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 GPT-5.6, Gemini 3.6 Flash, DeepSeek-V4, and Doubao-2.1 With the Formal EU-US Meniscus Rehabilitation 2024 Consensus: A Comparative Study at the Question and Statement Levels

Xiaohe Wang
University of Science and Technology of China
Dongjing Fang
University of Science and Technology of China
Liang Xu
University of Science and Technology of China
Yingming Wang
University of Science and Technology of China
Qichun Zhao
qichunzhao@ustc.edu.cn

University of Science and Technology of China

Research Article

Keywords: artificial intelligence, large language models, meniscus, rehabilitation, clinical consensus, sports medicine

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11193068/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models (LLMs) are increasingly used for clinical information retrieval in orthopaedics and sports medicine, but their value depends on fidelity to the relevant guideline or consensus. This study compared four LLMs with the formal 2024 ESSKA-AOSSM-AASPT EU-US Meniscus Rehabilitation Consensus in a realistic setting in which the consensus was named but its source text was not provided.

Methods

This consensus-referenced cross-sectional comparative study evaluated GPT-5.6, Gemini 3.6 Flash, DeepSeek-V4, and Doubao-2.1 using 37 clinical questions and 52 disaggregated recommendation statements. All models received identical instructions to answer strictly according to the consensus, without access to its PDF, full text, recommendation list, reference answers, or scoring criteria. Outcomes included question-level accuracy, strict statement-level concordance, at-least-partial coverage, error types, and potentially hazardous omissions. Paired differences were analyzed using Cochran's Q test and exact McNemar tests with Holm correction.

Results

Question-level accuracy differed among models (Cochran's $Q = 9.09$, $p = 0.028$): DeepSeek-V4 achieved 25/37 (67.6%), GPT-5.6 24/37 (64.9%), Gemini 3.6 Flash 23/37 (62.2%), and Doubao-2.1 16/37 (43.2%); no pairwise comparison remained significant after Holm correction. Strict statement-level concordance differed more markedly (Cochran's $Q = 27.53$, $p < 0.001$): GPT-5.6 achieved 25/52 (48.1%), Gemini 3.6 Flash 22/52 (42.3%), DeepSeek-V4 10/52 (19.2%), and Doubao-2.1 7/52 (13.5%). Errors primarily involved weight-bearing and range-of-motion restrictions, procedure-specific rehabilitation, return-to-sport criteria, outcome measures, and referral criteria.

Conclusions

LLM performance was heterogeneous when the consensus was named but not supplied. Question-level correctness did not ensure complete coverage of procedure-specific restrictions. LLMs may provide educational support, but individualized rehabilitation decisions require verification against the original consensus and clinical instructions.

1. Background

Artificial intelligence, particularly large language models (LLMs), is increasingly used in medical education, patient counseling, information retrieval, and clinical decision support [1–4]. However, clinical question answering requires more than fluent language generation. Responses must reflect the correct reference standard, the specific procedural context, the level of evidence, contraindications, and key safety-related constraints. Previous evaluations in medicine and surgery have shown heterogeneous LLM performance, with common problems including insufficient information, overstatement, and recommendations that conflict with clinical guidelines [5–10]. Guideline- or consensus-referenced evaluations are therefore needed to assess overall accuracy, completeness, and potential clinical risk.

Studies in orthopaedics and sports medicine have similarly shown that LLM responses may be readable and directionally helpful but remain inconsistent in completeness, guideline specificity, and procedural detail [7, 8]. This limitation is particularly important because recommendations in musculoskeletal care often depend on the specific intervention, biological constraints on tissue healing, and the stage of rehabilitation. A common real-world scenario is that users know the

name of a consensus document but have not uploaded the original article or recommendation list. The present study was designed to test this specific scenario.

Meniscal injuries are common in sports medicine, and treatment philosophy has shifted from routine resection toward meniscal preservation, repair, and individualized rehabilitation [11–14]. Degenerative meniscal lesions are generally managed initially with nonoperative treatment, whereas decisions for traumatic tears depend on tear pattern, symptoms, patient needs, and concomitant injuries [12, 15–19]. Rehabilitation recommendations differ substantially after partial meniscectomy, meniscal repair, and meniscal reconstruction. Incorrect or incomplete guidance may lead to premature weight-bearing, excessive knee flexion, rotational loading, or an overly early return to sport, particularly after repair or reconstruction [20, 21].

The formal EU-US Meniscus Rehabilitation 2024 Consensus developed by ESSKA, AOSSM, and AASPT provides detailed guidance on rehabilitation after meniscectomy, meniscal repair, and meniscal reconstruction, as well as prevention, nonoperative treatment, return to sport, patient-reported outcome measures (PROMs), performance-based measures (PBMs), and referral criteria [20, 21, 22]. Using this consensus as the reference standard, the present study compared GPT-5.6, Gemini 3.6 Flash, DeepSeek-V4, and Doubao-2.1 under an explicit source-document-not-provided prompt. By combining question-level accuracy, statement-level concordance, module-stratified analyses, and an assessment of safety-related errors, the study aimed to evaluate both the clinical usefulness and the potential risks of LLM-generated responses.

2. Methods

Study Design

This was a consensus-referenced cross-sectional comparative study of text responses generated by four LLMs to questions concerning meniscus rehabilitation. The study did not involve patients, human specimens, or animal experiments. The study process followed principles of transparency, reproducibility, and data traceability in medical research.

The methodology was informed by previous studies comparing LLMs with clinical guidelines or consensus documents: investigators extracted clinical questions from an authoritative reference document, submitted identical questions to different AI models, recorded responses verbatim, and submitted deidentified responses to independent reviewers for blinded assessment [5].

