

Quality and Consistency of Four Large Language Models in Answering AAOS Guideline-Based Questions on Knee Osteoarthritis and Total Knee Arthroplasty: A Blinded Multi-Rater Benchmark Study in Chinese

Handong Zhang

Shandong Provincial QianFoShan Hospital

Shenghua Zhao

Linzi District Maternal and Child Health Hospital (Qidu Hospital)

Bingyu Yan

Shandong Provincial QianFoShan Hospital

Huaqiang Sun

Shandong Provincial QianFoShan Hospital

Peng Wang

18263825807@163.com

Shandong Provincial QianFoShan Hospital

Research Article

Keywords: large language model, knee osteoarthritis, total knee arthroplasty, evidence-based evaluation, repeated-response consistency

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11009605/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

Large language models (LLMs) are increasingly used to obtain information on knee osteoarthritis (OA) and total knee arthroplasty (TKA), but evidence-based answer quality and consistency across repeated outputs require separate evaluation.

Methods

We included 24 standardized questions derived from American Academy of Orthopaedic Surgeons (AAOS) guidelines: eight on non-surgical knee OA and 16 on TKA surgical management. Four LLMs were tested in three independent sessions per question, yielding 288 answers anonymously rated by eight experts. Relevance, accuracy, understandability, and safety were scored from 1 to 5; their equally weighted mean was the primary outcome. Three-response consistency was rated separately. Quality outcomes were analyzed with maximum-likelihood mixed-effects models, whereas consistency used paired Friedman and Wilcoxon tests. Pairwise comparisons were Holm-adjusted, with paired t-test sensitivity analyses. Inter-rater reliability was evaluated using absolute-agreement intraclass correlation coefficients (ICCs).

Results

All ratings were complete. The overall model effect for overall quality was significant (Wald $\chi^2=78.30$, $df = 3$, $P < 0.001$). Adjusted means were 4.527 for GPT, 4.311 for DeepSeek, 3.866 for ERNIE, and 3.675 for Kimi. GPT exceeded DeepSeek by 0.216 points (95% CI 0.040–0.392; Holm-adjusted $P = 0.032$ in the mixed model), whereas paired t-test and Wilcoxon sensitivity analyses yielded $P = 0.055$ and $P = 0.069$, respectively. GPT scored highest for relevance, accuracy, and safety, whereas DeepSeek scored highest for understandability. Consistency means were 4.750, 4.490, 4.198, and 4.047, respectively (Friedman $\chi^2=45.471$, $P < 0.001$). The eight-rater average-measure ICCs were 0.966 for overall quality and 0.852 for consistency. The model-by-question-category interaction was not significant ($P = 0.805$).

Conclusions

The four LLMs differed in answer quality and repeated-output consistency in this benchmark, with advantages varying by evaluation domain. The overall-quality difference between GPT and DeepSeek was sensitive to the inferential method and should not be interpreted as a fixed absolute ranking. These findings reflect expert-rated textual performance and do not establish benefits for patient understanding or clinical outcomes.

Background

Information on knee OA and TKA spans non-surgical treatment, preoperative decision-making, perioperative management, and postoperative rehabilitation. In addition to knowing available treatment options, patients need to understand indications, limitations, risks, and expected recovery. A systematic review of online patient-education materials for total joint arthroplasty found that readability often exceeded recommended levels [1]. LLMs offer a natural-language interface for obtaining information continuously, and studies have evaluated their responses to common TKA questions, preoperative counseling needs, and postoperative concerns [2–5]. However, fluent or readable responses do not necessarily ensure that recommendations, evidentiary boundaries, and risk statements are accurate.

Previous work has expanded from general orthopaedic decision-making to multi-model comparisons in knee OA and TKA [6–9]. Using questions from the World Expert Meeting on Arthroplasty (WEMA), Pujol et al. performed blinded expert evaluation of four models with an adapted QUEST framework, demonstrating the feasibility of multi-model, evidence-oriented, multidimensional benchmarking [10]. The remaining need is therefore not simply another single-score ranking, but an evaluation of whether model advantages are consistent across quality domains and whether core conclusions and key risk statements remain stable across independent sessions under clearly defined language, platform, and prompting conditions. Chinese-language use and repeated-output stability remain insufficiently characterized in a reproducible benchmark.

We therefore constructed 24 standardized questions from AAOS guidelines on non-arthroplasty management of knee OA and surgical management of knee OA [11,12] and compared three independent Chinese-language responses from GPT, DeepSeek, ERNIE, and Kimi. The primary outcome was an equally weighted overall-quality score comprising relevance, accuracy, expert-perceived understandability, and safety. Repeated-response consistency was evaluated as an independent secondary outcome, and we explored whether model differences varied by question category. The aim was to characterize domain-specific performance and robustness rather than to assume a uniform ranking across all dimensions.

Methods

Study design and reporting framework

This was a cross-sectional, multi-model, guideline-based benchmark study using blinded evaluation by multiple experts. The study materials were LLM-generated text responses and did not involve patient treatment or clinical intervention. The workflow comprised prespecification of the question set and rating criteria, collection of three responses under standardized conditions, anonymization and randomized presentation, independent expert rating, data verification, and statistical comparison. Reporting was informed by the TRIPOD-LLM and CHART recommendations on model identification, prompting and query procedures, human evaluation, and analytical transparency [13,14].

