Assessing the methodologic quality of systematic reviews using generative large language models

Bowen Yao[1,2], Onuralp Ergun[1,2], Maylynn Ding[2], Carly D. Miller[1], Vikram M. Narayan[3], Philipp Dahm[1,2]

[1]Minneapolis VA Healthcare System, Minneapolis, MN, United States; [2]Department of Urology, University of Minnesota, Minneapolis, MN, United States; [3]Department of Urology, Emory University, Atlanta, GA, United States

Cite as: Yao B, Ergun O, Ding M, et al. Assessing the methodologic quality of systematic reviews using generative large language models. *Can Urol Assoc J* 2025 August 28; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.9243

Published online August 28, 2025

Corresponding author: Dr. Philipp Dahm, Department of Urology, University of Minnesota, Minneapolis, MN, United States; pdahm@umn.edu

***

ABSTRACT

Introduction: We aimed to evaluate whether generative large language models (LLMs) can accurately assess the methodologic quality of systematic reviews (SRs).

Methods: A total of 114 SRs from five leading urology journals were included in the study. Human reviewers graded each of the SRs in duplicates, with differences adjudicated by a third expert. We created a customized GPT "Urology AMSTAR 2 Quality Assessor" and graded the 114 SRs in three iterations using a zero-shot method. We performed an enhanced trial focusing on critical criteria by giving GPT detailed, step-by-step instructions for each of the SRs using chain-of-thought method. Accuracy, sensitivity, specificity, and F1 score for each GPT trial was calculated against human results. Internal validity among three trials were computed.

Results: GPT had an overall congruence of 75%, with 77% in critical criteria and 73% in non-critical criteria when compared to human results. The average F1 score was 0.66. There was a high internal validity at 85% among three iterations. GPT accurately assigned 89% of studies

KEY MESSAGES

- Generative artificial intelligence (GPT) can reliably assess the methodologic quality of systematic reviews in urology, achieving high accuracy compared to human expert reviewers.
- Providing step-by-step instructions (chain-of-thought method) significantly improves GPT's accuracy and consistency.
- While AI substantially enhances efficiency in assessing systematic reviews, it should complement, not replace, expert human judgment.

into the correct overall category. When given specific, step-by-step instructions, congruence of critical criteria improved to 91%, and overall quality assessment accuracy to 93%.
Conclusions: GPT showed promising ability to efficiently and accurately assess the quality of SRs in urology.

INTRODUCTION

Systematic reviews (SRs) play a critical role in synthesizing research findings, informing clinical guidelines, and influencing patient care and treatment outcomes.[1] There is an increasing amount of SRs published every year that are often of low methodological quality.[2] Several tools have been developed to assess the methodological quality of SRs, including the A MeaSurement Tool to Assess systematic Reviews (AMSTAR) and its successor instrument, AMSTAR 2.[3] AMSTAR 2 has been shown to have moderate inter-rater reliability and is widely used to assess the quality of SRs.[4] It assesses 16 criteria of methodological quality and then rates the quality of a SR by 4 overall quality categories, namely 'high', 'moderate', 'low' and 'critically low'. This requires significant prior knowledge and experience in how to conduct SRs and is also time-consuming.[4-5] Recent advancements in artificial intelligence (AI), particularly in natural language processing (NLP), have opened new avenues in medical research. Large language models (LLMs) such as GPT, generative pre-trained transformer, have shown potential in various applications, including data analysis and interpretation.[6] These models can process large volumes of data rapidly, offering efficiency that could significantly benefit medical literature analysis.[7-9] There are several reported limitations to this method, including the lack of transparency in methodology, hallucination risk, cost of usage, accessibility and equity. [10-12]

The application of AI in SRs, especially in methodological assessment, is a relatively new area of exploration. Several studies have reported using LLMs to explore the feasibility of using these tools for literature screening and quality assessment in the process of generating a SR.[13-16] A formal assessment of the feasibility and utility of using LLMs to assess methodological quality of SRs using AMSTAR 2 or similar methods is currently lacking.

