Research Synthesis Methods (2026), **17**: 1054–1056
doi:10.1017/rsm.2026.10073

Research Synthesis Methods

LETTER TO THE EDITOR

Five methodological considerations for validating LLMs in risk of bias assessment

Vihaan Sahu

Georgian National University SEU, GE
Email: vsahu@seu.edu.ge

Received: 22 December 2025; **Revised:** 22 December 2025; **Accepted:** 14 January 2026; published online 5 February 2026

To the Editors,

I read with interest the recent article by Eisele-Metzger and colleagues,[1] which provides a timely investigation into the use of the large language model (LLM) Claude 2 for assessing the risk of bias (RoB) in randomized controlled trials. The exploration of LLMs to address the substantial resource burden of systematic reviews is of great importance.[2] However, to ensure this line of inquiry yields robust and actionable insights, several methodological considerations warrant deeper discussion. I wish to raise five key critiques grounded in the broader literature on RoB reliability and LLM evaluation, which suggest opportunities to strengthen future validation studies and better define the practical role of LLMs in evidence synthesis.

1. Methodological concerns with practical implications: The imperfect reference standard

The study's use of existing Cochrane RoB assessments as the reference standard is pragmatic but problematic, as human assessments themselves show low reliability. Minozzi et al.[3] reported an overall inter-rater reliability (IRR) kappa of just 0.16 for RoB 2 among experienced reviewers, with domain-specific values ranging from slight ($\kappa = 0.04$) to moderate ($\kappa = 0.45$). Similarly, Armijo-Olivo et al.[4] found very poor agreement ($\kappa = 0.02$) between Cochrane and external reviewer panels. Consequently, when Eisele-Metzger et al. report a Cohen's κ of 0.22 between Claude and the Cochrane reference, it is unclear whether this reflects poor model performance or inherent noise in the "gold standard."[3,4] What proportion of disagreements stemmed from human inconsistency versus genuine model error? For review teams, this is not an academic concern; reconciling disagreements is a major time sink.[3,5] A more valid approach would compare the LLM against a consensus from a panel of experts using a prespecified, piloted protocol, as demonstrated in more rigorous LLM evaluations.[6] While the authors acknowledge the limitations of their reference standard in their discussion, this limitation could have been more thoroughly addressed in the study design through the use of a consensus-based reference standard.

2. Missed opportunity: LLMs as augmentation rather than replacement

While the authors rightly conclude that Claude cannot currently replace human assessors and appropriately recommend further investigation into "other models of support than providing stand-alone RoB judgements,"[1] a more immediate and practical application may be as an augmentation tool within a

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/rsm.2026.10073 Published online by Cambridge University Press

hybrid workflow. Gartlehner et al.[7] demonstrated that an LLM used for data extraction achieved 96.3% accuracy compared to humans, concluding that "integration of LLMs. . . should for now, only be done in the form of semi-automation." This aligns with evidence of the immense resource demands of reviews, which take a mean of 67 weeks and 5 team members to complete.[8] Building on the authors' opening for alternative support models, I encourage exploring Claude's output as a first-pass assessment, flagging trials or domains for focused human expert scrutiny. Could the authors conduct a secondary analysis simulating this workflow? How much human time could be saved if the LLM's judgments were used as one of two independent assessments, with humans resolving discrepancies?[7,8]

3. Information asymmetry and its impact on valid conclusions

The evaluation may have systematically disadvantaged the LLM due to residual information asymmetry. Although Claude was provided with compressed protocols or register entries, this process and the exclusion of supplementary materials may not equate to the comprehensive information synthesis performed by Cochrane reviewers, who routinely consult a wider array of sources in their original, uncompressed form.[9] This violates the principle of information parity crucial for fair comparison. Cierco Jimenez et al.[10] note that ML tool performance is heavily dependent on the information they can access, while the PRISMA statement emphasizes comprehensive information gathering as foundational for a valid review.[11] The qualitative differences in information access may have underestimated Claude's true capability. Did the authors audit which disagreements were attributable to information limitations? Future evaluations must ensure LLMs and human raters operate on identical information sets to draw valid conclusions about comparative performance.[9,11]

4. Inadequate exploration of performance discrepancies across studies and domains

The authors note variability in Claude's performance across studies but do not deeply investigate the causes. Heterogeneity in human–model agreement is well documented. Lai et al.[6] find LLM performance varied significantly by RoB domain, and Hartling et al.[12] show that agreement between human reviewers is influenced by trial characteristics (e.g., intervention type, funding source). Eisele-Metzger et al. report a domain-level kappa of 0.31 for "missing outcome data" but only 0.10 for "selection of the reported result."[1] What explains this threefold difference? Was performance poorer for industry-funded trials or subjective outcomes, as seen in human reliability studies?[12] A qualitative analysis of cases with high versus low agreement could reveal whether errors cluster around specific methodological complexities or reporting ambiguities, guiding targeted improvements in prompt engineering or model training.

