Journal of Orthopaedic Surgery

Journal of Orthopaedic Surgery
Vol. 34(2): 1–9
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10225536261459518
journals.sagepub.com/home/osj
Sage

Automating methodological quality assessment in orthopedic systematic reviews using large language models

Yu-Jui Huang[1], Kai-Cheng Chang[2], Ying-Chen Kuo[3], Cheng-Chen Tai[4] and Cheng-Pang Yang[1]

Abstract

Background: Systematic reviews represent the foundation of evidence-based orthopedic practice, yet their methodological rigor relies heavily on accurate and consistent methodological quality assessment. This step remains time-consuming, labor-intensive, and prone to subjectivity. Recent advances in large language models (LLMs) suggest potential for automating parts of evidence synthesis.

Purpose: This study examined whether LLMs can perform AMSTAR-1–based methodological quality assessment evaluations in orthopedic systematic reviews with accuracy comparable to human experts.

Methods: Ten sports medicine knee reviews were analyzed using three LLMs—GPT-4o, GPT-5, and GPT Consensus—and their binary responses were compared against expert AMSTAR-1 ratings from a published umbrella review (110 decisions). An external validation set of four reviews published between 2022 and 2025 was included to assess generalizability and safeguard against information leakage.

Results: Agreement with human reviewers reached 87% for GPT-4o, 89% for GPT-5, and 90% for GPT Consensus; all models achieved 84% agreement in the validation set. Concordance was strongest for structured, explicitly reported domains such as a priori design, literature search, and study characteristics, and lowest for judgment-based items including grey literature inclusion, publication bias, and conflict of interest.

Conclusions: LLMs cannot yet replace human reviewers, they can serve as reliable adjunct tools to enhance efficiency, transparency, and reproducibility in systematic review workflows within orthopedic research.

Keywords

sports, artificial intelligence, evidence-based, large language model

Received: 26 January 2026; revised: 28 May 2026; accepted: 30 May 2026

[1]Department of Orthopedic Surgery, Linkou Chang Gung Memorial Hospital, Taoyuan City, Taiwan
[2]Department of Pharmacy, Linkou Chang Gung Memorial Hospital, Taoyuan City, Taiwan
[3]Department of Physical Medicine and Rehabilitation, Shin Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan
[4]Medical Library, Department of Medical Education, Linkou Chang Gung Memorial Hospital, Taoyuan, Taiwan

Corresponding author:
Cheng-Pang Yang, Department of Orthopedic Surgery, Linkou Chang Gung Memorial Hospital, No. 5, Fuxing St, Guishan Dist, Taoyuan City 333, Taiwan.
Email: ronnie80097@gmail.com

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Introduction

Systematic reviews serve as the cornerstone of evidence-based clinical practice by synthesizing research findings across multiple studies to inform healthcare decisions. In orthopedics, where treatment options are often complex and patient outcomes can vary widely, the conclusions drawn from high-quality systematic reviews directly influence clinical decision-making and treatment planning.[1,2]

A critical component of systematic review methodology is the methodological quality assessment, which evaluates the presence of systematic errors in included studies (e.g., systematic review, randomized controlled trials). However, conducting methodological quality assessment is often time-consuming, subjectivity, requires domain-specific expertise, and is especially burdensome in orthopedics, where study heterogeneity and reporting inconsistencies are common.[3,4]

The integration of LLMs in evidence synthesis is rapidly shifting from citation screening toward automated methodological quality appraisal.[5,6] While current models demonstrate high agreement in objective domains, persistent challenges remain in subjective judgments, such as publication bias.[7] However, a critical evidence gap persists regarding the performance of the latest iterations—specifically GPT-5—within the specialized field of orthopedic surgery. Furthermore, few studies have rigorously addressed the risk of 'information leakage' in AI benchmarking, a gap that the present study seeks to address.

Given the rapidly evolving capabilities of large language models (LLMs) such as ChatGPT, there is growing interest in their potential to assist or automate components of