LLMs can be deployed using various prompting strategies which influence their reasoning style and accuracy. One commonly used approach is the *zero-shot* method, in which the model is given a task without any prior examples, relying entirely on its pretrained knowledge to generate responses.[17] This approach mirrors real-world applications where users expect generalizable outputs from foundation models with no to minimal input customization. In contrast, the *chain-of-thought* method enhances model performance by guiding it through a structured, step-by-step reasoning process, closely mimicking human problem-solving workflows.[18] This technique has been shown to improve the model's logical coherence and accuracy, especially in complex, multi-step tasks such as mathematical problem-solving and clinical reasoning.[18] In the context of systematic review methodology appraisal, these approaches

allow researchers to compare the model's ability to evaluate studies based solely on general instruction versus a more guided, logic-driven assessment. Prior studies that investigated the use of LLM such as GPT mostly utilized the zero-shot method.[13-16] We aim to use both methods, and this comparison is particularly important given the structured nature of AMSTAR 2. Human reviewers often have to rely on an algorithmic tool to help decide the assessment for certain criteria, with the aid of the guidance document.[4-5] Clear, deliberate and sequential evaluation may help improve the assessment using AI in a similar fashion.

In this study we explore the ability of ChatGPT (Open AI, San Francisco, CA) to assess the methodological quality of SRs in urology. The goal is to determine the consistency of ChatGPT's outputs and congruence of ChatGPT's assessments with those made by human experts using the AMSTAR 2 criteria for each of the criteria and the overall quality assessment of the literature.

METHODS

This study utilized data from a previously published study.[19] In brief, in this protocol-driven study, we searched all SRs related to questions of therapy and prevention from five leading urology journals (BJU International, European Urology, The Journal of Urology, Urology, and World Journal of Urology), between Jan 1, 2019, and June 30, 2021. Human reviewers working in pairs (MD, JJ, and GAA) screened all articles and settled discrepancies by consultation with the senior expert (PD). Data extraction was performed in duplicate in similar fashion, using a standardized form that was previously described that guides the reviewer to score the manuscript using the AMSTAR 2 criteria. Any differences between reviewers were discussed with and adjudicated by the senior expert (PD). Overall SR quality was assessed (critically low, low, moderate, high) for each of the included studies, based on results from seven critical and nine non-critical domains on AMSTAR 2.

In the current study, we used ChatGPT to perform independent scoring of included manuscripts using the AMSTAR 2 criteria. We utilized custom GPT builder and created a GPT 4 based tool, "Urology AMSTAR 2 Quality Assessor", using a zero-shot method. No prior information was given to the GPT builder from prior publication or human scoring results. We first provided "Urology AMSTAR 2 Quality Assessor" with the 16 criteria of AMSTAR 2, then fed the system a PDF file of each evaluable manuscript, and prompted the model to rate these SRs, in 3 separate iterations to assess for internal validity.

The last iteration of GPT grading was compared to human grading, and congruence, sensitivity, specificity, false positives, and false negatives were computed for each criterion. Chi square test was used to compare the overall compliance rate for each criterion between AI and human results. F1 score is a commonly used statistical test to evaluate the performance of artificial intelligence prediction models and was calculated for GPT grading using precision and recall. This metric was chosen because it balances both false positives and false negatives, which is particularly important in imbalanced datasets where certain criteria may be infrequently met. A

higher F1 score indicates better agreement between AI and human assessments across each individual criterion.

We calculated the positive and negative likelihood ratio using the sensitivity and specificity for each criterion, and using the pretest probability reported in our prior study, we calculated the post-test odds and probability for each criteria using the data from our initial trial. The overall quality category was calculated for each article based on the results from each iteration, and agreement was also calculated.

We then utilized chain-of-thought method of learning for each criterion similar to the steps that human reviewers used with the standardized form (see supplemental material for details). This was performed for each of the crucial criteria for all included studies.

All data analysis were performed using R (R Foundation for Statistical Computing, Vienna, Austria), Matlab (The Mathworks, Natick, MA) and Excel (Microsoft Corporation, Redmond, WA).

RESULTS

The prior study identified 563 references from 2019 to 2021 and ultimately included 114 studies unique SRs (please see prior study for details). As discussed previously, only 6 (2.3%) and 9 (3.5%) SRs, achieved a “high” (no critical weakness; up to one non-critical weakness) or “moderate” (no critical weakness; more than one non-critical weakness) confidence rating in overall quality.

Initial AI assessment of methodological quality

Each of the 16 AMSTAR criteria, and the results from the initial AI assessment are shown in Table 1. Similar rates of agreement were seen for criterion 1 (PICO), 7 (excluded studies details), 9 (RoB assessment), 10 (source of funding), 12 (Impact of RoB on meta-analysis), 13 (account for RoB when interpreting), 14 (discuss heterogeneity) and statistically significant variation was seen for the other criteria using chi square test.

