International Journal of

Digital Rehabilitation & Therapy

e-ISSN:3062-4088

https://ndpapublishing.com/index.php/ijdrt

How Accurate Are Public Perceptions? A Comparative Analysis of Artificial Intelligence Responses on Physical Therapy for Gonarthrosis

Bilgehan Kolutek Ay*[1] and Mustafa Tuna[2]

[1] *Sütçü İmam University Faculty of Medicine, Department of Physical Medicine and Rehabilitation. Kahramanmaraş, Türkiye*

[2] *Şanlıurfa Training and Research Hospital, Department of Physical Medicine and Rehabilitation, Şanlıurfa, Türkiye, mustafa5tuna@gmail.com*

* *Corresponding author*: *bilgehankolutek@gmail.com*

Cite this study: Kolutek Ay, B. & Tuna, M. (2026). How Accurate Are Public Perceptions? A Comparative Analysis of Artificial Intelligence Responses on Physical Therapy for Gonarthrosis. Int. J. Digital Rehabilitation & Therapy, (3), 1-6.

https://doi.org/10.67015/ijdrt.342

Research Article

Received: 07 April 2026
Revised: 05 May 2026
Accepted: 08 June 2026
Published: 07 August 2026

Keywords

Artificial Intelligence
Patient Education
Physical Therapy

ABSTRACT

Purpose: This study aimed to compare the content quality, reliability, readability, and structure of responses from different artificial intelligence (AI) models to the most frequently asked public questions on physical therapy for gonarthrosis. **Methods:** Ten highly frequent questions were collected from Google Trends and Google's "People Also Ask" section in Turkiye (2022–2025) using the lay term "physical therapy in knee osteoarthritis." Questions were submitted directly to ChatGPT, Google Gemini, and DeepSeek without additional prompts to simulate actual user behavior. Content reliability was assessed using mDISCERN (0–5) and overall quality with the Global Quality Score (GQS, 1–5). Readability was calculated via the Ateşman Turkish Readability Index (AOI), and structural features were analyzed by sentence count. Two independent raters performed evaluations, with disagreements resolved by a third rater (ICC=0.84). Inter-model comparisons were conducted using the Friedman test, followed by Wilcoxon signed-rank tests with Bonferroni correction ($p<0.016$). **Finding:** Significant differences were observed in mDISCERN ($\chi^2(2)=9.80$, $p=0.007$; $W=0.49$) and GQS ($\chi^2(2)=11.10$, $p=0.004$; $W=0.56$), with Gemini outperforming ChatGPT in reliability ($p=0.005$) and quality ($p=0.006$). ChatGPT showed higher readability than Gemini ($p=0.005$) and DeepSeek ($p=0.013$). DeepSeek responses were significantly longer ($p=0.001$). mDISCERN and GQS were strongly correlated ($r=0.827$, $p<0.001$), while quality and readability were negatively correlated. **Conclusion:**AI responses on physical therapy for gonarthrosis differ in quality and readability. Gemini excelled in reliability and overall quality, whereas ChatGPT offered higher readability. Response length did not reflect information quality. These findings emphasize the importance of AI model selection for patient education and clinical communication.

1. Introduction

Knee osteoarthritis (gonarthrosis) is a common degenerative joint disease that significantly affects patients' mobility, function, and quality of life. Physical therapy is widely recommended as a non-pharmacological intervention to alleviate pain, improve joint function, and enhance patient outcomes [1]. With the increasing accessibility of digital information, patients often seek health-related guidance online, raising concerns about the reliability, quality, and readability of information available to the public [2].

Patient-oriented questions obtained from search engines, such as those highlighted by Google Trends, reflect the real-world information-seeking behavior of the general population regarding knee osteoarthritis and its management. These questions provide insight into the topics that are most frequently of interest to the public, offering a valuable data source for evaluating Artificial Intelligence responses [3,4]. However, because such content is not always produced or verified by healthcare professionals, systematic evaluation of AI-generated responses is essential. Ensuring professional oversight is particularly important in healthcare contexts, where misleading or

incomplete information may impact patient understanding, decision-making, and adherence to treatment recommendations [5].

This study aims to systematically evaluate and compare the content quality, reliability, readability, and structural characteristics of AI responses to the most frequently asked public questions about physical therapy in gonarthrosis. By assessing these aspects, we seek to provide insights into the suitability and limitations of AI tools in patient education and clinical guidance.

