

Predicting Expert-Rated Quality of Large Language Model Responses to Orthopaedic Patient Questions Using Computational Intelligence

Volkan Kızılkaya

dr_volkankizilkaya@hotmail.com

Private Metrolife Hospital

Murat Aslanyürek

Kırklareli University

Canan Aslanyürek

Kırklareli University

Ali Levent

Private Metrolife Hospital

Mehmet Melih Asoğlu

Private Metrolife Hospital

Research Article

Keywords: Large Language Models, Computational Intelligence, Support Vector Regression, Natural Language Processing, Sentence-BERT, Clinical Quality, Patient Education

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11174522/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

Objective

This study aimed to compare the clinical quality of responses provided by four current large language models (LLMs) to orthopaedic patient questions based on expert evaluations and to investigate whether expert-rated response quality could be predicted using natural language processing (NLP), semantic representation, and machine learning methods.

Materials and Methods

Fifty patient-oriented questions covering anterior cruciate ligament injuries ($n = 15$), meniscal tears ($n = 15$), and total knee arthroplasty ($n = 20$) were submitted to ChatGPT, Claude, DeepSeek, and Gemini, yielding 200 unique responses. All responses were independently evaluated by three orthopaedic specialists for Accuracy, Completeness, and Misinformation, generating 600 expert evaluations. Models were compared using paired statistical methods overall and at the clinical-category and question levels. NLP features were extracted from the responses, and semantic representations and question-response similarities were calculated using SBERT. An RBF-SVR model combining NLP, semantic, and explainable textual features was developed to predict response-level quality scores obtained by averaging the ratings of the three experts. To prevent data leakage, the four LLM responses to the same clinical question were kept within the same cross-validation fold, and final performance was evaluated using nested grouped cross-validation.

Results

The four LLMs achieved generally high Accuracy (3.607–3.827/4) and Completeness (3.633–3.780/4) scores and low Misinformation scores (0.093–0.127/2). Descriptively, Claude had the highest mean Accuracy (3.827 ± 0.288), whereas DeepSeek had the highest Completeness (3.780 ± 0.283) and the lowest Misinformation score (0.093 ± 0.243). However, no overall statistically significant differences were detected among the models for Accuracy ($p = 0.070$), Completeness ($p = 0.361$), or Misinformation ($p = 0.451$). An overall model effect for Accuracy was observed in the TKA category ($p = 0.041$), but no pairwise comparison remained significant after Holm correction. Seven of the 13 lower-performing questions at the question level belonged to the TKA category. In the experts' free-text evaluations, the most frequently identified issue was unnecessary, overly detailed, or off-topic content, observed in 72 different responses. Responses generated by different LLMs to the same questions showed high semantic similarity (mean pairwise SBERT cosine similarity: 0.813–0.850). For Accuracy prediction, the final RBF-SVR model achieved MAE = 0.261, reducing absolute error by 11.6% compared with the mean-prediction baseline, and showed a weak but significant rank association with observed scores ($p = 0.202$; $p = 0.004$); however, R^2 was -0.161 . For Completeness prediction, MAE = 0.294 and $R^2 = -0.106$ were obtained. Ablation analyses showed that combining NLP, question-response semantic similarity, and explainable textual features reduced prediction error.

Conclusion

Current LLMs generally produce responses of high clinical quality to orthopaedic patient questions, although performance may vary by clinical topic and, particularly, at the individual-question level. High inter-model semantic similarity is not a direct indicator of clinical accuracy. Combining NLP, semantic, and explainable textual features provided a measurable signal for predicting expert Accuracy scores; however, negative R^2 values indicate that automated quality prediction remains limited in the present dataset. These findings suggest that LLMs may serve as

supportive tools in orthopaedic patient education and that machine-learning-based methods may contribute to the future automated pre-screening of the quality of clinical LLM responses.

1. INTRODUCTION

Large language models (LLMs) have attracted increasing interest in healthcare because they can interpret questions posed in natural language and transform complex medical information into fluent, understandable, user-tailored text. Particularly following the widespread adoption of ChatGPT, the potential of LLMs in areas such as patient education, access to clinical information, and support for patient-physician communication has been intensively investigated. Recent systematic reviews have shown that LLMs can provide accurate and adequate responses to patient questions in many settings, while performance may vary substantially according to the model used, clinical domain, question type, and evaluation method [1, 2]. This variability makes systematic assessment of clinical quality dimensions such as accuracy, completeness, and safety, in addition to linguistic fluency, essential when considering LLM outputs for clinical use. Orthopaedics is a clinical field in which patient education is important for diagnosis, treatment options, surgical expectations, and rehabilitation. Studies examining the use of LLMs in orthopaedics have increased rapidly in recent years, addressing applications including answering clinical questions, patient education, assessment of orthopaedic knowledge, and decision support [2, 3]. However, substantial methodological heterogeneity is evident in the current literature. Systematic reviews of LLM use in orthopaedics have reported considerable differences among studies in model versions, question sets, evaluation scales, and performance metrics [2, 3]. Moreover, as newer models emerge, findings from studies evaluating a single LLM are increasingly limited in their ability to reflect the relative performance of current models. Studies of patient-oriented orthopaedic questions suggest that LLMs can generally demonstrate promising performance, although consistent superiority in clinical quality is not always observed. ChatGPT has been reported to provide basic and evidence-based information for patient questions related to total knee arthroplasty (TKA), although some responses required additional clarification regarding clinical nuances [4]. Similarly, another evaluation of TKA-related questions reported that response quality varied according to the nature of the question [5]. This heterogeneity persists in more recent multi-model studies. For example, in patient questions concerning hip fractures, ChatGPT-4.5, Gemini 2.0 Flash, and DeepSeek-V3-FW generally demonstrated high accuracy and comprehensiveness, but showed different performance characteristics across question domains [6]. These findings indicate the importance of evaluating not only the overall mean performance of LLMs but also variability at the clinical-category and individual-question levels. Another challenge in the clinical evaluation of LLM responses is that high linguistic or semantic similarity does not necessarily indicate high clinical quality. A response may be well structured and semantically related to the question while still being incomplete, unnecessarily detailed, or clinically misleading. Recent orthopaedic literature has shown that LLM responses may perform differently across dimensions such as accuracy, completeness, readability, and patient usability [6, 7]. A recent systematic review and meta-analysis of hip arthroscopy and femoroacetabular impingement reported high accuracy but more variable completeness and highlighted important methodological limitations in question selection and evaluation methods across existing studies [7]. Clinical evaluation of LLMs should therefore incorporate complementary quality dimensions rather than relying on a single performance measure. Although most existing studies have focused on evaluating LLM responses using experts or clinical guidelines, whether expert-assigned clinical quality scores can be automatically predicted from measurable features extracted from response text has received less attention. Natural language processing (NLP) features can quantify structural properties such as response length, linguistic complexity, and readability, whereas semantic representation methods such as Sentence-BERT allow the semantic relationship between a question and its response to be modeled numerically. Combining these features with machine-learning methods may provide a potential approach for automated pre-screening of clinically lower-quality responses. However, because distributing responses to the same clinical question across training and test sets may cause data leakage, strict question-level grouped cross-validation strategies are required [8]. This study aimed to compare the clinical quality of patient-oriented

responses generated by ChatGPT, Claude, DeepSeek, and Gemini for anterior cruciate ligament injuries, meniscal tears, and total knee arthroplasty using evaluations by three independent orthopaedic specialists, and to examine how performance varied across clinical categories and individual questions. Second, we investigated whether expert-rated Accuracy and Completeness scores could be predicted using basic NLP features, SBERT-based semantic features, and explainable textual features extracted from LLM responses. Thus, in addition to comparing the clinical performance of current LLMs in orthopaedic patient education, this study evaluated an exploratory machine-learning framework for automated pre-screening of clinical response quality.

2. MATERIALS AND METHODS

2.1. Study Design and Dataset Construction

This study was designed to compare the clinical quality of responses provided by large language models (LLMs) to questions commonly asked by orthopaedic patients and to investigate whether response quality based on expert evaluations could be predicted using natural language processing and machine-learning methods. The study comprised two main analytical components. In the first, responses generated by different LLMs were evaluated by orthopaedic specialists for accuracy, completeness, and misinformation, and the clinical performance of the models was compared. In the second, the ability of linguistic and semantic features extracted from LLM responses to predict expert-assigned quality scores was investigated using machine-learning methods. A total of 50 patient-oriented orthopaedic questions covering three clinical categories were used: anterior cruciate ligament (ACL) injuries, meniscal tears, and total knee arthroplasty (TKA). The distribution was 15 questions on ACL injuries, 15 on meniscal tears, and 20 on TKA. Each question was submitted to four LLMs: ChatGPT, Gemini, Claude, and DeepSeek. Thus, four responses were obtained for each question, yielding a dataset of $50 \text{ questions} \times 4 \text{ LLMs} = 200$ unique LLM responses. All 200 responses were independently evaluated by three orthopaedic specialists. Because each expert evaluated every response, the total number of expert evaluations was $200 \text{ responses} \times 3 \text{ experts} = 600$. Accordingly, the data structure consisted of 50 unique clinical questions, 200 unique LLM responses, and 600 individual expert evaluations. This hierarchical structure was preserved throughout the analyses. In particular, the 600 expert evaluations were not treated as 600 distinct LLM responses; rather, they represented independent evaluations of the same 200 responses by three experts. Therefore, potential dependence and data leakage arising from repeated observations belonging to the same question or response were specifically considered in the statistical analyses and machine-learning validation procedures.