AI assessment had difference of greater than 25% in compliance rate compared to human results for criterion 5 (duplicate study screening), 8 (describe study details), and 11 (appropriate statistical combination of results). AI were more lenient with criterion 5 and criterion 11, while being more strict with criterion 8.

When measuring the congruence of AI assessment compared to human results as gold standard, the results differed for each criterion, ranging from a moderate 50% to 96%. The average congruence for all 16 criteria was 75%. The critical domains had congruence of 77%, and non-critical domains had 73%. (Table 2)

Each criterion had different patterns of sensitivity, specificity and congruence. Criterion 1 (PICO), 7 (Excluded studies details), 10 (Source of funding), 13 (Account for RoB when interpreting) had congruence greater than 80%. Meanwhile Criterion 8 (Included studies details) had congruence lower than 60%. The F1 score for prediction of each criterion also had a wide

range from 0.14 to 0.97, with a mean of 0.66, which is expected given the wide range of prevalence with different criteria.

Internal validity of AI generated Answers

When the same 114 studies were given to GPT in 3 separate iterations, there were some variances as expected from the randomness of the generative AI algorithms. Internal agreement is defined as all 3 iterations resulting in the same evaluation of the criterion for the specific study. The mean internal agreement for all criteria was 85%. Only Criterion 12 (Impact of RoB) and 15 (assessing publication bias) had agreements lower than 75%.

Overall quality assessment

GPT accurately assigned 89% of studies into the correct overall category (high, moderate, low, and critically low) compared to human assessment. GPT assigned a higher overall category to 3.5% of studies compared to human grading, and a lower overall category to 7% of all studies.

Results from chain-of-thought method

When given specific instructions similar to human workflow, GPT provided results with better congruence in several crucial criteria. (Table 3) The congruence for crucial criteria ranged from 88.9% to 95.6%, averaging 92.7%. Using the new assessments, GPT accurately categorized 93.0% of studies into the right overall quality. The results showed excellent sensitivity and specificity, as well as positive and negative predictive values. The average F1 score was 0.898.

Pre and post test probabilities

The pre and post-test probabilities for critical and non-critical criteria using the prediction models are demonstrated in Figures 1 and 2. Using the zero shot GPT based prediction model, we were able to demonstrate an average negative post-test probability of 0.079, and positive post-test probability of 0.882 for critical criteria. For non-critical criteria the model achieved negative post-test probability of 0.095 and positive post-test probability of 0.876. Only 4 out of 16 criteria (8, 11, 12, 15) had a negative post-test probability of greater than 10%. Only Criteria 11 and 12 had positive post-test probability of less than 80%.

DISCUSSION

Statement of principal findings

In this study, we present an innovative exploration of the use of generative AI to assess the methodological quality of SRs in urology. The results indicate promising capabilities of AI to efficiently and accurately evaluate the methodological quality of SRs. When given only the 16 AMSTAR 2 criteria, using zero-shot method, GPT performed moderately well assessing the quality of SRs. This was significantly enhanced when giving specific instructions and streamlining the workflow to that that of human reviewers in a chain-of-thought fashion. The overall quality assessment was accurate.

Strengths and weaknesses of the study

This AI-focused study builds on our research team's extensive experience in longitudinally assessing the methodological quality of SRs published in core urology journals. This allowed us to ensure that each of the 16 AMSTAR criteria were interpreted consistently and provided us a set of pre-test probabilities to apply our findings to. All human ratings, which served as the reference (or "gold") standard were established through consensus by two independent reviewers.

There are limitations to this study. First, its focus on urologic literature only; while we would expect that GPT would also perform well when applied to a set of urology-related SR published in other journals with higher or lower methodological reporting standards, this needs to be demonstrated in future studies. Second, it is important to note that the congruence between GPT and human results varies considerably by criterion. For example, GPT performed well in determining whether there was an *a priori*, registered protocol, most commonly in PROSPERO, a free, web-based registry specifically for SR, since it was able to search for specific keywords. Meanwhile, it was challenged in identifying information embedded within figures or supplemental documents, which was relevant to criterion 7 which relates to transparency in which studies were excluded at the full-text stage and why. Third, this study focuses exclusively on the methodological quality of how these SRs were conducted; we recognize that there are many other issues why SRs may not provide trustworthy results (such as poorly framed clinical questions that do not address the actual information need) or redundancy (referring to unnecessary duplication of published SR). Fourth, hallucination is a well described risk when using AI to interpret scientific findings. In our case, the risk was found to be moderately low with a high internal consistency of 85% across three iterations of using GPT. Lastly, we acknowledge that for the ease of presentation the focus of the pretest probabilities was on the point estimate only, without considering some degree of imprecision, in particular the 95% confidence intervals of the calculated LR+ and LR-.