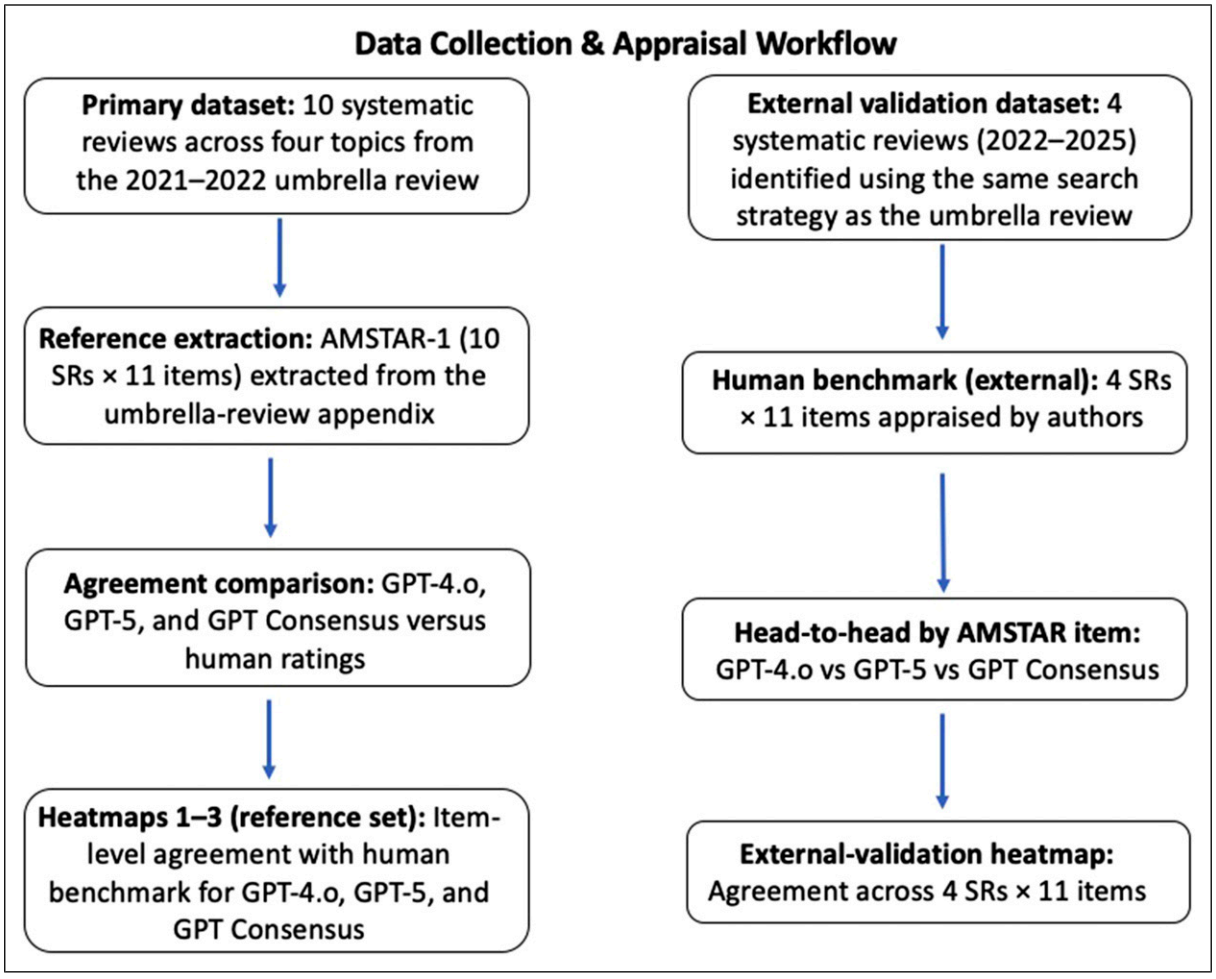


Figure 1. Flowchart outlining the data collection and methodological quality assessment appraisal workflow. The primary dataset included 10 systematic reviews (SRs) on orthopedic interventions from a 2021–2022 umbrella review, while an external validation set comprised 4 SRs published between 2022 and 2025. AMSTAR 1 ratings (11 items per SR) were extracted from the umbrella review appendix or independently completed by a clinical expert for the external set. Three large language models (LLMs)—GPT-4o, GPT 5, and GPT Consensus—were used to assess methodological quality assessment, and their outputs were compared to human benchmarks. Agreement metrics were analyzed at the item level, with comparative heatmaps generated for both reference and validation sets.

evidence synthesis, being as third methodological quality assessor, while maintaining accuracy and reducing human workload. Our goal in this study was to explore whether ChatGPT could reliably assess methodological quality assessment in orthopedic systematic reviews, thereby improving efficiency without compromising rigor.[8,9]

Methods

Study design

This study employed a comparative design to evaluate the performance of large language models (LLMs) in conducting systematic review methodological quality assessment using the AMSTAR 1 checklist. AMSTAR-1 was selected as the appraisal tool to ensure direct comparability with our human benchmark, a landmark umbrella review by Blom et al.[10] Furthermore, as our included reviews focused on Level 1 evidence (RCTs), AMSTAR-1 remains a validated and robust instrument for this purpose. Its 11-item binary format also provides a clear, structured framework ideal for this initial benchmarking of LLM capabilities. The primary objective was to assess how accurately LLMs could replicate expert-level methodological quality assessment in the context of orthopedic systematic reviews. Model performance was examined across two datasets: a core dataset derived from a highly cited umbrella review (Common elective orthopaedic procedures and their clinical-effectiveness: umbrella review of level 1 evidence[7]) and an independent external validation set designed to test generalizability. Flowchart outlining the data collection and methodological quality appraisal workflow was showed in Figure 1.