5. Temporal validity and the rapid evolution of LLMs

The study evaluates a specific version of Claude 2, but LLM capabilities are evolving at an unprecedented pace. Singhal et al.[13] document rapid improvements in clinical knowledge encoding, while Chen et al.[14] demonstrate that GPT-3.5 and GPT-4 exhibited significant behavioral drift, sometimes improving, sometimes degrading over just a few months. This poses a major challenge for evaluation reproducibility.[14] Weidinger et al.[15] argue for frameworks for ongoing evaluation of AI systems. The findings of Eisele-Metzger et al. are thus a snapshot of a moving target. How can the field develop evaluation frameworks that remain relevant? I propose that validation studies incorporate mechanisms for continuous re-evaluation, report model version and date explicitly, and use benchmarks that can track progress or regression over time.[14,15]

6. Conclusion

The work by Eisele-Metzger et al.[1] is a valuable step in exploring AI for evidence synthesis. However, to translate this promise into reliable practice, the methodology must mature. Fundamental issues

https://doi.org/10.1017/rsm.2026.10073 Published online by Cambridge University Press

include the use of noisy human reference standards, a narrow focus on replacement over augmentation, information asymmetry in evaluations, and inadequate scrutiny of performance heterogeneity. Most critically, the rapid evolution of LLMs threatens the temporal validity of single-point assessments.

I propose a coordinated approach to LLM evaluation for RoB assessment with five components: (1) standardized benchmarks using expert consensus reference standards[3,6]; (2) evaluation frameworks that assess augmentation potential and workflow integration, not just replacement capability[7,8]; (3) protocols ensuring strict information parity between human and AI assessments[9–11]; (4) detailed, qualitative analyses of factors influencing performance heterogeneity across domains and trial types[6–12]; and (5) mechanisms for continuous evaluation to track performance as models evolve.[13–15]

The ultimate goal in evidence synthesis must be to leverage technology to produce high-quality, timely evidence for healthcare decisions. By addressing these methodological considerations, we can ensure that the integration of LLMs into systematic review methodology is both rigorous and transformative.

Author contributions. Conceptualization and writing—review and editing: V.S.

Competing interest statement. The author declares that no competing interests exist.

Data availability statement. None declared.

Funding statement. The author declares that no specific funding has been received for this article.

References

[1] 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(3): 491–508. https://doi.org/10.1017/rsm.2025.12

[2] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2): e012545.

[3] Minozzi S, Cinquini M, Gianola S, et al. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *J Clin Epidemiol*. 2020;126: 37–44.

[4] Armijo-Olivo S, Ospina M, da Costa BR, et al. Poor reliability between Cochrane reviewers and blinded external reviewers when applying the Cochrane risk of bias tool in physical therapy trials. *PLoS One*. 2014;9(5): e96920.

[5] Minozzi S, Dwan K, Borrelli F, et al. Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. *J Clin Epidemiol*. 2022;141: 99–105.

[6] Lai H, Ge L, Sun M, et al. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Netw Open*. 2024;7(5): e2412687.

[7] Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods*. 2024;15(4): 576–589.

[8] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7(2): e012545.

[9] Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc*. 2015;23(1): 193–201.

[10] Cierco Jimenez R, Lee T, Rosillo N, et al. Machine learning computational tools to assist the performance of systematic reviews: a mapping review. *BMC Med Res Methodol*. 2022;22(1): 322.

[11] Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *PLoS Med*. 2021;18(3): e1003583.

[12] Hartling L, Hamm MP, Milne A, et al. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs. *J Clin Epidemiol*. 2013;66(9): 973–981.

[13] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972): 172–180.

[14] Chen L, Zaharia M, Zou JY. How is ChatGPT's behavior changing over time? arXiv:2307.09009. Posted online July 18, 2023. https://arxiv.org/abs/2307.09009

[15] Weidinger L, Mellor JFJ, Rauh M, et al. Ethical and social risks of harm from language models. arXiv:2112.04359. Posted online December 8, 2021. https://arxiv.org/abs/2112.04359

Cite this article: Sahu, V. Five methodological considerations for validating LLMs in risk of bias assessment. *Research Synthesis Methods*. 2026;17: 1054–1056. https://doi.org/10.1017/rsm.2026.10073

https://doi.org/10.1017/rsm.2026.10073 Published online by Cambridge University Press