Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Concordance of large language model responses with BACPAR evidence-based guidelines in amputee and prosthetic rehabilitation: a cross-sectional study

Yasin Ekinci

yasin.ekinci@sbu.edu.tr

Sağlık Bilimleri Üniversitesi

Muharrem Gökhan Beydağı

Fırat University

Research Article

Keywords: Amputation, Artificial Limbs, Rehabilitation, Artificial Intelligence, Large Language Models, Practice Guidelines as Topic

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11458036/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Concordance of large language model responses with BACPAR evidence-based guidelines in amputee and prosthetic rehabilitation: a cross-sectional study

AUTHORS: Yasin EKİNCİ [a], Muharrem Gökhan BEYDAGI [b]

AUTHORS' AFFILIATIONS:
[a] Sağlık Bilimleri Üniversitesi, Faculty of Gülhane Health Sciences, Department of Orthotics and Prosthetics, Ankara, Türkiye

[b] Fırat University, Faculty of Health Sciences, Department of Physical Therapy and Rehabilitation, Elazığ, Türkiye

Corresponding Author:

Yasin EKİNCİ
Address: Health Sciences University, Faculty of Gülhane Health Sciences, Department of Orthotics and Prosthetics, Gülhane Campus, 06018 Etlik/Keçiören, Ankara, Türkiye
E-mail: fztyasinekinci@gmail.com
ORCID: 0000-0001-6258-9234

Co-Author:

Muharrem Gökhan BEYDAGI
Address: Fırat University, Faculty of Health Sciences, Department of Physical Therapy and Rehabilitation, Elazığ, Türkiye
E-mail: mgbeydagi@firat.edu.tr
ORCID: 0000-0003-4655-1978

Abstract

Background: Large language models (LLMs) are increasingly used to access health information; however, it is unknown whether their responses align with evidence-based guidelines for amputation and prosthetic rehabilitation. This study aimed to assess the alignment of LLM responses with British Association of Chartered Physiotherapists in Amputee Rehabilitation (BACPAR) guideline recommendations and to compare model performance.

Methods: In this cross-sectional study, 18 open-ended clinical questions covering six sections of the BACPAR guidelines were formulated. The questions were posed to ChatGPT 5.2, Gemini 3, and DeepSeek-V3.2 once each on the same day, in new sessions, and using the exact same wording (54 responses). The responses, with model identities concealed, were independently scored by two expert physiotherapists using criteria for reliability, usefulness, guideline concordance, potential harm, and the modified DISCERN (mDISCERN) scale. ROUGE-L F1, semantic similarity, and Flesch-Kincaid Reading Ease (FKRE) scores were also calculated. The models were compared using Friedman and Bonferroni-corrected Wilcoxon tests, and relationships between the metrics were examined using Spearman's correlation.

Results: Inter-rater agreement was excellent (intraclass correlation coefficient, 0.928-1.000). All three models achieved success rates ranging from 82% to 93% on expert criteria, and no response was found to be clearly harmful. ChatGPT 5.2 scored marginally higher than Gemini 3 on mDISCERN (omnibus $p=0.050$; post-hoc $p=0.037$) and significantly higher than the other two models on ROUGE-L F1 ($p=0.001$). Semantic similarity was also higher for ChatGPT 5.2 and DeepSeek-V3.2 than for Gemini 3 ($p=0.008$). DeepSeek-V3.2 produced the most readable responses ($p \leq 0.002$), but the responses from all models fell within the "very difficult" readability range. When all responses were evaluated collectively, readability was negatively

correlated with reliability (r=-0.51), guideline concordance (r=-0.46), and information quality (r=-0.49). Automatic similarity metrics were only weakly to moderately correlated with expert scores.

Conclusions: The evaluated LLMs provided responses to questions regarding amputation and prosthetic rehabilitation that were largely guideline-concordant and safe. However, responses required advanced reading skills, and more readable responses tended to have lower clinical quality, suggesting that these tools should not be used without expert supervision, particularly in patient education. LLMs should be considered supportive tools that facilitate access to information in clinical practice.

Keywords: Amputation; Artificial Limbs; Rehabilitation; Artificial Intelligence; Large Language Models; Practice Guidelines as Topic

Background

Lower limb amputations are a serious clinical condition that can result from various causes in the adult population, such as trauma, vascular diseases, diabetes, and malignancies; they significantly affect an individual's functional capacity, independence, quality of life, and social life. Amputation management is a multidisciplinary and multifaceted process planned with consideration of the individual's amputation level, cause of amputation, etiology, accompanying comorbidities, and functional goals. In this process, appropriate prosthetic applications, rehabilitation programs, pain management, and psychosocial support approaches are evaluated in conjunction with wound care and medical treatments (1, 2).

During this process, it is crucial that clinical decisions or recommendations provided to patients align with evidence-based guideline recommendations developed through consensus among experts in the field. Guidelines play a crucial role in managing complex processes such as

lower-limb amputation rehabilitation. While various guidelines on this subject exist in the literature (3, 4), the "Evidence-based clinical guidelines for the physiotherapy management of adults with lower limb prostheses" by the British Association of Chartered Physiotherapists in Amputee Rehabilitation (BACPAR) stands out due to its comprehensive scope. The recommendations of BACPAR provide structured, evidence-based guidance for different patient profiles and additional clinical conditions in amputee rehabilitation (5).

In recent years, Large Language Models (LLMs) have been increasingly used in the healthcare sector to access information and support clinical decision-making. LLMs are advanced artificial intelligence systems trained on large amounts of text data to assist people in many different fields. Their ability to synthesize acquired information, explain complex concepts, and generate responses in an interactive manner has increased interest in the potential use of LLMs in the healthcare field. However, their use in the healthcare field requires a careful evaluation of the accuracy, reliability, and clinical appropriateness of the information they generate.

The accuracy of the responses generated by LLMs, their adherence to guidelines, and the risk of generating potentially harmful recommendations are of critical importance from the perspectives of clinical safety and ethics. There are studies in the literature that evaluate LLM responses using metrics such as guideline adherence, readability, and information quality in the context of various diseases and guidelines (6-8). However, in the field of amputee rehabilitation, the number of studies systematically evaluating LLM performance, particularly based on BACPAR recommendations, is limited (9). Therefore, it is necessary to evaluate the alignment of LLM responses with BACPAR recommendations using objective criteria to identify recommendations that may contain incorrect or incomplete information or pose potential harm.

The primary aim of this study was to evaluate the extent to which responses generated by Large Language Models (LLMs) regarding amputee rehabilitation adhere to the evidence-based

clinical recommendations of the BACPAR and to compare the performance of different LLMs. Secondary aims were to evaluate the perceived reliability and usefulness of LLM-generated responses by expert assessors, assess the level of potentially harmful content within the responses, and evaluate their information quality. In addition, we aimed to examine automated metrics, including readability and textual similarity, and to assess inter-rater agreement between expert evaluators.

Methods

This study was designed as a cross-sectional study. Ethical approval was obtained from the Health Sciences University Gülhane Ethics Committee for Non-Drug and Non-Device Interventional Studies (Decision No. 2026-184, dated 21/05/2026), and the study was conducted in accordance with the Declaration of Helsinki. As no human participants or patient data were involved and all questions were constructed as hypothetical, guideline-derived queries, informed consent was not required.