Data collection

Selection of primary dataset. The core dataset consisted of 10 systematic review articles selected from the 2021 publication titled "Common Elective Orthopaedic Procedures and Their Clinical Effectiveness: Umbrella Review of Level 1 Evidence." This umbrella review is recognized as one of the most comprehensive syntheses of high-quality orthopedic data, encompassing a range of elective procedures. For the purposes of this study, the focus was narrowed to interventions related to sports medicine knee surgery, given their global prevalence and clinical significance. Specifically, the included procedures were arthroscopic anterior cruciate ligament (ACL) reconstruction, arthroscopic meniscal repair, and arthroscopic partial meniscectomy.

The selection of these 10 studies was guided by both clinical relevance and the availability of expert-completed AMSTAR 1 assessments, which were provided in the appendix of the umbrella review. These expert ratings served as the reference standard for comparison with the LLM-generated assessments.

External validation and information-leakage safeguard. We prospectively assembled an external validation set to guard against information leakage. Because models with standard web-search capabilities could, in principle, retrieve the appendix of the umbrella review (Common elective orthopaedic procedures and their clinical effectiveness: umbrella review of level 1 evidence), in-sample agreement might be spuriously elevated. To evaluate robustness and generalizability beyond the primary dataset, we curated four out-of-sample systematic reviews (published 2022–2025) on the same knee procedures. These reviews were appraised by two researchers. Discrepancies were resolved through discussion or, if necessary, by a third independent senior expert (Dr Chang) to establish the final reference standard. This ensured that the LLMs were benchmarked against a highly reliable, consensus-based human evaluation. Performance on this set was used to verify that observed agreement reflected genuine AMSTAR 1 appraisal rather than answer retrieval.

We included studies published between January 2022 and May 2025 identified using the original umbrella-review search strategy. Searches of PubMed and Embase combined orthopaedic keywords with procedure-specific terms and applied filters for systematic reviews. Eligibility required full-text availability in PDF format and suitability for AMSTAR 1 scoring. Consistent with the reference set, inclusion was restricted to reviews of arthroscopic anterior cruciate ligament (ACL) reconstruction, arthroscopic meniscal repair, and arthroscopic partial meniscectomy.

LLM-based methodological quality assessment procedure

The OpenAI GPT series was selected as the primary platform for this study due to its status as the most prevalent LLM in medical and healthcare research. Recent bibliometric and scoping reviews by Su et al. and Maity et al. have identified the GPT family as the standard benchmark in medical diagnostics and academic applications, representing the vast majority of current literature in the field.[11,12] Prioritizing the GPT family ensures our findings remain aligned with existing medical AI benchmarks and allows for a clear longitudinal evaluation of advancements from GPT-4o to GPT-5.

To ensure reproducibility, specific model versions and access dates were documented for all evaluations. GPT-4o (accessed April 20, 2025) and GPT-5 (accessed January 12, 2026) were utilized via the standard ChatGPT interface. Additionally, the GPT Consensus model (accessed July 10, 2025) was evaluated; this specialized model utilizes the GPT-4o architecture as its underlying engine, integrated

with a customized scientific search interface for evidence synthesis.

All LLM evaluations were conducted via the standard ChatGPT web browser interface using default system settings to simulate real-world professional usage. We employed a consistent zero-shot prompting strategy, instructing the models to act as evidence-based medicine experts and appraise the meta-analysis papers using the 11-item AMSTAR-1 checklist. Specifically, the standardized prompt began with: 'Hi, I need you to become an expert of evidence-based medicine and help me review meta-analysis papers to examine the quality…', followed by the full-text content of each included study."

Each review was assessed independently by three different LLMs:

(1) GPT-4o, a general-purpose advanced model developed by OpenAI.
(2) GPT-5 Model, a proprietary language model designed for stepwise analytical reasoning.
(3) GPT Consensus Model, a research-optimized model hosted by Consensus.app, tailored for academic synthesis and evidence-based queries.

Each model received identical instructions: to evaluate the systematic review using the AMSTAR 1 checklist, providing binary yes/no judgments for each of the 11 items. To maintain ecological validity and simulate real-world application, no prompt engineering, examples, or templates were provided. Models were queried independently to avoid cross-contamination, and none had access to the human reference ratings. LLM-generated responses were exported and manually reviewed for binary score mapping.

Data analysis and agreement scoring

Model outputs were evaluated against the corresponding human expert assessments using a binary agreement system, where a score of 1 indicated concordance between the model and the human rating, and a score of 0 indicated a discrepancy. The primary dataset generated 110 comparison points per model (10 reviews × 11 checklist items), while the external validation set contributed an additional 44 comparison points (4 reviews × 11 items), resulting in 154 total model-human comparisons for each LLM.

For each model, we calculated the total number of agreements, the overall agreement percentage, item-level agreement rates to identify specific areas of model strength or weakness, and relative model performance comparisons. To account for chance agreement, Cohen's kappa coefficients were calculated. Additionally, the sensitivity and specificity of each LLM were determined. Importantly, no discrepancies were adjudicated or resolved post hoc, as the aim was to capture autonomous model performance in a real-world clinical evidence context.