Standardized question set and reference standard

Questions Q1–Q8 were derived from the AAOS Management of Osteoarthritis of the Knee (Non-Arthroplasty), 3rd Edition (2021), covering braces, topical and oral nonsteroidal anti-inflammatory drugs (NSAIDs), weight loss, exercise, neuromuscular training, self-management, and hyaluronic acid injection. Questions Q9–Q24 were derived from the AAOS Surgical Management of Osteoarthritis of the Knee guideline (2022), covering perioperative management, surgical techniques, and risk factors [11,12]. Questions were standardized clinical formulations based on guideline recommendations rather than verbatim excerpts. Q23 (smoking) and Q24 (simultaneous bilateral TKA) corresponded to consensus options in the guideline, and their evidentiary status was explicitly reflected in the rating anchors. For each question, the guideline section, recommendation strength, reference conclusion, and key evaluation criteria were prespecified for accuracy and safety scoring but were not provided to the models. The complete benchmark and evidence anchors are provided in Additional file 1.

Models and testing environment

Testing was performed in August 2026. The evaluated systems were ChatGPT (OpenAI; GPT-5.6 Luna), DeepSeek (DeepSeek; standard non-thinking web mode), ERNIE (Baidu; ERNIE 5.1, standard/non-thinking mode), and Kimi (Moonshot AI; Kimi K2.6, Standard mode). All four were tested through their official consumer web interfaces using standard text chat. Advanced reasoning, deep research, Agent functions, plugins, APIs, third-party knowledge bases, file uploads, and image input were not actively enabled, and user-controllable web search was kept off. Where the underlying model version of a consumer interface could not be independently verified, contemporaneous API model names were not used to infer it.

Standardized prompt and three-round response collection

Testing was conducted in Chinese. The standardized prompt, translated into English, was: “Please answer the following question accurately and clearly based on current reliable medical evidence; if risks or contraindications are involved, please also state the important precautions.” The corresponding question was then entered. The same prompt and question wording were used across all four platforms. Models were not informed of the AAOS source, reference answers, recommendation strength, or scoring criteria, and no role prompt, chain-of-thought request, or response-length constraint was added. Each model answered each question in three independent sessions. Every round used a new conversation with no retained context; no follow-up, correction, additional prompting, or “regenerate” function was used, and only the first complete response from each round was retained. This produced 4 models × 24 questions × 3 rounds = 288 responses.

Anonymization, randomized presentation, and blinding

A researcher who did not participate in rating removed model names, product self-identification, and other direct source identifiers. Basic formatting was standardized and random codes were assigned, without altering medical content, wording order, information volume, risk statements, or response structure. Responses were presented in randomized order. Raters were unaware of model identity and did not have access to other experts' scores. Model identity was revealed only after scoring and data verification were complete.

Expert evaluation and scoring

Eight professionals with experience in knee disorders, joint arthroplasty, or perioperative management independently evaluated all responses. Relevance, accuracy, understandability, and safety were rated on 1–5 Likert scales. Relevance reflected whether the answer directly addressed the clinical question; accuracy reflected consistency of medical content, recommendation direction, and applicable boundaries with the prespecified reference standard; understandability reflected expert-perceived clarity, organization, and ease of comprehension; and safety reflected avoidance of potentially harmful advice and provision of necessary risk information. Operational anchors are shown in Table 1. In addition, the three responses generated by the same model for the same question were assessed as one set and rated from 1 to 5 for consistency of core conclusions, recommendation direction, and key risk information. This measure represented repeated-output stability, not correctness, and was not included in the overall-quality score. Patient comprehension was not directly measured.

Table 1
Evaluation framework and scoring criteria

Outcome/domain	Operational definition	Scoring	Role in analysis
Relevance	Directly addresses the clinical question and remains focused on the requested information.	1 = clearly off-topic; 3 = partly relevant; 5 = highly relevant	Component of overall quality
Accuracy	Medical content, recommendation direction, and applicable boundaries are consistent with the prespecified AAOS evidence anchor.	1 = major errors; 3 = generally acceptable; 5 = highly accurate	Component of overall quality
Understandability	Expert-perceived clarity, organization, and ease of understanding.	1 = very difficult; 3 = moderate; 5 = very clear	Component of overall quality
Safety	Avoids potentially harmful or misleading advice and provides important risk/contraindication information when relevant.	1 = clearly unsafe; 3 = generally safe; 5 = very safe with clear risk boundaries	Component of overall quality
Overall quality	Arithmetic mean of relevance, accuracy, understandability, and safety for the same response.	Range 1–5	Primary outcome
Response consistency	Consistency of core conclusions, recommendation direction, and key risk information across the three independent responses to the same question.	1 = markedly inconsistent; 3 = broadly consistent; 5 = nearly identical in core conclusions and key risks	Independent secondary stability outcome
Note: Response consistency was evaluated separately and was not included in the overall quality score. Understandability reflects expert-perceived clarity rather than patient-tested comprehension.			

Outcomes and analysis units

The primary outcome was overall quality, defined as the arithmetic mean of relevance, accuracy, understandability, and safety assigned by the same expert to the same response. The four component domains and repeated-response consistency were secondary outcomes; the model-by-question-category interaction was exploratory. The four models generated 288 responses, each rated by eight experts, yielding 2,304 expert-level response-rating records and 9,216 individual quality-domain scores. Repeated-response consistency was assessed for 96 model-question units, yielding 768 expert consistency scores. Primary quality analyses used the eight-rater mean for each response as the analysis unit. Consistency analyses used the eight-rater mean for each model-question unit while retaining the paired/correlated structure within questions.

