

Assessing the Quality, Safety, and Readability of Large Language Model Responses from ChatGPT and Gemini for Knee Osteoarthritis Patient Education

Dr. V. Habibullah¹, Dr. Mubasheera Begum², Dr. K. Nizamuddin³, Dr. Dua Harmain⁴, Dr. Siddheesh Rajpurohit⁵, Dr. K. S. Mohammed Shahid⁶

¹ Professor, Department of Unani Pharmacology, Government Unani Medical College, Affiliated to The Tamil Nadu Dr. M.G.R. Medical University, Chennai, Tamil Nadu, India.

Email: urnetdoctor@gmail.com

ORCID: 0000-0002-4702-2325

² Professor, Department of General Medicine (Unani), Government Unani Medical College, Affiliated to The Tamil Nadu Dr. M.G.R. Medical University, Chennai, Tamil Nadu, India.

Email: muskaan.m@gmail.com

ORCID: 0000-0003-2242-5897

³ Research Scholar, Department of General Medicine (Unani), Government Unani Medical College, Affiliated to The Tamil Nadu Dr. M.G.R. Medical University, Chennai, Tamil Nadu, India.

Email: Nizamuddin8962@gmail.com

ORCID: 0009-0008-9745-6958

⁴ Research Scholar, Department of General Medicine (Unani), Government Unani Medical College, Affiliated to The Tamil Nadu Dr. M.G.R. Medical University, Chennai, Tamil Nadu, India.

Email: duaharmain76@gmail.com

ORCID: 0009-0009-8711-6033

⁵ Lead Medical Writer, Pharmacology, Meril Life Sciences Pvt. Ltd., Mumbai, Maharashtra, India.

Email: rajpurohitsiddheesh@gmail.com

ORCID: 0000-0002-2853-1415

⁶ Research Scholar, Department of General Medicine (Unani), Government Unani Medical College, Affiliated to The Tamil Nadu Dr. M.G.R. Medical University, Chennai, Tamil Nadu, India.

Email: ksmohammedshahid@gmail.com

ORCID: 0009-0000-1530-9914

Abstract

Background: Knee osteoarthritis is a common chronic musculoskeletal disorder, and ongoing patient education underpins its non-surgical management. Increasingly, patients turn to AI-powered conversational platforms to make sense of their symptoms, medications, exercises, and surgical options. It remains unclear whether the information these platforms generate is accurate, complete, and clear enough to support sound self-management decisions.

Objective: To compare the quality, safety, and readability of responses generated by ChatGPT and Gemini for knee osteoarthritis patient education.

Methods: Thirty-six patient-relevant questions were drawn from seven major patient education websites (AAOS, ACR, Arthritis Foundation, NHS, CDC, Mayo Clinic, and Cleveland Clinic) and supplemented with clinician-derived and culturally relevant queries, including questions on Unani treatment. Each was entered separately into ChatGPT (GPT-4o, May 2025) and Gemini (Gemini 1.5 Pro, May 2025) with no additional prompting. Five clinicians with more than ten years of experience in musculoskeletal medicine rated every response independently on a five-point Likert scale across five domains (accuracy, completeness, clarity, safety, and trustworthiness), blinded to platform identity. Inter-rater reliability was assessed with ICC(A,1) for single measures and ICC(A,k) for average measures, and readability with ten validated indices. Mann-Whitney U tests with Bonferroni correction (threshold $p < 0.0033$ for 15 comparisons) were applied throughout.

Results: ChatGPT scored significantly higher than Gemini on accuracy (19.69 ± 1.69 vs. 17.83 ± 2.14 ; $U = 307$, $p = 0.0001$, $r = 0.45$) and completeness (18.08 ± 1.76 vs. 16.19 ± 2.52 ; $U = 337$, $p = 0.0004$, $r = 0.41$), both differences surviving Bonferroni correction. Clarity, safety, and trustworthiness did not differ significantly. The average-measures ICC(A,k) improved on ICC(A,1), reaching moderate reliability for completeness (Gemini: 0.567) and approaching it for accuracy (ChatGPT: 0.421). For readability, ChatGPT produced higher Fry

Readability Formula scores (11.90 ± 1.52 vs. 10.70 ± 1.62 ; $p = 0.003$), the only readability difference to survive correction. Response word counts were similar (Gemini: 66.3 ± 13.3 ; ChatGPT: 67.1 ± 18.5 words; $p = 0.796$).

Conclusion: For knee osteoarthritis patient education, ChatGPT showed higher clinician-rated accuracy and completeness than Gemini under the current evaluation framework, while Gemini held a modest readability advantage on the indices examined. Before either platform's output is used in patient-facing educational materials, it should be reviewed by a qualified health professional..

Keywords Knee osteoarthritis, ChatGPT, Gemini, large language model, artificial intelligence, patient education, readability, safety, health communication..

Introduction

Knee osteoarthritis affects more than 500 million people worldwide and ranks among the leading causes of chronic pain and functional disability, the knee being the joint most often involved [1,2]. An aging population, rising obesity, and increasingly sedentary lifestyles continue to push that burden higher [3]. Managing the condition asks a great deal of the patient over time: keeping up with exercise, controlling weight, modifying daily activity, using analgesics judiciously, and eventually weighing surgical options. How well patients manage these tasks depends closely on the quality of the education they receive [4,5].

Clinical guidelines from the American College of Rheumatology and the Osteoarthritis Research Society International place patient education and self-management support among the first-line interventions [4,5]. In practice, though, a time-pressured consultation allows only so much depth and consistency. Patients fill the gaps by searching online, where quality is uneven at best. In South Asian and traditional-medicine settings they also look for information on Unani and other integrative approaches, a layer of complexity that mainstream health-information sources rarely cover.