Open-Ended Question Development Process

To evaluate the performance of large language models (LLMs) in amputee and prosthetic rehabilitation, a set of open-ended clinical questions was developed on the basis of the Evidence Based Clinical Guidelines for the Physiotherapy Management of Adults with Lower Limb Prostheses (3rd Edition, 2020) published by the British Association of Chartered Physiotherapists in Amputee Rehabilitation (BACPAR) (5). The primary objective in generating these questions was to transform guideline recommendations into queries resembling those encountered in daily clinical practice, to elicit detailed and explanatory responses from the models, and to obtain comprehensive answers suitable for both expert evaluation and computer-aided text analysis (10).

The BACPAR document organizes its recommendations into six thematic sections - the multidisciplinary team; prosthetic knowledge; assessment; the prosthetic rehabilitation program; patient education; and discharge, maintenance and participation - and comprises more than seventy individually graded recommendations together with eighteen good practice points. Submitting every recommendation to the models was neither methodologically necessary nor desirable, because a substantial proportion of these items are not suitable for conversion into discrete, independently verifiable open-ended questions: many are oriented toward local service implementation rather than direct patient care, several within the larger sections (notably the prosthetic rehabilitation program and patient education sections) are closely overlapping restatements of a single underlying principle, and others are compound or conditional statements for which no single, unambiguous "gold-standard" answer can be defined. Retaining such items would have introduced redundant content, inflated the apparent sample while weakening the one-to-one correspondence between each question and its reference answer, and compromised the objectivity of the guideline-concordance scoring.

Accordingly, candidate recommendations were screened against predefined eligibility criteria - namely, that the recommendation expressed a single, unambiguous, directly testable clinical directive; that it could be mapped to one complete and exact reference answer; that it did not substantially duplicate the content of an already-selected item; and that it was not answerable by a simple yes/no response. To preserve the representativeness of the guideline as a whole and to prevent the numerically larger sections from dominating the item pool, selection was further balanced so that all six sections were proportionally represented rather than oversampling any single domain. This process yielded a final set of eighteen questions spanning the complete thematic breadth of the guideline (Table 1), with each question anchored to a distinct, verbatim BACPAR recommendation.

The resulting sample size is consistent with, and in most cases exceeds, that of comparable studies evaluating the concordance of large language models with clinical practice guidelines, which have typically employed between ten and twenty guideline-derived items. It also reflects a deliberate balance between content coverage and methodological rigor: because each of the 54 model responses (18 questions × 3 models) was reproduced verbatim and independently scored by two experts across five separate instruments, a larger item pool would have substantially increased rater burden and the attendant risk of fatigue-related scoring drift, thereby threatening the integrity of the blinded expert evaluation that constitutes the core of this study. Eighteen questions therefore represented the point at which full thematic coverage of the guideline could be achieved without exceeding the limits of a rigorous, fully blinded, dual-rater assessment protocol.

The final question set comprised 18 open-ended questions covering all six sections of the guideline (Table 1). Questions were phrased from two complementary perspectives: the majority were framed from a clinician's standpoint, while four items (Q12, Q13, Q15, Q18) were deliberately phrased in lay, patient-facing language in order to test whether model performance was sensitive to the register of the query. The verbatim BACPAR recommendation corresponding to each question was designated as the reference ("gold-standard") answer; this reference text served both as the comparator for all automated similarity metrics and as the basis for the expected-element checklist used in the guideline-concordance rating. The full question set and the BACPAR recommendation to which each question maps are presented in Table 2.

Large Language Models and Response Collection Procedure

Three LLMs were evaluated: ChatGPT 5.2 (Model A), Gemini 3 (Model B), and DeepSeek-V3.2 (Model C). All models were accessed through their official public web interfaces. ChatGPT 5.2 and Gemini 3 were accessed using paid subscription accounts, whereas

DeepSeek-V3.2, which is offered free of charge through its web interface, was accessed using a standard account. All models were used with their default settings. Each model was presented with the same 18 open-ended clinical questions, yielding a total of 54 responses (18 questions × 3 models).

Because LLMs undergo frequent updates, which may affect reproducibility over time, each question was submitted to each model on a single occasion, on 6 July 2026, without custom instructions, system prompts or plug-ins. Each question was submitted in exactly the same wording, in English, in a newly initiated session for each model, and the following instruction was appended to every query: "Please provide your response without taking previous conversations into account." This standardized procedure prevented carry-over effects between sessions. Responses were recorded verbatim, without any alteration, editing, or truncation. To prevent bias during evaluation, model names were removed and responses were assigned codes (Model A/B/C); these codes were kept concealed from the raters until all expert evaluations were completed.

Evaluation Process of Responses

All model-generated responses were evaluated using a blinded expert-rating protocol together with a set of automated text-analysis metrics. After response collection, model names were removed, each response was coded, and the response order was randomized; model identity remained concealed until scoring was complete. The expert-based evaluation instruments and the scoring framework are summarized in Table 3.

Expert-based evaluation

Two expert clinicians (a physiotherapist specializing in amputee and prosthetic rehabilitation with 10 years of experience and a physiotherapist with 7 years of experience in the field), both

experienced in amputee and prosthetic rehabilitation, independently scored all 54 responses. Neither rater had access to the other's scores at any point during the evaluation. Expert-based evaluation comprised five instruments: reliability and usefulness (7-point Likert scales), guideline concordance (0-5), potential harm (0-2), and the modified DISCERN (mDISCERN) instrument (0-7). For all subsequent analyses, the mean of the two raters' scores was used as the composite score for each response.

Reliability and usefulness were rated on 7-point Likert scales adapted from the framework established by Uz and Umay (11). Reliability (1=not at all reliable; 7=very reliable) assessed medical accuracy, evidence basis, and the absence of misleading information, and represented the overall accuracy and clinical appropriateness of the response. Usefulness (1=not useful at all; 7=very useful) evaluated the contribution of the response to clinical decision-making, its applicability, clarity, and scope.

Guideline concordance (0-5) was assessed against a predefined list of expected elements derived from the corresponding BACPAR recommendation for each question. Raters assigned a score between 0 (not adherent) and 5 (fully adherent) on the basis of the presence or absence of the core recommendation and the correct prioritization of its constituent elements.

Potential harm (0-2) was evaluated according to the probability that the response could harm the patient through incorrect or incomplete guidance: 0=harmless (clinically non-risky content); 1=partially harmful (content that may pose a risk under certain conditions or that lacks adequate safety warnings); 2=clearly harmful (contraindicated, dangerous, or misleading content). Risky recommendations, ignored contraindications, and unsafe advice regarding residual limb care, prosthetic loading, falls, and complication management were considered in this assessment.

mDISCERN was used to rate the quality of health information in each response, taking into account clarity, the emphasis placed on sources and evidence, the balanced presentation of treatment options, the disclosure of uncertainties, and the explanation of risks. Each response was rated on a 0-7 scale, with higher scores indicating higher information quality. One rater applied the mDISCERN instrument in half-point increments where a response was judged to fall between two adjacent anchor points.