Statistical analysis

Analyses were performed in Python 3.12.14 using NumPy 2.3.5, SciPy 1.17.0, and pandas 2.2.3. Table 2 summarizes mean $\pm$ sample standard deviation (SD; denominator $n - 1$) across individual expert rating records: 576 records per model for each of the five quality outcomes and 192 per model for response consistency. Inferential analyses used the expert-averaged units defined above and did not treat repeated ratings of the same response as independent observations. All tests were two-sided.

Overall quality and the four component domains were analyzed with maximum-likelihood (ML) linear mixed-effects models. Model and testing round were categorical fixed effects. Clinical question was modeled with a random intercept and three model-specific random slopes using GPT as the reference, with an unstructured covariance matrix for the four random coefficients; residuals were assumed to be independent and homoscedastic. Models were fitted using the marginal covariance likelihood and cross-checked with the closed-form ML solution available for this balanced design. Adjusted marginal means were averaged equally across the three rounds. The overall model effect was tested using a Wald χ^2 test ($df = 3$). Six prespecified pairwise comparisons were Holm-adjusted within each outcome. Reported 95% confidence intervals (CIs) for individual estimates or contrasts were pointwise normal-approximation intervals and were not multiplicity-adjusted.

Repeated-response consistency was analyzed with the 24 questions as paired blocks using the Friedman test with tie correction, with Kendall's W reported as an effect-size measure. Pairwise comparisons used a normal approximation to the Wilcoxon signed-rank test, excluding zero differences (Wilcox method), assigning average ranks to tied absolute differences, and applying no continuity correction; the six comparisons were Holm-adjusted within the outcome. Pairwise differences were first restored to their known score precision to prevent floating-point noise from splitting theoretical ties. Paired t -tests ($df = 23$) were used as a parametric sensitivity analysis.

Inter-rater reliability was assessed using two-way random-effects, absolute-agreement ICCs, reporting both single-rater ICC(A,1) and eight-rater average-measure ICC(A,8) [16]. The 288 responses were the rating targets for quality outcomes, and the 96 model-question units were the targets for consistency. Percentile 95% CIs were estimated using a question-cluster bootstrap. In each of 20,000 resamples, 24 questions were sampled with replacement, retaining all models, rounds, and eight raters for each selected question. A fixed random seed was used to ensure reproducibility. This resampling retained the eight raters in the present study and therefore reflects uncertainty from question sampling.

For exploratory analysis, Q1–Q8 were classified as non-surgical knee OA questions and Q9–Q24 as TKA surgical-management questions, and a model-by-question-category interaction was tested. Overall quality and the four component domains were also analyzed in question-level paired sensitivity analyses using Friedman/Wilcoxon tests and paired t -tests. Unless otherwise stated, $P < 0.05$ was considered statistically significant.

Results

Rating completion and descriptive statistics

Each model generated 72 responses, and all 288 responses were included. All 2,304 expert-level quality-rating records and 768 consistency-rating records were complete, with all scores ranging from 1 to 5. Descriptive results are shown in Table 2. Mean overall-quality scores were 4.527 for GPT, 4.311 for DeepSeek, 3.866 for ERNIE, and 3.675 for Kimi; DeepSeek had the highest mean understandability score.

Inter-rater reliability

For overall quality, the single-rater ICC was 0.779 (95% CI 0.727–0.816), and the eight-rater average-measure ICC was 0.966 (0.955–0.973). The average-measure ICCs for relevance, accuracy, understandability, and safety were 0.905, 0.902, 0.892, and 0.903, respectively. The average-measure ICC for repeated-response consistency was 0.852 (0.778–0.902) (Table 2).

Table 2
Descriptive performance of the four LLMs and inter-rater reliability

Outcome	GPT	DeepSeek	ERNIE	Kimi	ICC(A,1)	ICC(A,8)
Overall quality	4.53 ± 0.41	4.31 ± 0.41	3.87 ± 0.44	3.68 ± 0.58	0.779 (0.727–0.816)	0.966 (0.955–0.973)
Relevance	4.71 ± 0.47	4.28 ± 0.59	3.89 ± 0.65	3.78 ± 0.72	0.544 (0.476–0.600)	0.905 (0.879–0.923)
Accuracy	4.66 ± 0.51	4.26 ± 0.59	3.89 ± 0.62	3.65 ± 0.75	0.536 (0.463–0.594)	0.902 (0.873–0.921)
Understandability	4.17 ± 0.65	4.49 ± 0.57	3.91 ± 0.58	3.68 ± 0.72	0.507 (0.439–0.559)	0.892 (0.862–0.910)
Safety	4.57 ± 0.55	4.22 ± 0.60	3.76 ± 0.61	3.59 ± 0.74	0.537 (0.462–0.597)	0.903 (0.873–0.922)
Response consistency	4.75 ± 0.43	4.49 ± 0.51	4.20 ± 0.54	4.05 ± 0.51	0.419 (0.305–0.535)	0.852 (0.778–0.902)

Values are mean ± sample SD across individual expert rating records (576 per model for quality outcomes; 192 per model for response consistency). Inferential analyses use the expert-averaged units defined in Methods. ICCs use two-way random effects and absolute agreement. Percentile 95% CIs use 20,000 question-cluster bootstrap resamples retaining the eight raters; a fixed random seed was used for reproducibility.