2. Materials and Methods

2.1. Study design

This study was conducted as a cross-sectional observational investigation, in which AI-generated responses to patient-oriented questions about physical therapy for knee osteoarthritis were evaluated. The study was designed in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational and cross-sectional research [6].

2.2. Generation of patient-oriented questions

To capture real-world information needs regarding knee osteoarthritis, patient-oriented questions were collected using Google Trends (Google LLC, USA) and the "People Also Ask" feature of the Google search engine in Türkiye, covering the period from 2022 to 2025. All searches were performed in Turkish in incognito mode to reduce personalized search effects and individual biases. The raw pool of questions was initially filtered by removing duplicates, semantically overlapping items, questions not directly related to knee osteoarthritis, and overly technical terms that would not reflect common patient language. Two clinical experts subsequently reviewed the selected questions for relevance to patients and clinical representativeness, and revisions were made based on consensus. The final set consisted of 10 frequently asked questions, which were then submitted to AI models in Turkish to evaluate their responses.

2.3. Generations of AI responses

In this study, three publicly accessible large language models were evaluated on December 1, 2025: ChatGPT® (GPT-5 Mini, OpenAI), Gemini® 3 Flash (Google), and DeepSeek V3® [7-9]. To ensure temporal consistency, all queries were submitted on the same day. Each question was presented directly to the models without additional prompts, instructions, or role definitions, simulating real-world user behavior. A total of 10 patient-oriented questions were posed in Turkish to each model, resulting in 30 responses overall (10 questions × 3 models). Responses were collected in their raw form and saved as digital text files, labeled according to the generating model.

2.4. Content evaulations

The generated responses were systematically evaluated in terms of content reliability, overall quality, readability, and structural characteristics. Content reliability was assessed using the modified DISCERN (MD) tool, which consists of five questions scored between 0 and 5, with higher scores indicating greater reliability [10]. The overall quality of the responses was evaluated using the Global Quality Score (GQS), a 5-point Likert scale measuring the usefulness and flow of health-related information [11].

Readability was assessed using the Ateşman Turkish Readability Index (ARI), which is specifically adapted for Turkish texts and provides a quantitative measure of text comprehensibility, with higher scores indicating easier readability [12]. Structural characteristics of the responses were evaluated based on sentence count, providing an indicator of response length and complexity. All evaluations were independently performed by two reviewers experienced in the clinical field (BKA, MT). In cases of disagreement, a third reviewer was consulted to reach consensus (AKÖ). This multi-parameter evaluation approach enabled a comprehensive and multidimensional assessment of AI-generated health information.

2.5. Statistical analysis

Variables were presented as descriptive statistics using mean ± standard deviation values. Differences among responses generated by the three AI models for the same questions were analyzed using the Friedman test. For variables with statistically significant differences, pairwise comparisons were performed using the Wilcoxon signed-rank test. To control for type I error in multiple comparisons, Bonferroni correction was applied, and the adjusted level of statistical significance was set at $p < 0.016$.

Effect sizes for the Friedman test were calculated using Kendall's W, while effect sizes for pairwise comparisons conducted with the Wilcoxon signed-rank test were calculated as r ($Z/\sqrt{N}$). Effect

size interpretations were based on established thresholds, with values of 0.10–0.29 considered small, 0.30–0.49 moderate, and ≥0.50 large effects.

3. Results

Descriptive statistics and comparative analyses of MD, GQS, ARI, and Sentence Count (SC) are presented in Table 1.

Table 1. Descriptive and comparative analysis of AI responses

Variables	ChatGPT	Gemini	DeepSeek	Friedman χ^2 (df=2)	p	Kendall's W	Significant Post-hoc Comparisons (Wilcoxon, Bonferroni)
mDISCERN	2.20±0.42	3.00±0.00	2.50±0.53	9.80	0.007	0.49	Gemini > ChatGPT (Z=−2.83; p=0.005; r=0.89)
GQS	2.80±0.92	4.20±0.42	3.50±0.71	11.10	0.004	0.56	Gemini > ChatGPT (Z=−2.74; p=0.006; r=0.87)
RI	78.64±6.46	64.76±8.13	71.99±3.08	14.60	0.001	0.73	ChatGPT > Gemini (Z=−2.80; p=0.005; r=0.89); ChatGPT > DeepSeek (Z=−2.50; p=0.013; r=0.79); DeepSeek > Gemini (Z=−2.60; p=0.009; r=0.82)
SC	27.10±19.12	40.50±16.32	70.60±17.69	14.31	0.001	0.72	DeepSeek > ChatGPT (Z=−2.60; p=0.009; r=0.82); DeepSeek > Gemini (Z=−2.81; p=0.005; r=0.89)