Data processing and storage were carried out in Microsoft Excel, while comparative visualizations were generated using Google colab.

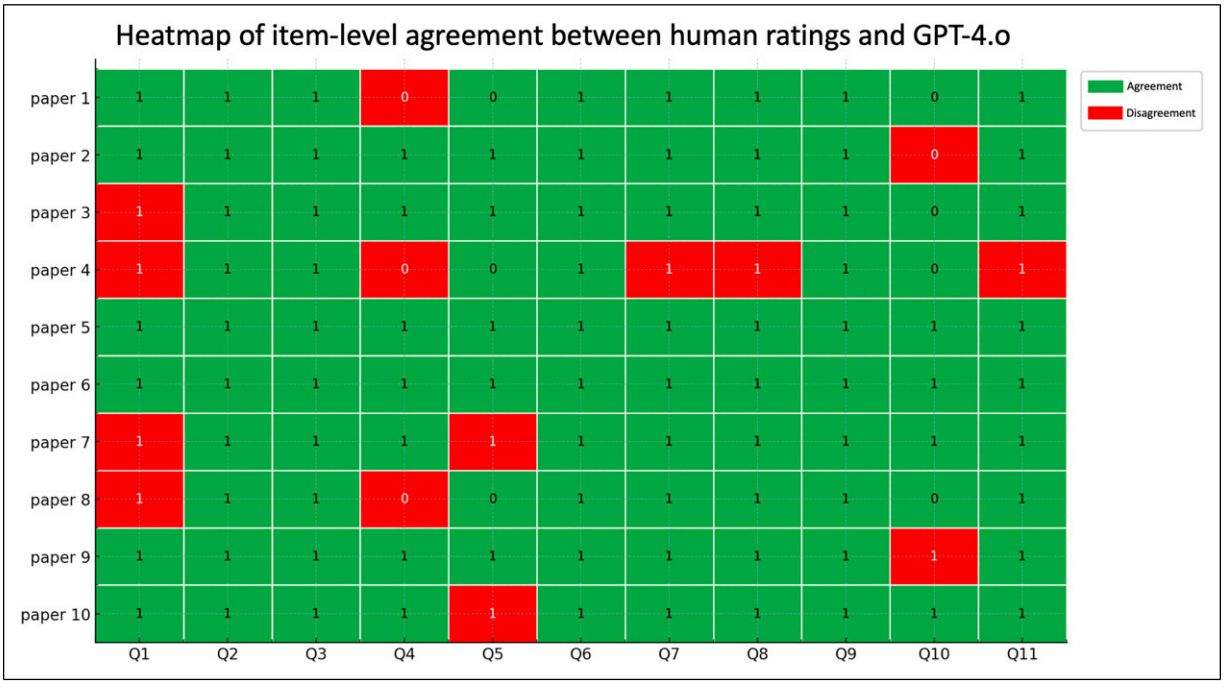


Figure 2. Heatmap of item-level agreement between human ratings and GPT-4.0 across 10 systematic reviews using the AMSTAR 1 checklist. Green indicates agreement; red indicates disagreement.

Results

This study examined the capacity of three large language models—GPT-4o, GPT 5, and GPT Consensus—to evaluate methodological quality assessment in orthopedic systematic reviews using the AMSTAR 1 checklist. Each model was tasked with rendering binary judgments (yes/no) across 11 checklist items for 10 systematic reviews, yielding 110 model-human comparison points per model.

Overall agreement with human reviewers

All three LLMs demonstrated a high degree of alignment with human expert assessments. GPT-4o achieved agreement in 96 out of 110 checklist items (87%) (Figure 2), while the GPT 5 (Figure 3) slightly outperformed it with 98 agreements (89%). The GPT Consensus Model achieved the highest overall performance, agreeing with human ratings on 99 items (90%) (Figure 4). These results suggest that LLMs are capable of producing reliable methodological quality assessment, particularly in domains where the source documents contain structured, explicitly reported information.

The diagnostic performance of GPT4o, Consensus, and GPT 5.0 was evaluated against the expert gold standard across 110 cases. The results for sensitivity, specificity, accuracy, and inter-rater reliability are summarized in Table 1. GPT-4o showed moderate agreement (87.27% accuracy, $\kappa = 0.431$) with high sensitivity (95.70%) but limited specificity (41.18%). The Consensus group reached the highest accuracy (90.91%) and sensitivity (96.77%; $\kappa = 0.615$), while GPT-5 provided the most balanced and reliable performance ($\kappa = 0.635$, substantial agreement), yielding the highest specificity (82.35%) at 89.09% accuracy.