Primary outcome: overall quality

Model type had a significant overall effect on overall quality (Wald $\chi^2=78.30$, $df = 3$, $P < 0.001$). Adjusted overall-quality scores were 4.527 (95% CI 4.394–4.661) for GPT, 4.311 (4.189–4.434) for DeepSeek, 3.866 (3.729–4.003) for ERNIE, and 3.675 (3.471–3.880) for Kimi (Table 3; Fig. 1). After Holm adjustment, GPT exceeded DeepSeek (difference 0.216, 95% CI 0.040–0.392, $P = 0.032$), ERNIE, and Kimi; DeepSeek exceeded ERNIE and Kimi (both $P < 0.001$), whereas the ERNIE–Kimi difference was not statistically significant ($P = 0.154$). In sensitivity analyses using the 24 questions as paired units, the overall-quality difference between GPT and DeepSeek did not reach significance after Holm adjustment (paired t-test $P = 0.055$; Wilcoxon $P = 0.069$), whereas the directions of the main comparisons of GPT and DeepSeek versus ERNIE and Kimi were retained. Thus, GPT had the highest mean overall quality, but its overall advantage over DeepSeek was sensitive to the inferential method.

Table 3
Adjusted overall quality and prespecified pairwise comparisons

Model/comparison	Estimate or difference	95% CI	Holm-adjusted P (ML)	Holm-adjusted P (paired t)	Holm-adjusted P (Wilcoxon)
GPT	4.527	4.394–4.661	—	—	—
DeepSeek	4.311	4.189–4.434	—	—	—
ERNIE	3.866	3.729–4.003	—	—	—
Kimi	3.675	3.471–3.880	—	—	—
GPT – DeepSeek	0.216	0.040–0.392	0.032	0.055	0.069
GPT – ERNIE	0.661	0.455–0.868	< 0.001	< 0.001	< 0.001
GPT – Kimi	0.852	0.629–1.075	< 0.001	< 0.001	< 0.001
DeepSeek – ERNIE	0.445	0.261–0.629	< 0.001	< 0.001	0.002
DeepSeek – Kimi	0.636	0.375–0.896	< 0.001	< 0.001	0.003
ERNIE – Kimi	0.191	-0.071–0.452	0.154	0.176	0.241

Overall ML model effect: Wald $\chi^2=78.30$, $df = 3$, $P < 0.001$. CIs are pointwise 95% normal-approximation intervals without multiplicity adjustment. P values are Holm-adjusted within the six comparisons for each outcome. Wilcoxon P values use the tie-corrected normal approximation, Wilcox zero handling, and no continuity correction.

Four quality domains

The overall model effect was significant for all four component domains (all $P < 0.001$; Table 4; Fig. 2). GPT had the highest adjusted means for relevance, accuracy, and safety (4.712, 4.658, and 4.571, respectively), whereas DeepSeek had the highest understandability score (4.488). Compared with DeepSeek, GPT scored 0.401 points higher for accuracy (95% CI 0.223–0.579; Holm-adjusted $P < 0.001$) and 0.349 points higher for safety (0.168–0.530; $P < 0.001$). DeepSeek scored 0.319 points higher than GPT for understandability (0.129–0.510; $P = 0.003$). ERNIE and Kimi did not differ significantly in any of the four domains in ML pairwise comparisons. These key domain-specific differences were preserved in question-level sensitivity analyses: Wilcoxon Holm-adjusted P values were 0.004 and 0.003 for GPT versus DeepSeek in accuracy and safety, respectively, and 0.021 for DeepSeek versus GPT in understandability. Full pairwise results are provided in Additional file 1.

Table 4
Adjusted domain-specific performance and response consistency

Outcome	GPT	DeepSeek	ERNIE	Kimi	Overall test
Relevance	4.712 (4.595–4.829)	4.278 (4.137–4.418)	3.892 (3.744–4.040)	3.783 (3.572–3.994)	Wald $\chi^2=114.60$; $P < 0.001$
Accuracy	4.658 (4.537–4.779)	4.257 (4.136–4.378)	3.894 (3.754–4.034)	3.648 (3.437–3.858)	Wald $\chi^2=108.54$; $P < 0.001$
Understandability	4.168 (3.999–4.337)	4.488 (4.354–4.622)	3.913 (3.783–4.043)	3.682 (3.473–3.892)	Wald $\chi^2=55.40$; $P < 0.001$
Safety	4.571 (4.431–4.712)	4.222 (4.103–4.341)	3.764 (3.618–3.910)	3.589 (3.384–3.793)	Wald $\chi^2=108.30$; $P < 0.001$
Response consistency	4.750	4.490	4.198	4.047	Friedman $\chi^2=45.471$; $P < 0.001$

Domain values are adjusted marginal means (95% CI) from the mixed model. Response consistency values are question-level means; Kendall’s $W = 0.632$. Full pairwise comparisons and sensitivity analyses are provided in Additional file 1.