Reference Standard and Evaluation Dataset

The reference standard was the formal EU-US Meniscus Rehabilitation 2024 Consensus developed by ESSKA, AOSSM, and AASPT, including Parts I and II [20, 21]. Part I primarily addressed rehabilitation after meniscectomy, meniscal repair, and meniscal reconstruction and yielded 19 questions and 29 statements. Part II primarily addressed prevention of meniscal injuries, nonoperative treatment, return to sport, PROMs/PBMs, and referral criteria and yielded 18 questions and 23 statements.

The investigators extracted 37 clinical questions from the consensus for the primary question-level analysis. These questions were further disaggregated into 52 recommendation statements for the statement-level analysis. The 37 questions and 52 statements were assigned to eight rehabilitation modules according to consensus topics: rehabilitation after arthroscopic partial meniscectomy; rehabilitation after meniscal repair; rehabilitation after meniscal reconstruction, transplantation, or scaffold implantation; prevention of meniscal injuries; nonoperative rehabilitation of acute traumatic meniscal tears; nonoperative rehabilitation of degenerative meniscal lesions; return to sport; and PROMs/PBMs and referral criteria (Supplementary 1).

AI Models and Prompting Procedure

Four LLMs were evaluated: GPT-5.6, Gemini 3.6 Flash, DeepSeek-V4, and Doubao-2.1. The models were accessed between August 3 and August 15, 2026. GPT-5.6 was accessed through ChatGPT [23]; Gemini 3.6 Flash through Gemini [24];

DeepSeek-V4 through DeepSeek Chat [25]; and Doubao-2.1 through Doubao [26].

All models were used in independent, single-turn sessions. To minimize carryover of context, each of the 37 clinical questions was submitted in a new conversation. No prior chat history was used, and no additional clarification was provided after the initial prompt. The 2024 ESSKA-AOSSM-AASPT consensus text, reference answers, scoring criteria, and statement-level reference list were not uploaded or otherwise provided to any model.

The standardized prompt prefix was: “Please answer strictly according to the 2024 ESSKA-AOSSM-AASPT Formal EU-US Meniscus Rehabilitation Consensus (Part I: rehabilitation after meniscus surgery; Part II: prevention, non-operative treatment and return to sport). No consensus PDF or source text is attached. Do not use other guidelines unless clearly marked as outside this consensus. Give a concise, clinically applicable answer and explicitly mention uncertainty or evidence grade when appropriate.” Each clinical question was appended verbatim after this prefix.

All models were used with their default generation settings. For web-based interfaces, temperature, top-p, maximum token length, random seed, and other sampling parameters could not be configured by users and were therefore unavailable to or uncontrolled by the investigators. The investigators did not enable any additional deep-research, browsing, or manually activated tool-use functions. No file, URL, PDF, source-text excerpt, or recommendation statement was uploaded during prompting.

Each question was submitted once to each model. Responses were copied immediately after generation and preserved verbatim, without manual correction, regeneration, summarization, or postprocessing (Supplementary table 2). Model names, access routes, question identifiers, and complete responses were recorded in a structured spreadsheet. The investigators then replaced model names with anonymized labels A, B, C, and D, randomized response order, and submitted the deidentified responses to reviewers for assessment.

Scoring Method

Two reviewers independently evaluated the AI responses, and disagreements were adjudicated by a third senior expert. In the primary question-level analysis, each response was classified as accurate or inaccurate. An accurate response was defined as one that included the principal consensus recommendation, did not contradict the consensus, and appropriately acknowledged uncertainty or evidence grade when evidence was limited or individualized judgment was required. An inaccurate response was defined as one that omitted a key consensus point, particularly a Grade A or B recommendation; provided advice contrary to the consensus; presented low-level evidence or expert opinion as a strong recommendation; or included information that could lead to an incorrect rehabilitation decision.

Inaccurate responses were further tagged as containing insufficient information and/or overstatement. Insufficient information was defined as a response that was directionally correct but failed to cover a key rehabilitation restriction, range-of-motion (ROM) limitation, referral criterion, return-to-sport criterion, or individualized condition. Overstatement was defined as an absolute, inappropriate, or consensus-inconsistent recommendation in a setting where evidence was limited, no clear recommendation existed, or individualized judgment was required.

For the statement-level analysis, reviewers assessed whether each model correctly covered each of the 52 consensus recommendation statements. Each model–statement pair was classified as correct (C), partial (P), missing (M), or incorrect (I). Correct coverage required all essential elements of the reference statement to be represented without material omission or contradiction. Partial coverage indicated incomplete coverage without a material contradiction. Missing coverage indicated that the recommendation was not addressed. Incorrect coverage indicated a material contradiction or an erroneous recommendation. Where the reference statement specified alternatives, inclusion of one eligible alternative was sufficient for that component. Strict concordance included C only, whereas at-least-partial coverage included C and P. The overall study design is shown in Fig. 1.

Potentially Hazardous Errors

The following potentially hazardous errors were also recorded separately:

- Recommending immediate full weight-bearing after meniscal root repair;
- Recommending early deep squatting, jumping, or rotational weight-bearing after repair of a radial or root tear;
- Determining return to sport solely on the basis of time after meniscal repair while disregarding ROM, effusion, strength, and neuromuscular control;
- Routinely recommending arthroscopic surgery for degenerative meniscal lesions without an initial trial of nonoperative treatment;
- Failing to recommend referral or omitting key referral criteria in the presence of persistent pain, recurrent stiffness, effusion, mechanical symptoms, neurological symptoms, suspected infection, or suspected deep vein thrombosis; and
- Confusing rehabilitation protocols after meniscectomy, meniscal repair, and meniscal reconstruction.

Statistical Analysis

Categorical variables are presented as frequencies and percentages. Exact binomial 95% confidence intervals (CIs) were calculated for primary question-level accuracy, strict statement-level concordance, and at-least-partial coverage. Question-level accuracy and strict statement-level concordance were compared among the four models using Cochran's Q test. When the overall test was statistically significant, pairwise comparisons were performed using the exact McNemar test with Holm-Bonferroni correction for multiple comparisons.