2.2. Orthopaedic Questions and Generation of Large Language Model Responses

OpenAI ChatGPT (GPT-5.6 Sol), Google Gemini (Gemini 3.8 Flash), Anthropic Claude (Claude Sonnet 5), and DeepSeek (DeepSeek-V4.1-Flash) were used to generate the responses. Responses were obtained through the web-based user interfaces of the models. To standardize data collection, Python-based Playwright browser automation was developed. The four models were run in separate browser processes, and data collection was synchronized at the question level so that the same clinical question was submitted to the models within as similar a time window as possible. Each question was submitted in an independent fresh-chat session, and only the first valid response generated by the model was recorded. The responses were not edited or modified for content by the researchers.

Please answer the following orthopaedic question as if explaining it to a patient without a medical background:
[Question]

Data collection and submission of the standardized prompts to the four language models were performed on August 15, 2026. In the event of transient connection or platform errors, the automation system repeated the relevant request after the predefined waiting period. One response from each of the four models was obtained for every question, producing 200 unique responses. Responses were identified using Question_ID and Response_ID; the question, clinical category, prompt used, full response text, source-line information, and response hash values were retained for data-integrity checks.

2.3. Expert Evaluation

The 200 responses obtained from the four language models were independently evaluated by three orthopaedic specialists. During evaluation, the experts were blinded to the identity of the language model that generated each response, and each expert evaluated all 200 responses, yielding 600 individual expert evaluations. Clinical quality was assessed using three criteria: Accuracy, defined as the degree to which the response was medically correct; Completeness, defined as the extent to which necessary and relevant information for the patient was covered; and Misinformation, defined as the degree to which clinically incorrect or misleading information was present. Accuracy and Completeness were scored from 0 to 4, and Misinformation from 0 to 2. Higher Accuracy and Completeness scores indicated higher quality, whereas higher Misinformation scores indicated more incorrect or misleading content. Experts could also enter free-text explanations and evaluation notes when considered necessary. These comments were not used to calculate quantitative scores but were retained to review evaluation records and interpret scoring differences. After completion of the evaluations, records were checked for completeness of all three expert scores for each Response_ID, duplicate entries, and compliance with the predefined score ranges. These checks confirmed that three independent expert evaluations were available for all 200 responses. Optional free-text comments were also examined descriptively and classified into four categories: unnecessary/overly detailed or off-topic content, incorrect/misleading information, incomplete/insufficient content, and specific clinical correction. Because a comment could belong to more than one category, the categories were not treated as mutually exclusive. As comments were optional, this analysis was used to characterize issues highlighted by experts rather than to compare model error rates.

2.4. Inter-Rater Reliability

Inter-rater reliability among the three orthopaedic specialists was evaluated using the original individual expert scores. Agreement was calculated separately for Accuracy, Completeness, and Misinformation. Because the scores were ordinal, ordinal Krippendorff's alpha (α) was used. In addition, because the same three experts evaluated all responses, intraclass correlation coefficients (ICC) were calculated. Using a two-way random-effects, absolute-agreement approach, both single-rater reliability ICC(2,1) and reliability of the mean of the three raters ICC(2,3) were reported. Selection and reporting of the ICC model followed the methodological framework described by Shrout and Fleiss [9]. To examine temporal stability of the expert evaluations, the same three orthopaedic specialists re-evaluated all 200 LLM responses two weeks after the initial assessment. Accuracy, Completeness, and Misinformation scores from the first and second evaluation periods were compared using the Wilcoxon signed-rank test. Repeat evaluations were used solely to assess temporal stability and were not included as additional observations in the primary performance analyses or machine-learning models. Scores between the two time points were compared only as a supportive analysis of temporal stability. The primary statistical analyses and machine-learning models were based on the 600 expert evaluations obtained during the first evaluation period.

2.5. Statistical Comparison of the Clinical Performance of the Language Models

The clinical performance of ChatGPT, Gemini, Claude, and DeepSeek was evaluated separately for Accuracy, Completeness, and Misinformation. Descriptive statistics, including mean, standard deviation, median, and

interquartile range, were calculated for each model. Because all four language models answered the same 50 questions, the paired data structure was taken into account in model comparisons. Overall differences among models were evaluated using the Friedman test, which does not require an assumption of normality, and effect size was reported using Kendall's W [10]. Pairwise model comparisons were performed using the Wilcoxon signed-rank test, with Holm adjustment of p values to control the probability of type I error arising from multiple comparisons [11]. Statistical significance was set at two-sided $p < 0.05$, and p values were interpreted together with effect sizes. To examine whether model performance varied by clinical topic, analyses were also performed separately for ACL injury, meniscal tear, and total knee arthroplasty categories. Accuracy, Completeness, and Misinformation scores of the four models were compared within each clinical category.

2.6. Clinical Category- and Question-Level Analyses

To examine whether language-model performance varied by clinical topic, results were additionally analyzed at the level of ACL injury, meniscal tear, and total knee arthroplasty (TKA) categories. Category-specific performance differences were evaluated by calculating Accuracy, Completeness, and Misinformation scores for each model within each category. Because model performance could vary not only between clinical categories but also between individual questions, a question-level analysis was conducted. For each Question_ID, the mean Accuracy, Completeness, and Misinformation scores across the four language models, as well as the between-model standard deviation and range of Accuracy, were calculated. This allowed identification of questions for which overall performance was lower and between-model differences were greater. The 25th percentile (Q1) of the distribution of question-level mean Accuracy was used as a descriptive threshold to identify relatively lower-performing questions. Questions at or below this value were classified as lower-performing questions. This threshold was not considered a clinically validated cutoff but was used as a descriptive measure to identify relatively difficult questions within the dataset. Finally, relationships among question-level Accuracy, Completeness, Misinformation, and between-model performance variability were examined using Spearman's rank correlation coefficient (ρ), which permits assessment of monotonic associations without requiring parametric distributional assumptions [12].

2.7. Extraction of Natural Language Processing Features

To quantitatively represent the linguistic and structural characteristics of the language-model responses, natural language processing (NLP)-based features were extracted from the 200 unique LLM responses. Feature extraction was performed at the unique-response level, thereby avoiding duplication caused by the same response being evaluated by three different experts. For each response, Word Count, Sentence Count, Character Count, Average Sentence Length, Average Word Length, and Lexical Diversity were calculated. Lexical diversity was defined as the ratio of unique words to the total number of words. These features were used to represent basic linguistic properties such as response length, structural complexity, and diversity of word use. Flesch Reading Ease and Flesch-Kincaid Grade Level were also calculated to assess linguistic complexity from a patient perspective. Higher Flesch Reading Ease values indicate easier-to-read text, whereas Flesch-Kincaid Grade Level represents the approximate educational level required to understand the text [13, 14]. This process yielded eight basic NLP features for each response. These features were subsequently evaluated as an independent feature group in the machine-learning analyses and formed the basis of combined feature sets used to examine the added contribution of semantic and explainable textual features to prediction performance.

2.8. SBERT-Based Semantic Representation and Question-Response Similarity

To represent not only surface linguistic characteristics but also the semantic content of the responses, a Sentence-BERT (SBERT)-based sentence-embedding approach was used [15]. The sentence-transformers/all-MiniLM-L6-v2

model was used for semantic representation, and each of the 200 unique LLM responses was represented by a 384-dimensional dense vector. The resulting embedding vectors were normalized and used in subsequent semantic-similarity and machine-learning analyses. To assess how semantically similar the responses generated by four different language models were for the same clinical question, cosine similarity was calculated between normalized SBERT vectors. Because there were six pairwise comparisons among four models for each question, a total of 300 model-pair similarity values were obtained across 50 questions. This analysis was performed descriptively to examine the semantic proximity of responses produced by different LLMs to the same clinical question. Inter-model semantic similarity was not considered a direct indicator of clinical accuracy. Because the semantic representation of the response alone might not be sufficient for predicting expert evaluations, each clinical question was also represented using the same SBERT model as a normalized 384-dimensional embedding vector. Cosine similarity between each question and its corresponding LLM response was then calculated to derive the Question-Response Cosine Similarity feature. Element-wise absolute differences between question and response embedding vectors were also calculated as an alternative representation of the semantic relationship between question and response. These semantic representations were evaluated in different feature combinations during machine learning. Thus, we investigated whether not only the semantic content of the response text itself but also its semantic relationship with the clinical question contributed to prediction of expert-assigned clinical quality scores.