Repeated-response consistency

Mean consistency scores were 4.750 for GPT, 4.490 for DeepSeek, 4.198 for ERNIE, and 4.047 for Kimi. The Friedman test showed a significant overall difference ($\chi^2=45.471$, $df = 3$, $P < 0.001$; Kendall’s $W = 0.632$). Holm-adjusted Wilcoxon comparisons showed higher consistency for GPT than DeepSeek ($P =$

0.019), ERNIE, and Kimi (both $P < 0.001$), and for DeepSeek than ERNIE ($P = 0.004$) and Kimi ($P < 0.001$). ERNIE and Kimi did not differ significantly ($P = 0.148$).

Exploratory and sensitivity analyses

The model-by-question-category interaction was not statistically significant (likelihood-ratio $\chi^2=0.99$, $df = 3$, $P = 0.805$), providing no evidence that relative model performance systematically differed between non-surgical knee OA and TKA surgical-management questions. Descriptive question-level results by category are provided in Additional file 1; this interaction test should not be interpreted as proof of equivalence. Friedman omnibus tests were significant for all six outcomes (all $P < 0.001$). Complete nonparametric omnibus tests and all 36 pairwise sensitivity comparisons are provided in Additional file 1.

Discussion

This study compared the quality and three-round consistency of four LLMs when answering Chinese-language questions derived from AAOS guidelines. The principal finding was that model advantages varied across evaluation domains rather than following a uniform hierarchy. GPT had the highest mean overall quality and higher relevance, accuracy, and safety scores, whereas DeepSeek had higher expert-perceived understandability. The mean overall-quality difference between GPT and DeepSeek was 0.216 points and reached the significance threshold in the ML analysis, but was not consistently confirmed by paired t-test or Wilcoxon sensitivity analyses. This difference therefore does not support a robust claim of overall superiority, and no prespecified threshold of clinical importance was available to interpret its practical magnitude. Compared with the study by Pujol et al., which evaluated four models using 20 WEMA questions and adapted QUEST domains, the present study adds a Chinese-language setting and an explicit repeated-output consistency assessment [10]. Because the studies differed in question sets, language, rating scales, and web-search settings, absolute scores should not be compared directly.

Accuracy, safety, and understandability should be regarded as related but non-equivalent properties. Gultekin et al. likewise observed that advantages in information completeness and clarity did not fully coincide when comparing enhanced modes for TKA education, although their operating conditions differed from ours and do not imply the same underlying model behavior [17]. A knee-OA study by Wu et al. also suggested that LLM responses do not invariably align with evidence-based recommendations [18]. Our benchmark included conditional conclusions such as “not recommended for routine use,” “no clear advantage demonstrated,” and “may be considered in selected patients.” Recasting these boundaries as absolute recommendations or prohibitions may create clinically important distortion. However, we did not independently code error types across all raw responses or examine associations between response length and score. The present data therefore do not establish that GPT scored higher because it better preserved guideline qualifiers, nor that DeepSeek sacrificed accuracy for brevity. They support separate evaluation of these dimensions.

Repeated-response consistency has a different meaning from single-response quality. Previous TKA and arthroplasty studies have examined response consistency or reliability under repeated querying [19,20]. Here, three independent sessions for the same model-question pair were treated as a separate evaluation unit. GPT and DeepSeek showed higher consistency than the other two models in the paired analyses, but stable output is not necessarily correct output: a model may reproducibly repeat an erroneous conclusion or consistently omit important risk information. Research on LLM-generated references further highlights the distinction between apparent textual credibility and verifiability [21]. Repeated-output consistency should therefore complement, rather than replace, accuracy and safety. Our consistency score was also based on subjective expert judgment, and three repetitions cannot represent long-term stability across model versions or time.

From an application perspective, these findings are better suited to prioritizing targets for subsequent validation than to determining immediate clinical deployment. Early studies of TKA counseling and knee-OA patient education have begun to examine LLMs in real communication settings [22,23], but our study did not include patients and did not measure comprehension, adherence, functional recovery, or adverse events. Expert-rated textual performance therefore cannot be assumed to translate into clinical benefit. Recommendations involving treatment boundaries—such as NSAIDs, perioperative glycemic control, tranexamic acid, and selection for simultaneous bilateral TKA—still require professional verification. The model-by-question-category interaction of $P = 0.805$ does not prove that performance is unaffected by question type; with only eight non-surgical OA questions and 16 surgical-management questions, power to exclude meaningful interactions was limited. Future studies should use real patient questions in clinician-supervised information-support settings and evaluate whether textual advantages translate into better understanding and decision quality.

This study has several limitations. The benchmark comprised only 24 AAOS guideline-derived questions, limiting generalizability beyond the sampled topics and the tested model versions. The outcomes were based on expert ratings, and the composite scale lacks external validation and cannot substitute for patient comprehension or clinical outcomes. Some comparisons were sensitive to the inferential method, some analyses were added post hoc, and multiplicity was not controlled globally across all secondary outcomes. External question sets, repeated testing at different time points, and studies involving actual users are needed. In addition, quality comparisons were based on the mean ratings of the eight experts in this study; the inferential framework did not incorporate sampling uncertainty for a broader population of raters. With only 24 questions, ML normal-approximation results should therefore be interpreted together with the paired question-level sensitivity analyses.

Conclusions

In 24 Chinese-language questions derived from AAOS guidelines and under the expert-rating conditions of this study, the four LLMs differed in answer quality and repeated-output consistency. GPT had the highest mean overall quality and performed well in relevance, accuracy, and safety, whereas DeepSeek had higher expert-perceived understandability. The overall-quality difference between GPT and DeepSeek

was sensitive to the statistical method, and no significant overall-quality difference was detected between ERNIE and Kimi. Evaluation should therefore report domain-specific performance, effect sizes, uncertainty, and repeated-output consistency rather than reduce results to a fixed absolute model ranking. These findings do not establish benefits for patient understanding or clinical outcomes.

