Original Paper

Evaluating the Accuracy of Large Language Models in Risk-of-Bias Assessment Using Version 2 of the Cochrane Risk-of-Bias Tool for Randomized Trials: Exploratory Feasibility Study

Yu-Ju Lai[1,2], BS; Shen-Hua Lin[1], MS; Jen-Wei Liu[1], PhD

[1]Department of Pharmacy, Fu Jen Catholic University Hospital, Fu Jen Catholic University, New Taipei City, Taiwan
[2]Graduate Institute of Clinical Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

Corresponding Author:

Jen-Wei Liu, PhD
Department of Pharmacy
Fu Jen Catholic University Hospital, Fu Jen Catholic University
No.69, Guizi Rd., Taishan Dist.
New Taipei City 24352
Taiwan
Email: jerryljw@gmail.com

Abstract

Background: Large language models (LLMs) have the potential to improve the efficiency of evidence synthesis, but their reliability in performing complex tasks such as risk-of-bias (ROB) assessment in randomized controlled trials (RCTs) remains unclear.

Objective: This study aimed to evaluate whether LLMs can reliably assess ROB in RCTs using version 2 of the Cochrane ROB tool for randomized trials (ROB 2).

Methods: This study was conducted between December 28, 2024, and February 28, 2025, in adherence to American Association for Public Opinion Research reporting guidelines. Twenty-nine RCTs were selected from published Cochrane systematic reviews across diverse medical fields. We developed a structured prompt engineering framework that transformed ROB 2 decision trees into logical rules for the LLM. Each RCT was independently evaluated twice by ChatGPT, with Cochrane review authors' assessments serving as the reference standard for comparison. The main outcomes were the accuracy and consistency of ROB 2 assessments at both the domain and trial levels, evaluated using accuracy, sensitivity, specificity, and F_1-score. Consistency between the repeated assessments was quantified using the Cohen κ and prevalence-adjusted, bias-adjusted κ.

Results: The LLM demonstrated a moderate aggregate domain accuracy of 73.1% (95% CI 64.7%-81.5%) in the first assessment and 75.9% (95% CI 66.3%-85.4%) in the second assessment. Domain-averaged sensitivity decreased from 61.4% (95% CI 48.1%-74.7%) to 53.4% (95% CI 41.7%-65.0%), whereas domain-averaged specificity increased from 75.8% (95% CI 65.3%-86.3%) to 81.1% (95%CI 67.1%-95%), indicating a conservative tendency in identifying a high ROB. Domain-level accuracy ranged from 62.1% to 87.9%, with the lowest accuracy observed in domain 1 and the lowest F_1-scores observed in domain 2. Consistency between repeated assessments was high, with a mean agreement of 89.0% (SD 7.5%), and Cohen κ values were 0.86, 0.39, 0.56, 0.84, and 0.85 in domains 1 to 5, respectively.

Conclusions: In this exploratory study, ChatGPT demonstrated moderate accuracy and high consistency in assessing ROB in RCTs using the ROB 2. However, its reliability diminished in complex scenarios requiring interpretation of implicit narratives or behavioral nuance. These findings suggest that LLMs may support methodological evaluations in systematic reviews by acting as automated screeners to reduce reviewer burden, but current implementation still requires expert oversight, particularly for trials involving subjective outcomes or nonstandard reporting.

J Med Internet Res 2026;28:e84915; doi: 10.2196/84915

Keywords: large language models; LLMs; risk of bias; version 2 of the Cochrane risk-of-bias tool for randomized trials; ROB 2; ChatGPT; randomized controlled trial; RCT; systematic reviews

Introduction

Systematic reviews play a foundational role in evidence-based medicine by synthesizing findings from primary studies to inform clinical decision-making, guideline development, and health policy [1]. As the volume of biomedical literature continues to grow rapidly, the demand for efficient, high-quality evidence synthesis has become more urgent. Among the key components of systematic reviews is the assessment of the risk of bias (ROB), particularly in randomized controlled trials (RCTs), which serve as the primary source of evidence for many clinical guidelines. The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) framework emphasizes that the certainty of evidence is closely tied to the ROB in included studies within a systematic review [2].

To support rigorous and transparent ROB assessments, the Cochrane Collaboration developed version 2 of its ROB tool for randomized trials (ROB 2), which evaluates bias at the outcome level across 5 domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported results [3]. The tool relies on signaling questions and structured algorithms to facilitate reproducible judgments. Despite its methodological strengths, applying the ROB 2 remains time and resource intensive, requiring trained reviewers and substantial manual effort.