Large language model platforms, chiefly ChatGPT and Gemini, are increasingly used as health-information sources because they answer natural-language questions in readable, conversational prose. Their uptake has run ahead of the evidence clinicians would normally rely on to recommend them or to advise caution. Earlier evaluations have flagged inconsistent factual accuracy, uneven adherence to current guidelines, a habit of stating information with more confidence than it warrants, and the occasional hallucination of plausible but incorrect content [6,7]. In knee osteoarthritis, poor education can translate into unsafe analgesic use, abandoned exercise, delayed specialist referral, and unrealistic surgical expectations, so these shortcomings carry direct clinical consequences.

Most evaluations of AI-generated medical information have concentrated on general medicine, oncology, and procedural topics, leaving disease-specific evidence for musculoskeletal conditions thin. Knee osteoarthritis education spans biomechanics, lifestyle change, pharmacology, intra-articular therapies, and surgery, demands that differ enough from those of acute or organ-specific diseases that findings from other specialties do not transfer reliably. We therefore set out to evaluate the quality, safety, and readability of large language model responses to knee osteoarthritis patient-education questions, comparing ChatGPT and Gemini as two of the most widely used platforms. The aim was not a simple head-to-head ranking but a characterisation of how well AI-generated content serves this particular condition. To keep that evaluation close to real practice, we built a structured question set that deliberately reached beyond standard Western guideline topics to the culturally specific and integrative-medicine queries, including Unani treatment, that patients in South Asian settings routinely raise.

Methods

Question Development and Categorisation

We built a structured question bank from the FAQ sections of seven major professional and patient-education bodies: the American Academy of Orthopaedic Surgeons (AAOS), American College of Rheumatology (ACR), Arthritis Foundation, National Health Service (NHS), Centers for Disease Control and Prevention (CDC), Mayo Clinic, and Cleveland Clinic. Google searches on the terms knee osteoarthritis, knee arthritis, osteoarthritis of knee, and knee OA patient education turned up supplementary material. The clinician evaluation team added questions drawn from everyday consultations, including queries about Unani treatment of knee OA, which comes up routinely in South Asian musculoskeletal practice, and about culturally common activities such as floor-level prayer and the use of Indian-style toilets. These additions were meant to improve ecological validity in South Asian settings; they were not analyzed as a separate subgroup.

Four clinicians reviewed the pool of 58 candidate questions independently. An item was dropped when at least two reviewers agreed it duplicated another, was too specialized for a patient audience, or was ambiguously worded. To check content validity informally, we confirmed that the surviving questions covered the main domains of knee osteoarthritis patient education. The 36 retained questions were sorted into ten educational domains: general understanding of the condition; symptoms and diagnosis; causes and risk factors; lifestyle modification and self-care; exercise and physiotherapy; pharmacological treatment, including medication safety

and warning signs; surgical management and escalation of care, including intra-articular injections and knee replacement; prognosis and disease progression; Unani and integrative medicine; and health-information resources and self-management support. Supplementary Appendix 1 gives the full question-to-domain mapping, and Figure 1 shows the selection process.

Figure 0. Question Selection Process

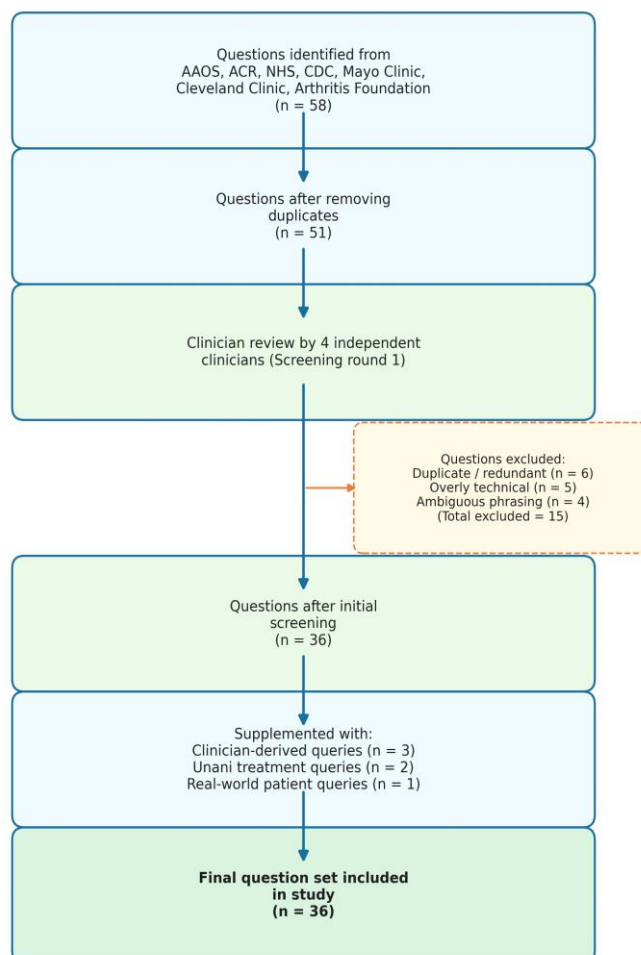


Figure 1. Question-selection flowchart. Fifty-eight candidate questions were identified, reviewed, and reduced to a final set of 36 across ten educational domains.

Fifty-eight questions were identified across seven patient education sources. After removing duplicates (n = 7) and clinician exclusion of overly technical or ambiguous items (n = 15), 36 questions were finalized across ten educational domains.

The final set of 36 questions was determined by considering three key factors. First, it was consistent with the number of questions used in similar methodological studies evaluating AI systems in musculoskeletal medicine [11]. Second, it ensured balanced representation across all ten assessment domains, with approximately three to four questions allocated to each domain. Third, it reflected a practical workload that the expert evaluation panel could complete while maintaining scoring accuracy and consistency. With five evaluators assessing 36 questions across five evaluation criteria for two AI platforms, the study generated a total of 1,800 individual ratings. This represented a feasible number of assessments that could be completed with a high level of quality within the available timeframe. A formal statistical power analysis was not conducted before the study, which should be acknowledged as a limitation. Future investigations are encouraged to incorporate prospective power calculations based on the anticipated effect sizes between the AI platforms being compared.

