4.76 $\pm$ 0.43	0.79	72
Clarity	4.98 $\pm$ 0.10	0.96	96
Completeness	4.26 $\pm$ 0.50	0.69	44
Clinical safety	4.78 $\pm$ 0.42	0.75	68
Linguistic complexity	5.00 $\pm$ 0.00	1.00	100
Reading ease	5.00 $\pm$ 0.00	1.00	100
Overall Mean	4.82 $\pm$ 0.28	0.83	76

Across all items, ChatGPT-5 demonstrated consistently high performance (Fig. 1).

Mean relevance scores were uniformly maximal (5.0) for all question–answer pairs, indicating that all responses directly addressed the respective patient questions. Clarity was similarly high, with a mean score of 4.98 $\pm$ 0.10, reflecting excellent readability and comprehensibility from a patient perspective.

Accuracy was rated highly, with a mean score of 4.76 ± 0.43 . While most responses were considered fully accurate, occasional minor reductions in score reflected limited detail or slight imprecision rather than substantial factual errors. Clinical safety was likewise high (mean 4.78 ± 0.42), with no responses identified as unsafe or misleading.

In contrast, completeness demonstrated greater variability and lower overall scores (mean 4.26 ± 0.50). Although many responses were considered comprehensive, several lacked quantitative information, detailed explanations, or individualized clinical context.

Linguistic complexity and reading ease achieved maximum scores across all items (both 5.00 ± 0.00), indicating consistently accessible and patient-oriented language.

The overall mean score across all Likert-based domains was 4.82 ± 0.28 (Table 4), reflecting uniformly high response quality.

Inter-rater reliability was excellent for relevance, linguistic complexity, and reading ease (ICC = 1.00), and remained high for clarity (ICC = 0.96) and overall scores (ICC = 0.83) (Table 4). Agreement was moderate for accuracy (ICC = 0.79) and clinical safety (ICC = 0.75), and lowest for completeness (ICC = 0.69). Exact agreement between raters ranged from 100% for relevance and linguistic domains to 44% for completeness, further highlighting the greater subjectivity and variability in assessing response depth (Table 4).

DISCUSSION

This study evaluated the quality and readability of ChatGPT-5's responses to common patient questions regarding DDH and PAO. Overall, ChatGPT-5 demonstrated strong performance across nearly all evaluated domains. Responses achieved maximal mean scores for relevance (5.0) and linguistic readability (5.0), indicating that questions were addressed directly and in language that is highly accessible for patients. Mean scores for clarity (4.98), accuracy (4.76), and clinical safety (4.78) were also rated very highly, with no responses classified as unsafe or misleading. Inter-rater reliability was excellent for relevance and readability (ICC = 1.00), high for clarity (ICC = 0.96) and overall scores (ICC = 0.83), and moderate for accuracy (ICC = 0.79) and clinical safety (ICC = 0.75). Completeness showed the greatest variability and the lowest mean score (4.26), largely due to missing quantitative details, limited depth, and a lack of individualized clinical context. This domain also demonstrated the lowest inter-rater agreement (ICC = 0.69), suggesting a degree of subjectivity in its assessment.

A central finding of this study is the improved readability of ChatGPT-5 compared to earlier model generations [1–4]. Previous versions, such as ChatGPT-4, often generated content exceeding recommended reading levels for patient education [1, 2, 5]. In contrast, ChatGPT-5 produced more accessible and patient-centered responses, in line with the National Institutes of Health (NIH) recommendations for patient education materials. This reflects progress in translating complex medical concepts into language suitable for a general audience.