2.9. Machine-Learning Prediction of Clinical Quality

Supervised machine-learning analyses were performed to investigate whether the expert-rated clinical quality of language-model responses could be predicted from textual and semantic features. Accuracy and Completeness scores were used as prediction targets. Because Misinformation scores were heavily concentrated at zero and showed marked imbalance in the dataset, Misinformation was not included as a continuous regression target in the primary machine-learning analysis.

For the machine-learning analyses, the scores assigned by the three experts to the same response were aggregated at the response level to create target values for the 200 unique LLM responses. Accuracy and Completeness targets were calculated as the arithmetic mean of the three experts' raw scores for each response. No consensus procedure, rescoring, or post hoc score modification was applied. Thus, each unique LLM response was represented as a single observation in the final machine-learning analysis.

Multiple feature groups were evaluated to examine the contribution of different levels of information to prediction of expert evaluations. The first feature group consisted of the eight basic NLP features defined in Section 2.7. In the second stage, the SBERT-based question-response cosine similarity defined in Section 2.8 was added. To represent formal and linguistic response characteristics in greater detail, explainable textual features were also extracted, including the counts and densities of numerical expressions, bulleted or numbered list structures, hedging expressions, clinical warning expressions, risk-related terms, and treatment-related expressions. These features were calculated using predefined rule-based word lists and regular-expression patterns applied consistently to all responses.

Initial model comparisons evaluated Ridge regression, Support Vector Regression (SVR), Random Forest, and Gradient Boosting algorithms [16–19]. Ridge regression was used as a linear regularized reference approach [16], whereas SVR was evaluated for modeling nonlinear patterns [17]. Random Forest and Gradient Boosting represented tree-based ensemble-learning methods [10, 11]. In initial analyses directly using high-dimensional SBERT embedding vectors, Principal Component Analysis (PCA) was applied for dimensionality reduction [20]. To prevent data leakage, PCA and feature standardization were performed only within the corresponding training data.

Following the initial comparisons of models and feature groups, the final prediction approach was defined as an SVR model with a radial basis function (RBF) kernel that jointly used eight basic NLP features, SBERT-based question-response cosine similarity, and 14 explainable textual features [17]. The final model therefore used 23 predictor variables. The C, epsilon, and gamma hyperparameters of the SVR model were not determined using the test data; hyperparameter selection was performed exclusively through inner cross-validation within the training data. Mean Absolute Error (MAE) was used as the primary performance metric for hyperparameter optimization.

Prediction performance was evaluated using MAE and Root Mean Squared Error (RMSE); the proportion of variance in the target variable explained by the model was assessed using R^2 , and the rank association between observed expert scores and model predictions was examined using Spearman's rank correlation coefficient (ρ). To determine whether the machine-learning models provided information beyond prediction of the central tendency of the target distribution, a training-fold mean baseline model was constructed that assigned the mean target value from each training fold as the prediction for the corresponding test samples. Performance of the final model was compared with this baseline approach.

2.10. Prevention of Data Leakage and Nested Grouped Cross-Validation

Because each of the same 50 clinical questions was answered by four different LLMs, distributing responses to the same question across training and test sets could result in data leakage and optimistic performance estimates. Therefore, Question_ID was used as the grouping variable in all cross-validation procedures, ensuring that the four LLM responses to the same question remained within the same fold. Question_ID and Response_ID were used only for grouping and record tracking and were not included as predictive features [8]. Nested grouped cross-validation was used for hyperparameter optimization and performance evaluation of the final RBF-SVR model [21, 22]. The outer loop used five-fold GroupKFold; in each fold, 160 responses from 40 questions were used for training and 40 responses from 10 previously unseen questions were used for testing. Within each outer training set, four-fold grouped cross-validation evaluated 96 hyperparameter combinations according to MAE: $C \in \{0.1, 0.5, 1, 2, 5, 10\}$, $\epsilon \in \{0.05, 0.10, 0.20, 0.30\}$, and $\gamma \in \{\text{scale}, 0.001, 0.01, 0.1\}$. Feature standardization and model training were performed within the same pipeline using only the relevant training data. Out-of-fold predictions from the five outer test folds were combined to calculate final MAE, RMSE, R^2 , and Spearman ρ values. A baseline model using only the target mean from the training data was also evaluated on the same outer folds. Although the nested structure separated hyperparameter selection from outer performance evaluation, it did not completely eliminate possible model-selection optimism because the final feature set and model family had been selected through prior experimental comparisons on the same dataset.

2.11. Ablation, Explainability, and Sensitivity Analyses

An ablation analysis was performed to evaluate the contribution of different feature groups to prediction performance. The SVR model was evaluated with four feature configurations: basic NLP features alone; NLP features plus question-response cosine similarity; explainable textual features alone; and a combined feature set including all features. This enabled assessment of the additional contribution of semantic and explainable textual features beyond basic NLP features. The ablation analysis was conducted as an exploratory analysis separate from the final generalizability evaluation. At this stage, five-fold GroupKFold based on Question_ID grouping was used, ensuring that the four LLM responses to the same clinical question remained in the same fold. Ablation results were obtained using single-level grouped cross-validation without nested hyperparameter optimization. Therefore, performance values from the ablation analysis were used to examine the relative contribution of feature groups and were not interpreted as estimates of final model performance. Final generalizability performance was determined using the nested grouped

cross-validation described in Section 2.10. Permutation feature importance analysis was applied to assess the contribution of variables in the final feature set to model predictions [23]. Feature values were independently permuted in the test data, the resulting change in model performance was measured, and the procedure was repeated 20 times for each feature. Importance values were used to explain contributions to predictive performance and were not interpreted as indicators of causal effects. While the primary machine-learning analysis used response-level target scores formed by averaging the three expert ratings for the 200 unique responses, additional analyses were also performed on the 600 evaluation-level records while retaining the original scores of all three experts to support interpretation of the findings. Different combinations of basic NLP features, response SBERT embeddings, and semantic features representing the question-response relationship were also compared. These were considered exploratory and sensitivity analyses for development of the primary model; final generalizability performance was based on the nested grouped cross-validation results described in Section 2.10.

3. RESULTS

3.1. Dataset and Inter-Rater Reliability

A total of 200 unique responses generated by four different LLMs for 50 clinical questions were evaluated. Fifteen questions concerned ACL injuries, 15 meniscal tears, and 20 total knee arthroplasty. Because each response was independently evaluated by three orthopaedic specialists, 600 individual expert evaluations were obtained. Complete evaluations by all three experts were confirmed for every response, with 50 responses per model. In the original individual expert scores, complete agreement among all three raters occurred in 56.0% of responses for Accuracy, 57.0% for Completeness, and 75.5% for Misinformation. The proportion of responses in which the maximum difference among raters did not exceed one point was 91.0%, 90.0%, and 96.0%, respectively. Rater differences of two or more points were observed in 18 responses (9.0%) for Accuracy, 20 (10.0%) for Completeness, and 8 (4.0%) for Misinformation. In coefficient-based analyses, ordinal Krippendorff $\alpha = 0.160$, ICC(2,1) = 0.191, and ICC(2,3) = 0.415 for Accuracy; $\alpha = 0.184$, ICC(2,1) = 0.220, and ICC(2,3) = 0.458 for Completeness; and $\alpha = 0.085$, ICC(2,1) = 0.079, and ICC(2,3) = 0.205 for Misinformation. Agreement rates and reliability coefficients for the raw expert scores are presented in Table 1.

Table 1
Inter-rater reliability of raw expert evaluations

Measure	Tam uyum, n (%)	≤ 1 puan fark, n (%)	≥ 2 puan fark, n (%)	Krippendorff α	ICC(2,1)	ICC(2,3)
Accuracy	112 (56,0)	182 (91,0)	18 (9,0)	0,160	0,191	0,415
Completeness	114 (57,0)	180 (90,0)	20 (10,0)	0,184	0,220	0,458
Misinformation	151 (75,5)	192 (96,0)	8 (4,0)	0,085	0,079	0,205

After a two-week interval, all 200 LLM responses were re-evaluated by the three experts. Comparison of the first and second evaluation periods showed no statistically significant systematic difference in Accuracy, Completeness, or Misinformation scores (Wilcoxon signed-rank test; all $p > 0.05$). This finding indicated no evident systematic shift in scoring over the two-week period.