Strengths and weaknesses in relation to other studies, discussing important differences in results

There are no published studies that have previously used large language models to assess the methodological quality of SRs. We are aware of a major project registered with the Open Science Framework with the title WISEST (WhIch Systematic Evidence Synthesis is best).[20] which aims to develop an automated clinical decision-support algorithm to choose amongst SRs on the same topic. The proposed tool uses both AMSTAR 2 as well as ROBIS (Risk Of Bias in Systematic Reviews) as another similar tool assessing SR quality and to date only preliminary data has been presented in abstract form.[21,22] Meanwhile, there are broad-based efforts underway to leverage the power of artificial intelligence in all aspects of SR production, including screening of the literature, data abstraction, risk of bias assessment and even analysis.[23] An artificial intelligence-based classifier for randomized controlled trials developed for the

Cochrane, which has a reputation for high quality, rigorous and trustworthy SRs was recently found to reduce manual study identification workload with a very low and acceptable risk of missing eligible RCTs.[24]

Meaning of the study: possible explanations and implications for clinicians and policymakers

Findings of this study have important implications for the future of SRs. Whereas in the past, highly trained human reviewers with appropriate expertise and experience would require 15 to 30 minutes to score SRs, GPT is able to complete this task in a matter of seconds.[25] The API and custom GPT function makes it easy to personalize the tool based on user preferences. By automating the evaluation process, AI could significantly enhance future reviewers' workflow when assessing the methodological quality of urologic SRs.

Potential users of such tools would include SR authors preparing their study for submission, journal editors screening submissions for methodological quality as well as guideline developers who frequently rely on published SRs. For example, the Canadian Urological Association (CUA) Guideline Committee that seeks to produce high-quality, evidence-based guidelines using the GRADE approach has limited resources and relies on published reviews which could be efficiently screened for quality using this tool.[26] Even more influential would be the use of these models by policy makers such as CMS contractors that also rely on published SRs to make influential determinations about coverage decisions.[27] Additionally, generative AI has the potential to enhance and supplement training for early career researchers, and clinicians, patients, and other persons that are not familiar with methodological assessment tools as an educational tool.

On the other hand, this study illuminates the limitations inherent in AI-driven assessments, such as the risk of inconsistency and the challenge of interpreting complex, nuanced methodological issues. The high congruence with human results does not imply perfect interpretation of methodological quality of every SR. These findings suggest that while AI can provide valuable support in SR quality assessment, it should be viewed as a complement to, rather than a replacement for, expert human judgment.

Cost and accessibility are other issues. Despite the publicity of "Urology AMSTAR 2 Quality Assessor," users need to subscribe to OpenAI's ChatGPT Plus plan to utilize the enhanced function. The API version of "Urology AMSTAR 2 Quality Assessor" costs about $10 per million input tokens, and $30 per million output tokens (each token is about 750 words).[28] For example, one round of going through the 114 studies in this project cost about $20. This can add up to a significant cost, especially when it is necessary to go through many publications. This reveals another potential barrier to quality tools, in addition to the "paywall" in scientific publishing.[29]

Since the conclusion of the training and testing of models used in current study, more LLMs have been made available, including LLaMa 4, Gemini 2.0, Deepseek, and GPT 4.5. They have varying efficacy and cost and could be utilized to develop similar models to those used in

this study. Whether using the same or different LLM, they can behave differently over time due to changes in the background algorithm.[30] The exact input and sequence of questions also impact the performance of the model, demonstrated by the improved results with our chain-of-thought model.[31]

Unanswered questions and future research

Looking forward, integrating AI into SR processes holds substantial promise for enhancing efficiency and possibly the quality of SRs. Future iterations of AI models, coupled with advancements in NLP and machine learning, could offer more nuanced and context-aware assessments. Additionally, exploring the integration of AI in other stages of SR preparation, such as literature searching and data extraction, could further revolutionize the field.