Abbreviations

AAOS
American Academy of Orthopaedic Surgeons
CI
confidence interval
CR
cruciate retaining
ICC
intraclass correlation coefficient
LLM
large language model
ML
maximum likelihood
NSAID
nonsteroidal anti-inflammatory drug
OA
osteoarthritis
PS
posterior stabilized
SD
standard deviation
TKA
total knee arthroplasty.

Declarations

Author Contribution

HZ: conceptualization, methodology, investigation, data curation, statistical analysis, visualization, and drafting of the original manuscript. SZ: investigation, data curation, professional evaluation, and review and revision of the manuscript. BY: investigation, professional evaluation, data verification, and review and revision of the manuscript. HS: conceptualization, methodology, supervision, and review and revision of the manuscript. PW: conceptualization, methodology, supervision, project administration, and review and revision of the manuscript. All authors are accountable for their contributions and reviewed and approved the final manuscript.

Data Availability

The standardized question set, evidence anchors, and supplementary statistical results are provided with this article. De-identified expert rating data, analysis code, and anonymized large language model responses are available from the corresponding author on reasonable request, subject to journal policy.

Use of generative AI in manuscript preparation

During manuscript revision, ChatGPT (OpenAI) was used to assist with language editing, analysis-code drafting, statistical checking, and revision of the presentation of results. The tool was not used to generate or replace the original expert ratings in the present analysis. The authors reviewed and verified the final statistical implementation, interpretation, and manuscript content and take responsibility for the scientific accuracy and integrity of the submitted work.

References

1. Karimi AH, Shah AK, Hecht CJ 2nd, et al. Readability of Online Patient Education Materials for Total Joint Arthroplasty: A Systematic Review. *J Arthroplasty*. 2023;38(7):1392–1399. doi:10.1016/j.arth.2023.01.032.
2. Zhang S, Liao ZQG, Tan KLM, et al. Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement. *Knee Surg Relat Res*. 2024;36(1):15. doi:10.1186/s43019-024-00218-5.
3. Mika AP, Mulvey HE, Engstrom SM, et al. Can ChatGPT Answer Patient Questions Regarding Total Knee Arthroplasty? *J Knee Surg*. 2024;37(9):664–673. doi:10.1055/s-0044-1782233.
4. Taylor WL 4th, Cheng R, Weinblatt AI, et al. An Artificial Intelligence Chatbot is an Accurate and Useful Online Patient Resource Prior to Total Knee Arthroplasty. *J Arthroplasty*. 2024;39(8 Suppl 1):S358-S362. doi:10.1016/j.arth.2024.02.005.
5. Bains SS, Dubin JA, Hameed D, et al. Use and Application of Large Language Models for Patient Questions Following Total Knee Arthroplasty. *J Arthroplasty*. 2024;39(9):2289–2294. doi:10.1016/j.arth.2024.03.017.
6. Agharia S, Szatkowski J, Fraval A, et al. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: an analysis of ChatGPT 3.5, ChatGPT 4, and Bard. *J Orthop*. 2024;50:1–7. doi:10.1016/j.jor.2023.11.063.
7. Gurbuz S, Bahar H, Yavuz U, et al. Comparative Efficacy of ChatGPT and DeepSeek in Addressing Patient Queries on Gonarthrosis and Total Knee Arthroplasty. *Arthroplast Today*. 2025;33:101730. doi:10.1016/j.artd.2025.101730.
8. Lee H, Shin JS, Seo YJ, et al. A Comparative Analysis of GPT-3.5, GPT-4, GPT-4 Omni, Gemini Advanced, and Gemini 1.5 in Answering Total Knee Replacement-Related Questions. *Orthop J Sports Med*. 2026;14(1):23259671251402988. doi:10.1177/23259671251402988.