GQS: Global Quality Score; RI: Readability Index; SC: Sentence Count. Data are presented as mean ± standard deviation. Comparisons among models were performed using the Friedman test, and for statistically significant results, Wilcoxon signed-rank tests with Bonferroni correction were applied (adjusted significance level $p < 0.016$). Effect sizes were reported as Kendall's W for Friedman tests and r ($Z/\sqrt{N}$) for pairwise comparisons.

Among the models, Gemini demonstrated the highest mean MD score (3.00 ± 0.00), followed by DeepSeek (2.50 ± 0.53) and ChatGPT (2.20 ± 0.42). Notably, Gemini scores were consistent across all questions. Similarly, Gemini achieved the highest GQS values (4.20 ± 0.42), whereas DeepSeek (3.50 ± 0.71) and ChatGPT (2.80 ± 0.92) showed lower scores, with greater variability observed in ChatGPT responses.

In contrast, ChatGPT exhibited the highest readability (78.64 ± 6.46), followed by DeepSeek (71.99 ± 3.08) and Gemini (64.76 ± 8.13). DeepSeek generated the longest responses in terms of sentence count (40.50 ± 16.32), whereas ChatGPT produced shorter responses (27.10 ± 19.12), with substantial variability across models.

Significant differences were found among models for MD ($\chi^2(2)=9.80$, $p=0.007$, $W=0.49$) and GQS ($\chi^2(2)=11.10$, $p=0.004$, $W=0.56$). Post-hoc analyses revealed that Gemini provided significantly higher MD ($p=0.005$, $r=0.89$) and GQS ($p=0.006$, $r=0.87$) scores compared to ChatGPT, while other pairwise comparisons were not significant after Bonferroni correction.

Readability also differed significantly ($\chi^2(2)=14.60$, $p=0.001$, $W=0.73$), with ChatGPT outperforming both Gemini ($p=0.005$, $r=0.89$) and DeepSeek ($p=0.013$, $r=0.79$). DeepSeek responses were also more readable than Gemini ($p=0.009$, $r=0.82$). Sentence count showed significant variation ($\chi^2(2)=14.31$, $p=0.001$, $W=0.72$), with DeepSeek producing longer responses than both ChatGPT ($p=0.009$, $r=0.82$) and Gemini ($p=0.005$, $r=0.89$).

As shown in Table 2, Spearman correlation analysis demonstrated a strong positive correlation between mDISCERN and GQS scores ($r = 0.827$, $p < 0.001$). The Readability Index was weakly and negatively correlated with both mDISCERN ($r = −0.393$, $p = 0.032$) and GQS ($r = −0.369$, $p = 0.045$), whereas Sentence Count was not significantly associated with either score (both $p > 0.05$).

Table 2. Results of spearman correlation analysis between variables

		mDISCERN	GQS	RI	SC
mDISCERN	Spearman's rho	—	0.827	−0.393	0.303
	p	—	<0.001	0.032	0.103
GQS	Spearman's rho	0.827	—	−0.369	0.283
	p	<0.001	—	0.045	0.129

GQS: Global Quality Score; RI: Readability Index; SC: Sentence Count.

4. Discussion

This study aimed to compare the responses of different AI models to public questions regarding physical therapy in gonarthrosis in terms of content quality, reliability, readability, and structural characteristics. Our findings indicate that AI-generated health content varies depending on the model, with each model presenting specific strengths and limitations.

The Gemini model achieved the highest scores in both MD and GQS and delivered responses in a highly uniform structure. This aligns with the literature reporting that next-generation AI models tend to provide higher quality content [13,14]. However, the template-like nature of Gemini's responses may limit its ability to offer personalized information. ChatGPT, on the other hand, produced shorter and more comprehensible responses; according to the ARI, its responses scored 78.4, corresponding to a 5th–6th grade level, making them easily understandable for the general public [15] . Nevertheless, ChatGPT's quality scores were more variable compared to Gemini, likely due to the model generating adaptive responses depending on the question. The literature suggests that simpler content improves readability but may pose a risk of superficial information due to lack of detail [13,16]

DeepSeek provided longer and more detailed responses, with a significantly higher average sentence count compared to the other models. While this information density may enhance content quality, it also risks overwhelming users in a clinical context. Additionally, the absence of a physiotherapy expert perspective in DeepSeek's responses suggests that its training data may reflect healthcare system-specific differences. This observation aligns with previous reports indicating heterogeneous performance across AI systems, with response quality and structure being model-dependent [14,17].