In validation dataset (Figure 5), we finally extracted 4 studies meet inclusion criteria for methodological quality assessment. Across the 44 AMSTAR items in this validation set, each model achieved identical agreement rates of 37 out of 44 (84%). The consistency between the primary and validation datasets supports the reliability and generalizability of these models for methodological quality assessment across varied topics and timeframes.

Domain-level agreement patterns

Analysis of individual AMSTAR items revealed variation in model performance depending on the nature of the checklist criterion. The highest levels of agreement (≥90%) were observed in domains characterized by objective and well-defined reporting. For example, Item 1 (a priori design) was consistently recognized when protocols were clearly stated, while Item 3 (comprehensive literature search) saw strong agreement due to the transparent presentation of search strategies. Similarly, Item 6 (reporting of included study characteristics) achieved high agreement when the primary studies explicitly detailed key elements such as population, intervention, comparator, and outcomes.

In contrast, lower agreement rates (<80%) emerged in more inferential or subjectively reported domains. Item 4 (inclusion of grey literature) presented challenges, as LLMs struggled to detect whether non-indexed sources had been considered, particularly when such searches were only implied. Item 10 (assessment of publication bias) also proved problematic, largely because many reviews lacked explicit reporting on this dimension, leaving models to infer compliance. Item 11 (conflict of interest) further highlighted

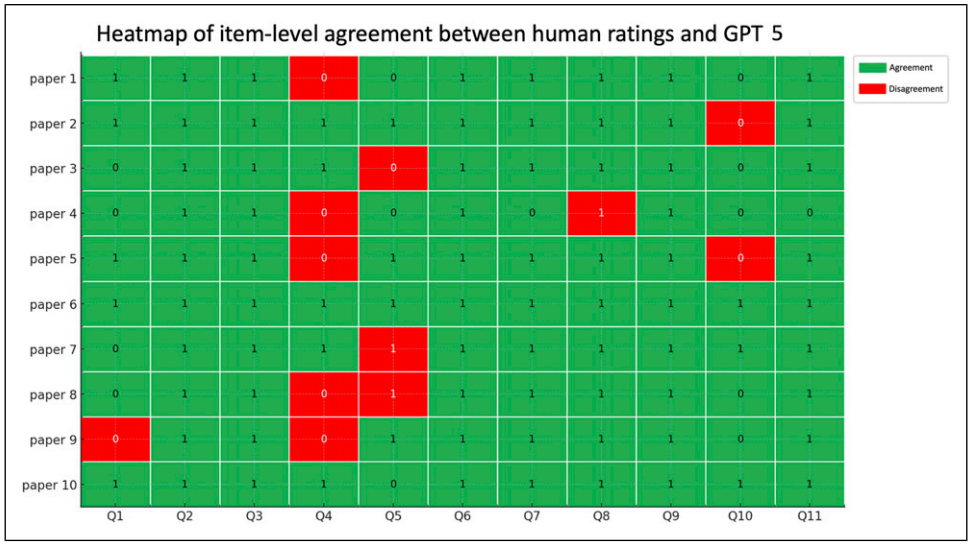


Figure 3. Heatmap of item-level agreement between human ratings and the GPT 5 across 10 systematic reviews using the AMSTAR 1 checklist. Green indicates agreement; red indicates disagreement.

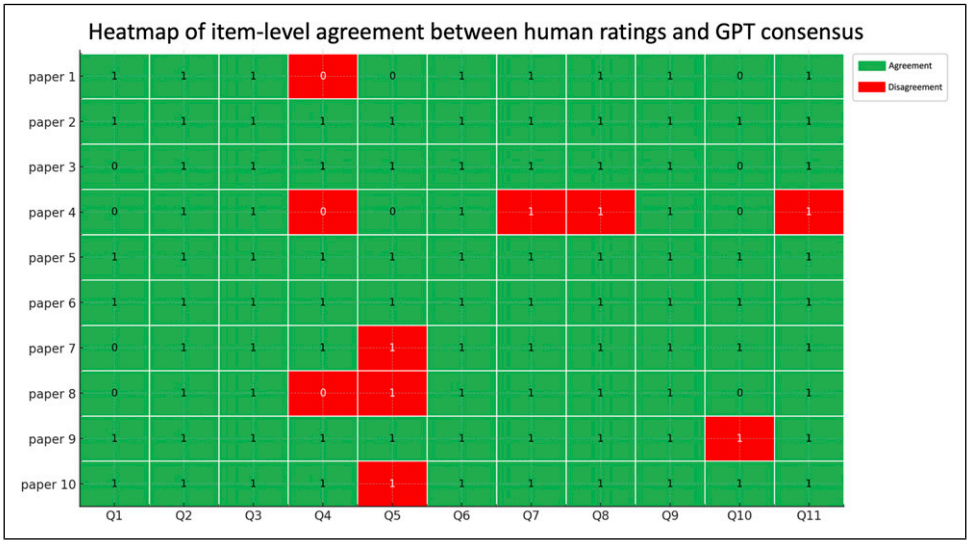


Figure 4. Heatmap of item-level agreement between human ratings and the GPT Consensus model across 10 systematic reviews using the AMSTAR 1 checklist. Green indicates agreement; red indicates disagreement.

discrepancies, as funding and disclosure statements were often ambiguously phrased or relegated to appendices, making consistent identification difficult for the models.