In addition to improved readability, ChatGPT-5 demonstrated strong levels of factual accuracy and clinical safety, consistent with findings from earlier model generations [1–3, 6]. Previous studies evaluating ChatGPT-4 reported generally satisfactory responses to DDH- and PAO-related patient queries [1]. Similarly, Gaddis et al. [2] reported high accuracy and comprehensiveness rates (98.3% rate as excellent or satisfactory) for ChatGPT-4.0 responses to PAO-related questions, while Özbek et al. [6] found that ChatGPT-4.0 delivered at least satisfactory, and often excellent, responses to patient questions regarding hip arthroscopy. Our findings indicate that ChatGPT-5 maintains a high level of factual correctness, with minor reductions in scores primarily due to limited detail rather than incorrect information. The absence of unsafe or misleading responses is consistent with prior analyses reporting no major safety concerns across GPT models in orthopedic contexts [3]. These results indicate that factual reliability and clinical safety have been maintained across model iterations, supporting the potential role of LLMs as tools for patient education.

However, completeness remained the most variable domain and showed the lowest mean scores, consistent with previous studies [1, 2, 6, 8]. Lower scores were mainly associated with missing quantitative data, limited detail in explanations, and insufficient clinical personalisation. Ramadanov et al. [3], in a systematic review and meta-analysis, identified completeness as the most inconsistent domain across ChatGPT generations. Despite generally high levels of accuracy and safety, variability in response depth and comprehensiveness remains a central limitation of LLMs. Our study further supports this observation.

Despite these limitations ChatGPT-5 demonstrates high readability, overall accurate information delivery, and strong clinical safety. From a clinical perspective, these findings suggest that ChatGPT-5 may serve as a valuable adjunctive tool for patient education by providing relevant, accurate and clinically safe information. However, AI-generated responses should support, but not replace, physician consultations and shared decision-making processes. This is particularly important in orthopedic care, where nuanced discussions regarding surgical indications, rehabilitation, risks, and long-term outcomes are essential for informed patient understanding and individualized treatment planning. The lack of individualized clinical judgment remains an important limitation of current LLMs.

CONCLUSION

In conclusion, ChatGPT consistently provides relevant and safe information as a supportive tool for patient education in orthopedics. Compared with earlier versions, ChatGPT-5 demonstrates notable improvements in readability and patient-centered communication. Although completeness remains variable and dependent on question complexity, the overall performance of ChatGPT-5 suggests increasing usefulness in enhancing patient understanding and supporting shared decision-making in conditions such as DDH and PAO.

Declarations

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

GENERATIVE AI STATEMENT

Generative artificial intelligence (ChatGPT-5, OpenAI, San Francisco, CA, USA) was evaluated as the object of investigation in this study. Generative AI tools were additionally used for language editing and improvement of English grammar and style. The authors reviewed and edited all content and take full responsibility for the final manuscript.

Clinical trial number

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Author 1: Conceptualization, study design, data collection, data analysis, data interpretation, manuscript drafting, manuscript revision, and corresponding author responsibilities. Author 2: Data interpretation, critical manuscript revision, and approval of the final manuscript. Author 3: Data interpretation, critical manuscript revision, and approval of the final manuscript. Author 4: Study supervision, data interpretation, critical manuscript revision, and approval of the final manuscript. Author 5: Conceptualization, study design, independent content evaluation, data interpretation, manuscript revision, and approval of the final manuscript. All authors have read and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

The authors have no acknowledgements to declare.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request. The evaluated material consisted of ChatGPT-generated question–answer pairs and reviewer assessment data generated specifically for this study.