Large language models (LLMs), with their advanced capabilities in natural language understanding and reasoning, have been proposed as tools to automate complex tasks in medical text analysis [4]. Preliminary research suggests that LLMs may be able to interpret trial reports and mimic human judgment in various evaluative tasks. However, whether LLMs can perform structured ROB assessments in line with established tools such as the ROB 2 remains an open question [5]. To address this gap, we conducted a study to evaluate the feasibility, accuracy, and consistency of using LLMs to perform ROB assessments for RCTs guided by a structured prompt based on the ROB 2 framework.

Methods

Overview

This study was conducted between December 28, 2024, and February 28, 2025, in adherence to the American Association for Public Opinion Research reporting guidelines [6].

For this study, we used the GPT-o3-mini-high model via the web interface. OpenAI's GPT-o3-mini has been optimized for science, technology, engineering, and mathematics reasoning, with medium reasoning effort matching the performance of GPT-o1 in math, coding, and science while delivering faster responses [7]. Due to the technical constraints of the ChatGPT web interface, all assessments were performed using the model's default settings as manual adjustments to parameters such as temperature and top_p are not supported in this environment. This model was used to systematically assess the ROB based on the ROB 2 in RCTs using its reasoning capabilities to simulate the evaluation process conducted in systematic reviews.

Ethical Considerations

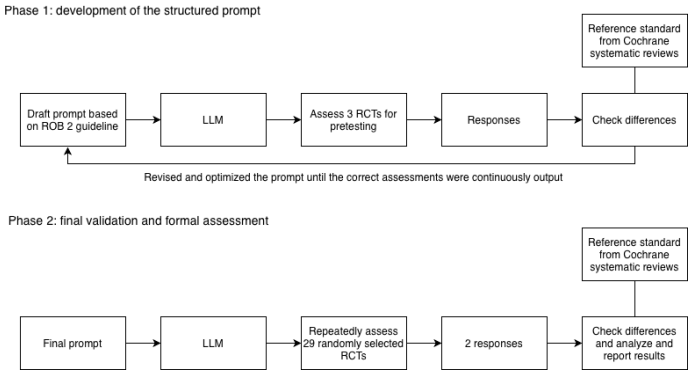
We used only publicly available published literature and LLMs without involving any personal or identifiable private data. This study followed the "Scope of Human Research Projects Exempt from Institutional Review Board Review" (1010265075) [8] issued by the Ministry of Health and Welfare, Taiwan. As this study relied solely on legally and publicly disclosed information, it was exempt from ethical review or informed consent requirements.

Model Training and Prompt Development

Prompt development followed the logic and structure outlined in the ROB 2 guidance document [9]. Systematic prompt engineering techniques were developed to create structured prompts, allowing ChatGPT to perform tasks.

We used 3 RCTs as pilot data, which were excluded from the final analytic sample. An iterative refinement process was used referencing published systematic reviews as a reference standard. On the basis of the framework by Lai et al [5], we defined the LLMs' character, provided foundational definitions of the ROB 2 domains, and systematically transformed the official ROB 2 decision trees into logical rules within the prompt. When model-generated ROB judgments differed from the original assessments, prompts were revised and tested until alignment with expert assessments was consistently achieved. Figure 1 shows the main study process, and the final prompt can be found in Multimedia Appendix 1.

Figure 1. Flow diagram of the main study process. LLM: large language model; RCT: randomized controlled trial; ROB 2: version 2 of the Cochrane risk-of-bias tool for randomized trials; SR: systematic review.



Selection of Sample

We conducted a search of the PubMed database to identify Cochrane systematic reviews that used the ROB 2 for assessing ROB in RCTs. Two reviewers (YJL and JWL) independently screened the full texts of the retrieved systematic reviews to determine eligibility. Reviews were excluded if they did not provide detailed ROB assessments, such as an ROB table. From the eligible 178 Cochrane systematic reviews, we first randomized the identified reviews using numbers generated via Google Sheets and sequentially screened them until a pooled baseline of over 100 RCTs was established. Second, using the same random number generation method, we selected 30 RCTs from this pool for analysis to mitigate potential clustering effects. These 30 RCTs were distributed across 11 reviews. Of the 30 RCTs, 1 (3.3%) was excluded due to incomplete reporting of ROB assessment, resulting in a final analytic sample of 29 (96.7%) RCTs. Notably, one of the included trials, that by Maher et al [10], was retracted after our study period [11]. Because its Cochrane assessment contributed to our reference standard, we retained this trial in the primary analytic sample but conducted a sensitivity check to evaluate its impact.