3.2. Comparison of Overall Clinical Performance of the LLMs

The overall performance of the four LLMs across the 50 clinical questions was compared for Accuracy, Completeness, and Misinformation. Mean Accuracy scores were 3.607 ± 0.533 for ChatGPT, 3.827 ± 0.288 for Claude, 3.767 ± 0.388 for DeepSeek, and 3.787 ± 0.314 for Gemini. Completeness scores were 3.633 ± 0.496 , 3.753 ± 0.403 , 3.780 ± 0.283 ,

and 3.740 ± 0.458 , respectively; Misinformation scores were 0.127 ± 0.201 , 0.120 ± 0.231 , 0.093 ± 0.243 , and 0.107 ± 0.218 . Descriptive results and overall comparison tests are presented in Table 2.

Table 2
Overall clinical performance of the four LLMs and Friedman test results

Measure	ChatGPT	Claude	DeepSeek	Gemini	Friedman χ^2	p	Kendall's W
Accuracy	$3,607 \pm 0,533$	$3,827 \pm 0,288$	$3,767 \pm 0,388$	$3,787 \pm 0,314$	7,060	0,070	0,047
Completeness	$3,633 \pm 0,496$	$3,753 \pm 0,403$	$3,780 \pm 0,283$	$3,740 \pm 0,458$	3,204	0,361	0,021
Misinformation	$0,127 \pm 0,201$	$0,120 \pm 0,231$	$0,093 \pm 0,243$	$0,107 \pm 0,218$	2,638	0,451	0,018

3.3. Model Performance by Clinical Category

To examine variation in model performance by clinical topic, results were evaluated separately for ACL injury, meniscal tear, and total knee arthroplasty (TKA) categories. Accuracy, Completeness, and Misinformation scores by category and model are presented in Table 3.

Table 3
Expert evaluation scores according to clinical category and LLM

Clinical Category	LLM	Accuracy	Completeness	Misinformation
ACL Injury	ChatGPT	3.844 ± 0.213	3.867 ± 0.211	0.111 ± 0.163
	Claude	3.800 ± 0.211	3.667 ± 0.454	0.133 ± 0.169
	DeepSeek	3.733 ± 0.382	3.756 ± 0.367	0.089 ± 0.198
	Gemini	3.756 ± 0.367	3.822 ± 0.305	0.089 ± 0.198
Meniscal Tear	ChatGPT	3.689 ± 0.367	3.689 ± 0.367	0.067 ± 0.138
	Claude	3.911 ± 0.153	3.822 ± 0.375	0.111 ± 0.241
	DeepSeek	3.844 ± 0.213	3.822 ± 0.248	0.067 ± 0.187
	Gemini	3.822 ± 0.278	3.622 ± 0.700	0.156 ± 0.248
TKA	ChatGPT	3.367 ± 0.700	3.417 ± 0.639	0.183 ± 0.253
	Claude	3.783 ± 0.394	3.767 ± 0.391	0.117 ± 0.271
	DeepSeek	3.733 ± 0.491	3.767 ± 0.244	0.117 ± 0.311
	Gemini	3.783 ± 0.311	3.767 ± 0.308	0.083 ± 0.213

In category-level Friedman analyses, no statistically significant differences among models were detected in the ACL group for Accuracy ($p = 0.876$), Completeness ($p = 0.500$), or Misinformation ($p = 0.568$). Similarly, overall differences among models were not significant for meniscal questions in Accuracy ($p = 0.241$), Completeness ($p = 0.473$), or Misinformation ($p = 0.532$). In the TKA category, however, an overall difference among models was detected for Accuracy ($\chi^2(3) = 8.233$, $p = 0.041$; Kendall's W = 0.137). TKA Completeness ($p = 0.090$) and Misinformation ($p = 0.343$) were not statistically significant. In pairwise comparisons of TKA Accuracy scores, no model pair reached statistical significance after Holm correction (all adjusted $p \geq 0.117$). Therefore, the overall difference in the TKA category could

not be attributed to a specific model pair. The distribution of mean expert Accuracy scores by clinical category and model is shown in Fig. 1.

3.4. Question-Level Performance and Between-Model Variability

To evaluate variation in model performance across individual clinical questions, mean Accuracy, Completeness, and Misinformation scores across the four LLMs were calculated for each question. The 25th percentile of the question-level mean Accuracy distribution was 3.667 and the 75th percentile was 3.917. According to the predefined descriptive threshold, 13 questions with mean Accuracy ≤ 3.667 were identified as lower-performing. Seven of these questions were in the TKA category, three in ACL, and three in the meniscal category. The lowest mean Accuracy was observed for Q44 in the TKA category, “How long do patients stay in the hospital after knee replacement?” For this question, mean Accuracy was 2.417, Completeness 2.833, and Misinformation 0.917. Model-specific Accuracy scores for Q44 were 2.000 for ChatGPT, 2.667 for Claude, 2.000 for DeepSeek, and 3.000 for Gemini. Mean expert Accuracy scores for all 50 questions and the lower-performing questions identified according to the 25th percentile are shown in Fig. 2.

A total of 123 free-text comments were recorded in the expert evaluations and were associated with 88 different LLM responses (44.0%). Responses receiving at least one expert comment numbered 27/50 (54.0%) for ChatGPT, 21/50 (42.0%) for DeepSeek, 21/50 (42.0%) for Gemini, and 19/50 (38.0%) for Claude. The most frequently identified issue was unnecessary, overly detailed, or off-topic content, observed in 72 different responses. Explicit comments regarding incorrect or misleading information were recorded for 10 responses, incomplete or insufficient content for four, and specific clinical corrections for 15. The distribution by model and comment category is shown in Table 4.

Table 4
Distribution of expert free-text comments according to LLM and comment category

LLM	Responses with any comment, n (%)	Unnecessary/overly detailed, n	Incorrect/misleading, n	Incomplete/insufficient, n	Specific clinical correction, n
ChatGPT	27 (54.0)	24	0	1	3
Claude	19 (38.0)	15	5	0	4
DeepSeek	21 (42.0)	18	3	0	4
Gemini	21 (42.0)	15	2	3	4
Total	88 (44.0)	72	10	4	15

3.5. Semantic Similarity Analysis

In the SBERT-based semantic analysis, pairwise cosine similarities were calculated among the responses provided by the four LLMs to the same clinical question. Mean similarity values exceeded 0.81 for all model pairs. The highest mean response similarity was observed between Claude and Gemini (0.850 ± 0.074), followed by Claude-DeepSeek (0.843 ± 0.113) and ChatGPT-Claude (0.831 ± 0.071). Semantic similarity results for model pairs are shown in Table 5.

Table 5
Pairwise semantic similarity of LLM responses to the same clinical questions

LLM Pair	n	Mean Cosine Similarity	SD	Median	Min	Max
ChatGPT–Claude	50	0.831	0.071	0.850	0.651	0.929
ChatGPT–DeepSeek	50	0.813	0.108	0.841	0.220	0.928
ChatGPT–Gemini	50	0.815	0.083	0.845	0.606	0.929
Claude–DeepSeek	50	0.843	0.113	0.864	0.144	0.953
Claude–Gemini	50	0.850	0.074	0.861	0.461	0.946
DeepSeek–Gemini	50	0.822	0.119	0.854	0.169	0.93

3.6. Machine-Learning Prediction of Clinical Quality Scores

In the final machine-learning approach, RBF-SVR models were constructed using 23 NLP, semantic, and explainable textual features jointly. Generalizability performance was evaluated using nested grouped cross-validation based on Question_ID, and all performance metrics were calculated from out-of-fold predictions obtained from the outer test folds for the 200 unique responses. For Accuracy prediction, the final SVR model achieved MAE = 0.261, representing an 11.6% lower absolute prediction error than the training-fold mean baseline (MAE = 0.295). RMSE was 0.428 and R^2 was -0.161, while a positive and statistically significant rank association was found between observed and predicted Accuracy scores (Spearman ρ = 0.202; p = 0.004). For Completeness prediction, the SVR model achieved MAE = 0.294, a 3.8% improvement over the baseline MAE = 0.306. However, RMSE = 0.439, R^2 = -0.106, and Spearman ρ = 0.076 (p = 0.283). Final nested grouped cross-validation results are shown in Table 6.