Stratified module-level analyses were performed to evaluate accuracy across rehabilitation topics. Because several modules contained only 2 to 13 questions or statements, module-level analyses were considered exploratory, and their p values were reported descriptively rather than used for confirmatory inference. Because insufficient information and overstatement could coexist within the same inaccurate response, error types were summarized descriptively. All tests were two-sided, and $p < 0.05$ was considered statistically significant in the primary question-level and statement-level analyses.

3. Results

Primary Question-Level Analysis

All four models returned responses to all 37 clinical questions. The primary question-level analysis demonstrated a statistically significant difference in overall accuracy among the four LLMs (Cochran's Q = 9.09, df = 3, p = 0.028). DeepSeek-V4 produced the greatest number of accurate responses (25/37, 67.6%; 95% CI, 50.2%–82.0%), followed by GPT-5.6 (24/37, 64.9%; 95% CI, 47.5%–79.8%), Gemini 3.6 Flash (23/37, 62.2%; 95% CI, 44.8%–77.5%), and Doubao-2.1 (16/37, 43.2%; 95% CI, 27.1%–60.5%) (Supplementary Table 3).

Although the overall test indicated differences among models, no pairwise exact McNemar comparison remained statistically significant after Holm-Bonferroni correction. In the unadjusted analysis, the largest difference was between DeepSeek-V4 and Doubao-2.1: DeepSeek-V4 was accurate and Doubao-2.1 inaccurate for 12 questions, whereas the opposite pattern occurred for 3 questions (unadjusted p = 0.035; Holm-adjusted p = 0.211) (Table 2).

Question-level results indicated that models generally recognized the overall direction of consensus recommendations. However, this binary question-level metric could not adequately capture missing procedure-specific details, omitted restrictions, or errors in evidence grading, necessitating further statement-level analysis.

Table 1
Primary question-level analysis: overall accuracy and error types across the four models

Model	GPT-5.6	Gemini 3.6 Flash	DeepSeek-V4	Doubao-2.1
Accurate	24	23	25	16
Total	37	37	37	37
Accuracy, %	64.9	62.2	67.6	43.2
Exact 95% CI, %	47.5–79.8	44.8–77.5	50.2–82.0	27.1–60.5
Inaccurate	13	14	12	21
Insufficient information	11	8	8	13
Overstatement	5	9	6	16
Potentially hazardous errors or major safety omission	3	4	3	4

Note: CIs are exact binomial 95% confidence intervals. Insufficient information and overstatement could coexist within the same inaccurate response; therefore, the sum of these two columns may exceed the number of inaccurate responses.

Table 2
Pairwise question-level comparisons (exact McNemar test)

Comparison	n10/n01	Unadjusted p	Holm-adjusted p
GPT-5.6 vs Gemini 3.6 Flash	5/4	1.000	1.000
GPT-5.6 vs DeepSeek-V4	3/4	1.000	1.000
GPT-5.6 vs Doubao-2.1	12/4	0.077	0.327
Gemini 3.6 Flash vs DeepSeek-V4	3/5	0.727	1.000
Gemini 3.6 Flash vs Doubao-2.1	9/2	0.065	0.327
DeepSeek-V4 vs Doubao-2.1	12/3	0.035	0.211

Note: n10/n01 indicates the number of questions for which the first model was accurate and the second was inaccurate / the number for which the first model was inaccurate and the second was accurate.

Statement-Level Analysis

Statement-level analysis demonstrated more pronounced differences in strict statement-level concordance among the four models (Cochran's Q = 27.53, df = 3, $p < 0.001$). GPT-5.6 correctly covered the largest number of statements (25/52, 48.1%; 95% CI, 34.0%–62.4%), followed by Gemini 3.6 Flash (22/52, 42.3%; 95% CI, 28.7%–56.8%), DeepSeek-V4 (10/52, 19.2%; 95% CI, 9.6%–32.5%), and Doubao-2.1 (7/52, 13.5%; 95% CI, 5.6%–25.8%) (Table 3; Supplementary Table 4).

This pattern was not entirely consistent with the primary question-level analysis. DeepSeek-V4 had the numerically highest question-level accuracy but relatively low strict statement-level concordance, suggesting that its responses often captured the general direction of a question without fully covering specific recommendation conditions, restrictions, evidence grades, or time frames.

When correct and partial coverage were combined as at-least-partial coverage, GPT-5.6 covered 43/52 statements (82.7%; 95% CI, 69.7%–91.8%), Gemini 3.6 Flash 40/52 (76.9%; 95% CI, 63.2%–87.5%), DeepSeek-V4 39/52 (75.0%; 95% CI, 61.1%–86.0%), and Doubao-2.1 29/52 (55.8%; 95% CI, 41.3%–69.5%). Thus, several models could partially identify relevant consensus topics, but complete coverage of weight-bearing restrictions, ROM limitations, return-to-sport criteria, and referral conditions remained limited.

Pairwise statement-level comparisons showed that, after Holm correction, statistically significant differences remained between GPT-5.6 and DeepSeek-V4, GPT-5.6 and Doubao-2.1, Gemini 3.6 Flash and DeepSeek-V4, and Gemini 3.6 Flash and Doubao-2.1 (Table 4). The difference between GPT-5.6 and Gemini 3.6 Flash was not statistically significant.