Study Design Overview

The overall study methodology from question sourcing to statistical analysis is summarized in Figure 2.

Figure 1. Study Workflow

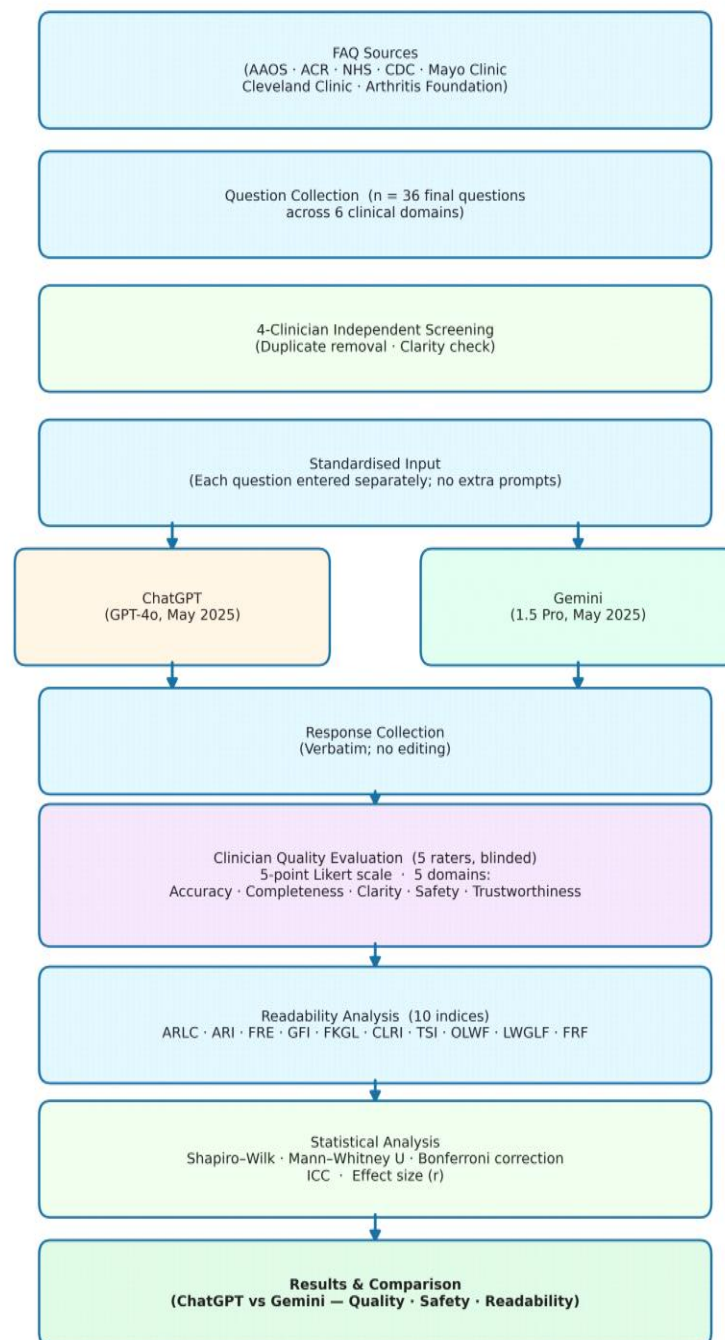


Figure 2. Study workflow. Overview of the full methodology from question sourcing to statistical analysis.

Each question was entered separately into ChatGPT and Gemini in independent sessions, without additional prompting. Responses were evaluated by five blinded raters across five quality domains and supplemented by an automated readability assessment using ten validated indices.

Response Generation

All 36 questions were submitted individually to ChatGPT (GPT-4o, OpenAI Inc.; accessed May 2025) and Gemini (Gemini 1.5 Pro, Google LLC; accessed May 2025), with each query initiated in a new chat session. No predefined system prompts, role instructions, follow-up questions, or additional clarifications were provided. Instead, each platform received only the original question as input. The generated responses were recorded exactly as produced and stored separately for each platform in a structured data spreadsheet. This approach was intended to replicate the typical experience of a non-technical user seeking information independently through these AI systems.

Quality Evaluation

Five clinicians with expertise in musculoskeletal medicine, including four faculty members and one Head of Department, each possessing a minimum of ten years of clinical experience, evaluated the AI-generated responses. The assessments were conducted independently, with the evaluators blinded to the identity of the platforms. To ensure anonymity, the responses were identified as **System A** and **System B** and presented in a randomized order. The independent ratings provided by all five evaluators were incorporated into the statistical analyses. Every response was assessed using a five-point Likert scale across the five domains described in Table 1. The evaluation rubric was prepared by the authors based on established quality principles from the DISCERN and HONcode frameworks, together with evidence from previous studies evaluating AI for patient education. However, the rubric was not independently validated, and no operational anchor examples for each Likert scale point were developed before the evaluation. These two aspects are addressed later in the **Limitations** section.

Table 1. Scoring rubric for the five quality domains.

Domain	Construct assessed	Score 1	Score 3	Score 5
Accuracy	Scientifically and medically correct	Mostly incorrect	Partially correct	Completely accurate
Completeness	Covers all key points for patient education	Major gaps present	Partly comprehensive	Fully comprehensive
Clarity	Understandable to a non-specialist patient	Very difficult	Moderately clear	Clear and accessible
Safety	Free from harmful or misleading advice	Potentially harmful	Some concern	Clearly safe
Trustworthiness	Aligned with current clinical guidelines	Contradicts guidelines	Partially aligned	Fully aligned

The five domain scores (each 1–5) from all five raters were summed per question, giving a maximum possible score of 25 per domain. All analyses used these independent ratings rather than any consensus rating. Where the methods mention "consensus discussion," it refers only to a post-scoring step for flagging ambiguous items for later methodological reflection; it did not alter any recorded score used in the analyses.