Table 6
Nested grouped cross-validation performance of the final RBF-SVR model

Target	Model	MAE	MAE Improvement	RMSE	R^2	Spearman ρ	p
Accuracy	Mean baseline	0.295	—	0.402	-0.021	-0.191	0.007
Accuracy	Final RBF-SVR	0.261	11.6%	0.428	-0.161	0.202	0.004
Completeness	Mean baseline	0.306	—	0.420	-0.011	-0.148	0.036
Completeness	Final RBF-SVR	0.294	3.8%	0.439	-0.106	0.076	0.283

Inspection of the out-of-fold prediction distributions showed that observed mean expert scores were largely concentrated in the upper part of the scale and that model predictions similarly clustered within a narrow high-score range. For Accuracy, MAE = 0.261, RMSE = 0.428, R^2 = -0.161, and Spearman ρ = 0.202 (p = 0.004); for Completeness, MAE = 0.294, RMSE = 0.439, R^2 = -0.106, and Spearman ρ = 0.076 (p = 0.283). The distributions of observed and predicted values are shown in Fig. 3. Despite the low MAE values, the negative R^2 values appear to reflect the limited target variance and ceiling effect caused by expert scores being concentrated near the upper end of the scale.

3.7. Ablation and Explainability Analyses

An ablation analysis was performed on the SVR model to examine the contribution of different feature groups to prediction of expert evaluations. With basic NLP features alone, MAE = 0.277 and R^2 = -0.043 were obtained for Accuracy. Adding question-response cosine similarity to the NLP features reduced MAE to 0.274, although explained variance remained negative (R^2 = -0.016). Using explainable textual features alone yielded a positive R^2 (0.042), while

the model combining all feature groups produced the lowest MAE (0.270) and highest R^2 (0.090). A similar pattern was observed for Completeness prediction. R^2 was -0.105 with basic NLP features and increased to 0.030 with explainable features alone. The model combining NLP features, question-response semantic similarity, and explainable features achieved the lowest MAE (0.284), highest R^2 (0.074), and strongest rank correlation (Spearman $\rho = 0.293$; $p < 0.001$). Ablation results are presented in Table 7. These ablation findings are exploratory results based on single-level Question_ID-grouped cross-validation and were evaluated separately from the final model generalizability performance obtained using nested grouped cross-validation.

Table 7
Ablation analysis of feature groups using SVR

Target	Feature Set	MAE	RMSE	R^2	Spearman ρ	p
Accuracy	NLP	0.277	0.406	-0.043	0.118	0.095
Accuracy	NLP + Q-R similarity	0.274	0.401	-0.016	0.111	0.117
Accuracy	Explainable features	0.275	0.389	0.042	0.202	0.004
Accuracy	NLP + Q-R similarity + Explainable	0.270	0.379	0.090	0.230	0.001
Completeness	NLP	0.293	0.439	-0.105	0.138	0.052
Completeness	NLP + Q-R similarity	0.292	0.430	-0.060	0.116	0.100
Completeness	Explainable features	0.292	0.411	0.030	0.240	< 0.001
Completeness	NLP + Q-R similarity + Explainable	0.284	0.402	0.074	0.293	< 0.001

Using all feature groups together in the ablation analysis yielded the lowest MAE values for both target variables. With the combined feature set, MAE decreased to 0.270 for Accuracy and 0.284 for Completeness. The change in prediction error as feature groups were added is shown in Fig. 4.

4. DISCUSSION

In this study, responses generated by ChatGPT, Claude, DeepSeek, and Gemini to orthopaedic patient questions were evaluated by three orthopaedic specialists for Accuracy, Completeness, and Misinformation. In addition, we investigated whether expert-rated clinical quality scores could be predicted using machine learning based on natural language processing, SBERT-based semantic features, and explainable textual features. The principal findings showed that all four LLMs generally achieved high Accuracy and Completeness and low Misinformation scores, with no statistically significant overall difference among models. Performance nevertheless varied by clinical category and, particularly, at the individual-question level. Responses generated by different models to the same questions also showed high semantic similarity. In the machine-learning analysis, combining linguistic, semantic, and explainable features reduced absolute prediction error for Accuracy relative to the baseline model, although negative R^2 values indicated that automated prediction of expert evaluations remains challenging. Mean Accuracy scores of 3.61–3.83/4, Completeness scores of 3.63–3.78/4, and low Misinformation scores indicate that the current LLMs evaluated in this study can generally produce clinically high-quality responses to orthopaedic patient questions. However, no statistically significant overall differences among models were found for Accuracy, Completeness, or Misinformation. Accordingly, descriptive differences in mean scores should not be interpreted as evidence of overall clinical superiority of a particular model. This finding is broadly consistent with recent orthopaedic literature. For example, a study comparing ChatGPT-4.5, Gemini 2.0 Flash, and DeepSeek-V3-FW for patient questions related to hip fractures found similar accuracy across the three models, with 96.7% of responses rated as “Good” or “Excellent” [6]. Çakmur et al.,

comparing ChatGPT-4o, Gemini 1.5 Flash, and DeepSeek-V3 for meniscal injuries, likewise reported that all three models generated high-quality patient-education content but that model performance varied according to the quality dimension assessed. Gemini performed better for reliability, whereas ChatGPT showed better readability, and there was no significant difference among models in the overall quality score [24]. Similarly, Hsu et al., who examined concordance of ChatGPT, Gemini, and Claude responses with the 2024 AAOS guideline for acute isolated meniscal pathology, found no significant overall difference among models in guideline concordance [25]. These results support evaluating LLM performance together with the clinical topic and quality dimension rather than using a single “best model” framework. Previous studies of total knee arthroplasty also show that LLMs can provide highly accurate answers to patient questions while performance varies by question type. In a study of 50 frequently asked TKA questions, 88% of ChatGPT responses were classified as accurate [26]. Magruder et al. reported that ChatGPT performed better on TKA questions related to well-defined clinical guidelines and less well on questions concerning pain and functional outcomes [5]. Overall inter-rater reliability in that study was 0.33. Considered together with our category- and question-level findings, these results suggest that orthopaedic LLM performance depends not only on the model used but also on clinical content and contextual sensitivity. In the present study, no overall significant model differences were found in the ACL and meniscal categories, whereas a significant overall model effect for Accuracy was detected in TKA questions. However, because no pairwise model comparison remained significant after Holm correction, the TKA finding appears to reflect between-model variability within this category rather than superiority of a specific model. Question-level analyses supported this interpretation. Seven of the 13 questions classified as lower-performing were in the TKA category. The lowest mean Accuracy was observed for Q44, which asked about hospital length of stay after knee replacement. Because length of stay depends on factors such as the healthcare system, hospital protocols, patient comorbidities, perioperative pathways, and same-day discharge practices, a single universal answer is difficult to provide. Lower performance on such questions may therefore reflect not only limitations in model knowledge but also the inherently context-dependent nature of the question. Question-level relationships also provided important findings. The strong positive association between Accuracy and Completeness ($p = 0.848$) indicates that clinically accurate responses generally included the necessary information more comprehensively. The strong negative association between Accuracy and Misinformation ($p = -0.764$) was in the expected direction. More notably, the strong negative association between Accuracy and between-model variability ($p = -0.831$) indicates that on difficult questions, not only did average performance decrease, but model responses also diverged more substantially. Reporting only overall mean scores for each model may therefore conceal clinically important failure patterns, and future LLM evaluations should also consider question-level performance and between-model variability.

Another important finding was the structure of agreement among expert evaluations. Although the maximum difference among experts did not exceed one point for approximately 90% of responses for Accuracy and Completeness, Krippendorff α and ICC values remained low. Similarly, coefficient-based reliability for Misinformation was limited despite high percentage agreement. A major explanation may be the pronounced ceiling effect in the score distributions: most Accuracy and Completeness scores were concentrated at the upper end of the scales, whereas Misinformation scores clustered around zero. Even when raters assign identical or very similar scores, low variance in the target variable can reduce coefficient-based reliability measures. Magruder et al. likewise reported low inter-rater reliability of 0.33 in their evaluation of TKA questions, suggesting that similar problems may arise when experts evaluate LLM responses [5]. In contrast, Çakmur et al. reported ICC values above 0.90 across evaluation domains in their meniscal study [24]. These differences across studies show that the scoring scale, question set, experts' evaluation approach, and score distribution may substantially affect inter-rater reliability. It is therefore methodologically important that inter-rater reliability in the present study was reported using the original individual expert scores without score merging, rescoring, or consensus procedures. Although machine-learning targets were constructed as the arithmetic mean of three independent expert scores, thereby reducing the influence of a single rater, limited reliability in the underlying evaluations may have constrained measurement precision of the targets and