Table 3
Statement-level analysis: strict concordance and at-least-partial coverage

Model	GPT-5.6	Gemini 3.6 Flash	DeepSeek-V4	Doubao-2.1
Strictly correct	25	22	10	7
Total	52	52	52	52
Strict concordance, %	48.1	42.3	19.2	13.5
Exact 95% CI, %	34.0–62.4	28.7–56.8	9.6–32.5	5.6–25.8
Partial	18	18	29	22
At-least-partial	43	40	39	29
At-least-partial coverage, %	82.7	76.9	75.0	55.8
Exact 95% CI, %	69.7–91.8	63.2–87.5	61.1–86.0	41.3–69.5
Missing or incorrect	9	12	13	23

Note: Categories are correct (C), partial (P), missing (M), and incorrect (I). Strict concordance includes C only. At-least-partial coverage includes C and P. Missing or incorrect coverage includes M and I. CIs are exact binomial 95% confidence intervals.

Table 4
Pairwise comparisons of strict statement-level concordance (exact McNemar test)

Comparison	n10/n01	Unadjusted p	Holm-adjusted p
GPT-5.6 vs Gemini 3.6 Flash	10/7	0.629	1.000
GPT-5.6 vs DeepSeek-V4	15/0	< 0.001	< 0.001
GPT-5.6 vs Doubao-2.1	19/1	< 0.001	< 0.001
Gemini 3.6 Flash vs DeepSeek-V4	16/4	0.012	0.035
Gemini 3.6 Flash vs Doubao-2.1	17/2	< 0.001	0.003
DeepSeek-V4 vs Doubao-2.1	7/4	0.549	1.000

Note: n10/n01 indicates the number of recommendation statements for which the first model was strictly correct and the second was not / the number for which the first model was not strictly correct and the second was strictly correct.

Module-Level Results

Module-level question accuracy varied across rehabilitation topics (Table 5). GPT-5.6 performed well for return-to-sport questions (4/4, 100.0%) and rehabilitation after arthroscopic partial meniscectomy (4/5, 80.0%). Gemini 3.6 Flash performed well in prevention of meniscal injury (2/2, 100.0%) and showed relatively balanced performance across several modules. DeepSeek-V4 performed well for questions involving meniscal reconstruction, transplantation, or scaffold implantation (5/6, 83.3%) and nonoperative rehabilitation of acute traumatic meniscal tears (4/5, 80.0%), although these results did not translate into similarly high strict statement-level concordance. Doubao-2.1 had low accuracy in the partial-meniscectomy, meniscal-repair, and nonoperative traumatic-tear modules.

Module-level analysis of strict statement-level concordance further revealed clinically meaningful differences (Table 6). GPT-5.6 achieved the highest strict concordance for rehabilitation after arthroscopic partial meniscectomy (8/10, 80.0%) and

return to sport (4/4, 100.0%). Gemini 3.6 Flash showed moderate performance in the partial-meniscectomy and meniscal-repair modules but low concordance in the reconstruction module. DeepSeek-V4 frequently provided directionally reasonable answers but omitted details required for strict statement-level concordance. Doubao-2.1 had no strictly correct statements in the meniscal-repair, nonoperative acute traumatic-tear, nonoperative degenerative-lesion, or injury-prevention modules.

Table 5
Module-level question accuracy

Module	Questions	GPT-5.6	Gemini 3.6 Flash	DeepSeek-V4	Doubao-2.1
Rehabilitation after arthroscopic partial meniscectomy	5	4/5 (80.0%)	3/5 (60.0%)	3/5 (60.0%)	1/5 (20.0%)
Rehabilitation after meniscal repair	8	4/8 (50.0%)	4/8 (50.0%)	5/8 (62.5%)	2/8 (25.0%)
Rehabilitation after meniscal reconstruction, transplantation, or scaffold implantation	6	4/6 (66.7%)	4/6 (66.7%)	5/6 (83.3%)	4/6 (66.7%)
Prevention of meniscal injury	2	1/2 (50.0%)	2/2 (100.0%)	2/2 (100.0%)	1/2 (50.0%)
Nonoperative rehabilitation of acute traumatic meniscal tears	5	3/5 (60.0%)	3/5 (60.0%)	4/5 (80.0%)	1/5 (20.0%)
Nonoperative rehabilitation of degenerative meniscal lesions	4	3/4 (75.0%)	3/4 (75.0%)	2/4 (50.0%)	2/4 (50.0%)
Return to sport	4	4/4 (100.0%)	3/4 (75.0%)	3/4 (75.0%)	3/4 (75.0%)
PROMs/PBMs and referral criteria	3	1/3 (33.3%)	1/3 (33.3%)	1/3 (33.3%)	2/3 (66.7%)

Note: Because several modules contained only a small number of questions, module-level analyses were exploratory.

Table 6
Module-level strict statement concordance

Module	Statements	GPT-5.6	Gemini 3.6 Flash	DeepSeek-V4	Doubao-2.1
Rehabilitation after arthroscopic partial meniscectomy	10	8/10 (80.0%)	6/10 (60.0%)	2/10 (20.0%)	2/10 (20.0%)
Rehabilitation after meniscal repair	13	5/13 (38.5%)	5/13 (38.5%)	3/13 (23.1%)	0/13 (0.0%)
Rehabilitation after meniscal reconstruction, transplantation, or scaffold implantation	6	3/6 (50.0%)	1/6 (16.7%)	2/6 (33.3%)	1/6 (16.7%)
Prevention of meniscal injury	2	0/2 (0.0%)	1/2 (50.0%)	0/2 (0.0%)	0/2 (0.0%)
Nonoperative rehabilitation of acute traumatic meniscal tears	7	1/7 (14.3%)	3/7 (42.9%)	0/7 (0.0%)	0/7 (0.0%)
Nonoperative rehabilitation of degenerative meniscal lesions	6	3/6 (50.0%)	3/6 (50.0%)	0/6 (0.0%)	0/6 (0.0%)
Return to sport	4	4/4 (100.0%)	2/4 (50.0%)	3/4 (75.0%)	2/4 (50.0%)
PROMs/PBMs and referral criteria	4	1/4 (25.0%)	1/4 (25.0%)	0/4 (0.0%)	2/4 (50.0%)
Note: Because several modules contained only a small number of statements, module-level analyses were exploratory.					
Potentially Hazardous Errors and Risk Omissions					

Potentially hazardous errors were less frequent than general inaccuracies but were clinically important because they involved safety-critical rehabilitation decisions. GPT-5.6 and DeepSeek-V4 each had 3 potentially hazardous errors or risk omissions, whereas Gemini 3.6 Flash and Doubao-2.1 each had 4 (Table 1).