Safety was defined in advance by four criteria adapted from the DISCERN and HONcode principles: (1) suggesting that medical consultation could be delayed or was unnecessary; (2) encouraging unsupervised starting or stopping of medications; (3) recommending therapies without solid clinical evidence and without appropriate qualification; and (4) containing statements likely to cause disproportionate anxiety. A response meeting any one of these criteria was eligible for a safety score of 1–2.

Inter-Rater Reliability

Inter-rater reliability was assessed with a two-way mixed-effects intraclass correlation coefficient based on absolute agreement, and we report both single-measure ICC(A,1) and average-measure ICC(A,k). ICC(A,1) reflects how reliable a single rater's assessment is; ICC(A,k) reflects the reliability of the mean across all five raters and is the better statistic when the score used in analysis is itself an aggregate [8]. We interpreted values against the Koo and Li (2016) thresholds: ICC < 0.50 = poor, 0.50–0.74 = moderate, 0.75–0.90 = good, > 0.90 = excellent.

Readability Evaluation

Each AI-generated response was run through an online multi-index readability tool, which computed ten indices: Average Reading Level Consensus (ARLC), Automated Readability Index (ARI), Flesch Reading Ease (FRE), Gunning Fog Index (GFI), Flesch–Kincaid Grade Level (FKGL), Coleman–Liau Readability Index (CLRI), Text Standard Index (TSI), Original Linsear Write Formula (OLWF), Linsear Write Grade Level Formula (LWGLF), and Fry Readability Formula (FRF). Of these, the Flesch Reading Ease and Flesch–Kincaid Grade Level are the two most widely reported in medical readability research and serve as our primary reference metrics; the remaining indices are reported for completeness and cross-comparison. For the grade-level indices, lower scores mean simpler text; for FRE, higher scores mean greater accessibility. As a benchmark, most health-literacy guidance recommends that patient-facing material be written at roughly a sixth-to-eighth grade level, so grade-level scores above that range indicate text likely to exceed the reading ability of a substantial share of the general patient population. Word counts served as a proxy for response length.

Statistical Analysis

Data were managed using Microsoft Excel and analyzed using SPSS version 27. The Shapiro–Wilk test confirmed a non-normal distribution for most variables. Group comparisons used the Mann–Whitney U test; U statistics, Z scores, effect sizes ($r = |Z|/\sqrt{N}$), and 95% confidence intervals for r (computed via Fisher z-transformation; Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale: Erlbaum; 1988) are reported alongside the p-values. To account for the 15 simultaneous comparisons (five quality domains and ten readability indices), we applied a Bonferroni correction, giving an adjusted threshold of $p < 0.0033$. Uncorrected p-values are also reported. All tests were two-tailed.

Ethical Approval

The study involved the evaluation of AI-generated text only and did not involve human subjects, identifiable patient data, or patient interaction. Therefore, formal institutional ethics committee clearance was not required.

Results

Quality Domain Evaluation

ChatGPT outscored Gemini on accuracy and completeness; both differences survived Bonferroni correction and reached medium-to-large effect sizes. Clarity, safety, and trustworthiness did not differ significantly between the platforms. Table 2 and Figure 3 give the full results.

Table 2. Quality domain scores: Gemini vs. ChatGPT (n = 36 questions per platform).

Domain	Gemini Mean SD	±	ChatGPT Mean SD	±	U	Z	p-value	r (95% CI)	Significance
Accuracy	17.83 2.14	±	19.69 1.69	±	307	-3.840	0.0001	0.453 [0.25, 0.62]	** Bonferroni
Completeness	16.19 2.52	±	18.08 1.76	±	337	-3.503	0.0004	0.413 [0.20, 0.59]	** Bonferroni
Clarity	16.64 1.62	±	17.19 2.03	±	544	-1.177	0.234	0.139 [-0.10, 0.36]	ns
Safety	16.50 1.90	±	15.86 1.71	±	764	1.312	0.184	0.155 [-0.08, 0.37]	ns
Trustworthiness	16.78 1.76	±	17.53 1.73	±	478	-1.915	0.052	0.226 [-0.01, 0.44]	ns

Scores = sum of the five raters' Likert ratings (1–5 scale) per question; maximum = 25. ** $p < 0.0033$ (Bonferroni-corrected threshold for 15 tests). ns = not significant. Effect size r : small ≥ 0.10 , medium ≥ 0.30 , and large ≥ 0.50 .

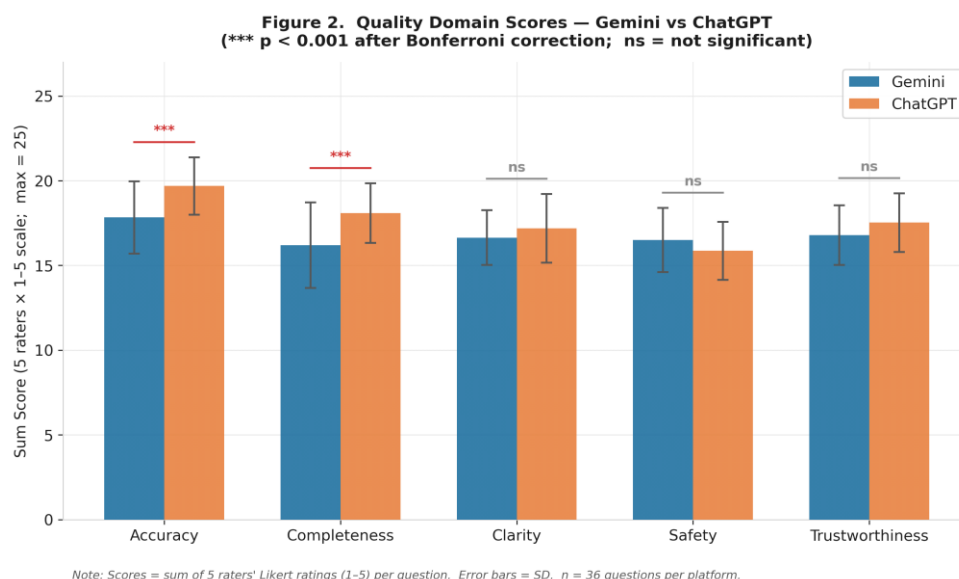


Figure 3. Quality domain scores: Gemini vs. ChatGPT ($n = 36$ questions per platform). Clustered bar chart of mean summed ratings across the five domains.