These patterns align with known characteristics of LLM reasoning: strong performance in extracting explicit textual information, but diminished accuracy when dealing with implicit content, judgment-based evaluation, or non-standardized phrasing.

Comparative model behavior

Although the overall agreement rates across models were similar, subtle behavioral differences were noted. Among them, the GPT Consensus Model exhibited slightly superior performance, suggesting potential utility in academic quality appraisal workflows. GPT-4o tended to provide more elaborate justifications, often flagging uncertainty when source material was ambiguous. In contrast, the GPT 5 adopted a more categorical approach, issuing firm binary judgments even in borderline cases. The GPT Consensus Model, optimized for academic applications, demonstrated a more conservative assessment style, which may have contributed to its slightly higher alignment with expert judgments.

Table 1. The diagnostic performance of GPT4o, Consensus, and GPT 5.0.

Model	GPT 4.o	GPT consensus	GPT 5.0
Sensitivity	95.70%	96.77%	90.32%
Specificity	41.18%	58.82%	82.35%
Accuracy	87.27%	90.91%	89.09%
Cohen's kappa (κ)	0.431	0.615	0.635
Interpretation	Moderate	Substantial	Substantial

Discussion

This study evaluated the performance of three large language models (LLMs)—GPT-4o, GPT 5, and GPT Consensus—in conducting AMSTAR 1-based methodological quality assessment across systematic reviews in orthopedic sports medicine. All models demonstrated strong alignment with human expert ratings, with agreement ranging from 87% to 90%. These findings underscore the emerging utility of LLMs in evidence synthesis and methodological appraisal, particularly within high-output clinical fields such as orthopedics.

Our results represent a notable advancement over earlier work. For example, Barsby et al. (2025) reported a broader range of accuracy (39%–93%) in earlier iterations of GPT-3.5 and GPT-4, suggesting inconsistency in earlier model performance (Barsby & Hume, 2025). In contrast, the present study demonstrated a tighter performance band (84%–90%), likely reflecting improvements in model architecture, fine-tuning, and reasoning capabilities. As newer versions such as GPT-5.2 are released with optimizations for scientific reasoning, we anticipate further gains in accuracy and reliability. These technological trends suggest a promising trajectory for integrating LLMs into systematic review workflows.

LLMs as effective aids in literature methodological quality assessment

Our findings reinforce growing consensus around the potential of LLMs to enhance and expedite methodological quality assessments. Nashwan and Jaradat highlighted that LLMs could streamline systematic reviews by enabling consistent and rapid quality appraisal when paired with

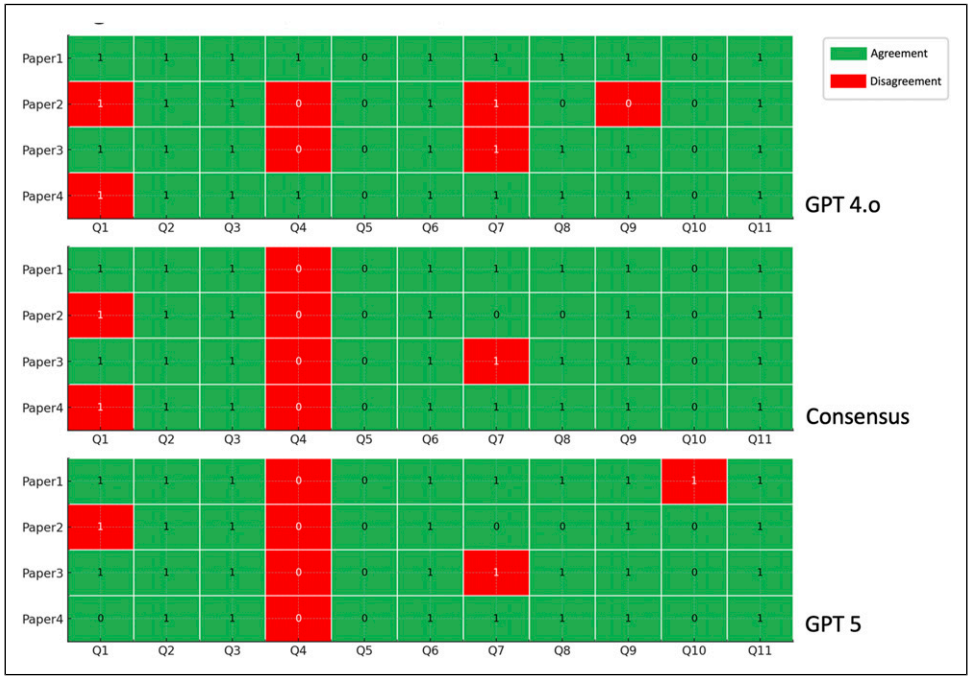


Figure 5. External validation heatmaps showing item-level agreement between human ratings and each language model (GPT-4o, GPT Consensus, and GPT 5) across four systematic reviews using the AMSTAR 1 checklist. Each row represents one paper, and each column corresponds to an AMSTAR item (Q1–Q11). Green indicates agreement; red indicates disagreement. All three models demonstrated high consistency, with most disagreements occurring in less explicitly reported domains (e.g., Q1, Q4, Q9, and Q10).

