SWAR 68: AI-assisted quality appraisal with large language models

Objective of this SWAR
(1) To develop and propose a structured prompt specifically designed for applying AMSTAR in the quality appraisal of systematic reviews and meta-analyses.
(2) To conduct a dynamic evaluation of three LLMs, assessing their accuracy, efficiency and consistency in AMSTAR-based quality appraisal across multiple model versions and iterations.

Study area: Statistical Analysis, Quality appraisal
Sample type: Publications
Estimated funding level needed: Low

Background
Systematic reviews and meta-analyses are increasingly recognized as essential evidence sources for policy-making, clinical guidelines, and health technology assessments.[1,2] As their number has grown more than twenty-fold over the past two decades, the workload involved in identifying, appraising, and synthesizing eligible studies has increased substantially.[3,4] Among these tasks, methodological quality appraisal remains particularly critical and labor-intensive.[5]

To assess the methodological quality of systematic reviews and meta-analyses, standardized critical appraisal tools such as A MeaSurement Tool to Assess Systematic Reviews (AMSTAR) are widely used and have demonstrated good validity.[6,7,8] However, quality appraisal largely relies on subjective human judgment, making the process labor-intensive and potentially prone to bias.[9,10] Recent advances in large language models (LLMs) suggest strong potential for automating appraisal tasks, with emerging studies demonstrating feasibility in risk of bias and methodological assessments using established tools (e.g., Cochrane ROB tool, AMSTAR 2).[11,12,13]

LLMs show considerable potential for automating labour-intensive stages of evidence synthesis, thereby enhancing the efficiency, timeliness and accuracy of evidence for decision-making. However, robust comparative evidence across leading LLMs is limited, particularly regarding their performance with established appraisal tools such as AMSTAR. Systematic comparisons of widely used models (e.g., ChatGPT, Claude, DeepSeek) are therefore warranted, given their rapid evolution and increasing adoption in evidence workflows. Moreover, the lack of structured, adjudicated benchmarking frameworks and the fast pace of model development highlight the need for dynamic evaluation approaches to assess accuracy, efficiency and consistency over time.

This Study Within a Review (SWAR) [14] will use a dataset from six ongoing systematic reviews conducted by McMaster University and Lanzhou University, for which independent ratings from two human raters and a consensus rating are available, using the AMSTAR tool for quality appraisal. Prompts have been drafted using a pilot workflow with the following steps: (1) Parallel quality appraisal of the same studies by multiple LLMs using an initial draft prompt; (2) Comparison of LLMs outputs with the human consensus reference to identify discrepancies; and (3) iterative refinement and optimization of prompts based on identified differences until consistent AI outputs were obtained. The methodological quality appraisal will be conducted independently by three large language models. The appraisal results generated by the models will be extracted and standardized for subsequent analysis with a dynamic evaluation comparing outcomes across different models and across multiple iterations of each model over time. This will be followed by preparation of the final files containing the completed quality appraisal results.

Interventions and Comparators
Intervention 1: Data extraction using ChatGPT5.2.
Intervention 2: Data extraction using Claude 4 (Sonnet variants).
Intervention 3: Data extraction using DeepSeek V3.2.

Index Type: Quality appraisal

Method for Allocating to Intervention or Comparator:
Non-Random

Outcome Measures
Primary: Consistency (Kappa coefficient); and accuracy (Precision & Recall & F1 score).
Secondary: Efficiency (Time saved).

Analysis Plans
We will compare the results of AI-assisted quality appraisal with the adjudicated human reference standard and calculate accuracy, recall, precision, F1 score and time saved. All proportion estimates will be accompanied by 95% confidence intervals using the Kappa coefficient exact binomial method to obtain more robust estimates of binomial outcomes.

Possible Problems in Implementing This SWAR
None anticipated.

References

1. Moat KA, Lavis JN, Wilson MG, et al. Twelve myths about systematic reviews for health system policymaking rebutted. Journal of Health Services Research & Policy 2013;18(1):44-50.
2. Haby MM, Chapman E, Clark R, et al. What are the best methodologies for rapid reviews of the research evidence for evidence-informed decision making in health policy and practice: a rapid review. Health Research Policy and Systems 2016;14(1):83.
3. Hoffmann F, Allers K, Rombey T, et al. Nearly 80 systematic reviews were published each day: Observational study on trends in epidemiology and reporting over the years 2000-2019. Journal of Clinical Epidemiology 2021;138:1–11.
4. Marques-Cruz M, Pinto F, Vieira RJ, et al. Use of artificial intelligence to support the assessment of the methodological quality of systematic reviews. Journal of Clinical Epidemiology 2025;187:111944.
5. Gartlehner G, Kugley S, Crotty K, et al. Artificial Intelligence–Assisted Data Extraction With a Large Language Model: A Study Within Reviews. Annals of Internal Medicine 2025;178(12):1763-71.
6. Seehra J, Pandis N, Koletsi D, Fleming PS. Use of quality assessment tools in systematic reviews was varied and inconsistent. Journal of Clinical Epidemiology 2016;69:179–84.e5.
7. Stone JC, Barker TH, Aromataris E, et al. From critical appraisal to risk of bias assessment: clarifying the terminology for study evaluation in JBI systematic reviews. JBI Evidence Synthesis 2023;21:472-7.
8. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Medical Research Methodology 2007;7(1):10.
9. Zhou Y, Hu H. Can AI assess literature like experts? An entropy-based comparison of ChatGPT-4o, DeepSeek R1, and human ratings. Frontiers in Research Metrics and Analytics 2025;10:1684137.
10. Giummarra MJ, Lau G, Grant G, Gabbe BJ. A systematic review of the association between fault or blame-related attributions and procedures after transport injury and health and work-related outcomes. Accident Analysis and Prevention 2020;135:105333.
11. Lai H, Ge L, Sun M, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA network open 2024;7(5):e2412687.
12. Forero DA, Abreu SE, Tovar BE, Oermann MH. 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. Journal of the American Medical Informatics Association 2025;32(9):1471-6.
13. Zhou Y, Hu H. Can AI assess literature like experts? An entropy-based comparison of ChatGPT-4o, DeepSeek R1, and human ratings. Frontiers in Research Metrics and Analytics 2025;10:1684137.
14. Devane D, Burke NN, Treweek S, et al. Study within a review (SWAR). Journal of Evidence-Based Medicine 2022;15(4):328-32.

Publications or presentations of this SWAR design

Examples of the implementation of this SWAR

People to show as the source of this idea: Zixuan Gan
Contact email address: ganzx2024@lzu.edu.cn
Date of idea: 04/02/2026
Revisions made by: Zixuan Gan
Date of revisions: 12/02/2026