Scores represent the summed Likert ratings of five evaluators per question (scale 1–5; maximum score = 25). Error bars = SD. ** $p < 0.0033$ (Bonferroni-corrected threshold); ns = not significant.

Inter-Rater Reliability

The single-measure ICC(A,1) results demonstrated poor reliability across all five quality domains for both AI platforms, with values ranging from 0.002 to 0.208. These findings indicate minimal absolute agreement among individual raters when assessing the responses. In comparison, the average-measure ICC(A,k), which provides a more appropriate estimate of the reliability of the combined scores from all five raters used in the analysis, showed considerable improvement over the single-rater measurements. Among the evaluated domains, Completeness for Gemini achieved moderate reliability, with an ICC(A,k) value of 0.567 (95% CI: 0.30–0.75). Accuracy for ChatGPT showed a value approaching the moderate reliability category but remained below the defined threshold, with an ICC(A,k) of 0.421 (95% CI: 0.07–0.67). The remaining domains continued to demonstrate poor reliability. Specifically, the ICC(A,k) values for clarity, safety, and trustworthiness indicated poor to negligible agreement among the evaluators. Complete ICC results are provided in Table 3.

Table 3. Inter-rater reliability: ICC(A,1) and ICC(A,k) for all the quality domains.

Domain	ICC(A,1) Gemini	95% CI	ICC(A,k) Gemini	95% CI	ICC(A,1) ChatGPT	95% CI	ICC(A,k) ChatGPT	95% CI
Accuracy	0.115	[0.01,0.27]	0.395	[0.04,0.65]	0.127	[0.01,0.29]	0.421	[0.07,0.67]
Completeness	0.208	[0.08,0.38]	0.567	[0.30,0.75]	0.099	[0.00,0.24]	0.353	[0.01,0.62]
Clarity	0.048	[-0.04,0.19]	0.201	[-0.27,0.54]	0.130	[0.03,0.28]	0.428	[0.13,0.66]

Safety	0.002	[-0.07,0.13]	0.012	[-0.52,0.42]	0.047	[-0.01,0.15]	0.197	[-0.07,0.47]
Trustworthiness	0.043	[-0.04,0.17]	0.183	[-0.23,0.51]	0.091	[0.01,0.22]	0.335	[0.03,0.59]

$ICC(A,1)$ = single-measure intraclass correlation, absolute agreement. $ICC(A,k)$ = average-measures ICC across the five raters. Interpretation (Koo & Li, 2016): < 0.50 = poor, $0.50-0.74$ = moderate, $0.75-0.90$ = good, > 0.90 = excellent. $k = 5$ raters.

Implications of Low Inter-Rater Reliability

The poor single-measure $ICC(A,1)$ values across all domains are a central finding, and they belong in the results rather than tucked away among the limitations. A few factors likely contributed. No formal rater-calibration exercise preceded scoring, and calibration rounds anchored to exemplars at each scale level usually improve the ICC considerably. The five raters, though all experienced clinicians, came from different subspecialties and may have read the anchors differently, particularly for subjective constructs such as clarity and trustworthiness. The responses themselves also sat in a narrow quality band: when most cluster between 3 and 4 on a five-point scale, within-rater variance can outweigh between-question variance and mathematically suppress the ICC.

In practical terms, the non-significant results for clarity, safety, and trustworthiness cannot be read as evidence of equivalence; they may instead reflect insufficient measurement precision. It is worth being exact about what the two ICC estimates say here. Agreement was acceptable only at the aggregate level, where the average-measures $ICC(A,k)$ reached moderate values in a few domains, while agreement between individual raters, captured by $ICC(A,1)$, stayed weak throughout. That split is a familiar feature of subjective clinical scoring: experienced clinicians reading the same free-text response through slightly different anchors will disagree on a single rating even when their pooled judgement is fairly stable. The accuracy and completeness differences, which survived correction for multiple comparisons, are more robust to this concern, because their between-platform effect sizes ($r = 0.45$ and 0.41) were large enough to surface despite the rater noise. Researchers running similar evaluations should add a calibration phase and report both $ICC(A,1)$ and $ICC(A,k)$ as primary methodological outcomes. A chance-corrected agreement statistic such as Krippendorff's alpha would also serve as a useful robustness check alongside the ICC.

Readability Analysis

Following Bonferroni adjustment, eight out of the ten evaluated readability indices demonstrated no statistically significant variation between the two AI platforms. ChatGPT responses showed significantly higher Fry Readability Formula scores compared with Gemini (11.90 ± 1.52 vs. 10.70 ± 1.62 ; $p = 0.003$), representing the only readability measure that remained significant after correction. This finding indicates that ChatGPT outputs exhibited slightly higher textual complexity according to this specific readability assessment. The Coleman-Liau Readability Index showed a significant difference when evaluated using the unadjusted significance level (13.18 ± 2.99 vs. 11.30 ± 3.50 ; $p = 0.017$); however, this difference was not statistically significant after Bonferroni correction.

The number of words generated per response was comparable between the two platforms (Gemini: 66.3 ± 13.3 words; ChatGPT: 67.1 ± 18.5 words; $p = 0.796$), suggesting that the observed readability variation was associated with differences in sentence formation and vocabulary selection rather than response length. Overall, both platforms produced responses that exceeded the recommended readability standards for patient education materials. Most readability indices indicated grade-level estimates within the eleventh to thirteenth grade range, while the average Flesch Reading Ease scores for Gemini (55.2) and ChatGPT (48.5) were categorized within the "fairly difficult" range. Therefore, although a measurable statistical difference existed between the platforms, the magnitude of this difference was minor compared with the common limitation that both systems produced content above the recommended sixth- to eighth-grade reading level. The complete readability analysis for both platforms is presented in Table 4 and Figure 4.