Automated text-based metrics

Automated metrics were computed on the verbatim English responses in Python (v3.12), using the corresponding BACPAR recommendation for each question as the reference text.

Semantic similarity between each model-generated response and its reference answer was quantified using distributed word-vector representations, an approach that represents text as dense embeddings shown to balance semantic expressiveness with computational efficiency (12, 13). Each response and its corresponding reference text were converted into a 300-dimensional document embedding by mean-pooling the pre-trained GloVe word vectors implemented in the spaCy en_core_web_lg language model (v3.8), and semantic similarity was computed as the cosine similarity between the two embeddings. Values approaching 1 indicate greater semantic proximity to the reference recommendation, whereas values approaching 0 indicate lower proximity. A static mean-pooled embedding approach was preferred over a transformer-based sentence encoder because it is fully deterministic, locally reproducible, and independent of externally hosted model weights. It should be noted, however, that mean-pooled static embeddings compress fine-grained syntactic and compositional information and therefore yield systematically higher absolute cosine values than contextual sentence encoders; accordingly, semantic similarity values were interpreted comparatively between models rather than as absolute measures of semantic equivalence.

ROUGE-L was used to evaluate the surface-level resemblance between model outputs and the reference recommendations. This metric quantifies the longest common subsequence (LCS) between two texts, thereby capturing sentence-level structural and lexical congruence (14, 15). ROUGE-L F1 scores were calculated using the rouge-score package with stemming enabled, where ROUGE-L F1=$(1 + \beta^2) \times P \times R / (R + \beta^2 \times P)$, with P=LCS/m, R=LCS/n, and β set to 1. Higher scores indicate greater lexical correspondence with the reference text.

The linguistic accessibility of the generated responses was assessed using the Flesch-Kincaid Reading Ease (FKRE) score (16). The FKRE is derived from average sentence length and average syllable count per word (FKRE=206.835 - 1.015 × [total words / total sentences] - 84.6 × [total syllables / total words]); higher scores indicate text that is easier to read (17). Readability analyses were performed using the textstat package in Python.

Inter-rater reliability

Because the expert-rated scores exhibited a pronounced ceiling effect, with the great majority of responses receiving scores at or near the upper end of each scale, inter-rater agreement was quantified using three complementary indices rather than a single coefficient. First, the intraclass correlation coefficient (ICC) was computed using a two-way random-effects, absolute-agreement, single-rater model [ICC (2,1)], with 95% confidence intervals; ICC values were interpreted as poor (<0.50), moderate (0.50-0.75), good (0.75-0.90), and excellent (>0.90) (18). Second, because chance-corrected agreement coefficients such as Cohen's kappa are known to be paradoxically deflated when the marginal distribution of ratings is highly skewed, Gwet's first-order agreement coefficient (AC1) was calculated, as it is substantially more robust to prevalence and marginal imbalance (19); AC1 values were interpreted according to the Landis and Koch benchmarks (0.61-0.80=substantial; 0.81-1.00=almost perfect) (20). Third, exact percentage agreement and within-1-point percentage agreement between the two raters

were calculated for each instrument, as this descriptive index is directly interpretable and unaffected by the base-rate problem. All three indices were computed across the full set of 54 responses.

Statistical Analysis

Statistical analyses were performed using Python (v3.12; scipy, statsmodels, pingouin) and IBM SPSS Statistics version 27 (IBM Corp., Armonk, NY, USA). The normality of the data was examined using both visual methods (histograms and probability plots) and analytical tests (Shapiro-Wilk and Kolmogorov-Smirnov). Continuous data were reported as mean ± standard deviation (SD). Because all expert-based scores deviated significantly from normality (Shapiro-Wilk $p<0.05$ for every expert-rated metric in all three models) and the design involved three dependent measurements obtained on the same 18 questions, non-parametric procedures were applied throughout.

Comparisons of model performance were conducted using the Friedman test. For variables showing a statistically significant omnibus difference ($p<0.05$), post-hoc pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction for the three pairwise contrasts within each metric. Relationships between assessment criteria were analysed using Spearman's correlation coefficient (r), both within each model (n=18) and in the pooled sample (n=54). To control multiple comparisons, the Benjamini-Hochberg procedure was employed to apply a false discovery rate (FDR) correction. Correlation coefficients were interpreted according to predefined thresholds as negligible (0-0.29), weak (0.30-0.49), moderate (0.50-0.69), strong (0.70-0.89), and very strong (0.90-1.00) (21). In all analyses, the level of statistical significance was set at $p<0.05$.

Generative AI tools were used to assist with language editing and drafting of the manuscript, as detailed in the Acknowledgements.

Results

Comparison of LLM Scores Across Evaluation Outcomes

Agreement between the two independent raters was excellent across all five expert-based instruments (Table 4, Table 5). ICC(2,1) values ranged from 0.928 to 1.000, Gwet's AC1 indicated almost perfect agreement for four of the five instruments (0.859-1.000) and substantial agreement for mDISCERN (0.677), and within-1-point agreement reached 100% for all instruments. Given this level of concordance, the mean of the two raters' scores was used as the composite score in all subsequent analyses.

The performance of the three models across the eight evaluation metrics is summarized in Table 4. Descriptively, Model A obtained the highest score on every expert-based metric: mDISCERN 6.50 ± 0.66 (93% of the maximum attainable score), reliability 6.33 ± 0.89 (90%), usefulness 6.33 ± 0.89 (90%), and guideline concordance 4.44 ± 0.68 (89%). Overall performance was high for all three models, with success rates ranging from 82% to 93% across the expert-rated dimensions, indicating that all evaluated LLMs produced largely guideline-consistent content in response to the BACPAR-derived questions.

The Friedman test revealed a statistically significant between-model difference for mDISCERN ($\chi^2(2)=6.00$, $p=0.050$) and for reliability ($\chi^2(2)=6.08$, $p=0.048$). Post-hoc analysis confirmed that Model A achieved significantly higher mDISCERN scores than Model B (Bonferroni-corrected $p=0.037$), whereas the contrasts between Model A and Model C ($p=0.322$) and between Model B and Model C ($p=1.000$) were not significant. For reliability, in contrast, the omnibus difference was not confirmed by any pairwise contrast after Bonferroni correction (A

vs B, p=0.071; A vs C, p=0.774; B vs C, p=0.707), and this finding should therefore be interpreted with caution. No significant between-model differences were observed for usefulness ($\chi^2(2)=4.96$, p=0.084), guideline concordance ($\chi^2(2)=1.86$, p=0.394), or potential harm ($\chi^2(2)=0.50$, p=0.779).

As the omnibus mDISCERN difference was situated at the conventional significance threshold (p=0.050), this result should be regarded as marginal and interpreted together with the corroborating post-hoc contrast (A > B, p=0.037) rather than as a strong omnibus effect.

With respect to safety, all three models displayed a favourable profile, with safety rates of 92% for Model A, 89% for Model B, and 92% for Model C. Critically, no response from any model was rated as clearly harmful (a score of 2) by either rater; the small number of responses assigned a score of 1 were concentrated in the same questions across models, reflecting incomplete rather than dangerous content.