9. Gök Ü, Pamuk Ç, Serttaş MF, et al. ChatGPT-4.0 or DeepSeek-V3? Comparative analysis of answers to the most frequently asked questions by total knee replacement candidate patients. *Medicine (Baltimore)*. 2025;104(34):e43951. doi:10.1097/MD.00000000000043951.
10. Pujol O, Ferrer R, Coelho A, et al. Four general-purpose large language models (ChatGPT-5, Claude 4, Grok 4 and Gemini 2.5) show comparable performance in specialised total knee arthroplasty clinical questions. *Knee Surg Sports Traumatol Arthrosc*. 2026. doi:10.1002/ksa.70544.
11. American Academy of Orthopaedic Surgeons. Management of Osteoarthritis of the Knee (Non-Arthroplasty), 3rd Edition: Evidence-Based Clinical Practice Guideline. Rosemont, IL: AAOS; 2021. Published August 31, 2021. <https://www.aaos.org/oak3cpg>. Accessed 11 September 2026.
12. American Academy of Orthopaedic Surgeons. Surgical Management of Osteoarthritis of the Knee: Evidence-Based Clinical Practice Guideline. Rosemont, IL: AAOS; 2022. Published December 2, 2022. <https://www.aaos.org/smoak2cpg>. Accessed 11 September 2026.
13. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31(1):60–69. doi:10.1038/s41591-024-03425-5.
14. The CHART Collaborative. Reporting Guideline for Chatbot Health Advice Studies: The CHART Statement. *JAMA Netw Open*. 2025;8(8):e2530220. doi:10.1001/jamanetworkopen.2025.30220.
15. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7(1):258. doi:10.1038/s41746-024-01258-7.
16. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15(2):155–163. doi:10.1016/j.jcm.2016.02.012.
17. Gultekin O, Hirschmann MT, Arıkan Hİ, et al. Evaluating deepresearch and deepthink in total knee arthroplasty patient education: ChatGPT-4o excels in comprehensiveness, Deepseek R1 leads in clarity and readability of orthopedic information. *Jt Dis Relat Surg*. 2026;37(2):470–476. doi:10.52312/jdrs.2026.2645.
18. Wu H, Yao S, Bao H, et al. ChatGPT-4.0 and DeepSeek-R1 does not yet provide clinically supported answers for knee osteoarthritis. *Knee*. 2025;56:386–396. doi:10.1016/j.knee.2025.06.007.
19. Magruder ML, Rodriguez AN, Wong JCJ, et al. Assessing Ability for ChatGPT to Answer Total Knee Arthroplasty-Related Questions. *J Arthroplasty*. 2024;39(8):2022–2027. doi:10.1016/j.arth.2024.02.023.
20. Wright BM, Bodnar MS, Moore AD, et al. Is ChatGPT a trusted source of information for total hip and knee arthroplasty patients? *Bone Jt Open*. 2024;5(2):139–146. doi:10.1302/2633-1462.52.BJO-2023-0113.R1.
21. Schwartzman JD, Shaath MK, Kerr MS, et al. ChatGPT is an Unreliable Source of Peer-Reviewed Information for Common Total Knee and Hip Arthroplasty Patient Questions. *Adv Orthop*. 2025;2025:5534704. doi:10.1155/aort/5534704.
22. Liu J, Daher M, Laperche J, et al. ChatGPT versus expert arthroplasty surgeons in total knee arthroplasty patient counseling. *Knee*. 2025;55:12–17. doi:10.1016/j.knee.2025.03.005.

23. Yang Y, Lin J, Zhang J. The effects of ChatGPT on patient education of knee osteoarthritis: a preliminary study of 60 cases. *Int J Surg.* 2025;111(12):9753–9756.
doi:10.1097/JS9.0000000000002494.