Table 4. Readability index comparison: Gemini vs. ChatGPT (n = 36 responses per platform).

Index	Gemini Mean $\pm$ SD	ChatGPT Mean $\pm$ SD	p-value	Significance
ARLC	11.10 $\pm$ 4.30	11.58 $\pm$ 2.72	0.574	ns
ARI	13.53 $\pm$ 9.40	13.53 $\pm$ 4.30	1.000	ns
FRE †	55.19 $\pm$ 21.48	48.47 $\pm$ 18.33	0.158	ns
GFI	11.81 $\pm$ 4.31	11.35 $\pm$ 3.33	0.615	ns
FKGL	11.32 $\pm$ 7.28	10.93 $\pm$ 3.56	0.777	ns
CLRI	11.30 $\pm$ 3.50	13.18 $\pm$ 2.99	0.017	* uncorrected
TSI	10.13 $\pm$ 4.66	9.66 $\pm$ 3.09	0.613	ns
OLWF ‡	72.17 $\pm$ 12.85	69.67 $\pm$ 11.00	0.378	ns
LWGLF	11.18 $\pm$ 5.56	10.52 $\pm$ 4.02	0.567	ns
FRF	10.70 $\pm$ 1.62	11.90 $\pm$ 1.52	0.003	** Bonferroni

† FRE: higher = easier-to-read. ‡ OLWF uses a different scale than the other indices. All other indices were grade level (lower = simpler text). ** $p < 0.0033$ (Bonferroni-corrected). * $p < 0.05$ (uncorrected). ns = not significant. Effect sizes (r) and 95% CI: ARLC 0.066 $[-0.17, 0.29]$; ARI 0.000 $[-0.23, 0.23]$; FRE 0.166 $[-0.07, 0.38]$; GFI 0.059 $[-0.18, 0.29]$; FKGL 0.033 $[-0.20, 0.26]$; CLRI 0.281 $[0.05, 0.48]$; TSI 0.060 $[-0.17, 0.29]$; OLWF 0.104 $[-0.13, 0.33]$; LWGLF 0.067 $[-0.17, 0.30]$; FRF 0.350 $[0.13, 0.54]$.

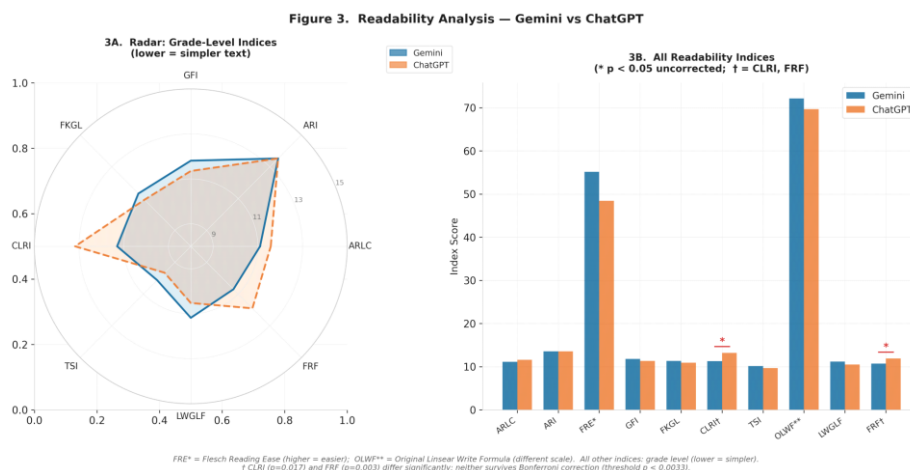


Figure 4. Readability analysis: Gemini vs. ChatGPT. Panel A: radar chart (grade-level indices); Panel B: bar chart (all ten indices).

Panel A: Radar chart of grade-level indices (lower = simpler). Panel B: Bar chart of all ten indices. * $p < 0.05$ uncorrected; ** $p < 0.003$ (Bonferroni-corrected, FRF only).

Effect Sizes Across All Comparisons

Figure 5 plots the effect sizes for all 15 comparisons as a forest plot. The two quality differences that survived correction were medium-to-large by Cohen's conventions (accuracy $r = 0.45$; completeness $r = 0.41$), and the

surviving readability difference was a medium effect (FRF $r = 0.35$). Every other comparison was small or negligible.