attainable out-of-sample predictive performance. Negative R^2 values should therefore be interpreted in the context of both limited target variance and measurement uncertainty. The low coefficient-based reliability measures also suggest that limited variability in expert evaluations and between-rater differences may have reduced the precision of the target scores. Although averaging the three independent expert scores partially reduced individual-rater variability, limited underlying rater reliability may still have constrained achievable model performance. SBERT-based analyses showed high semantic similarity among responses generated by different LLMs to the same clinical questions, suggesting that current models largely generate responses around similar clinical concepts. Mean cosine similarities of approximately 0.81–0.85 across model pairs support this interpretation. However, semantic similarity is not synonymous with clinical accuracy. A response may be highly similar semantically to the question or to another model response while remaining clinically incomplete, overly generalized, or partly incorrect. SBERT similarity was therefore treated as an auxiliary feature representing semantic proximity rather than as a direct indicator of clinical quality. This distinction is also relevant to recent findings on accuracy and readability. Bohn et al. compared ChatGPT-4 with AAOS OrthoInfo content for meniscal injuries and found similar accuracy but lower readability for ChatGPT responses [27]. ChatGPT responses corresponded to approximately a 13.8-grade reading level, compared with 9.8 for OrthoInfo. Thus, a clinically accurate or semantically relevant response is not automatically more appropriate for patients. Accuracy, completeness, misinformation, readability, and semantic characteristics are therefore better regarded as distinct quality dimensions in clinical LLM evaluation. Free-text expert comments likewise showed that patient-oriented response quality extends beyond accuracy and completeness. The issue most frequently identified by experts was unnecessary, overly detailed, or off-topic content rather than direct misinformation, noted in 72 different responses. Such comments were more frequent for ChatGPT responses, whereas explicit incorrect or misleading information comments were recorded for Claude, DeepSeek, and Gemini. Because comments were optional, however, these distributions should not be interpreted as direct comparisons of model performance or error rates. A central engineering contribution of this study was the investigation of whether expert-rated clinical quality scores could be automatically predicted from features extracted from LLM responses. In ablation analyses, basic NLP features alone yielded negative R^2 values for Accuracy and Completeness, whereas performance improved after adding explainable textual features. The approach combining NLP features, question-response semantic similarity, and explainable textual features yielded the lowest MAE for both target variables in the ablation analyses. Positive R^2 values observed in exploratory ablation analyses should not be interpreted as estimates of final model generalizability; in the stricter nested grouped cross-validation evaluation, out-of-sample R^2 values were negative. This suggests that surface features such as response length, sentence count, or readability alone are insufficient for clinical-quality prediction, whereas semantic relationships and structural characteristics may provide complementary information. In the stricter nested grouped cross-validation analysis, the final RBF-SVR model achieved MAE = 0.261 for Accuracy and reduced absolute prediction error by 11.6% relative to the training-fold mean baseline. A weak but statistically significant rank association was also observed between predicted and observed Accuracy ($p = 0.202$; $p = 0.004$). However, $R^2 = -0.161$ and RMSE remaining higher than the baseline are important limitations. For Completeness, MAE improved by only 3.8%, and the Spearman correlation was not significant. These findings do not indicate that the proposed approach can replace expert evaluation as an automated clinical scoring system. Rather, they suggest that the linguistic, semantic, and explainable characteristics of responses contain a limited but measurable predictive signal associated with expert Accuracy ratings. The observed-predicted distribution, in which expert scores were mostly concentrated between 3 and 4 while model predictions occupied a narrower high-score band, supports this interpretation. This ceiling effect may have enabled low MAE through predictions near the mean while making it difficult to distinguish the small number of lower-quality responses. Future studies should therefore use datasets with broader quality distributions. Inclusion of more clinical questions, different orthopaedic subspecialties, and responses spanning a wider range of quality may improve the discriminative ability of automated quality-prediction models. These results should also be interpreted in the context of the broader orthopaedic literature. In the 2026 systematic review and meta-analysis by Ramadanov et al. of patient questions on

hip arthroscopy, eight studies were evaluated and pooled accuracy was 88.6%, whereas completeness was lower at 83.8% and more variable across studies [7]. As in the present study, this finding suggests that evaluating LLM responses using a single quality metric may be insufficient. Even when current models show high overall performance, clinically meaningful differences may emerge for particular question types or quality dimensions.

This study has several limitations. First, the evaluation was limited to 50 clinical questions, four LLMs, and three clinical domains: ACL injuries, meniscal tears, and TKA. The findings cannot be directly generalized to other orthopaedic conditions or medical specialties. Second, because LLMs are updated rapidly, the results represent the model versions used during the data-collection period. Third, clinical quality was based on expert evaluation, and coefficient-based inter-rater reliability was low. The use of three independent experts and evaluation of inter-rater reliability using the original individual scores without consensus or rescoring allowed rater variability to be presented transparently. Nevertheless, low coefficient-based reliability indicates that measurement uncertainty in expert evaluations could not be fully eliminated. Fourth, a single standardized prompt and the first valid response were used for each model; different prompting strategies or repeated generations may yield different results. Sample size is also an important limitation for the machine-learning analysis. Although the dataset contained 600 expert evaluations, these represented evaluations of 200 unique LLM responses by three experts and cannot be considered 600 independent textual samples. Conducting the primary ML analysis at the level of the 200 unique responses and keeping the four model responses to the same question within the same cross-validation fold reduced data leakage, but the sample size remains limited. In addition, because the final model family and feature set were selected following prior experimental comparisons on the same dataset, nested cross-validation does not completely eliminate possible model-selection optimism. The proposed approach therefore requires validation in a larger, independent external dataset. Overall, the findings indicate that current LLMs can usually generate high-quality responses to orthopaedic patient questions, but performance may vary by question and clinical context. These systems are better considered tools that support rather than replace clinician-patient communication in patient education. Expert verification remains important, particularly for questions sensitive to patient and healthcare-system context, such as length of stay, complication risk, individual prognosis, or treatment choice. From an engineering perspective, the partial predictability of expert quality evaluations from textual and semantic features may provide a basis for future quality-control systems for automated pre-screening of LLM responses; however, current performance does not support use of such a system as an independent clinical decision-making or quality-assessment tool.

5. CONCLUSION

In this study, responses generated by ChatGPT, Claude, DeepSeek, and Gemini to patient questions concerning ACL injuries, meniscal tears, and total knee arthroplasty generally demonstrated high Accuracy and Completeness and low Misinformation. No statistically significant difference in overall clinical performance was found among the four LLMs; however, performance varied by clinical category and particularly by individual question. The overall model difference in Accuracy within the TKA category and the concentration of a substantial proportion of lower-performing questions in this category indicate that LLM evaluations should consider clinical topic- and question-level performance in addition to overall model averages. SBERT-based analyses showed that different LLMs generated largely semantically similar responses to the same clinical questions. However, semantic similarity was not equivalent to clinical accuracy and should not be used alone as an indicator of clinical quality. This finding supports the importance of clinical expert evaluation when assessing LLM responses. From an engineering perspective, we investigated whether expert-assigned clinical quality scores could be predicted from NLP features, question-response semantic similarity, and explainable textual features. Combining feature groups yielded the lowest prediction errors in the ablation analyses, and in the stricter nested grouped cross-validation analysis the final RBF-SVR model reduced MAE for Accuracy by 11.6% relative to the baseline model. Nevertheless, negative R^2 and weak rank correlation indicate that automated quality prediction

remains limited in the present dataset. The proposed approach should therefore be considered a method for automated pre-screening of LLM responses rather than a replacement for expert evaluation. In conclusion, current LLMs can generate responses capable of supporting clinician-patient communication in orthopaedic patient education, but expert verification remains important, particularly for context-sensitive questions. Machine-learning approaches based on linguistic, semantic, and explainable features developed and validated in larger independent datasets may contribute to future support systems for automated quality control of clinical LLM responses and early identification of potentially risky responses.

Declarations

Ethics approval and consent to participate

This study did not involve patients, patient data, human biological material, or clinical interventions. The study data consisted of responses generated by large language models and evaluations performed by the study investigators. Therefore, ethics committee approval and informed consent to participate were not required.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

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

Author Contribution

V.K. conceived and designed the study, contributed to the development of the clinical question set, participated in expert evaluation of the LLM responses, interpreted the clinical findings, and drafted and critically revised the manuscript. M.A. developed the computational methodology, performed the natural language processing, SBERT-based semantic analyses, machine-learning modelling, statistical analyses, and interpretation of the computational results, and contributed to manuscript preparation. C.A. contributed to data organization, methodological review, interpretation of the findings, and critical revision of the manuscript. A.L. and M.M.A. contributed to the development and clinical review of the orthopaedic questions, performed expert evaluation of the LLM responses, contributed to interpretation of the clinical findings, and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

During manuscript preparation, ChatGPT (OpenAI) was used for language editing, improvement of clarity, and assistance with manuscript formatting. All AI-assisted text was critically reviewed, revised, and verified by the authors,

who take full responsibility for the final content. The generative AI systems evaluated as study objects are described separately in the Methods section.

