Research Synthesis Methods (2026), **17**: 1057–1059
doi:10.1017/rsm.2026.10102

Research Synthesis Methods

LETTER TO THE EDITOR

Response to: Five methodological considerations for validating LLMs in risk of bias assessment

Angelika Eisele-Metzger[1,2], Judith-Lisa Lieberum[3], Markus Toews[1], Waldemar Siemens[1], Felix Heilmeyer[4], Christian Haverkamp[4], Daniel Boehringer[3] and Joerg J. Meerpohl[1,2]

[1]Institute for Evidence in Medicine, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[2]Cochrane Germany, Cochrane Germany Foundation, Freiburg, Germany
[3]Eye Center, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany
[4]Institute for Digitization in Medicine, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany

Corresponding author: Angelika Eisele-Metzger; Email: angelika.eisele-metzger@uniklinik-freiburg.de

Received: 29 April 2026; **Revised:** 29 April 2026; **Accepted:** 12 May 2026; published online 8 June 2026

Dear Editors,

We thank Vihaan Sahu[1] for the thoughtful letter concerning our article "Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2."[2] Below, we respond to the methodological considerations raised in the letter.

1. Imperfect reference standard—human RoB 2 assessments

We acknowledge that using published Cochrane RoB 2 judgments as a reference standard has its limitations, given the modest inter-rater reliability reported for the tool.[3,4] However, this limitation was already discussed in our article. As we suggested, future work could employ a purpose-built expert reference standard (e.g., from a panel of experts), although this would also require considerably more time and resources. Given the high methodological quality of most Cochrane reviews, the assessments by various Cochrane authors were arguably best suited for our project to establishing an external and feasible reference standard.

2. Augmentation rather than replacement

We agree that a "human-in-the-loop" or semiautomated approach may currently be the most practical way to harness the judgments made by a large language model (LLM). Our study did not test such a workflow, but we welcome further investigations. Plausible scenarios include using an LLM for an initial, rough pre-assessment or for flagging potentially critical sections of a report, which are then verified by human reviewers. Likewise, an LLM could assist with extracting or highlighting data relevant to the risk of bias assessment, as suggested in our discussion.

© 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.2026.10102 Published online by Cambridge University Press

3. Information parity between model and human reviewers

In our study, the LLM received the full-text article together with a compressed protocol or registry entry (when available). Human Cochrane reviewers, in contrast, could consult any supplementary material, contact study authors, and examine complete protocols. This consequently resulted in a slight imbalance in the information available to LLMs and human reviewers. However, providing every piece of ancillary information for the 100 randomized controlled trials we examined would have been unfeasible. Moreover, as described in our article, Claude struggled with very long protocols. Therefore, we implemented a refined approach by supplying a compressed version to ensure that the model assessed the study itself rather than an extensive protocol.

4. Exploration of performance heterogeneity

Claude's agreement with human judgments varied across RoB 2 domains, that is, highest Cohen's κ for Domain 3 ("missing data"), followed by Domain 4 ("outcome measurement"), and lowest for Domain 5 ("selective reporting") in our main analysis. This suggests that there is likely some heterogeneity in performance across the various RoB-2 domains. As described in our article, we conducted several additional analyses to investigate possible influencing factors, including, e.g., sensitivity analyses using two alternative prompts and one alternative model (Claude 3), subgroup analyses by intervention type (pharmacological vs. other—nonpharmacological, nonsurgical—interventions), and subgroup analyses by the availability of a published protocol or registry entry. Of course, these were not exhaustive. Although κ values fluctuated modestly in our additional analyses, none indicated consistently strong performance. We would therefore urge caution and advise against placing overly strong emphasis on individual κ values.

5. Temporal validity and rapid model evolution

We share the concern that evaluations of single LLMs can become outdated quickly in this rapidly evolving field, underscoring the importance of approaches that support reproducibility and continuous reassessments. Our study, conducted in February 2024, contributes an early evaluation within this dynamic landscape. Building on this, we emphasize that future validation studies should be strengthened by incorporating plans for reevaluation and by transparently reporting key details such as model version, run date, prompts, and data used to facilitate replication studies. The RAISE (Responsible Use of AI in Evidence SynthEsis) Statement[5,6] provides valuable guidance in this regard. Complementary initiatives, including the Digital Evidence Synthesis Tool Evaluations[7] and our currently ongoing living scoping review,[8] are actively advancing the coordination and tracking of evaluation efforts.

We appreciate Sahu's insightful appraisal and his constructive proposals. We look forward to further constructive discourse and collaboration with the systematic review community.

Competing interest statement. The authors declare that no competing interests exist.

Disclosure of AI use. We used the following AI assistance to prepare this work: GPT-OSS 120b via Open WebUI as a formulation aid for writing the letter. The Open WebUI service is operated by the University of Freiburg's data center and meets the highest standards of data protection and IT security. All content generated with the assistance of AI has been critically reviewed by the authors for accuracy and amended where necessary.

Data availability statement. Data sharing is not applicable to this article.

Funding statement. The authors declare that no specific funding has been received for this article.

References

[1] Sahu V. Five methodological considerations for validating LLMs in risk of bias assessment. *Res Synth Methods*. 2026; 1–3. https://doi.org/10.1017/rsm.2026.10073.

[2] 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.2026.10102 Published online by Cambridge University Press

[3] Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. 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] Minozzi S, Dwan K, Borrelli F, Filippini G. 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.
[5] Flemyng E, Noel-Storr A, Macura B, et al. Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. *Cochrane Database Syst Rev*. 2025(10).
[6] Thomas J, Hair K, Noel-Storr A, et al. Responsible Use of AI in Evidence Synthesis (RAISE) 2: building and evaluating AI evidence synthesis tools. Open Science Framework; 2026. https://doi.org/10.17605/OSF.IO/FWAUD.
[7] Bond M, Finnerty A, O'Mara-Eves A, et al. Digital Evidence Synthesis Tool Evaluations. EPPI Visualiser Database; 2024. https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=435.
[8] Eisele-Metzger A, Bond M, Gopal I, et al. Large language models to support conducting systematic reviews: protocol for a living scoping review. Open Science Framework; 2026. https://doi.org/10.17605/OSF.IO/7QHM2.

Cite this article: Eisele-Metzger A, Lieberum J.-L., Toews M, Siemens W, Heilmeyer F, Haverkamp C, Boehringer D, Meerpohl JJ. Response to: Five methodological considerations for validating LLMs in risk of bias assessment. *Research Synthesis Methods*. 2026;17: 1057–1059. https://doi.org/10.1017/rsm.2026.10102

https://doi.org/10.1017/rsm.2026.10102 Published online by Cambridge University Press