Figures

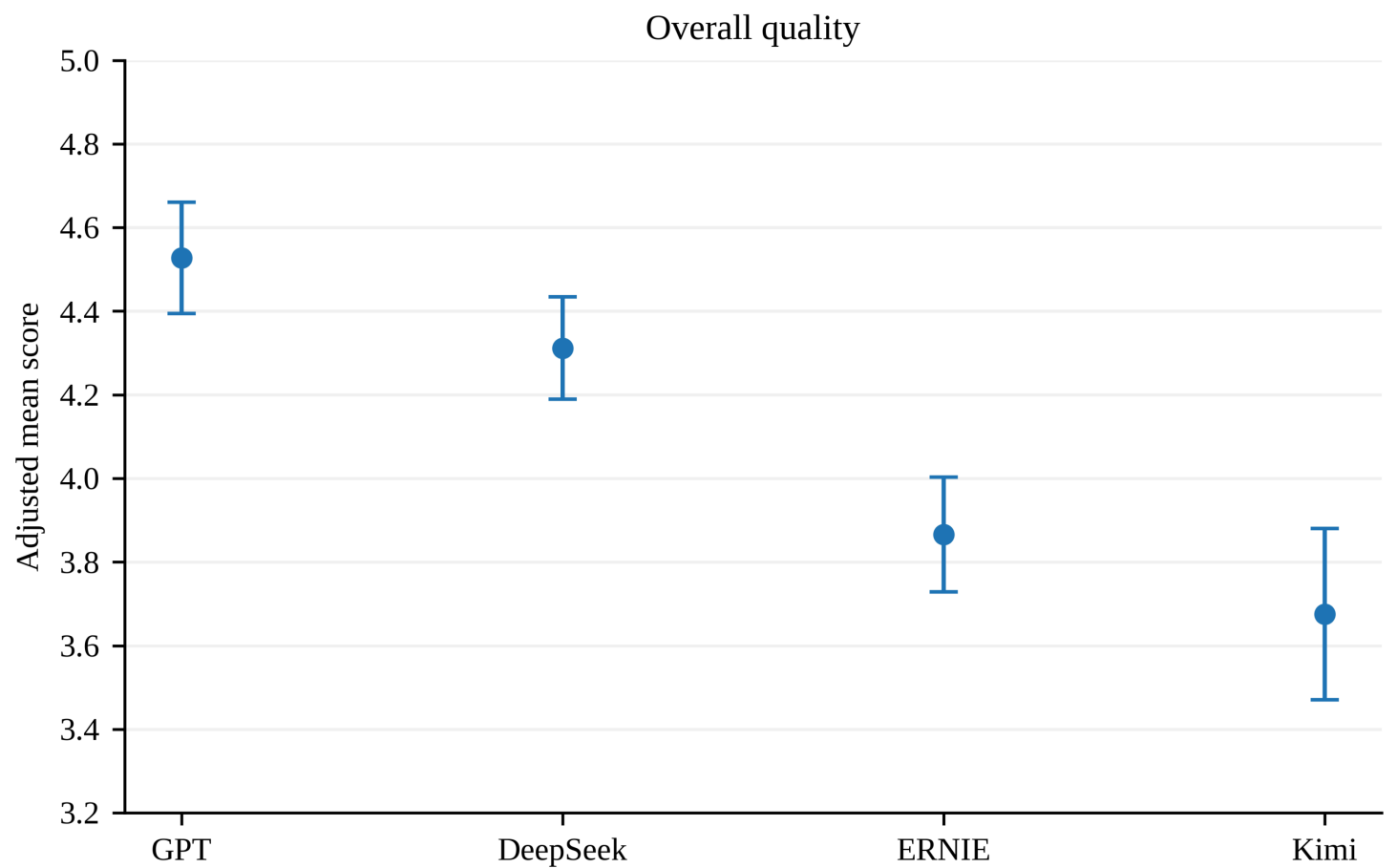


Figure 1

Adjusted overall quality scores across the four LLMs. Points represent adjusted marginal means from the mixed-effects model; error bars represent 95% confidence intervals.

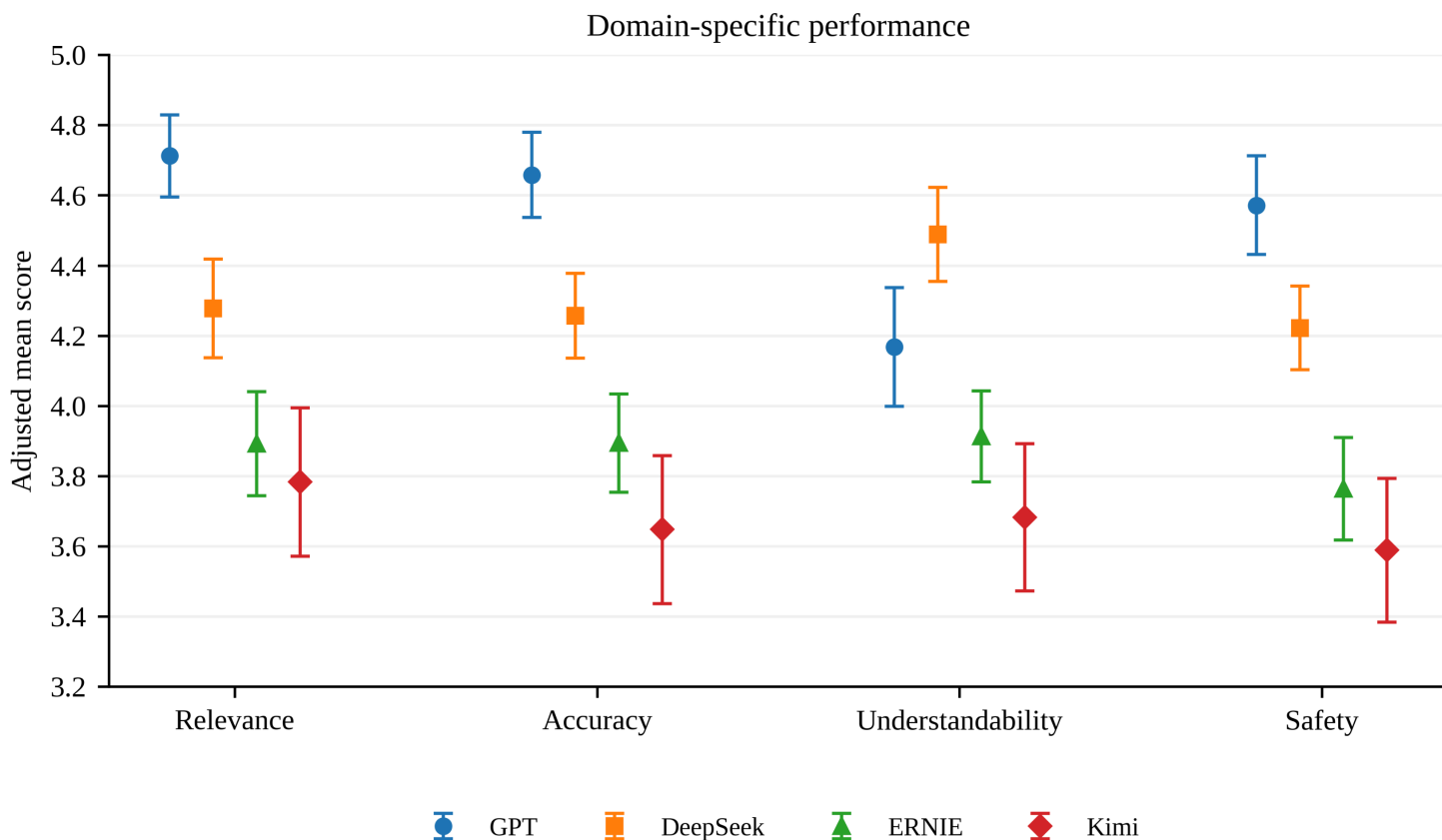


Figure 2

Domain-specific performance of the four LLMs. Points represent adjusted marginal means for relevance, accuracy, understandability, and safety; error bars represent 95% confidence intervals. Domains are discrete outcomes and are not connected by lines.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02UPLOADIFSUPPLEMENTARYFIELDAdditionalFile1.docx](#)