Availability of data and materials

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1. Jacobs, M.M.G., Oosterhoff, J.H.F., Agricola, R., van der Weegen, W.: Large language models versus healthcare professionals in providing medical information to patient questions: a systematic review. *Int. J. Med. Inf.* **209**, 106250 (2026). 10.1016/j.ijmedinf.2025.106250 PMID:41529641. PubMed <https://pubmed.ncbi.nlm.nih.gov/41529641/>
2. Zhang, C., Liu, S., Zhou, X., Zhou, S., Tian, Y., Wang, S., et al.: Examining the role of large language models in orthopedics: systematic review. *J. Med. Internet Res.* **26**, e59607 (2024). 10.2196/59607 PMID:39546795. PubMed <https://pubmed.ncbi.nlm.nih.gov/39546795/>
3. Mo, K., Lin, R., Dunn, E., Girgis, G., Fang, W., Walsh, J., et al.: Systematic review on large language models in orthopaedic surgery. *J. Clin. Med.* **14**(16), 5876 (2025). 10.3390/jcm14165876 PMID:40869701. PubMed <https://pubmed.ncbi.nlm.nih.gov/40869701/>
4. Mika, A.P., Mulvey, H.E., Engstrom, S.M., Polkowski, G.G., Martin, J.R., Wilson, J.M.: Can ChatGPT answer patient questions regarding total knee arthroplasty? *J. Knee Surg.* **37**(9), 664–673 (2024). 10.1055/s-0044-1782233 PMID:38442904. PubMed <https://pubmed.ncbi.nlm.nih.gov/38442904/>
5. Magruder, M.L., Rodriguez, A.N., Wong, J.C.J., Erez, O., Piuizzi, N.S., Scuderi, G.R., et al.: Assessing ability for ChatGPT to answer total knee arthroplasty-related questions. *J. Arthroplasty.* **39**(8), 2022–2027 (2024). 10.1016/j.arth.2024.02.023 PMID:38364879. PubMed <https://pubmed.ncbi.nlm.nih.gov/38364879/>
6. Zhang, Y., Huang, T., Liu, C., Miller, A.N., Yang, M., Harris, I.A., et al.: Comparative evaluation of large language models for hip fracture-related patient questions: DeepSeek-V3-FW, Gemini 2.0 Flash, and ChatGPT-4.5. *Digit. Health.* **12**, 20552076251412989 (2026). 10.1177/20552076251412989 PMID:41623725. PubMed <https://pubmed.ncbi.nlm.nih.gov/41623725/>
7. Ramadanov, N., Penchev, P., Voss, M., Heinz, M., Prill, R., Salzmann, M., et al.: ChatGPT provides accurate and safe responses to patient questions on hip arthroscopy, while completeness remains variable: a systematic review and single-arm meta-analysis. *Knee Surg. Sports Traumatol. Arthrosc.* **34**(6), 2236–2246 (2026). PMID:41979367. PubMed <https://pubmed.ncbi.nlm.nih.gov/41979367/>
8. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. *PLoS One.* **14**(11), e0224365 (2019). 10.1371/journal.pone.0224365 PMID:31697686. PubMed <https://pubmed.ncbi.nlm.nih.gov/31697686/>
9. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* **86**(2), 420–428 (1979)
10. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**(200), 675–701 (1937)
11. Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, 65–70. (1979)
12. Spearman, C.: The proof and measurement of association between two things. (1961)

13. Flesch, R.: A new readability yardstick. *J. Appl. Psychol.* **32**(3), 221–233 (1948)
14. Kincaid, J.P., Fishburne, R.P. Jr., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (1975)
15. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992). (2019), November
16. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics.* **12**(1), 55–67 (1970)
17. Drucker, H., Burges, C.J., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. *Advances in neural information processing systems*, 9. (1996)
18. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
19. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 1189–1232. (2001)
20. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci.* ;374(2065):20150202. (2016)
21. Varma, S., Simon, R.: Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7**, 91 (2006)
22. Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S.: Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.* **6**, 10 (2014)
23. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**(177), 1–81 (2019)
24. Çakmur, B.B., Koluman, A.C., Çiftçi, M.U., Aloğlu Çiftçi, E., Ziroğlu, N.: Gemini 1.5 Flash provides the most reliable content while ChatGPT-4o offers the highest readability for patient education on meniscal tears. *Knee Surg. Sports Traumatol. Arthrosc.* **34**(3), 1141–1149 (2026). 10.1002/ksa.70247 PMID:41451752. PubMed <https://pubmed.ncbi.nlm.nih.gov/41451752/>
25. Hsu, W.K., Chuang, H.C., Wang, Y.Y., Hsu, K.L., Kuan, F.C., Su, W.R., et al.: Concordance of ChatGPT, Gemini, Claude, and OpenEvidence with the 2024 AAOS guidelines on acute isolated meniscal pathology. *Knee.* ;61:104427. 10.1016/j.knee.2026.104427. PMID:41865615. PubMed: (2026). <https://pubmed.ncbi.nlm.nih.gov/41865615/>
26. Zhang, S., Liao, Z.Q.G., Tan, K.L.M., Chua, W.L.: Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement. *Knee Surg. Relat. Res.* **36**(1), 15 (2024). 10.1186/s43019-024-00218-5 PMID:38566254. PubMed <https://pubmed.ncbi.nlm.nih.gov/38566254/>
27. Bohn, C., Hand, C., Tannir, S., Ulrich, M., Saniei, S., Girod-Hoffman, M., et al.: American Academy of Orthopedic Surgeons' OrthoInfo provides more readable information regarding meniscus injury than ChatGPT-4 while information accuracy is comparable. *J ISAKOS.* ;11:100843. 10.1016/j.jisako.2025.100843. PMID:39988021. PubMed: (2025). <https://pubmed.ncbi.nlm.nih.gov/39988021/>

Figures

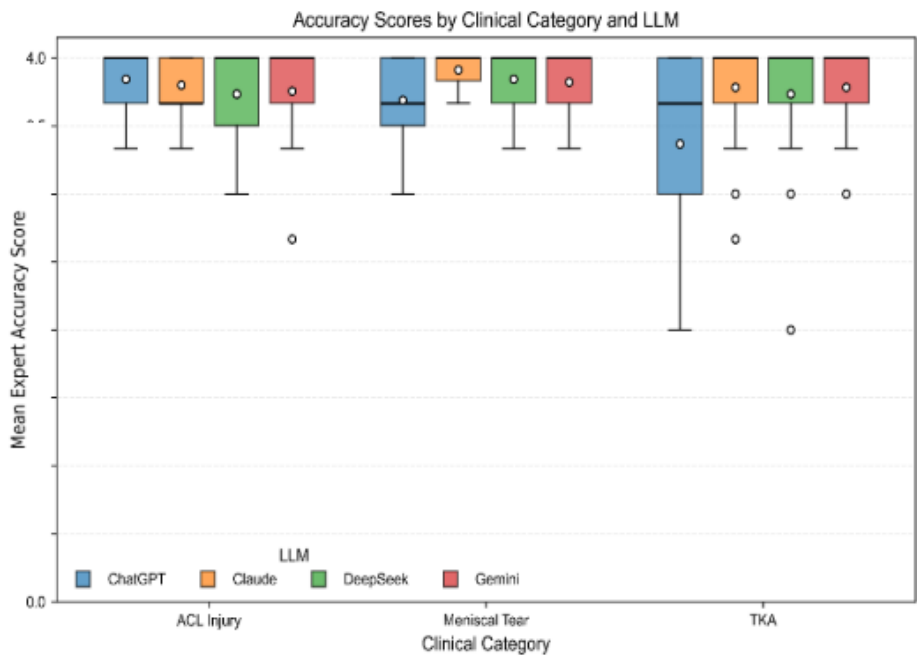


Figure 1

Accuracy scores across clinical categories and large language models.