human oversight.[13] Hasan et al. further proposed a framework for GPT-4 integration, reporting the highest model-human agreement in domains with objectively defined methodological quality assessment criteria.[14] Our findings echo these observations, particularly regarding domain-specific performance.

Domain-level agreement: Structured versus subjective items

The highest agreement was observed in AMSTAR items involving structured, clearly defined reporting—specifically, a priori design (Item 1), comprehensive literature search (Item 3), and reporting of study characteristics (Item 6). These items were frequently well-articulated in the reviews, making them more accessible for LLM interpretation. Prior evaluations, such as those conducted with Claude, have shown similar strengths in structured domains within methodological quality assessment assessments for randomized trials.[5]

Conversely, items requiring nuanced interpretation—such as grey literature inclusion (Item 4), publication bias assessment (Item 10), and conflict of interest disclosure (Item 11)—were more challenging. These items often involve subtle language cues or are embedded in appendices, leading to lower model agreement. This trend mirrors findings by Suster et al., who reported weaker LLM performance in bias domains that require judgment-based reasoning or inference.[6]

Model behavior and relative performance

Although all three models performed similarly overall, differences in behavior were apparent. GPT-4o produced more elaborate explanations but occasionally defaulted to uncertainty in ambiguous contexts. The GPT 5 Model offered assertive judgments across domains but showed greater variability when interpreting subjective criteria. The GPT Consensus Model, designed for academic synthesis, adopted a more conservative tone, which likely contributed to its slightly higher alignment with expert evaluations.

These patterns are consistent with prior research by Wang et al.,[15] who found that transformer-based models such as BERT performed well in structured preclinical methodological quality assessment but struggled with unstructured narrative inputs.[10]

Broader context in LLM-Based methodological quality assessment research

The results align with earlier machine learning research on methodological quality assessment. Marshall et al. pioneered the use of distant supervision to automate Cochrane-style methodological quality assessment, though these early

models lacked the generative capabilities of modern LLMs.[16] More recent discussions, such as those by Zeng, have shifted toward ethical considerations and the interpretability of LLMs in critical evidence tasks.[17] These perspectives highlight the importance of transparency and accountability in deploying LLMs for sensitive roles in evidence synthesis. Similarly, Viswanathan et al. emphasized the complexity and reproducibility challenges in methodological quality evaluations—areas where LLMs could offer standardization, if appropriately guided.[18]

Limitations

Several limitations must be acknowledged. The dataset used was relatively small, consisting of only 10 core papers and four additional validation reviews, all from a single orthopedic subspecialty. This constrains generalizability across broader clinical domains. Additionally, we deliberately avoided using structured prompts to simulate naturalistic use cases; however, this may have introduced variability in LLM responses, a concern highlighted by Barsby et al.[19]

Taken together, these findings indicate that while LLMs are not yet substitutes for human experts, they hold significant potential to augment systematic review workflows—especially in structured appraisal tasks such as AMSTAR-based methodological quality assessment.

Nonetheless, it is worth noting that large language models such as GPT are continuously evolving. With the ongoing release of more capable versions (e.g., GPT-5.2 and emerging reasoning models), we can reasonably expect improvements in both the accuracy and stability of automated methodological quality assessment. While current limitations persist, the trajectory of model development suggests that many of these challenges may diminish over time.

Conclusion

This study demonstrated that large language models—GPT-4o, GPT 5, and GPT Consensus—can reliably perform AMSTAR 1-based methodological quality assessment in systematic reviews of orthopedic interventions. While performance was strongest in domains characterized by clear and structured reporting, limitations persisted in more inferential or judgment-based domains. Although LLMs are not yet suitable as standalone reviewers, their utility as adjunct tools are evident. They offer substantial promise in enhancing the speed, consistency, and accessibility of methodological quality evaluations—particularly in high-volume fields like orthopedics.

Future research should focus on improving prompt standardization, expanding domain coverage, and developing integration frameworks that combine LLM efficiency with human oversight. Ultimately, while LLMs can democratize access to rigorous evidence synthesis, their adoption must be paired with transparency, caution, and continued reliance on expert judgment.

Author's note

Ying-Chen Kuo's current institutions, School of Medicine, Fu Jen Catholic University, New Taipei City, Taiwan; Graduate Institute of Sports Science, National Taiwan Sport University, Taoyuan, Taiwan.