CONCLUSIONS

This study marks a significant step forward in the intersection of LLMs and scientific publishing, offering valuable insights into the capabilities and limitations of using AI to assess the methodological quality of SRs. GPT based custom algorithm showed promising ability to efficiently and accurately assess the quality of SRs in urology. As AI technology continues to evolve, its role in enhancing the rigor and efficiency of SRs warrants further exploration, with the ultimate aim of bolstering evidence-based medicine and improving patient care outcomes.

REFERENCES

1. Murad MH, Montori VM, Ioannidis JP, et al. How to read a systematic review and meta-analysis and apply the results to patient care: Users' guides to the medical literature. *JAMA* 2014;312:171-9. https://doi.org/10.1001/jama.2014.5559
2. Ioannidis JP. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q* 2016;94:485-514. https://doi.org/10.1111/1468-0009.12210
3. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 2017;358:j4008. https://doi.org/10.1136/bmj.j4008
4. Shea BJ, Hamel C, Wells GA, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009;62:1013-20. https://doi.org/10.1016/j.jclinepi.2008.10.009
5. Perry R, Whitmarsh A, Leach V, et al. A comparison of two assessment tools used in overviews of systematic reviews: ROBIS versus AMSTAR-2. *Syst Rev* 2021;10:273. https://doi.org/10.1186/s13643-021-01819-x
6. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *Int J Qual Methods* 2023;22:16094069231211248. https://doi.org/10.1177/16094069231211248
7. Shah R, Chircu A. IoT and AI in healthcare: A systematic literature review. *Issues Inf Syst* 2018;19:33-41. https://doi.org/10.48009/3_iis_2018_33-41
8. Ali O, Abdelbaki W, Shrestha A, et al. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *J Innov Knowl* 2023;8:100333. https://doi.org/10.1016/j.jik.2023.100333
9. Locke S, Bashall A, Al-Adely S, et al. Natural language processing in medicine: A review. *Trends Anaesth Crit Care* 2021;38:4-9. https://doi.org/10.1016/j.tacc.2021.02.007
10. Cheshire WP Jr. Loopthink: A limitation of medical artificial intelligence. *Ethics Med* 2017;33:7-12.
11. Younis HA, Eisa TAE, Nasser M, et al. A systematic review and meta-analysis of artificial intelligence tools in medicine and healthcare: Applications, considerations, limitations, motivation and challenges. *Diagnostics (Basel)* 2024;14:109. https://doi.org/10.3390/diagnostics14010109
12. Salvagno M, Taccone FS, Gerli AG. Artificial intelligence hallucinations. *Crit Care* 2023;27:180. https://doi.org/10.1186/s13054-023-04473-y
13. Rathi H, Malik A, Behera DC, et al. P21 A comparative analysis of large language models (LLM) utilised in systematic literature review. *Value Health* 2023;26:S6. https://doi.org/10.1016/j.jval.2023.09.030
14. Landschaft A, Antweiler D, Mackay S, et al. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. *Int J Med Inform* 2024;189:105531. https://doi.org/10.1016/j.ijmedinf.2024.105531
15. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: A comparative analysis. *Syst Rev* 2024;13:219. https://doi.org/10.1186/s13643-024-02609-x