The negative correlation between readability and quality highlights differences arising from language adaptation and translation processes in AI models. More detailed and complex responses tend to increase quality scores but reduce public comprehension, whereas simpler and adaptive responses enhance readability but show greater variability in quality scores. These findings emphasize the importance of model selection in AI-based health information provision and clinical communication.

When compared with the literature, our findings are largely consistent with general trends. Wang et al. demonstrated that AI-assisted monitoring and rehabilitation systems can play an active role in clinical care and enable early intervention. Çiğdem & Deniz reported high accuracy of AI in diagnosing and predicting the prognosis of knee osteoarthritis [18,19]. Yoo et al. indicated that LLMs such as ChatGPT produce content that is presentation-wise appropriate but may still be complex for public comprehension [20]. Turkish studies on AI-generated osteoarthritis educational materials also revealed that content quality and readability are model-dependent [21].

4.1. Limitation of the study

The findings of this study should be interpreted in light of several methodological limitations. First, the AI-generated responses analyzed were obtained at a specific point in time using particular versions of each model. Given the rapidly evolving nature of large language models, responses produced at later dates or with updated versions may differ in content, structure, and quality. This limitation reflects the inherent characteristics of cross-sectional evaluations of AI-generated health information.

Second, in this study, the questions provided to the AI models were framed using a non-directive, zero-shot approach to approximate typical user behavior. Although this method was chosen to reflect how the general public might interact with AI, the study did not assess how alternative prompting techniques, more structured question formats, or iterative interactions could influence the quality, depth, or clarity of the responses.

Third, it should be noted that the risk of misinformation or hallucinated content is an inherent limitation of AI-generated outputs. Nevertheless, within the boundaries of the predefined evaluation criteria and the clinical assessment framework used in this study, no clear

instances of misinformation or hallucination were observed in the analyzed responses.

Finally, readability analyses relied on standard linguistic indices to estimate text complexity. These metrics may not accurately reflect real-world patient comprehension or the ability to apply the information in practical rehabilitation contexts. Additionally, language-specific characteristics of Turkish may influence readability scores, making them relatively more favorable than actual patient understanding.

5. Conclusion

In this context, our study demonstrates that Gemini provides high quality and consistent responses, ChatGPT offers more understandable and user-friendly outputs, and DeepSeek excels in information density but lacks an expert perspective. DeepSeek did not incorporate insights from physical medicine and rehabilitation (PMR) specialists, an important detail considering that in Turkiye, patient education and rehabilitation for gonarthrosis are typically conducted under the guidance of PMR experts. These findings indicate that AI-based systems have potential utility in patient education and clinical communication, yet their outputs must be reviewed and validated by healthcare professionals, especially when local clinical practices or specialist input are critical. The selection of models should consider differences in content quality, readability, and structural features to match the needs of users and patients. Future studies should evaluate AI responses' adaptation and clinical validity across different languages, cultural contexts, and healthcare systems to ensure safe and effective application.

Acknowledgement

Presentation: This work was presented orally at the 31st National Physical Medicine and Rehabilitation Congress, 2026.

Conflict of Interest

No conflict of interest is declared by tehe authors. In addition, no financial support was received.

Ethics Committee

This study did not involve human or animal subjects and used only publicly available AI outputs; therefore, ethics approval was not required

Author Contributions

Study Design, BKA, MT; Data Collection, BKA, MT; Statistical Analysis, BKA, MT; Data Interpretation, BKA, MT; Manuscript Preparation, BKA, MT; Literature Search, BKA, MT. All authors have read and agreed to the published version of the manuscript.