ORCID iD

Yu-Jui Huang https://orcid.org/0000-0003-1662-155X

Ethical considerations

Retrospective testing of already existing specimen is already covered by an ethics vote.

Author contributions

YJ-H and KC-C wrote the manuscript. YC-K, CC-T performed the statistics. CP-Y reviewed the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Each author certifies that he has no commercial associations.

Data Availability Statement

The dataset supporting the conclusions of this article is available from the corresponding author on reasonable request.

References

1. Gagnier JJ and Kellam PJ. Reporting and methodological quality of systematic reviews in the orthopaedic literature. *J Bone Joint Surg Am* 2013; 95: e771–e777. https://doi.org/10.2106/jbjs.L.00597
2. DiSilvestro KJ, Tjoumakaris FP, Maltenfort MG, et al. Systematic reviews in sports medicine. *Am J Sports Med* 2016; 44: 533–538. https://doi.org/10.1177/0363546515580290
3. Yamamoto N, Taito S, Miura T, et al. Epidemiology and reporting characteristics of systematic reviews in orthopedic journals: a meta-epidemiological study. *J Clin Med* 2023; 12: 20231110. https://doi.org/10.3390/jcm12227031
4. Lee D, Paul R, Lencer A, et al. Declining quality of systematic reviews in orthopaedic sports medicine: an updated systematic review. *Arthrosc Sports Med Rehabil* 2022; 4: e789–e795. https://doi.org/10.1016/j.asmr.2021.11.013
5. Scherbakov D, Hubig N, Jansari V, et al. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *J Am Med Inform Assoc* 2025; 32: 1071–1086. https://doi.org/10.1093/jamia/ocaf063

6. Forero DA, Abreu SE, Tovar BE, et al. Automated analyses of risk of bias and critical appraisal of systematic reviews (ROBIS and AMSTAR 2): a comparison of the performance of 4 large language models. *J Am Med Inform Assoc* 2025; 32: 1471–1476. https://doi.org/10.1093/jamia/ocaf117
7. Li Y, Yang F, Wu M, et al. Human-AI collaboration enhances the performance of large language models in risk of bias assessment. *BMC Med Res Methodol* 2026; 26: 37. https://doi.org/10.1186/s12874-025-02763-3
8. Eisele-Metzger A, Lieberum J-L, Toews M, et al. Exploring the potential of Claude 2 for risk of bias assessment: using a large language model to assess randomized controlled trials with RoB 2. *Res Synth Methods* 2025; 16: 491–508. https://doi.org/10.1017/rsm.2025.12
9. Šuster S, Baldwin T and Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Res Synth Methods* 2024; 15: 988–1000. https://doi.org/10.1002/jrsm.1749
10. Blom AW, Donovan RL, Beswick AD, et al. Common elective orthopaedic procedures and their clinical effectiveness: umbrella review of level 1 evidence. *Bmj* 2021; 374: n1511. https://doi.org/10.1136/bmj.n1511
11. Su H, Sun Y, Li R, et al. Large language models in medical diagnostics: scoping review with bibliometric analysis. *J Med Internet Res* 2025; 27: e72062. https://doi.org/10.2196/72062
12. Maity S and Saikia MJ. Large language models in healthcare and medical applications: a review. *Bioengineering (Basel)* 2025; 12: 631. https://doi.org/10.3390/bioengineering12060631
13. Nashwan AJ and Jaradat JH. Streamlining systematic reviews: harnessing large language models for quality assessment and risk-of-bias evaluation. *Cureus* 2023; 15: e43023. https://doi.org/10.7759/cureus.43023
14. Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med* 2024; 29: 394–398. https://doi.org/10.1136/bmjebm-2023-112597
15. Wang Q, Liao J, Lapata M, et al. Risk of bias assessment in preclinical literature using natural language processing. *Res Synth Methods* 2022; 13: 368–380. https://doi.org/10.1002/jrsm.1533
16. Marshall IJ, Kuiper J and Wallace BC. Automating risk of bias assessment for clinical trials. *IEEE J Biomed Health Inform* 2015; 19: 1406–1412. https://doi.org/10.1109/jbhi.2015.2431314
17. Zeng Y, Kairong L, Dong F, et al. Quantifying risk propensities of large language models: ethical focus and bias detection through role-play. *arXiv preprint arXiv:241108884* 2024.
18. Viswanathan M, Patnode CD, Berkman ND, et al. Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. *J Clin Epidemiol* 2018; 97: 26–34. https://doi.org/10.1016/j.jclinepi.2017.12.004
19. Barsby J, Hume S, Lemmey HA, et al. Pilot study on large language models for risk-of-bias assessments in systematic reviews: A(I) new type of bias? *BMJ Evid Based Med* 2025; 30: 71–74. https://doi.org/10.1136/bmjebm-2024-112990