Standardized risk-flag assessment showed that all four models only partially addressed early deep squatting, jumping, or rotational loading after radial/root repair. Coverage of referral criteria was also incomplete for all models. Gemini 3.6 Flash exhibited partial risk related to immediate full weight-bearing after root repair in one repair-related response, whereas Doubao-2.1 showed partial confusion among rehabilitation protocols for different procedures. These errors or omissions did not necessarily take the form of explicit hazardous recommendations; however, if applied without verification of the original consensus or procedure-specific details, they could result in unsafe practice (Table 7; Supplementary Table 5).

Table 7
Summary of potentially hazardous errors and risk omissions

Potentially hazardous error/risk omission	GPT-5.6	Gemini 3.6 Flash	DeepSeek-V4	Doubao-2.1
Immediate full weight-bearing after meniscal root repair	Not observed(Q07/Q13)	Partly present(Q13)	Not observed(Q07/Q13)	Not observed
Early deep squatting, jumping, or rotational loading after radial/root repair	Partly present(Q08)	Partly present(Q08)	Partly present(Q08)	Partly present(Q08)
Return-to-sport decisions based on time alone rather than functional criteria	Not observed(Q11/Q31/Q34)	Not observed(Q31)	Not observed(Q10/Q11/Q31/Q34)	Not observed(Q31)
Routine surgery for degenerative meniscal lesions without an initial trial of nonoperative treatment	Not observed(Q27/Q28)	Not observed(Q27/Q29)	Not observed(Q27/Q29)	Not observed(Q27)
Omission of referral criteria or failure to recommend referral	Partly present(Q05/Q37)	Partly present(Q05/Q37)	Partly present(Q05/Q37)	Partly present(Q05/Q37)
Confusion among rehabilitation protocols after meniscectomy, repair, and reconstruction	Not observed(overall assessment)	Partly present(Q13)	Not observed(overall assessment)	Partly present(Q14)

Note: “Present” indicates an explicit hazardous recommendation or a clear risk; “Partly present” indicates incomplete or ambiguous treatment that could lead to unsafe application; and “Not observed” indicates that no such error was identified in the assessed responses. This table presents descriptive results and should be interpreted together with the detailed question-level and statement-level assessments.

4. Discussion

This study evaluated four LLMs against the formal EU-US Meniscus Rehabilitation 2024 Consensus developed by ESSKA, AOSSM, and AASPT in a deliberately realistic scenario in which the consensus was named but the source document was not provided. The principal finding was not simply that overall accuracy differed among models; rather, the apparent clinical usefulness of their responses changed substantially according to the level of evaluation. At the question level, DeepSeek-V4 had the numerically highest accuracy, followed by GPT-5.6 and Gemini 3.6 Flash, whereas Doubao-2.1 showed less consistent performance. At the statement level, however, GPT-5.6 and Gemini 3.6 Flash demonstrated clearer advantages, whereas

DeepSeek-V4 declined substantially despite its stronger question-level performance. This divergence is central to interpretation of the findings: a response may sound clinically reasonable yet still fail to reproduce the procedure-specific detail required to implement a formal rehabilitation consensus.

The source-document-not-provided design was not merely a technical detail but the clinical scenario under investigation. In practice, clinicians, trainees, therapists, and patients may know the title of a consensus statement and ask an LLM to answer strictly according to the named document, while typically not uploading the PDF, recommendation tables, evidence grades, or source text. The updated standardized instruction used in this study attempted to constrain the models by explicitly naming the formal EU-US Meniscus Rehabilitation 2024 Consensus and instructing them not to use other guidelines unless clearly identified. Even under this stronger prompt, none of the four models achieved high strict statement-level concordance. Naming a reference document may therefore improve the structure of a response but does not ensure that the model can retrieve or faithfully reconstruct specific recommendation statements.

This distinction is particularly important in meniscus rehabilitation. The broad diagnosis of a “meniscal tear” does not correspond to a single rehabilitation pathway. Recommendations differ among isolated partial meniscectomy, meniscal repair, root or radial tear repair, meniscal reconstruction, scaffold-related procedures, and nonoperative treatment of traumatic or degenerative lesions. Early full weight-bearing and symptom-tolerated ROM are generally permitted after partial meniscectomy, whereas the meniscus may require procedure- and tear-pattern-specific restrictions after repair or reconstruction. General statements such as “progressive rehabilitation,” “criteria-based progression,” or “individualized return to sport” may therefore be directionally acceptable but remain clinically incomplete if they omit weight-bearing status, ROM limits, precautions regarding deep flexion, uncertainty concerning brace use, healing time frames, or functional criteria.

The discrepancy between question-level accuracy and strict statement-level concordance also explains why model rankings differed between analyses. DeepSeek-V4 often captured the general direction of a question and therefore scored relatively well at the question level. However, many of its responses failed to achieve strict statement-level concordance because they omitted specific restrictions, misstated evidence grades, or transformed conditional recommendations into more definitive protocols. GPT-5.6 and Gemini 3.6 Flash more often identified relevant consensus concepts and preserved uncertainty, resulting in higher statement-level concordance. Doubao-2.1 frequently generated confident and highly structured responses, but this structure often included unsupported timelines, fabricated evidence grades, or excessively specific rehabilitation steps. These findings indicate that fluency and apparent organization should not be mistaken for fidelity to a consensus.