16. Scherbakov D, Hubig N, Jansari V, et al. The emergence of large language models (LLM) as a tool in literature reviews: An LLM automated systematic review. *arXiv* 2024. https://doi.org/10.48550/arXiv.2409.04600
17. Xian Y, Schiele B, Akata Z. Zero-shot learning – the good, the bad and the ugly. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017:4582-91. https://doi.org/10.1109/CVPR.2017.328
18. Chain-of-thought prompting elicits reasoning in large language models. Accessed May 4, 2025. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C24&q=Chain-of-Thought+Prompting+Elicits+Reasoning+in+Large+Language+Models
19. Ding M, Johnson J, Ergun O, et al. Methodological quality of systematic reviews for questions of therapy and prevention published in the urological literature (2016–2021) fails to improve. *SIU J* 2023;4:415-22. https://doi.org/10.48083/WURA1857
20. Lunny C, Whitelaw S, Ferri N, et al. WISEST (WhIch Systematic Evidence Synthesis is besT) survey. *OSF*. Published September 3, 2023. Accessed April 6, 2025. https://osf.io/prac3
21. Buehn S, Mathes T, Prengel P, et al. The risk of bias in systematic reviews tool showed fair reliability and good construct validity. *J Clin Epidemiol* 2017;91:121-8. https://doi.org/10.1016/j.jclinepi.2017.06.019
22. Lunny C, Veroniki AA, Shea B, et al. Introduction to the WISEST (Which Systematic Evidence Synthesis is best) Project: Developing an automated clinical decision-support algorithm to choose amongst systematic review(s) on the same topic. In: *Cochrane Colloquium Abstracts*; 2023 Sep 4–6; London, UK.
23. Khalil H, Ameen D, Zarnegar A. Tools to support the automation of systematic reviews: A scoping review. *J Clin Epidemiol* 2022;144:22-42. https://doi.org/10.1016/j.jclinepi.2021.12.005
24. Thomas J, McDonald S, Noel-Storr A, et al. Machine learning reduced workload with minimal risk of missing studies: Development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol* 2021;133:140-51. https://doi.org/10.1016/j.jclinepi.2020.11.003
25. Bojcic R, Todoric M, Puljak L. Adopting AMSTAR 2 critical appraisal tool for systematic reviews: Speed of the tool uptake and barriers for its adoption. *BMC Med Res Methodol* 2022;22:104. https://doi.org/10.1186/s12874-022-01592-y
26. Ding M, Gandhi V, Gonzalez-Padilla DA, et al. Assessing the methodologic heterogeneity of Canadian Urological Association guidelines: Adoption of the GRADE approach (2018–2023). *Can Urol Assoc J* 2025;19(6). https://doi.org/10.5489/cuaj.8926
27. Dahm P, Oxman AD, Djulbegovic B, et al. Stakeholders apply the GRADE evidence-to-decision framework to facilitate coverage decisions. *J Clin Epidemiol* 2017;86:129-39. https://doi.org/10.1016/j.jclinepi.2017.02.019
28. OpenAI. ChatGPT (Mar 14 version) [large language model]. OpenAI. Published 2023. Accessed April 6, 2025. https://help.openai.com/en/articles/8554956-understanding-your-api-usage
29. James JE. Pirate open access as electronic civil disobedience: Is it ethical to breach the paywalls of monetized academic publishing? *J Assoc Inf Sci Technol* 2020;71:1500-4. https://doi.org/10.1002/asi.24351

30. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? *Harv Data Sci Rev* 2024;6(2). https://doi.org/10.1162/99608f92.5317da47
31. Leonardi F, Feldman P, Almeida M, et al. Contextual feature drift in large language models: An examination of adaptive retention across sequential inputs. *Preprint* 2024. https://doi.org/10.31219/osf.io/pu948

FIGURES AND TABLES

Figure 1. Pre and post-test probabilities of zero shot model for critical criteria. Black diamonds show the pretest probabilities of each criterion. Blue circles show the positive post-test probabilities, and red circles show the negative post-test probabilities. The blue and red arrows show the positive and negative likelihood ratios impact on post-test probabilities.