In contrast to the expert-rated dimensions, the automated linguistic metrics discriminated clearly between models. Model A achieved a significantly higher ROUGE-L F1 score (0.08 ± 0.04) than both Model B (0.04 ± 0.01) and Model C (0.04 ± 0.02) ($\chi^2(2)=17.33$, p<0.001; post-hoc A > B, p=0.001; A > C, p=0.001), indicating closer lexical and structural correspondence with the BACPAR reference text; the contrast between Models B and C was not significant (p=0.911). Semantic similarity also differed significantly between models ($\chi^2(2)=10.78$, p=0.005), with Model A (0.93 ± 0.02) and Model C (0.93 ± 0.02) scoring higher than Model B (0.92 ± 0.03) (A > B, p=0.008; C > B, p=0.008), although the absolute magnitude of these differences was small.

The most pronounced between-model difference was observed for readability ($\chi^2(2)=22.33$, p<0.001). Model C produced substantially more readable responses (FKRE 26.02 ± 17.53) than

both Model A (0.48 ± 27.12; p<0.001) and Model B (12.81 ± 19.36; p=0.002), whereas the difference between Models A and B did not survive Bonferroni correction (p=0.103). Nevertheless, all three models generated text falling within the "very difficult" readability band, corresponding to a college-graduate or professional reading level.

Taken together, Model A produced the outputs that were lexically and semantically closest to the BACPAR recommendations, achieved the highest information quality as measured by mDISCERN, and was descriptively superior on all remaining expert-rated dimensions, whereas Model C produced the most linguistically accessible responses.

Correlation Between Evaluation Metrics

Spearman correlation analyses are presented in Table 6. Within each model, the expert-based metrics were strongly to very strongly inter-correlated. Reliability, usefulness, guideline concordance and mDISCERN were positively associated with one another in all three models (r=0.86-1.00; all FDR-corrected p<0.001), and each was inversely related to potential harm (Model A, r=-0.71 to -0.72; Model B, r=-0.60 to -0.76; Model C, r=-0.59 to -0.63; all FDR-corrected p<0.05). Responses judged to be of higher informational quality and more guideline-consistent were therefore also judged to carry less potential for patient harm. The near-perfect association observed between reliability and usefulness (r=1.00 in Models A and C; r=0.97 in Model B) indicates that the expert panel did not empirically discriminate between these two constructs in the present sample; these two dimensions should consequently be regarded as largely redundant here.

Readability was inversely related to clinical quality. In Model B, FKRE correlated negatively with reliability (r=-0.62, FDR p=0.020) and guideline concordance (r=-0.58, FDR p=0.029), and positively with potential harm (r=0.57, FDR p=0.035); in Model C, FKRE correlated

negatively with mDISCERN (r=-0.60, FDR p=0.025). The same direction of association was observed in Model A but did not survive FDR correction (r=-0.49 to -0.51; FDR p=0.063-0.084). This pattern was confirmed in the pooled sample (n=54, Table 7), in which FKRE was moderately and inversely correlated with reliability (r=-0.51), usefulness (r=-0.49), mDISCERN (r=-0.49) and guideline concordance (r=-0.46) (all FDR-corrected $p \leq 0.001$), and positively correlated with potential harm (r=0.29, FDR p=0.038). More readable responses were thus, on average, lower clinical quality and of marginally greater potential harm.

In the pooled sample, both automated similarity metrics were weakly to moderately associated with the expert-based dimensions. ROUGE-L F1 correlated positively with reliability (r=0.45), usefulness (r=0.43), guideline concordance (r=0.35) and mDISCERN (r=0.32) (all FDR-corrected $p<0.05$), and semantic similarity correlated positively with reliability (r=0.35), usefulness (r=0.35), guideline concordance (r=0.32) and mDISCERN (r=0.33), and negatively with potential harm (r=-0.42) (all FDR-corrected $p<0.05$). The two automated similarity measures were themselves only weakly correlated with one another (r=0.35, FDR p=0.014). Within individual models, however, none of the correlations between the automated metrics and the expert-rated dimensions retained significance after FDR correction, reflecting the reduced statistical power available at n=18 per model; these within-model associations are therefore reported as exploratory.

Discussion

Large language models have become tools that both clinicians and patients are increasingly turning to for health information. In the fields of orthopedics and sports medicine as well, the use of LLMs to support the development of rehabilitation recommendations is on the rise (22). However, the extent to which the responses provided by these models align with evidence-based guidelines is not yet well understood, particularly in specialized fields such as amputation

and prosthetic rehabilitation. In studies in literature, responses provided by LLMs on rehabilitation-related topics-such as exercise in patients with ankylosing spondylitis, physical activity in children with cystic fibrosis, postoperative care following total hip arthroplasty, the management of sports injuries, and exercise programs for musculoskeletal disorders-have been analyzed (23-27). However, none of these have addressed amputation or prosthetic rehabilitation. In the field of amputation, Kumar and Pratihar recently developed an LLM-supported clinical decision support system for selecting prosthetic components in patients with lower extremity amputations, but the recommendations in this system were generated by specialized software, and the LLM was used solely to explain them, and the responses of general-purpose LLMs were not evaluated against clinical guidelines (28). This study is one of the first to evaluate three large language models (ChatGPT 5.2, Gemini 3, and DeepSeek-V3.2), for their concordance with the BACPAR evidence-based guidelines on the physiotherapy management of adults with lower limb prostheses.

ChatGPT 5.2 achieved significantly higher scores on the ROUGE-L F1 metric than the other two models. This indicates that ChatGPT 5.2's responses are closer to BACPAR recommendations at the word level. However, the scores were actually low across all models. The primary reason for this may be the comparison of short, single-sentence guideline recommendations with the models' lengthy responses. In particular, in previous studies that applied ROUGE-L to medical texts, the length of the generated output was limited to be close to that of the reference text (14, 15); in contrast, this study did not impose any length constraints on the models' responses. The fact that there was no difference in expert scores regarding guideline concordance also suggests that word-level similarity does not, on its own, reflect clinical content alignment. This finding underscores the necessity of human evaluation, in line with Tang et al., who demonstrated that automated metrics generally do not show a strong

correlation with summary quality and are inadequate for identifying factual inconsistencies or potential medical harm (14).

In terms of semantic similarity, ChatGPT 5.2 and DeepSeek-V3.2 achieved significantly higher scores than Gemini 3. However, the differences between the models were minimal. It is noteworthy that while semantic similarity was high, ROUGE-L scores were low. This finding suggests that the models largely reproduced the content of the guideline but expressed it in different wording rather than in the exact words of the guideline. Tang and colleagues also reported that summaries generated by LLMs were able to capture key information but were written using phrasing different from that of the reference text (14).

The most notable difference among the models was in readability, and DeepSeek-V3.2 produced responses with a significantly higher level of readability than the other two models (29, 30). Sari and colleagues, however, found no difference between the two models (26). It was assumed that the main reason for this was that the models were of different versions. However, the responses from all three models were at the “very difficult” readability level. Compared to other studies, the readability level was found to be lower in the present study (23-25). Other studies have targeted patient groups that do not exhibit significant changes in body biomechanics and generally seek information about exercise programs. The current study, however, included technical terms explaining walking biomechanics and prosthetic components. Because the responses were more complex, the readability level is considered to be low.