Insufficient information was the most common error pattern. In many responses, a model mentioned an appropriate general concept but omitted one or more details critical to implementation. For example, a response might state that rehabilitation after meniscal repair should incorporate both time and criteria without specifying the protection, restoration, and return-to-activity phases; mention return to sport without addressing effusion, ROM, quadriceps and hamstring strength, neuromuscular control, psychological readiness, or sport-specific demands; or recommend referral without listing persistent pain, recurrent stiffness or effusion, functional instability, mechanical symptoms, neurological symptoms, suspected infection, or suspected deep vein thrombosis. These omissions are clinically important because rehabilitation decisions are implemented through specific thresholds and restrictions rather than broad principles.

Overstatement was the second major failure pattern. Several responses presented expert-opinion or low-level-evidence recommendations as high-grade evidence, constructed fixed week-by-week protocols when the consensus emphasized criteria-based progression, or introduced unsupported differences between medial and lateral procedures. This was particularly evident for reconstruction, brace use, supervised outpatient rehabilitation versus home exercise, and return-to-sport timing. The problem was not merely academic. When an LLM converts uncertainty into certainty, users may be less likely to verify the original consensus or seek procedure-specific clarification from a surgeon or therapist. This is particularly concerning in postoperative settings, where tissue healing, repair stability, concomitant procedures, and surgeon-imposed restrictions should determine rehabilitation progression.

Potentially hazardous errors clustered in predictable high-risk domains. Weight-bearing and ROM recommendations after root, radial, or complex tear repair or reconstruction represented one such domain; deep loaded flexion, jumping, and rotational activity after repair represented another. Return-to-sport recommendations were also vulnerable to error: time-based advice alone may appear practical but is incomplete without objective knee function, strength, neuromuscular control, effusion status, subjective readiness, and sport-specific progression. Finally, incomplete referral criteria may delay reassessment of patients with persistent pain, recurrent swelling or stiffness, mechanical symptoms, neurological signs, suspected infection, suspected deep vein thrombosis, or failure to achieve expected milestones. These errors may not always appear as explicitly hazardous recommendations, but omission of safety conditions can itself lead to unsafe interpretation.

Compared with question-level accuracy alone, statement-level analysis provided a more conservative and clinically meaningful view of LLM performance. GPT-5.6, Gemini 3.6 Flash, and DeepSeek-V4 achieved at-least-partial coverage for a substantial proportion of statements, indicating that they generally recognized relevant topic domains. Strict concordance, however, was markedly lower, and even the best-performing model correctly covered fewer than half of the 52 recommendation statements. The gap between partial topic recognition and complete recommendation coverage has important implications for future medical LLM evaluations. Studies that score only overall response correctness may overestimate clinical reliability, particularly in guideline domains in which a single question contains multiple interdependent conditions.

These results are consistent with previous medical LLM evaluations showing that readable responses are not necessarily guideline-concordant. Orthopaedic and sports medicine applications add another layer of difficulty because recommendations are often procedure-specific, stage-specific, and dependent on surgical biology. Meniscus rehabilitation is illustrative: risk often arises not from one obviously incorrect statement but from compressing different procedures into a single generic rehabilitation pathway. If a model cannot distinguish meniscectomy, repair, and reconstruction or ignores tear morphology and concomitant procedures, it may generate advice that is appropriate in one context but inappropriate in another.

From a clinical perspective, the present findings support a limited but useful role for LLMs. They may help users understand terminology, identify major rehabilitation domains, prepare questions for discussion with clinicians, or generate preliminary educational summaries. However, they should not be used independently to prescribe postoperative meniscus rehabilitation or determine return to sport. A safer approach would require the model to actively request missing clinical information, including procedure type, tear pattern, root or radial involvement, repair stability, concomitant ACL reconstruction or osteotomy, postoperative time point, effusion, ROM, pain, strength testing, functional control, and surgeon-defined restrictions. Ideally, future systems should cite or display specific source recommendations and evidence grades rather than relying only on memory-based reconstruction of a named consensus.

The findings also have implications for prompt design. The explicit instruction used in this study was stronger than a general evidence-based medicine prompt because it named a specific consensus and warned the models not to use other guidelines. Nevertheless, without access to the source text, the improvement resulting from this constraint remained limited. Guideline-specific clinical questions may therefore require a retrieval-augmented workflow or direct document upload. Asking an LLM to answer strictly according to a named document is not equivalent to giving the model access to that document. Future studies should compare source-document-not-provided prompts, document-upload prompts, retrieval-augmented prompts, and prompts requiring explicit extraction of evidence grades to determine which workflow best improves fidelity and safety.

This study has several limitations. First, model outputs were collected at a single time point, and commercial LLMs may change over time. Second, sampling parameters could not be fully controlled in web-based systems, which may affect reproducibility. Third, scoring inevitably involved reviewer judgment, particularly for partial coverage and overstatement; however, the prespecified list of reference statements reduced arbitrariness. Fourth, the study evaluated textual concordance rather than patient outcomes, clinician behavior, or real-world rehabilitation adherence.

Despite these limitations, the study reflects a common and clinically relevant interaction pattern. Knowing and naming a consensus document without uploading the source text is a realistic way in which LLMs are used, and the present findings show that this workflow remains vulnerable to incomplete and overconfident responses. The practical message should therefore remain balanced: LLMs can provide readable and generally directionally helpful responses, but their outputs should be treated as preliminary educational material. Rehabilitation decisions that depend on tissue healing, procedure-specific restrictions, return-to-sport thresholds, or referral triggers should be verified against the original consensus and individualized postoperative instructions.