References

1. Ni, X., Hu, L., Zhang, X., Wang, Z., Yan, C., Peyrodie, L., & Hu, S. (**2024**). Physical therapy options for knee osteoarthritis: A review. *Medicine*, *103*(30), e38415. [PubMed] [CrossRef]
2. Hirani, R., Noruzi, K., Khuram, H., Hussaini, A. S., Aifuwa, E. I., Ely, K. E., & Etienne, M. (**2024**). Artificial intelligence and healthcare: a journey through history, present innovations, and future possibilities. *Life*, *14*(5), 557. [PubMed] [CrossRef]
3. Jia, X., Pang, Y., & Liu, L. S. (**2021**). Online health information seeking behavior: a systematic review. In *Healthcare,* 9(12)1740. [PubMed] [CrossRef]
4. Bachl, M., Link, E., Mangold, F., & Stier, S. (**2024**). Search engine use for health-related purposes: Behavioral data on online health information-seeking in Germany. *Health Communication*, *39*(8), 1651-1664. [PubMed] [CrossRef]
5. Hancı, V., Ergün, B., Gül, Ş., Uzun, Ö., Erdemir, İ., & Hancı, F. B. (**2024**). Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. *Medicine*, *103*(33), e39305. [PubMed] [CrossRef]
6. Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (**2007**). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet*, *370*(9596), 1453-1457. [PubMed] [CrossRef]
7. OpenAI. (**2025**). *ChatGPT (GPT-5 Mini) [Computer software]*. San Francisco, CA, USA: OpenAI. Retrieved December 1, 2025, from https://chat.openai.com
8. Google. (**2025**). *Gemini 3 Flash [Computer software]*. Mountain View, CA, USA: Google. Retrieved December 1, 2025, from https://gemini.google.com
9. DeepSeek AI. (**2025**). *DeepSeek Model [Computer software]*. San Francisco, CA, USA: DeepSeek AI. Retrieved December 1, 2025, from https://www.deepseek.ai
10. Charnock, D., Shepperd, S., Needham, G., & Gann, R. (**1999**). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of epidemiology and community health*, *53*(2), 105-111. [PubMed] [CrossRef]
11. Bernard, A., Langille, M., Hughes, S., Rose, C., Leddin, D., & Van Zanten, S. V. (**2007**). A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *American Journal*

of Gastroenterology, *102*(9), 2070-2077. [PubMed] [CrossRef]
12. Ateşman, E. (**1997**). Measurement of readability in Turkish. *Ankara University Tömer Language Journal*, 58, 71–74.
13. Hershenhouse, J. S., Mokhtar, D., Eppler, M. B., Rodler, S., Storino Ramacciotti, L., Ganjavi, C., & Cacciamani, G. E. (**2025**). Accuracy, readability, and understandability of large language models for prostate cancer information to the public. *Prostate Cancer and Prostatic Diseases*, *28*(2), 394-399. [PubMed] [CrossRef]
14. Marshall, R. F., Mallem, K., Xu, H., Thorne, J., Burkholder, B., Chaon, B., & Berkenstock, M. (**2024**). Investigating the accuracy and completeness of an artificial intelligence large language model about uveitis: an evaluation of ChatGPT. *Ocular İmmunology and İnflammation*, *32*(9), 2052-2055. [PubMed] [CrossRef]
15. Weiss, B. D. (**2007**). Help patients understand. *Manual for clinicians. AMA Foundation*.
16. Bange, M., Huh, E., Novin, S. A., Hui, F. K., & Yi, P. H. (**2019**). Readability of patient education materials from RadiologyInfo. org: has there been progress over the past 5 years?. *American Journal of Roentgenology*, *213*(4), 875-879. [PubMed] [CrossRef]
17. Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., & Tseng, V. (**2023**). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health*, *2*(2), e0000198. [PubMed] [CrossRef]
18. Wang, F., Wang, L., Zhong, L., Feng, J., & Wang, X. (**2025**). Artificial intelligence for knee osteoarthritis care and rehabilitation: a systematic review. *Pain Management Nursing*. [PubMed] [CrossRef]
19. Cigdem, O., & Deniz, C. M. (**2023**). Artificial intelligence in knee osteoarthritis: a comprehensive review for 2022. *Osteoarthritis Imaging*, *3*(3), 100161. [PubMed] [CrossRef]
20. Yoo, M., & Jang, C. W. (**2025**). Presentation suitability and readability of ChatGPT's medical responses to patient questions about on knee osteoarthritis. *Health Informatics Journal*, *31*(1), 14604582251315587. [PubMed] [CrossRef]
21. Özbek, İ. C. (**2025**). Evaluation of Artificial Intelligence-Supported Osteoarthritis Information Texts: Content quality and readability analysis. *Journal of Physical Medicine & Rehabilitation Sciences*, *28*(1), 21. [CrossRef]

Author(s) 2026. **Open Access** This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/ licenses/by/4.0/).