Application of ChatGPT for ROB Assessment

We used the GPT-o3-mini-high model for its strong reasoning capabilities, enabling consistent and accurate application of structured prompts in the ROB assessment process.

Each RCT was evaluated using a new ChatGPT conversation to ensure a clean contextual slate. The reviewers defined the primary outcome, provided domain-specific criteria, and uploaded the full trial text. ChatGPT then assessed ROB across the 5 domains in the ROB 2 framework, offering domain-level judgments with rationale (Multimedia Appendix 2). Each trial was assessed twice under identical conditions. In cases of system interruption, the session was discarded and repeated. A standardized interaction protocol was followed to maximize consistency and reproducibility.

Establishment of the Reference Standard

The criterion standard for comparison was the ROB assessment reported by the Cochrane systematic review authors using the ROB 2 (Table S1 in Multimedia Appendix 3). These Cochrane assessments were selected for their rigorous methodology and multidisciplinary collaboration, which minimize bias and ensure clinical diversity [12,13]. When apparent errors were identified (eg, typographical mistakes), authors were contacted for clarification. If no response was received, the original Cochrane systematic review assessments were retained as the reference standard to ensure objectivity. A sensitivity analysis was then performed to evaluate the impact of potential errors using data that were manually corrected based on research team consensus.

Discrepancy Categorization and Analysis

To evaluate the discrepancies between the LLM and the reference standard, we conducted an error analysis. Two researchers (YJL and JWL) independently categorized the discrepancies by comparing the LLM-generated rationales against the Cochrane reviewers' evidence. Any disagreements were resolved through consensus.

Discrepancies were categorized as data extraction differences if there was a fundamental mismatch in the explicit information or evidence retrieved from the RCT between the LLM and the Cochrane reviewers. Conversely, they were classified as judgment differences if the LLM accurately extracted the same key points as the human reviewers but applied a different logical interpretation leading to a conflicting judgment.

Statistical Analysis

The LLM was prompted to answer individual signaling questions using "yes," "probably yes," "no," "probably no," and "no information." Signaling responses were then operationally mapped to the official "low risk" and "high risk" categories following the ROB 2 guideline. ROB domain judgments were categorized as follows: "low risk" was defined as a negative outcome, whereas "some concerns" and "high risk" were grouped as a positive outcome. On the basis of this classification, we calculated true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

Model performance was evaluated using accuracy, sensitivity, specificity, precision, and F_1-score. For domain-level analyses, sensitivity and specificity were macroaveraged across domains. Metrics were defined as follows:

Accuracy = (TPs + TNs)/total number of assessments

Sensitivity = TPs/(TPs + FNs)

Specificity = TNs/(TNs + FPs)

Precision = TPs/(TPs + FPs)

F_1-score = 2 × (precision × sensitivity)/(precision + sensitivity)

To assess internal consistency, we computed the Cohen $\varkappa$ and the prevalence-adjusted, bias-adjusted (PABA) $\varkappa$:

Cohen $\varkappa = (P_o - P_e)/(1 - P_e)$

PABA $\varkappa = 2 \times P_o - 1$

In these equations, P_o is calculated as (number of agreements on positive + number of agreements on negative)/total number of assessments.

$P_e = [(P_1 \times P_2) + (N_1 \times N_2)]/(\text{total number of assessments})^2$

For trials in which the model provided identical ratings across both assessments, the Cohen $\varkappa$ was mathematically undefined due to zero variance, which resulted in a zero denominator in the $\varkappa$ calculation and was therefore reported as "Not available." The agreement thresholds in Table S2 in Multimedia Appendix 3 followed standard interpretations: $\varkappa$ values of 0.81 to 0.99 indicated near-perfect agreement. Additionally, exploratory subgroup analyses by clinical discipline were performed to investigate potential variability in model performance across different medical contexts. Trials were categorized based on the primary medical focus of the RCT and the scope of the parent Cochrane review from which the trial was extracted. In cases in which a trial potentially spanned multiple disciplines, the final categorization was determined through discussion and consensus between 2 authors (YJL and JWL). All statistical analyses were conducted using Google Sheets.

Results

Characteristics of the RCTs

The final dataset included 29 RCTs [10,14-41] extracted from 11 Cochrane systematic reviews [42-52]. These trials represented a spectrum of medical disciplines, including cardiology (n=5, 17.2%), psychology (n=6, 20.7%), infectious diseases (n=6, 20.7%), gastroenterology (n=5, 17.2%), pulmonology (n=4, 13.8%), bone health (n=1, 3.4%), and obstetrics and gynecology (n=2, 6.9%). Primary efficacy-related outcomes varied across studies, including binary outcomes (n=18, 62.1%), with most focusing on all-cause mortality and asthma exacerbation or treatment success. Continuous outcomes (n=11, 37.9%) mainly included pain, physical function, and quality of life. The trials were published between 2000 and 2024. All RCTs were published in English and assessed for ROB using the ROB 2, ensuring a standardized approach to bias evaluation.