References

1. Leopold VJ, Fahy S, Perka C, Goronzy J, Grammatopoulos G, Beaulé PE, Hardt S. Assessment of quality and readability of information provided by ChatGPT in relation to developmental dysplasia of the hip and periacetabular osteotomy. *J Hip Preserv Surg.* 2026;hnaf025. 10.1093/jhps/hnaf025.
2. Gaddis JM, Arellano E, Martin BC, Alsabawi Y, Salgado-Flores M, South C, Castaneda P, Wells JE. Assessing the accuracy and readability of ChatGPT 4.0's original and simplified responses to common patient questions regarding periacetabular osteotomy. *J Exp Orthop.* 2026;70457. 10.1002/jeo2.70457.
3. Ramadanov N, Penchev P, Voss M, Heinz M, Prill R, Salzmann M, Becker R, Osterberger T, Banke IJ. ChatGPT provides accurate and safe responses to patient questions on hip arthroscopy, while completeness remains variable: A systematic review and single-arm meta-analysis. *Knee Surg Sports Traumatol Arthrosc.* 2026;70396. 10.1002/ksa.70396.
4. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods.* 1996;1(1):30–46.
5. Flesch R. A new readability yardstick. *J Appl Psychol.* 1948;32(3):221–33. 10.1037/h0057532.
6. Özbek EA, Ertan MB, Kindan P, et al. ChatGPT Can Offer At Least Satisfactory Responses to Common Patient Questions Regarding Hip Arthroscopy. *Arthroscopy.* 2025;41:1806–27. <https://doi.org/10.1016/j.arthro.2024.08.036>.
7. Cotugna N, Vickery CE, Carpenter-Haefele KM, EVALUATION OF LITERACY LEVEL OF PATIENT, EDUCATION PAGES IN HEALTH-RELATED JOURNALS. *J Community Heal.* 2005;30:213–9. <https://doi.org/10.1007/s10900-004-1959-x>.

Figures

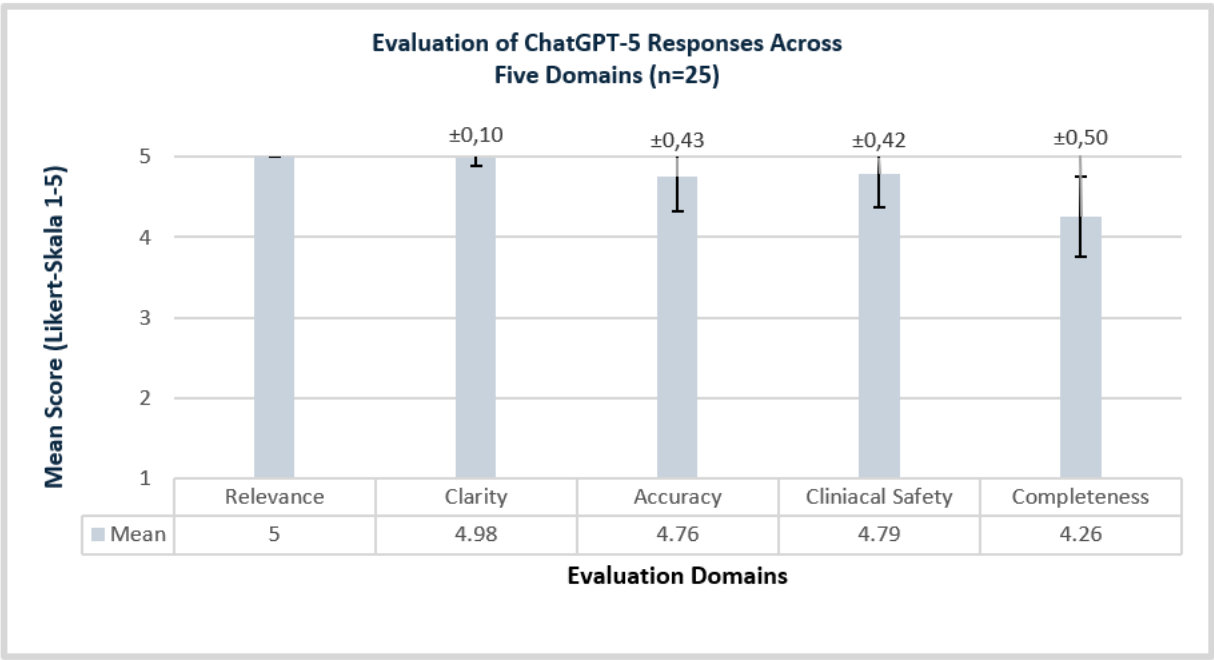


Figure 1

Mean evaluation scores of ChatGPT-5-generated responses to 25 common patient questions regarding developmental dysplasia of the hip (DDH) and periacetabular osteotomy (PAO) across five assessment domains. Responses were independently evaluated by two fellowship-trained hip preservation surgeons using a 5-point Likert scale (1 = poor, 5 = excellent). Error bars represent standard deviations.