In terms of information quality, ChatGPT 5.2 performed at a significantly — though marginally — higher level than Gemini 3 (omnibus $p=0.050$); unlike reliability, where no pairwise contrast survived correction, this advantage was confirmed in post-hoc testing ($p=0.037$). This finding appears to be consistent with the results reported by Skaik et al., who noted that ChatGPT

models provided higher-quality responses than Gemini in cases of shoulder instability (31). Gemini scored higher than ChatGPT in hip arthroplasty (24), while DeepSeek-V3 scored higher than ChatGPT in ankylosing spondylitis (26). These conflicting results indicate that model performance can vary depending on the clinical domain and the model version.

An overall difference in reliability was observed among the models; however, this difference was not confirmed in pairwise comparisons. Therefore, it is more appropriate to interpret ChatGPT 5.2's higher score as a trend rather than as definitive superiority. Previous studies using the same scale found ChatGPT's responses regarding rheumatic diseases to be of moderate-to-high reliability (11), while in ankylosing spondylitis, DeepSeek-V3 was reported to provide more reliable responses than ChatGPT-4 (26). The fact that all three models demonstrated high reliability in our study may reflect the improved performance of newer model versions or the fact that the guideline-derived questions had clear and well-defined answers.

No significant difference was found among the models in terms of usefulness, and the responses from all three models were found to be highly useful from a clinical perspective. Uz and Umay, who used the same scale, evaluated ChatGPT’s responses regarding rheumatic diseases as useful (11). Sari and colleagues, however, reported that DeepSeek-V3 provided more useful responses than ChatGPT-4 in ankylosing spondylitis (26). The lack of observed differences among the models in the current study may be explained by the ceiling effect in the expert ratings. Furthermore, the fact that the usefulness scores almost completely overlapped with the reliability scores suggests that the evaluators did not distinguish between these two dimensions in practice.

No significant difference was found among the models in terms of guideline concordance, and the responses from all three models were largely consistent with the BACPAR

recommendations. This result represents a more positive outcome than what is reported in the literature, which indicates that LLM recommendations are mostly only partially consistent with guidelines and that key elements are frequently omitted (22). This difference may stem from the fact that our questions were derived from principle-level recommendations with a single, clear answer, rather than detailed treatment plans.

No differences were found among the models in terms of potential harm, and no response was explicitly deemed as clearly harmful. This finding is consistent with the results reported by Gültekin et al., who noted that ChatGPT-4o and DeepSeek R1 received high safety scores for total knee arthroplasty and that no factual errors were detected (29). However, it is known that LLMs can present misinformation in a persuasive manner (14). Therefore, recommendations should be verified by an expert with regard to individual patient characteristics (22).

In all three models, the expert-based criteria were strongly correlated with one another and were all found to be inversely correlated with potential harm. This suggests that reliable and guideline-concordant responses are also safer. Readability, however, was inversely correlated with quality; that is, as responses became easier to read, their quality decreased. In Gemini 3, responses that were easier to read were less reliable and less concordant with guidelines. In DeepSeek-V3.2, on the other hand, the information quality of these responses was lower. A similar relationship was observed in ChatGPT 5.2, though it was not statistically significant.

A similar finding has been reported in previous studies: while ChatGPT's responses were of higher quality but harder to read, Gemini and DeepSeek provided responses that were easier to read (29, 31). However, these studies only demonstrated this difference between models. Our findings reveal that the same phenomenon is also observed within each model's own responses. In other words, even a single model may compromise on quality when simplifying its response. Automatic similarity metrics, however, showed no significant correlation with expert scores in

any of the models. This suggests that word or semantic similarity alone does not reflect clinical quality (14).

When the responses from the three models were evaluated together, the inverse relationship between readability and clinical quality became more pronounced. Responses that were easier to read received lower scores on all expert criteria and had a slightly higher potential for harm. Automatic metrics that were not found to be significant on a model-by-model basis were found to be weakly to moderately correlated with expert scores when evaluated collectively. This result is partially consistent with the finding by Gulden et al., who reported that ROUGE-L is correlated with human evaluation (15). However, the weak nature of this correlation indicates that these metrics cannot replace expert evaluation (14). Furthermore, the fact that responses more semantically similar to the guideline have a lower potential for harm suggests that proximity to the guideline's content is also important from a safety perspective.

The inter-rater reliability between the two expert raters was excellent across all criteria. These values are considerably higher than the levels of agreement reported in similar studies (23-25). This high level of agreement may be explained by the fact that the rating criteria were based on predefined expected elements for each question. However, the concentration of most scores at the upper end of the scale may have artificially inflated agreement. Nevertheless, the high Gwet's AC1 coefficient, which is more robust to ceiling effects, supports that the agreement was largely genuine. Agreement was relatively lower only for mDISCERN, which may be due to one of the raters using half-point scores on this scale.

This study has some limitations. The number of questions was limited, which reduced the statistical power of the model-based analyses. Each question was asked of the models only once. Given that LLM responses can vary depending on the wording of the question and over time, repeated queries could have strengthened the findings (22, 27). The questions were asked

only in English, and the evaluation was limited to two experts. Furthermore, the concentration of scores at the upper end of the scale made it difficult to identify differences between the models. Finally, the fact that the questions were derived from single, clear guideline recommendations does not reflect complex clinical situations that require individual decision-making.

Conclusions

In this study, ChatGPT 5.2, Gemini 3, and DeepSeek-V3.2 provided responses to questions derived from the BACPAR guidelines that were largely consistent with the guidelines, and none of the responses were found to be clearly harmful. While ChatGPT 5.2 stood out in terms of information quality and lexical similarity to the guideline text, DeepSeek-V3.2 produced the most readable responses. However, the responses from all models required advanced reading skills, and responses that were easier to read tended to have lower clinical quality. Automatic similarity metrics, meanwhile, showed only a limited correlation with expert evaluation. These findings suggest that LLMs can be used as supportive tools to facilitate access to information in amputation and prosthetic rehabilitation; however, they should not be used without expert oversight, particularly in patient education. Future studies should evaluate the models' performance in complex clinical cases, the consistency of their responses to repeated queries, and their ability to generate responses in language appropriate for the patient.

List of abbreviations

AC1: Gwet's first-order agreement coefficient

BACPAR: British Association of Chartered Physiotherapists in Amputee Rehabilitation

CI: Confidence interval

FDR: False discovery rate

FKRE: Flesch-Kincaid Reading Ease

GloVe: Global Vectors for Word Representation

ICC: Intraclass correlation coefficient

LCS: Longest common subsequence

LLM: Large language model

mDISCERN: Modified DISCERN

ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation – Longest common subsequence

SD: Standard deviation

Declarations

Ethics approval and consent to participate

The study was approved by the Health Sciences University Gülhane Ethics Committee for Non-Drug and Non-Device Interventional Studies (Decision No. 2026-184, 21 May 2026) and conducted in accordance with the Declaration of Helsinki. As no human participants or patient data were involved, informed consent was not required.

Clinical trial number

Clinical trial number: not applicable.

Consent for publication

Not applicable.

Availability of data and materials

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

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.