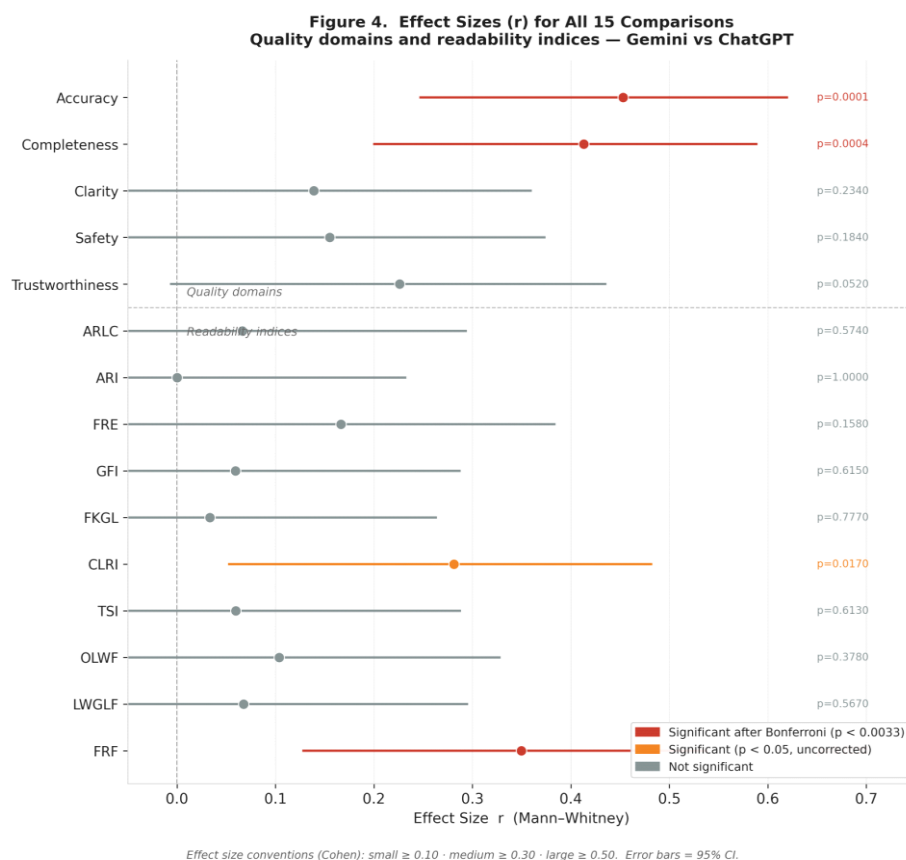


Figure 5. Forest plot of effect sizes (r) for all 15 comparisons. Quality domains and readability indices: Gemini vs. ChatGPT.

Red = significant after Bonferroni correction ($p < 0.0033$). Orange = significant at the uncorrected threshold only. Grey = not significant. Effect size conventions: small ≥ 0.10 , medium ≥ 0.30 , large ≥ 0.50 . Error bars represent 95% CI (Fisher z -transformation on r).

Discussion

ChatGPT received higher clinician ratings for accuracy and completeness than Gemini for knee osteoarthritis patient education, and both differences held up after correction for multiple comparisons, with medium-to-large effect sizes. These are the study's central findings, and they matter in practice: accuracy and completeness are probably the two attributes that most directly decide whether AI-generated content helps a patient manage a chronic condition or misleads them.

ChatGPT's accuracy advantage lines up with other disease-specific evaluations from the past two years. Comparing the two platforms on pulmonary artery catheter education, Asai et al. found ChatGPT in stronger agreement with procedural guidelines [11]. Marey et al., assessing cardiovascular imaging education across three platforms, reported the highest overall performance for ChatGPT-4, with Gemini adequate but lower [13]. In each case, as here, the subject area seems to matter: knee osteoarthritis, with its well-established evidence base for exercise, weight management, pharmacological sequencing, and surgical indications, may simply be well represented in GPT-4o's training data.

Completeness requires specific clinical consideration because effective self-management of knee osteoarthritis involves understanding and integrating multiple important aspects of care. Patients need to recognize that exercise should be maintained even when discomfort is present, that prolonged unsupervised use of analgesic medications may involve potential risks, that weight reduction can enhance the effectiveness of physiotherapy, and that surgical intervention is generally considered only after conservative treatment options have been fully attempted. A response may provide correct information regarding one individual aspect while still being incomplete if other essential components are omitted, making it technically accurate but less valuable for practical patient guidance.

The finding that ChatGPT achieved significantly higher completeness scores across 36 questions covering these diverse aspects indicates that it addressed a broader range of relevant information and did so with greater consistency compared with the other platform.

Clarity, safety, and trustworthiness showed no significant differences between the platforms, but caution is warranted before reading this as equivalence. The ICC data for these three domains were especially poor; the safety ICC(A,1) for Gemini sat near zero (0.002), a sign that the instrument simply lacked the sensitivity to detect differences that may well exist. A null result from an unreliable measurement is not the same as a true null effect. Studies probing these attributes will need to invest more in rater calibration, validated instruments such as DISCERN, and perhaps patient-centred comprehension testing, all of which would cut measurement noise.

The readability findings provide additional insight but do not substantially alter the overall practical interpretation. Among the evaluated measures, the Fry Readability Formula was the only index that remained statistically significant after Bonferroni correction, with ChatGPT responses demonstrating slightly greater complexity despite the similarity in response word counts between the platforms. One possible explanation is that more comprehensive responses may require increased linguistic complexity, including longer sentence structures and the use of more specialized terminology, in order to communicate additional clinical information effectively. However, this interpretation remains speculative, as the relationship between response completeness and linguistic complexity was not directly examined in this study and should therefore be interpreted cautiously. The difference observed in the Coleman–Liau Readability Index followed a similar pattern but did not remain significant after correction, meaning it does not provide separate supporting evidence for this interpretation.

If response complexity and completeness do move together, a practical question follows for deployment: which patients are better served by ChatGPT's richer but slightly harder output, and which by Gemini's simpler responses? Will et al., studying LLM simplification of patient education materials, found that aggressive readability reduction sometimes introduced factual errors [12], a useful caution when deciding whether Gemini's simpler text reflects real accessibility or quiet information loss. We did not measure information density directly, so this question is left for future work.

The inclusion of Unani medicine queries is a distinguishing feature of this study, reflecting the real-world information-seeking behavior of patients in South Asian clinical contexts. Questions 14–18 and 29 in the final question set address Unani treatment options, drug combinations, duration of therapy, and dietary guidance, topics that clinicians in India and similar settings meet in almost every osteoarthritis consultation. Which platform handles them better, with more accuracy, more epistemic humility, or more cultural sensitivity, is something this analysis cannot fully resolve and that future research should take up directly.

The poor single-measure ICC values across all domains are the study's most important methodological limitation. The likely contributors are the three set out in the Results: no calibration round before scoring, the different subspecialty backgrounds of the five raters, and a narrow quality band that compresses the between-question variance the ICC depends on. The more meaningful metric for the research question is ICC(A,k), which reached moderate reliability for completeness (0.567) and approached, without reaching, the moderate threshold for accuracy (0.421) in the two domains where between-platform differences survived correction, lending qualified support to the robustness of those findings.