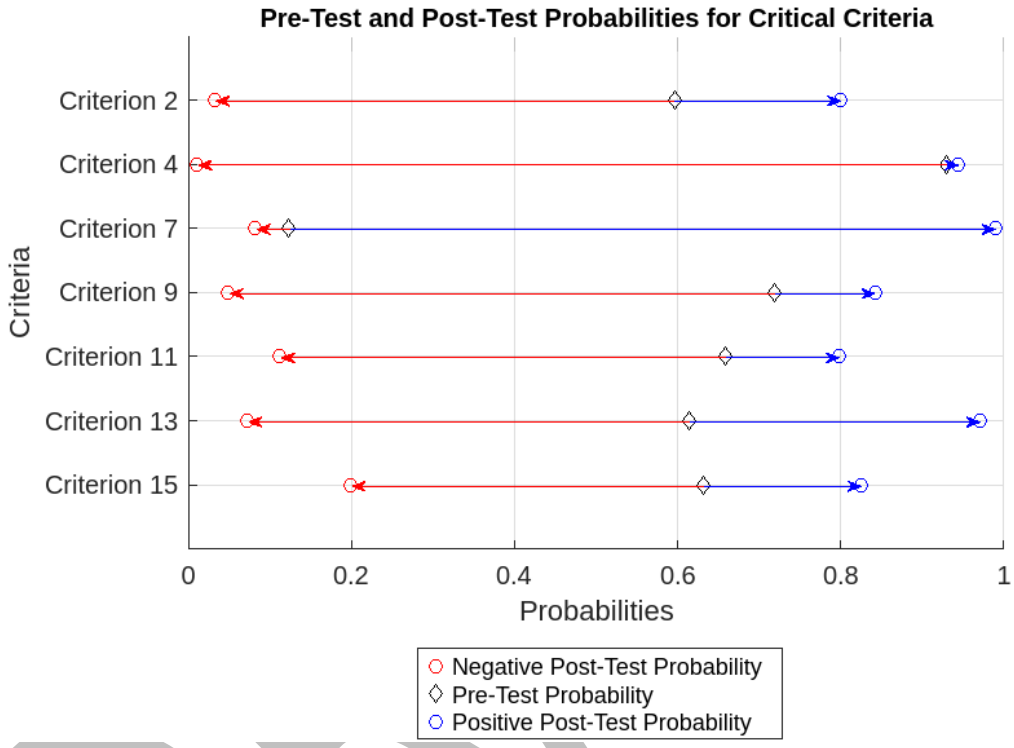

Figure 2. Pre and post-test probabilities of zero shot model for non-critical criteria. Black diamonds show the pretest probabilities of each criterion. Blue circles show the positive post-test probabilities, and red circles show the negative post-test probabilities. The blue and red arrows show the positive and negative likelihood ratios impact on post-test probabilities.

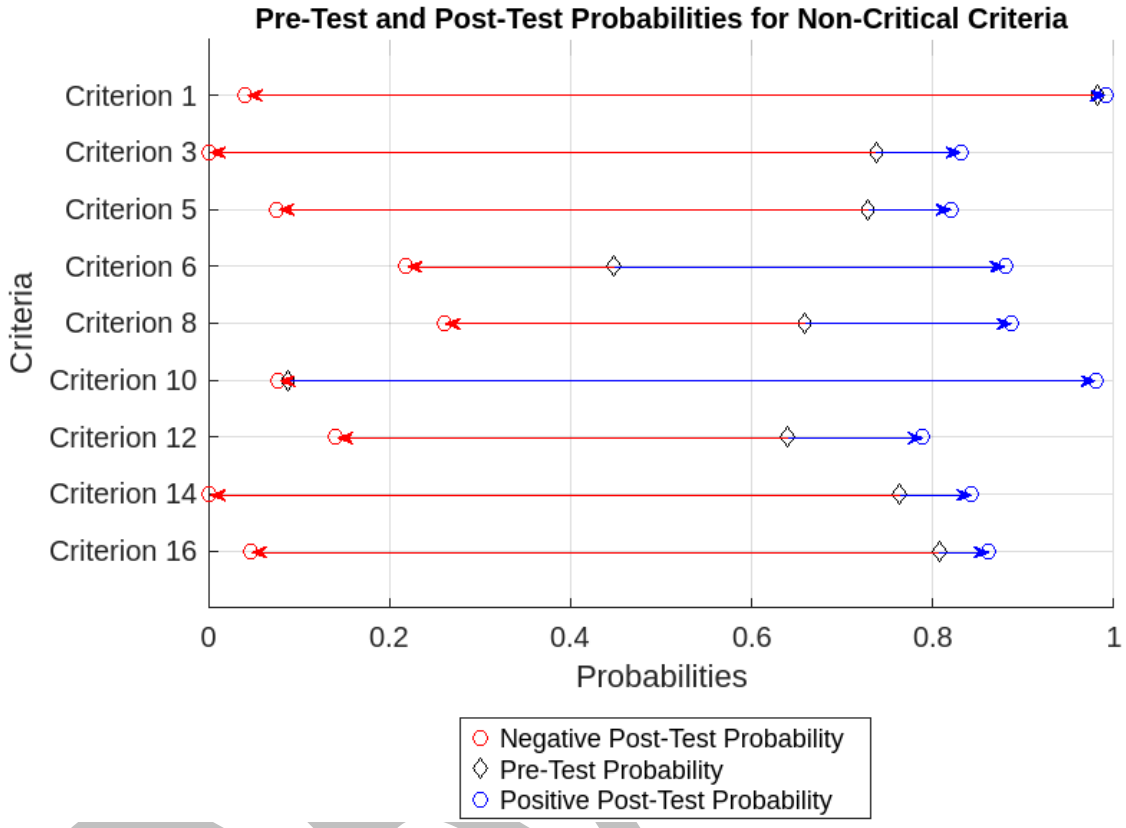

Table. 1 Description of each AMSTAR criterion, the keywords, and the percentage of compliance from current cohort with human and AI assessments, and p-value from Chi-squared test comparing compliance between human and AI assessments