Authors' contributions

YE: conceptualization, methodology, data collection, analysis, writing – original draft. MGB: expert evaluation, writing – review and editing. All authors read and approved the final manuscript.

Acknowledgements

During the preparation of this work, the authors used ChatGPT, Gemini, and Claude (Anthropic) to assist with improving language quality, refining sentence structure, translating certain sections into English, and drafting and revising portions of the manuscript text. After using these tools, the authors reviewed, verified, and edited all content as needed and take full responsibility for the content of the published article.

References

1. Pohjolainen T, Alaranta H, Wikström J. Primary survival and prosthetic fitting of lower limb amputees. Prosthet Orthot Int. 1989;13(2):63-9.

2. Esquenazi A, DiGiacomo R. Rehabilitation after amputation. J Am Podiatr Med Assoc. 2001;91(1):13-22.

3. Webster JB, Crunkhorn A, Sall J, Highsmith MJ, Pruziner A, Randolph BJ. Clinical practice guidelines for the rehabilitation of lower limb amputation: an update from the Department of Veterans Affairs and Department of Defense. Am J Phys Med Rehabil. 2019;98(9):820-9.

4. Demir Y, Aydemir K. Gülhane lower extremity amputee rehabilitation protocol: a nationwide, 123-year experience. Turk J Phys Med Rehabil. 2020;66(4):373-82.

5. British Association of Chartered Physiotherapists in Amputee Rehabilitation (BACPAR). Evidence based clinical guidelines for the physiotherapy management of adults with lower limb prostheses, 3rd ed. 2020. https://www.bacpar.org/resources/healthcare-professionals-resources/prosthetic-guidelines/. Accessed 13 Oct 2026.

6. Usen A, Kuculmez O. Evaluation of the performance of large language models in the management of axial spondyloarthropathy: analysis of EULAR 2022 recommendations. Diagnostics (Basel). 2025;15(12):1455.

7. Çelik NÇ, Kılınç EA. Assessment of ChatGPT's adherence to EULAR diagnostic criteria and therapeutic protocols for rheumatoid arthritis at two distinct time points, 14 days apart, utilizing binary and multiple-choice inquiries. Clin Rheumatol . 2025;44(6):2233-9.

8. Labinsky H, Nagler L-K, Krusche M, Griewing S, Aries P, Kroiß A, et al. Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: results of the Rheum2Guide study. Rheumatol Int. 2024;44(10):2043-53.

9. Lu S, Nietsch K, Duey A, Zaidat B, Ndjonko LCM, Shrestha N, et al. Management of lower extremity traumas: comparing appropriate use criteria ChatGPT recommendations. Am J Surg. 2025; 242:116229.

10. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc . 2023;30(7):1237-45.

11. Uz C, Umay E. "Dr ChatGPT": is it a reliable and useful source for common rheumatic diseases? Int J Rheum Dis. 2023;26(7):1343-9.

12. Fattoh IE, Kamal Alsheref F, Ead WM, Youssef AM. Semantic sentiment classification for COVID-19 tweets using universal sentence encoder. Comput Intell Neurosci. 2022; 2022:6354543.

13. Izzidien A, Fitz S, Romero P, Loe BS, Stillwell D. Developing a sentence level fairness metric using word embeddings. Int J Digit Humanit . 2023;5(2):95-130.

14. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med. 2023;6(1):158.

15. Gulden C, Kirchner M, Schüttler C, Hinderer M, Kampf M, Prokosch H-U, et al. Extractive summarization of clinical trial descriptions. Int J Med Inform. 2019; 129:114-21.

16. Kılıçoğlu MS, Yurdakul OV, Aydın T. Evaluation of Chat Generative Pre-trained Transformer's responses to frequently asked questions about psoriatic arthritis: a study on quality and readability. Ann Med Res. 2025;32(2):79-84 .

17. Solnyshkina M, Zamaletdinov R, Gorodetskaya L, Gabitov A. Evaluating text complexity and Flesch-Kincaid grade level. J Soc Stud Educ Res . 2017;8(3):238-48.

18. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155-63.

19. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(1):29-48.

20. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74 .

21. Mukaka MM. A guide to appropriate use of correlation coefficient in medical research. Malawi Med J. 2012;24(3):69-71.

22. Gołembiewski B, Mulski G, Manicka M, Baranowicz M, Mazurkiewicz W, Borecka Z, et al. Are large language models capable of generating safe and guideline-concordant rehabilitation recommendations in orthopaedics and sports medicine? A review. J Pre Clin Clin Res. 2026;20(2):65-70.

23. Çelik Z, Sari F. Evaluation of ChatGPT-4 responses on physical activity guidance in children with cystic fibrosis: reliability, quality, and readability. Eur J Pediatr. 2025;184(11):676.

24. Bozgeyik-Bağdatli S, Bazancir-Apaydin Z, Güney-Deniz H. Artificial intelligence in patient education: a comparative evaluation of chatbot reliability after total hip arthroplasty. Am J Phys Med Rehabil. 2026. doi:10.1097/PHM.0000000000003061.

25. Fu Y, Li H, You M, Wang L, Liu W, Zhou K, et al. Evaluation of large language models in generating physical exercise rehabilitation programs for musculoskeletal disorders across multiple clinical scenarios. Healthcare (Basel). 2026;14(15):2389.

26. Sari F, Çelik Z, Mirza Y. ChatGPT-4 vs. DeepSeek-V3: a comparative study of response quality, reliability, usefulness, and readability for exercise and rehabilitation strategies in patients with ankylosing spondylitis. Clin Rheumatol . 2026;45(1):187-95.

27. Kaya O, Huri G, Özbek EA, Gönder N, Demir İH, Dalkır KA. Large language models in sports injury care: a comparative expert evaluation of GPT-4o and GPT-5. BMC Sports Sci Med Rehabil. 2026; 18:232.

28. Kumar V, Pratihar DK. ProsthetiX-AI: an LLM-based clinical decision support system for evidence-based prosthetic recommendations. Health Inf Sci Syst. 2026;14(1):56.

29. Gultekin O, Hirschmann MT, Arıkan Hİ, Kilinc BE, Yilmaz B, Abul S, 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):470-6.

30. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025; 198:105871.

31. Skaik K, Omoseni S, Dagher D, Shah D, Fermín TM, Agostinone P, et al. ChatGPT models provide higher-quality but lower-readability responses than Google Gemini regarding anterior shoulder instability, with no added benefit of the orthopaedic expert plugin. Knee Surg Sports Traumatol Arthrosc. 2026;34(2):763-75.

Tables

Table 1. Mapping of BACPAR guideline domains and recommendations to the study questions.

Section	BACPAR guideline domain	Recommendations mapped	Question no.
1	The Multidisciplinary Team	1.1	Q1
2	Prosthetic Knowledge	2.1, 2.2, 2.9	Q2, Q3, Q4
3	Assessment	3.1, 3.2, 3.5	Q5, Q6, Q7
4	The Prosthetic Rehabilitation Programme	4.1.1, 4.2, 4.5, 4.6, 4.13, 4.22	Q8-Q13
5	Patient Education	5.1.1, 5.2.4, 5.4.2	Q14, Q15, Q16
6	Discharge, Maintenance and Participation	6.1, 6.5	Q17, Q18
Total	6 domains	18 recommendations	18 questions

Source: British Association of Chartered Physiotherapists in Amputee Rehabilitation (2020). Evidence based clinical guidelines for the physiotherapy management of adults with lower limb prostheses, 3rd Edition (5).