A further limitation, one the AI healthcare literature has only lately begun to confront, is that the data are tied to a moment in time. GPT-4o and Gemini 1.5 Pro are both updated continuously, often without announcement, in ways that shift their responses; the same question set put to the same platforms six months later could give noticeably different results. We report the exact model versions and access dates (GPT-4o, May 2025; Gemini 1.5 Pro, May 2025) precisely so the study can be replicated and anchored to a defined point in each platform's evolution.

The study shares the design limitations common to clinician-rating evaluations of AI output. We did not measure patient-level outcomes such as comprehension, recall, trust, or adherence, so our findings speak to the quality of the content as experts judge it, not to whether it actually helps patients. Clinical usefulness is therefore inferred rather than shown, and the obvious next step is a prospective study that puts these responses in front of patients and measures what they understand, retain, and act on, ideally against a clinician-authored comparator. The 36-question set is representative but cannot cover every aspect of knee osteoarthritis education. Evaluation was in English only, which limits relevance to multilingual settings. And only two platforms were assessed; other widely used models, including Copilot, Claude, and various open-source LLMs, fell outside our scope.

Taken together, the results support a cautious clinical reading: for knee osteoarthritis education, ChatGPT currently earns higher clinician-rated accuracy and completeness than Gemini, while Gemini scored comparably on safety and somewhat higher on accessibility. These readings describe performance within the present evaluation framework rather than fixed properties of either platform. Neither should be used without clinician oversight, and both conclusions deserve revisiting as the models keep evolving.

Strengths

This study assessed AI responses across five quality domains and ten readability indices at once, a broader frame than most earlier work, which has tended to look at accuracy or readability in isolation. A blinded, multi-rater design reporting both ICC(A,1) and ICC(A,k), an explicit Bonferroni correction, and pre-specified safety criteria together give a level of methodological transparency that remains uncommon in AI patient education research. Drawing the question bank from seven authoritative sources and adding culturally and clinically grounded items, including Unani medicine queries, also makes the findings more applicable to real, pluralistic patient populations than question sets built from Western academic guidelines alone.

Limitations

Reliability was the study's main methodological constraint. Between individual raters, agreement was weak across most quality domains (ICC(A,1)); only the average-measures ICC(A,k) reached moderate values, and only in a few domains. Consistency therefore held at the aggregate level but not at the level of a single clinician's rating, a pattern that reflects the wider difficulty of scoring free-text clinical content on subjective constructs. This limits how far the non-significant findings can be interpreted. Future studies should build in formal rater calibration and report average-measure ICC(A,k) as the primary reliability statistic whenever several raters contribute to a composite score. The scoring rubric was author-designed rather than independently validated, and it was applied without pre-specified anchor exemplars for each scale point; its domains were drawn from the established DISCERN and HONcode constructs, but adopting a formally validated instrument such as DISCERN or PEMAT, or piloting the rubric before use, would strengthen construct validity in later work. We did not perform a formal sample-size justification or power analysis. Evaluation was confined to English-language responses, two AI platforms, and a single time point, and all three constraints limit generalizability. Finally, clinical usefulness to patients was inferred from expert ratings rather than tested directly; prospective patient-based validation of comprehension and behaviour is the priority for future work.

Funding

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

Conflict of Interest

The authors declare that they have no competing interests or financial relationships that could have influenced the work reported in this paper.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors used an AI-based tool for language correction and to improve the readability of the manuscript. It was not used to generate scientific content, study design, data, analyses, or conclusions. The ChatGPT and Gemini responses evaluated in this study were collected as research data only and were not involved in writing the manuscript. The authors reviewed and edited all text and take full responsibility for the content of the publication..

References

1. World Health Organization (WHO). Osteoarthritis. Available from: <https://www.who.int/news-room/fact-sheets/detail/osteoarthritis> [Accessed 20 May 2026].
2. The GBD 2021 Osteoarthritis Collaborators. Global, regional, and national burden of osteoarthritis, 1990–2020, and projections to 2050. *Lancet Rheumatol.* 2023;5:e508–e522.
3. Liu Q, Niu J, Li H, et al. Global and regional prevalence, incidence, and risk factors of knee osteoarthritis in population-based studies. *EClinicalMedicine.* 2021;39:101082.
4. Kolasinski SL, Neogi T, Hochberg MC, et al. 2019 American College of Rheumatology/Arthritis Foundation guideline for the management of osteoarthritis of the hand, hip, and knee. *Arthritis Rheumatol.* 2020;72:220–233.
5. Bannuru RR, Osani MC, Vaysbrot EE, et al. OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. *Osteoarthritis Cartilage.* 2019;27:1578–1589.
6. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models in healthcare. *NPJ Digit Med.* 2023;6:120.
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel).* 2023;11:887.
8. Koo TK, Li MY. Guidelines for selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15:155–163.

9. Ayers JW, Zhu Z, Poliak A, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted on a public social media forum. *JAMA Intern Med.* 2023;183:589-596. doi:10.1001/jamainternmed.2023.1838
10. Dastani M, Rezaee R, Bagheri S, et al. Large language models' capabilities in responding to tuberculosis medical questions: testing ChatGPT, Gemini, and Copilot. *BMC Med Inform Decis Mak.* 2025.
11. Asai OG, Sabu N, Gondode P, Das S, Chauhan G. Comparative evaluation of popular Gen-AI chatbots in generating patient education material on pulmonary artery catheter insertion. *Ann Card Anaesth.* 2026;29:81–88.
12. . Will J, Gupta M, Zaretsky J, Dowlath A, Testa P, Feldman J. Enhancing the readability of online patient education materials using large language models: a cross-sectional study. *J Med Internet Res.* 2025;27:e69955.
13. Marey I, Abdalla W, Ali S, et al. Evaluating the accuracy and reliability of AI chatbots in patient education on cardiovascular imaging: a comparative study of ChatGPT, Gemini and Copilot. *Egyptian J of Radiology and Nuclear Medicine.* 2025.