5. Conclusions

In this explicit scenario in which the consensus was named but the source document was not uploaded, concordance of GPT-5.6, Gemini 3.6 Flash, DeepSeek-V4, and Doubao-2.1 with the formal EU-US Meniscus Rehabilitation 2024 Consensus developed by ESSKA, AOSSM, and AASPT was heterogeneous. DeepSeek-V4 had the numerically highest question-level accuracy, whereas GPT-5.6 had the highest strict statement-level concordance. The discrepancy between question-level and statement-level performance indicates that directionally correct LLM responses may still omit clinically important procedure-specific restrictions.

LLMs may support education and preliminary knowledge retrieval in meniscus rehabilitation but cannot replace clinical judgment or verification of the original consensus. Particular caution is warranted for weight-bearing and ROM restrictions, repair and reconstruction protocols, return-to-sport criteria, PROMs/PBMs, and referral indications.

Abbreviations

AASPT
American Academy of Sports Physical Therapy
ACL
anterior cruciate ligament
AOSSM
American Orthopaedic Society for Sports Medicine
CI
confidence interval
DML
degenerative meniscal lesion
ESSKA
European Society of Sports Traumatology,Knee Surgery and Arthroscopy
LLM
large language model
PBM
performance-based measure
PROM
patient-reported outcome measure
PWB
partial weight-bearing
ROM
range of motion
WB
weight-bearing.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human participants, human data, human tissue, human specimens, or animals.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Additional file 1

Description

Supplementary Tables 1–5 provide the clinical questions and reference statements, raw model responses, question-level scoring, statement-level coverage, and safety-related assessments.

Funding

Declaration

This work was supported by the Doctor Research Start-up Fund (Grant No. RC2024067).

Author Contribution

XW and DF contributed equally to the conceptualization and design of the study, methodology development, data collection, formal analysis, visualization, and preparation of the original manuscript. LX contributed to data collection, validation, interpretation of the results, and critical revision of the manuscript. YW contributed to the study design, methodology, supervision, and critical revision of the manuscript. QZ supervised the study, contributed to project administration and interpretation of the findings, and critically revised the manuscript. All authors read and approved the final manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this article and its additional file.

References

1. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. 10.1038/s41586-023-06291-2.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198. 10.1371/journal.pdig.0000198.
3. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. 10.2196/45312.
4. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29:1930–40. 10.1038/s41591-023-02448-8.

5. Lin CR, Chen YJ, Tsai PA, et al. Multiple large language models versus clinical guidelines for postmenopausal osteoporosis: a comparative study of ChatGPT-3.5, ChatGPT-4.0, ChatGPT-4o, Google Gemini, Google Gemini Advanced, and Microsoft Copilot. Arch Osteoporos. 2025;20:120. 10.1007/s11657-025-01587-4.
6. Aksoy E. The performance of artificial intelligence in one anastomosis gastric bypass surgery: comparative efficacy of ChatGPT-4.0, ChatGPT-Omni, and, Gemini AI. Obes Surg. 2025;35:1469–1475. 10.1007/s11695-025-07794-9
7. Yang JW, Ardavanis KS, Slack KE, et al. Chat Generative Pretrained Transformer (ChatGPT) and Bard: artificial intelligence does not yet provide clinically supported answers for hip and knee osteoarthritis. J Arthroplasty. 2024;39:1184–90. 10.1016/j.arth.2024.01.029.
8. Mejia MR, Arroyave JS, Saturno M, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American Spine Society guideline comparison. Neurospine. 2024;21:149–58. 10.14245/ns.2347052.526.
9. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023;95:104770. 10.1016/j.ebiom.2023.104770.
10. Deng L, Wang T, Yangzhang, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg. 2024;110:1941–50. 10.1097/JS9.0000000000001066.
11. Beaufils P, Becker R, Kopf S, et al. Surgical management of degenerative meniscus lesions: the 2016 ESSKA meniscus consensus. Knee Surg Sports Traumatol Arthrosc. 2017;25:335–46. 10.1007/s00167-016-4407-4.
12. Kopf S, Beaufils P, Hirschmann MT, et al. Management of traumatic meniscus tears: the 2019 ESSKA meniscus consensus. Knee Surg Sports Traumatol Arthrosc. 2020;28:1177–94. 10.1007/s00167-020-05847-3.
13. Snoeker BAM, Bakker EWP, Kegel CAT, et al. Risk factors for meniscal tears: a systematic review including meta-analysis. J Orthop Sports Phys Ther. 2013;43:352–67. 10.2519/jospt.2013.4295.
14. Thorlund JB, Juhl CB, Ingelsrud LH, et al. Risk factors, diagnosis and non-surgical treatment for meniscal tears: evidence and recommendations: a statement paper commissioned by the Danish Society of Sports Physical Therapy (DSSF). Br J Sports Med. 2018;52:557–65. 10.1136/bjsports-2017-098429.
15. Sihvonen R, Paavola M, Malmivaara A, et al. Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. N Engl J Med. 2013;369:2515–24. 10.1056/NEJMoa1305189.
16. Kise NJ, Risberg MA, Stensrud S, et al. Exercise therapy versus arthroscopic partial meniscectomy for degenerative meniscal tear in middle aged patients: randomised controlled trial with two year follow-up. BMJ. 2016;354:i3740. 10.1136/bmj.i3740.
17. van de Graaf VA, Noorduyn JCA, Willigenburg NW, et al. Effect of early surgery vs physical therapy on knee function among patients with nonobstructive meniscal tears: the ESCAPE randomized clinical trial. JAMA. 2018;320:1328–37. 10.1001/jama.2018.13308.
18. Wijn SRW, Hannink G, Østerås H, et al. Arthroscopic partial meniscectomy vs non-surgical or sham treatment in patients with MRI-confirmed degenerative meniscus tears: a systematic review and meta-analysis with individual participant data from 605 randomised patients. Osteoarthritis Cartilage. 2023;31:557–66. 10.1016/j.joca.2023.01.002.
19. Zhang Y, Cao W, Cao Q, et al. Comparative effects on pain arising from injury to the knee meniscus in adults: a systematic review and network meta-analysis. Clin Rehabil. 2021;35:801–11. 10.1177/0269215520976649.
20. Pujol N, Giordano AO, Wong SE, et al. The formal EU-US Meniscus Rehabilitation 2024 Consensus: an ESSKA-AOSSM-AASPT initiative. Part I-Rehabilitation management after meniscus surgery (meniscectomy, repair and reconstruction). Knee Surg Sports Traumatol Arthrosc. 2025;33:3002–13. 10.1002/ksa.12674.
21. Prill R, Ma CB, Wong SE, et al. The formal EU-US Meniscus Rehabilitation 2024 Consensus: an ESSKA-AOSSM-AASPT initiative. Part II-Prevention, non-operative treatment and return to sport. Knee Surg Sports Traumatol Arthrosc. 2025;33:3014–24. 10.1002/ksa.12689.