AMSTAR 2 criteria	Keywords	Compliance human assessment	Compliance AI assessment	p
1. Did the research questions and inclusion criteria for the review include the components of PICO (participants, intervention, comparator, outcomes)?	PICO	0.991	0.982	0.563
2. Did the report of the review contain an explicit statement that the review methods were established prior to the conduct of the review and did the report	A priori Protocol	0.605	0.816	<0.001

justify any significant deviations from the protocol?				
3. Did the review authors explain their selection of the study designs for inclusion in the review?	Selection of study design	0.746	0.947	<0.001
4. Did the review authors use a comprehensive literature search strategy?	Search strategy	0.930	0.991	0.018
5. Did the review authors perform study selection in duplicate?	Duplicate screening	0.684	0.974	<0.001
6. Did the review authors perform data extraction in duplicate?	Duplicate data extraction	0.368	0.140	<0.001
7. Did the review authors provide a list of excluded studies and justify the exclusions?	Excluded studies details	0.105	0.211	0.059
8. Did the review authors describe the included studies in adequate detail?	Included studies details	0.605	0.263	<0.001
9. Did the review authors use a satisfactory technique for assessing the risk of bias (RoB) in individual studies that were included in the review?	RoB assessment	0.728	0.798	0.214
10. Did the review authors report on the sources of funding for the studies included in the review?	Source of funding	0.105	0.123	0.678
11. If meta-analysis was performed, did the review authors use appropriate methods for statistical combination of results?	Statistical combination of results	0.509	0.939	<0.001
12. If meta-analysis was performed, did the review authors assess the potential impact of RoB in individual studies on the results of the meta-analysis or other evidence synthesis?	Impact of RoB on meta-analysis	0.482	0.570	0.186
13. Did the review authors account for RoB in individual studies when	Account for RoB	0.237	0.254	0.759

interpreting/discussing the results of the review?	when interpreting			
14. Did the review authors provide a satisfactory explanation for, and discussion of, any heterogeneity observed in the results of the review?	Discuss heterogeneity	0.763	0.763	1.00
15. If they performed quantitative synthesis, did the review authors carry out an adequate investigation of publication bias (small study bias) and discuss its likely impact on the results of the review?	Assess publication bias	0.351	0.149	<0.001
16. Did the review authors report any potential sources of conflict of interest, including any funding they received for conducting the review?	Conflict of interest	0.789	0.947	<0.001

Table 2. Statistical performance of AI model when compared to human results, with sensitivity, specificity, congruence, F1 score using the 3rd iteration of zero shot, and internal agreements among three iterations

Criteria	Sensitivity	Specificity	Congruence	F1 score	Internal agreement
1	0.9576	0.9909	0.9462	0.9730	0.96
2	0.9750	0.7568	0.6022	0.7791	0.89
3	1.0000	0.7972	0.7097	0.8543	0.95
4	0.9907	0.9417	0.9247	0.9633	0.99
5	0.9348	0.7941	0.6667	0.8090	0.90
6	0.7500	0.8977	0.6129	0.2414	0.96
7	0.9113	0.9904	0.9032	0.1429	0.77
8	0.6792	0.9130	0.4516	0.3871	0.92
9	0.9574	0.8209	0.7312	0.8495	0.82
10	0.9180	0.9811	0.8817	0.2500	0.90
11	0.9022	0.7721	0.6237	0.7101	0.75
12	0.8737	0.7669	0.6022	0.6667	0.45
13	0.9250	0.9722	0.8925	0.7500	0.89
14	1.0000	0.8143	0.7419	0.8700	0.77

15	0.7934	0.8318	0.6237	0.4110	0.74
16	0.9592	0.8462	0.7742	0.8776	0.96

Table 3. Model performance using chain-of-thought methods, with congruence, sensitivity, specificity, F1 score

Criteria	Keywords	Congruence	Sensitivity	Specificity	F1
2	A priori Protocol	0.9211	0.9538	0.8776	0.9323
4	Search strategy	0.9298	0.9375	0.5000	0.9633
7	Excluded studies details	0.9298	1.0000	0.9259	0.6000
9	RoB assessment	0.8860	0.9600	0.7436	0.9172
11	Statistical combination of results	0.9474	0.9367	0.9714	0.9610
13	Account for RoB when interpreting	0.9211	0.9079	0.9737	0.9452
15	Assess publication bias	0.9561	0.9718	0.9302	0.9650