Table 2. Open-ended clinical questions based on the BACPAR 2020 recommendations.

No.	Open-ended question	BACPAR rec.	Key elements expected in the reference answer
Q1	Within amputee rehabilitation services, what designated professional role carries overall responsibility for physiotherapy care coordination?	1.1	A physiotherapist specialising in amputee rehabilitation is accountable for managing physiotherapy care
Q2	In amputee prosthetic rehabilitation, what theoretical understanding concerning prosthetic componentry and systemic impact is expected from the physiotherapist?	2.1	Knowledge of prosthetic componentry theory and of the effects of prosthetic rehabilitation on the remaining body systems
Q3	In amputee gait training, what specific domains of knowledge are required to ensure safe and effective re-education?	2.2	Understanding of physiological and prosthetic gait principles and of the physical and biomechanical factors affecting them
Q4	What is the role of the physiotherapist in prosthesis provision?	2.9	Contribution to decisions on prosthetic provision and prescription, informed by musculoskeletal function, cognition and exercise tolerance
Q5	During comprehensive assessment in amputee management, what formal documentation standard applies to physical examination and cardiovascular evaluation?	3.1	Written evidence of full physical examination and of previous and present function, including cardiovascular status
Q6	Within amputee management, what is mandated regarding fall risk during the assessment phase?	3.2	Falls-risk assessment must be documented

Q7	In amputee rehabilitation planning, how are treatment priorities and rehabilitation objectives formally structured?	3.5	A problem list and treatment plan with agreed goals, formulated in partnership with the patient
Q8	Following prosthesis provision in amputee rehabilitation, what timing principle governs the initiation of prosthetic rehabilitation?	4.1.1	Prosthetic rehabilitation should begin as early as possible to optimise clinical outcomes
Q9	In amputee prosthetic gait rehabilitation, what overarching biomechanical objective defines optimal rehabilitation outcomes?	4.2	Establishment of an energy-efficient gait modelled on normal physiological walking patterns
Q10	Within amputee rehabilitation programmes, how must therapeutic exercise interventions be structured in relation to strength, flexibility, and joint mobility?	4.5	A personalised programme combining specific strengthening and stretching with maintenance/improvement of joint mobility
Q11	In amputee functional retraining, how is progression of task complexity formally governed?	4.13	Functional skills of increasing complexity taught within the patient's limits
Q12	What should a physiotherapist do if they detect unusual movement in an amputee who has recently started using a new prosthesis?	4.6	Identification of altered movement strategies; efficient prosthetic control taught via postural control, weight transfer, proprioception, mental practice and exercise
Q13	From whom should I seek help regarding the feelings I experience related to my missing limb?	4.22	The physiotherapist, alongside the multidisciplinary team, contributes to management of phantom sensation/pain
Q14	In amputee management, what essential information must be provided concerning the operational capabilities of the prosthesis?	5.1.1	Patients/carers must be informed about the prosthesis, its functions and its limitations
Q15	Is there anything I need to do for my residual limb while continuing to use the prosthesis?	5.2.4	Information on residual-limb skin care, sweat management, and problems arising from poor hygiene or inadequate/overzealous skin care
Q16	In amputee management, what expectation must be communicated regarding the relationship between amputation level and anticipated functional mobility?	5.4.2	Patients/carers must be made aware that amputation level affects expected function and mobility
Q17	After discharge from routine physiotherapy in amputee rehabilitation, what structured review requirement applies within the first year?	6.1	A process for regular review using outcome measures within 12 months of discharge from regular physiotherapy

Q18	What actions should be taken if conditions change during prosthesis use?	6.5	Access to further assessment and treatment/intervention when the individual's circumstances change

Q12, Q13, Q15 and Q18 were deliberately phrased in lay, patient-facing language; the remaining questions were phrased from a clinician's perspective. The verbatim BACPAR recommendation corresponding to each question was used as the reference ("gold-standard") answer for all automated similarity metrics.

Table 3. Expert-based evaluation instruments and scoring framework used in this study.

Instrument / metric	Construct assessed	Scale	Interpretation (higher score=…)
Reliability (7-point Likert)	Clinical trustworthiness of the response if used alone	1-7	More clinically reliable; fewer critical errors
Usefulness (7-point Likert)	Practical utility for amputee rehabilitation decision-making	1-7	More actionable and clinically helpful
Guideline concordance	Agreement with the expected key elements of the corresponding BACPAR recommendation	0-5	More guideline-consistent content coverage
Potential harm	Presence of potentially harmful advice, ignored contraindications, or unsafe suggestions	0-2	0=no harm; 2=higher potential harm (worse safety)
Modified DISCERN (mDISCERN)	Quality of health information (clarity, balance, evidence support, decision support)	0-7	Higher information quality; better decision support

Each response was scored independently by two expert raters; the mean of the two scores was used in all analyses. Reliability and usefulness scales were adapted from Uz and Umay (11). Inter-rater agreement is reported in the footnote to Table 4 and in Table 5.

Table 4. Comparison of Model A, Model B, and Model C scores across evaluation outcomes.

Score	Model A Mean ± SD [%]	Model B Mean ± SD [%]	Model C Mean ± SD [%]	$\chi^2(2)$	p	Significant post-hoc contrasts
ROUGE-L F1	0.08 ± 0.04	0.04 ± 0.01	0.04 ± 0.02	17.33	<0.001	A > B (0.001); A > C (0.001)
Semantic similarity	0.93 ± 0.02	0.92 ± 0.03	0.93 ± 0.02	10.78	0.005	A > B (0.008); C > B (0.008)
FKRE	0.48 ± 27.12	12.81 ± 19.36	26.02 ± 17.53	22.33	<0.001	C > A (<0.001); C > B (0.002)
mDISCERN	6.50 ± 0.66 [93%]	6.10 ± 0.86 [87%]	6.19 ± 0.98 [88%]	6.00	0.050	A > B (0.037)
Reliability	6.33 ± 0.89 [90%]	5.78 ± 1.25 [83%]	6.06 ± 1.10 [87%]	6.08	0.048	Omnibus sig.; no pairwise contrast survived Bonferroni

Usefulness	6.33 ± 0.89 [90%]	5.83 ± 1.28 [83%]	6.06 ± 1.10 [87%]	4.96	0.084	-
Guideline concordance	4.44 ± 0.68 [89%]	4.11 ± 1.01 [82%]	4.17 ± 1.00 [83%]	1.86	0.394	-
Potential harm	0.17 ± 0.38 [92%]	0.22 ± 0.43 [89%]	0.17 ± 0.38 [92%]	0.50	0.779	-

Model A: ChatGPT 5.2; Model B: Gemini 3; and Model C: DeepSeek-V3.2.
Values in the "p" column represent Friedman test results; post-hoc comparisons used the Wilcoxon signed-rank test with Bonferroni correction (p-values in the final column are Bonferroni-corrected). Success rate (%)=(mean / maximum) × 100 (max 7 for reliability, usefulness, mDISCERN; 5 for concordance). For potential harm, safety rate=[1 - (mean/2)] × 100. Expert-based scores are the mean of two independent raters; inter-rater agreement was excellent (ICC(2,1)=0.928-1.000; within-1-point agreement=100%).