22. Silvers-Granelli HJ, Bizzini M, Arundale A, et al. Does the FIFA 11 + injury prevention program reduce the incidence of ACL injury in male soccer players? Clin Orthop Relat Res. 2017;475:2447–55. 10.1007/s11999-017-5342-5.
23. OpenAI. ChatGPT. https://chatgpt.com/. Accessed 15 Aug 2026.
24. Google, Gemini. https://gemini.google.com/. Accessed 15 Aug 2026.
25. DeepSeek. DeepSeek Chat. https://chat.deepseek.com/. Accessed 15 Aug 2026.
26. ByteDance. Doubao. https://www.doubao.com/chat/. Accessed 15 Aug 2026.

Figures

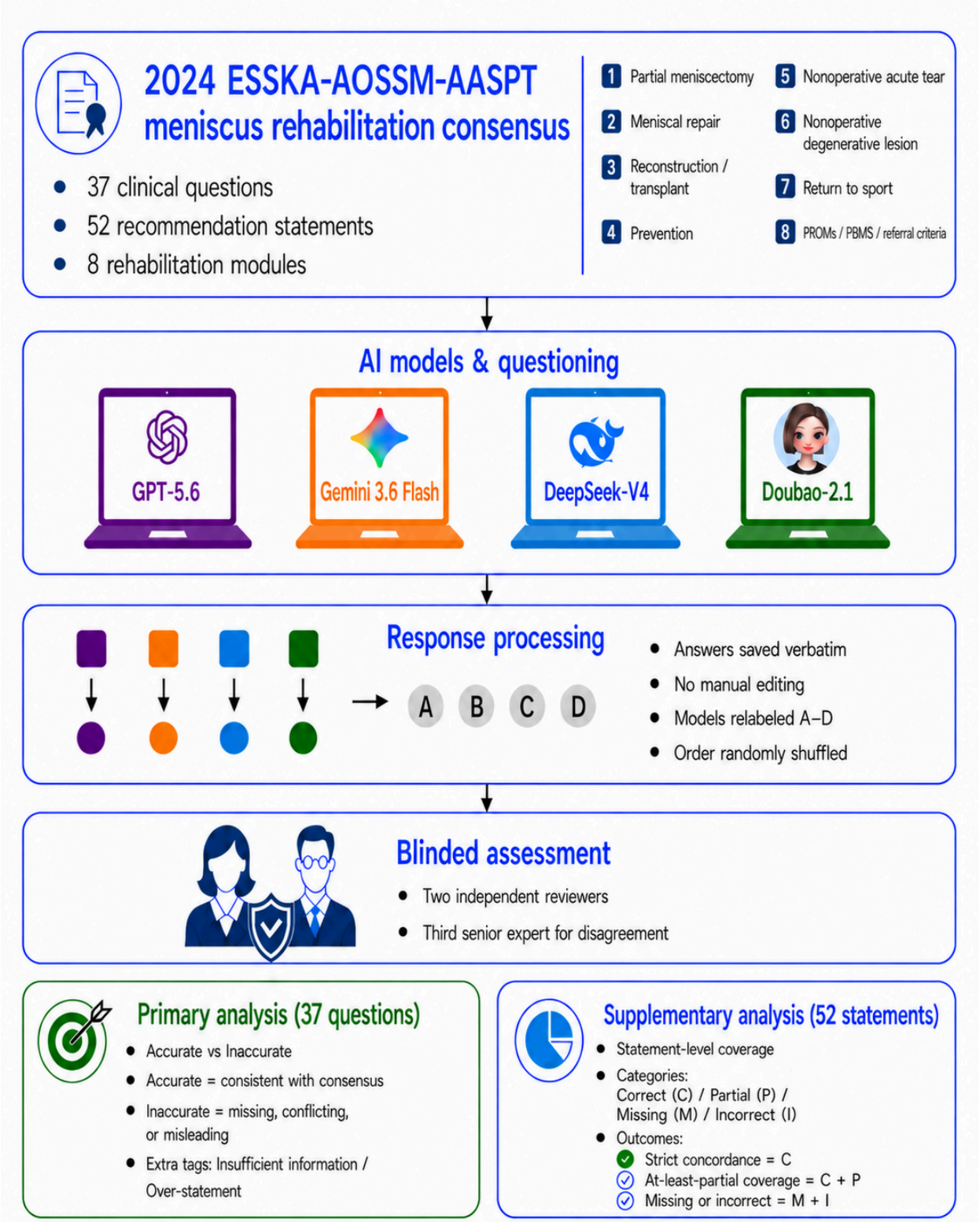


Figure 1

Overall study design and workflow.

Supplementary Files

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

- 9.27SupplementarymaterialsENGLISH.docx