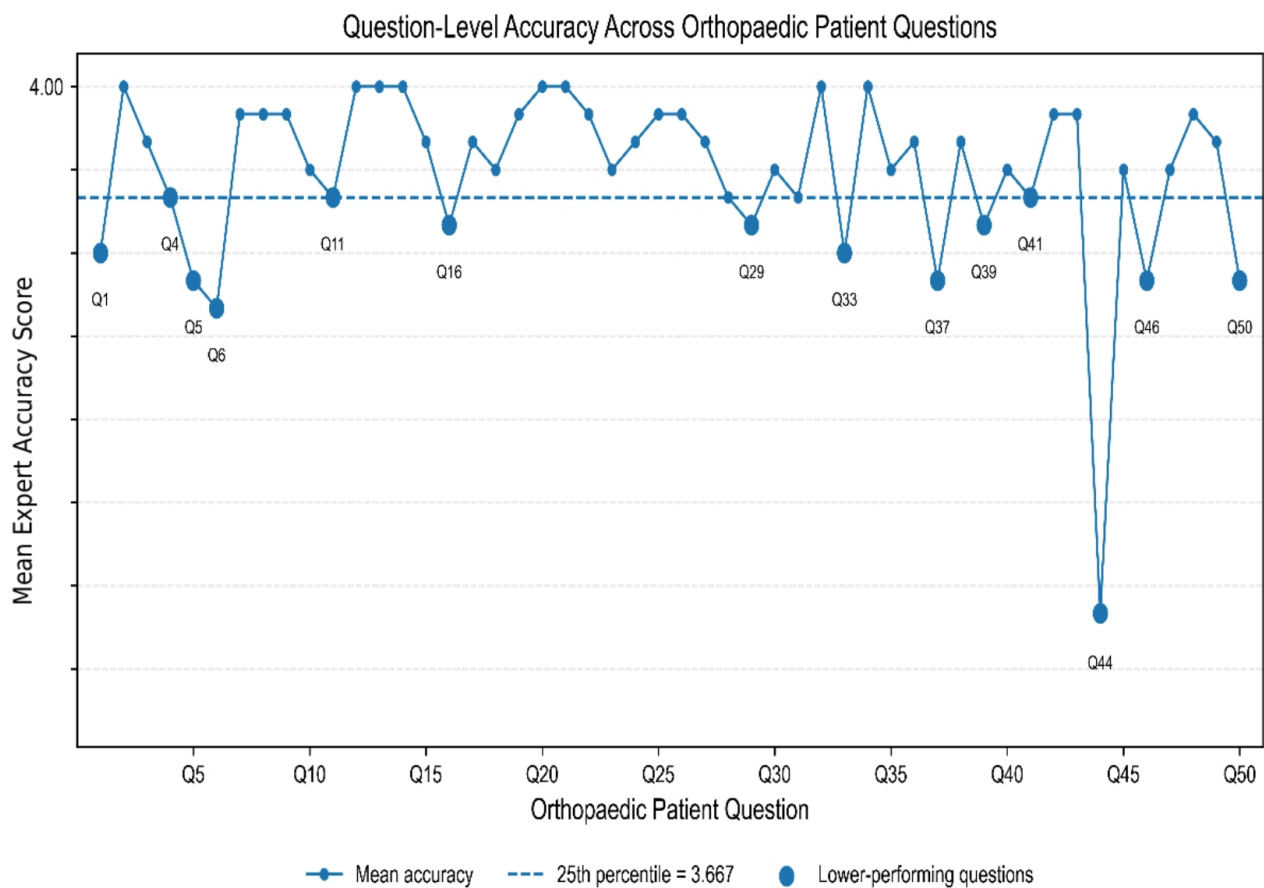


Figure 2

Question-level Accuracy across orthopaedic patient questions.

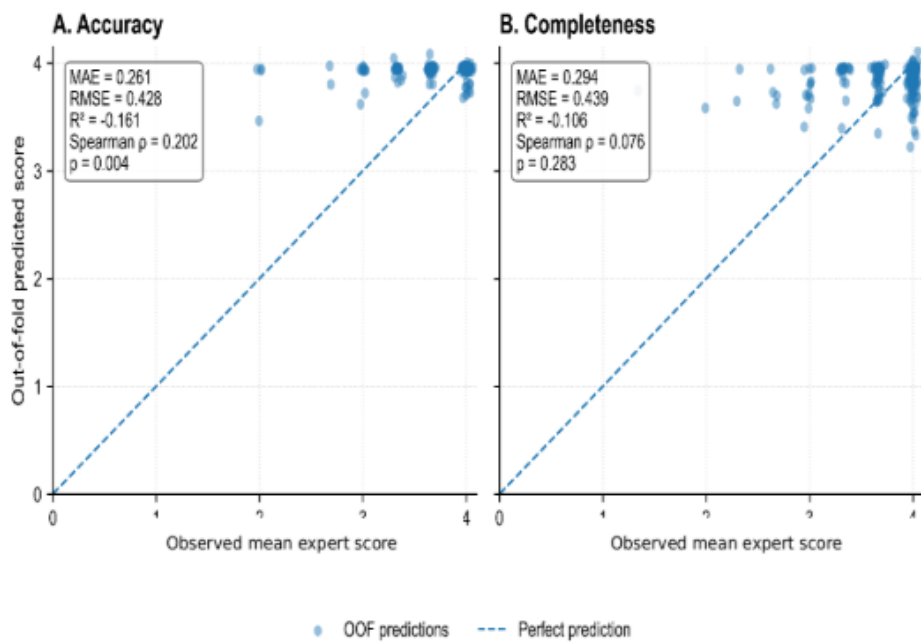


Figure 3

Observed versus out-of-fold predicted mean expert scores using the final RBF-SVR model.

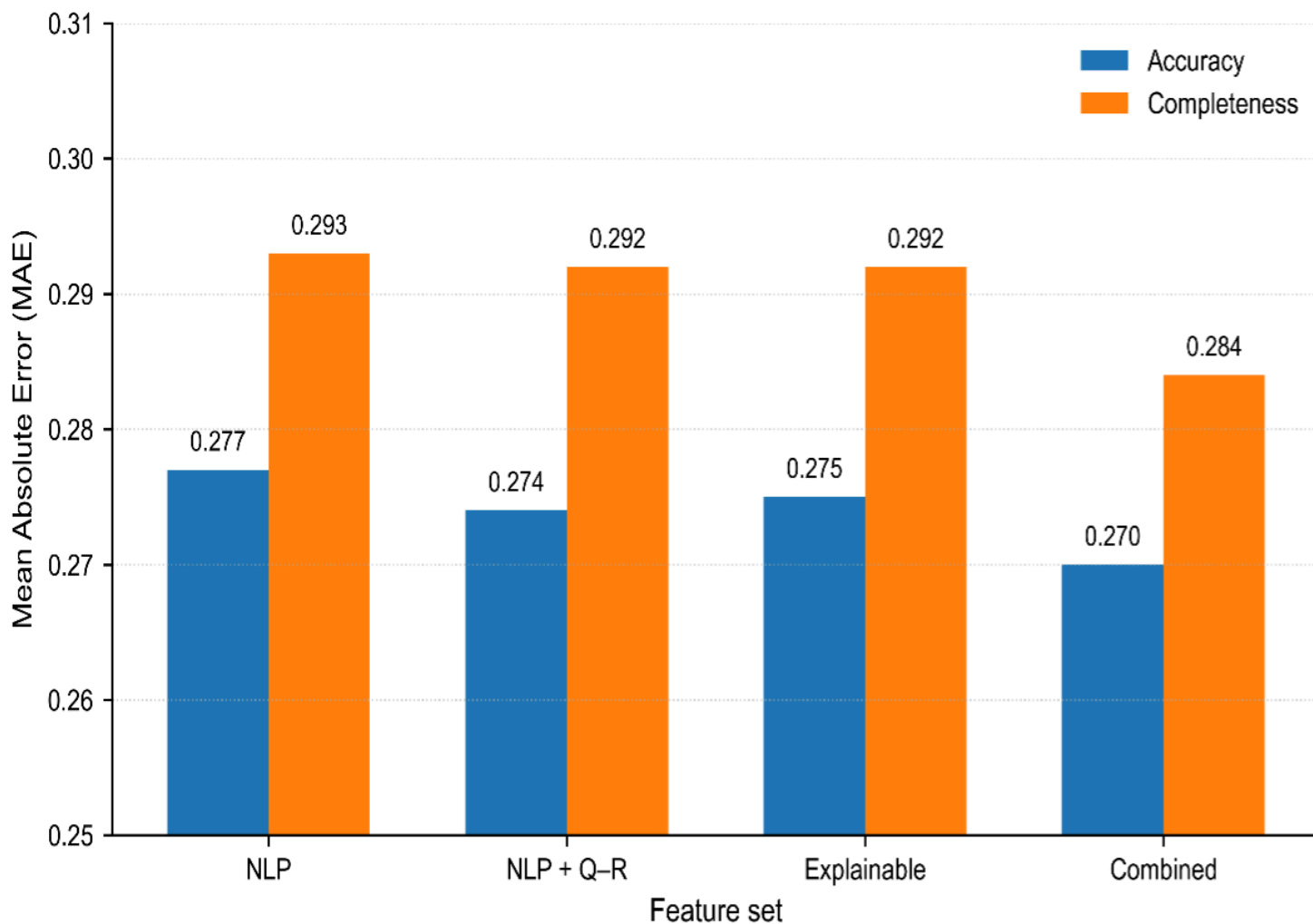


Figure 4

Ablation analysis of feature sets for expert quality score prediction.