Table 5. Inter-rater reliability between the two expert raters (n=54 responses).

Instrument	ICC(2,1)	95% CI	Gwet's AC1	Exact agree.	Within-1-pt agree.
Reliability	0.955	0.92-0.97	0.866	88.9%	100%
Usefulness	0.955	0.92-0.97	0.867	88.9%	100%
Guideline concordance	0.935	0.88-0.96	0.859	88.9%	100%
Potential harm	1.000	1.00-1.00	1.000	100%	100%
mDISCERN	0.928	0.88-0.96	0.677	74.1%	100%

ICC: intraclass correlation coefficient (two-way random effects, absolute agreement, single rater); AC1: Gwet's first-order agreement coefficient. For AC1 and exact agreement, mDISCERN scores were rounded to the nearest integer, as one rater applied half-point increments.

Table 6. Spearman's correlation coefficients (r) and FDR-corrected p-values between evaluation metrics for each large language model.

Metric	Model		Sem. sim.	FKRE	Reliab.	Useful.	Concord.	Harm	mDISCERN
ROUGE-L F1	Model A	r	0.36	-0.00	0.38	0.38	0.41	-0.16	0.34
		p	0.199	0.990	0.175	0.175	0.150	0.541	0.223
	Model B	r	0.27	-0.32	0.41	0.39	0.35	-0.40	0.16
		p	0.338	0.252	0.150	0.164	0.220	0.156	0.541
	Model C	r	0.18	-0.21	0.46	0.46	0.19	-0.33	0.32
		p	0.509	0.448	0.100	0.100	0.493	0.241	0.252
Semantic sim.	Model A	r	-	-0.16	0.43	0.43	0.40	-0.50	0.39
		p	-	0.541	0.129	0.129	0.153	0.070	0.164

	Model B	r	-	-0.40	0.32	0.34	0.31	-0.52	0.26
		p	-	0.153	0.252	0.223	0.263	0.063	0.345
	Model C	r	-	-0.32	0.20	0.20	0.20	-0.30	0.25
		p	-	0.252	0.476	0.476	0.474	0.272	0.378
FKRE	Model A	r	-0.16	-	-0.51	-0.51	-0.49	0.22	-0.51
		p	0.541	-	0.063	0.063	0.084	0.448	0.063
	Model B	r	-0.40	-	-0.62*	-0.54	-0.58*	0.57*	-0.48
		p	0.153	-	0.020	0.050	0.029	0.035	0.091
	Model C	r	-0.32	-	-0.43	-0.43	-0.41	0.16	-0.60*
		p	0.252	-	0.129	0.129	0.150	0.541	0.025
Reliability	Model A	r	0.43	-0.51	-	1.00*	0.99*	-0.72*	0.99*
		p	0.129	0.063	-	<0.001	<0.001	0.003	<0.001
	Model B	r	0.32	-0.62*	-	0.97*	0.94*	-0.75*	0.86*
		p	0.252	0.020	-	<0.001	<0.001	0.001	<0.001
	Model C	r	0.20	-0.43	-	1.00*	0.87*	-0.59*	0.92*
		p	0.476	0.129	-	<0.001	<0.001	0.026	<0.001
Usefulness	Model A	r	0.43	-0.51	1.00*	-	0.99*	-0.72*	0.99*
		p	0.129	0.063	<0.001	-	<0.001	0.003	<0.001
	Model B	r	0.34	-0.54	0.97*	-	0.97*	-0.76*	0.90*
		p	0.223	0.050	<0.001	-	<0.001	0.001	<0.001
	Model C	r	0.20	-0.43	1.00*	-	0.87*	-0.59*	0.92*
		p	0.476	0.129	<0.001	-	<0.001	0.026	<0.001
Guideline concord.	Model A	r	0.40	-0.49	0.99*	0.99*	-	-0.72*	0.98*
		p	0.153	0.084	<0.001	<0.001	-	0.003	<0.001
	Model B	r	0.31	-0.58*	0.94*	0.97*	-	-0.71*	0.88*

		p	0.263	0.029	<0.001	<0.001	-	0.003	<0.001
	Model C	r	0.20	-0.41	0.87*	0.87*	-	-0.61*	0.92*
		p	0.474	0.150	<0.001	<0.001	-	0.022	<0.001
Potential harm	Model A	r	-0.50	0.22	-0.72*	-0.72*	-0.72*	-	-0.71*
		p	0.070	0.448	0.003	0.003	0.003	-	0.003
	Model B	r	-0.52	0.57*	-0.75*	-0.76*	-0.71*	-	-0.60*
		p	0.063	0.035	0.001	0.001	0.003	-	0.025
	Model C	r	-0.30	0.16	-0.59*	-0.59*	-0.61*	-	-0.63*
		p	0.272	0.541	0.026	0.026	0.022	-	0.016

*Model A: ChatGPT 5.2; Model B: Gemini 3; and Model C: DeepSeek-V3.2. All p-values are corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate procedure. * FDR-corrected p<0.05. Expert-based scores represent the mean of the two independent raters. Sem. sim.=semantic similarity; FKRE=Flesch-Kincaid Reading Ease; mDISCERN=modified DISCERN.*

Table 7. Pooled-sample Spearman correlations (n=54).

Metric pair	r	FDR-corrected p
ROUGE-L F1 - Semantic similarity	0.35	0.014
ROUGE-L F1 - FKRE	-0.33	0.020
ROUGE-L F1 - Reliability	0.45	0.001
ROUGE-L F1 - Usefulness	0.43	0.002
ROUGE-L F1 - Guideline concordance	0.35	0.014
ROUGE-L F1 - Potential harm	-0.26	0.065
ROUGE-L F1 - mDISCERN	0.32	0.022
Semantic similarity - FKRE	-0.22	0.107
Semantic similarity - Reliability	0.35	0.014
Semantic similarity - Usefulness	0.35	0.014
Semantic similarity - Guideline concordance	0.32	0.022
Semantic similarity - Potential harm	-0.42	0.002
Semantic similarity - mDISCERN	0.33	0.018
FKRE - Reliability	-0.51	<0.001
FKRE - Usefulness	-0.49	<0.001
FKRE - Guideline concordance	-0.46	0.001
FKRE - Potential harm	0.29	0.038
FKRE - mDISCERN	-0.49	<0.001
Reliability - Usefulness	0.98	<0.001
Reliability - Guideline concordance	0.93	<0.001
Reliability - Potential harm	-0.68	<0.001
Reliability - mDISCERN	0.92	<0.001

Usefulness - Guideline concordance	0.95	<0.001
Usefulness - Potential harm	-0.68	<0.001
Usefulness - mDISCERN	0.93	<0.001
Guideline concordance - Potential harm	-0.68	<0.001
Guideline concordance - mDISCERN	0.92	<0.001
Potential harm - mDISCERN	-0.63	<0.001