Accuracy

Aggregate Domain and Overall Accuracy

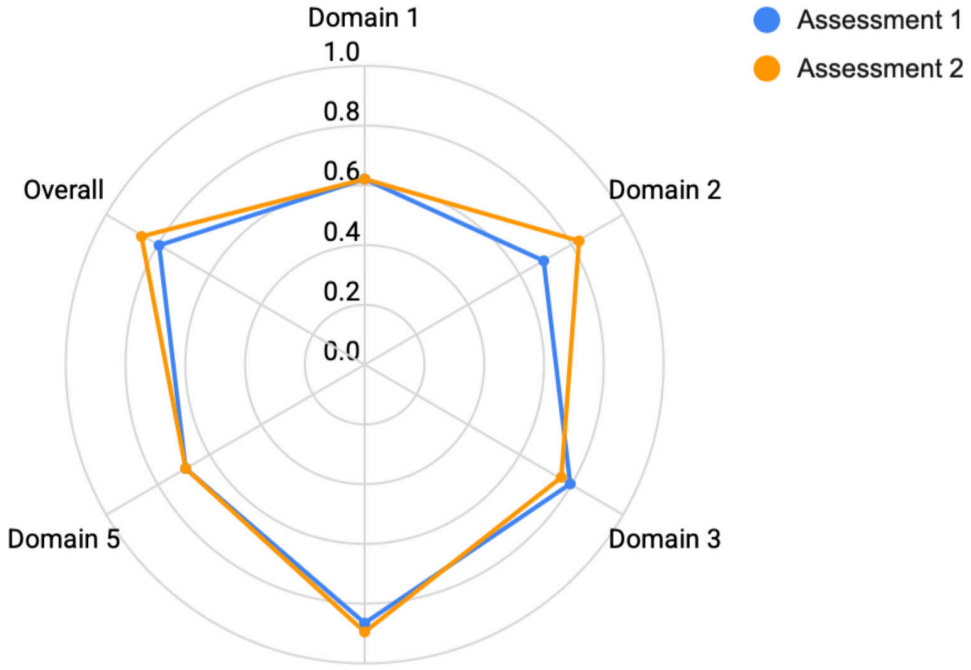
The LLM underwent 2 independent assessments for ROB evaluation, as shown in Table S3 in Multimedia Appendix 3. The aggregate domain accuracy (Table 1 and Figure 2) was comparable between assessments at 73.1% (95% CI 64.7%-81.5%) for the first and 75.9% (95% CI 66.3%-85.4%) for the second, reflecting a marginal improvement in accuracy in the second assessment (relative difference [RD] 2.8%, 95% CI –3.1% to 8.6%).

Table 1. Domain-specific accuracy of assessments.

	TPs[a], n	TNs[b], n	FPs[c], n	FNs[d], n	Accuracy, %	Sensitivity, %	Specificity, %	Precision, %	F_1-score, %
Domain 1									
Assessment 1	5	13	9	2	62.1	71.4	59.1	35.7	47.6
Assessment 2	5	13	9	2	62.1	71.4	59.1	35.7	47.6
Domain 2									
Assessment 1	3	17	5	4	69.0	42.9	77.3	37.5	40.0
Assessment 2	3	21	1	4	82.8	42.9	95.5	75.0	54.6
Domain 3									
Assessment 1	4	19	5	1	79.3	80.0	79.2	44.4	57.1
Assessment 2	2	20	4	3	75.9	40.0	83.3	33.3	36.4
Domain 4									
Assessment 1	2	23	2	2	86.2	50.0	92.0	50.0	50.0
Assessment 2	2	24	1	2	89.7	50.0	96.0	66.7	57.1
Domain 5									
Assessment 1	5	15	6	3	69.0	62.5	71.4	45.5	52.6
Assessment 2	5	15	6	3	69.0	62.5	71.4	45.5	52.6

[a]TP: true positive.
[b]TN: true negative.
[c]FP: false positive.
[d]FN: false negative.

Figure 2. Radar chart of accuracy in domains.



In contrast, the overall ROB accuracy, derived from the final overall ROB judgment for each trial, was 79.3% (23/29) in the first assessment and 86.2% (25/29) in the second.

Sensitivity, reflecting the ability to detect high-risk judgments, was higher for the first assessment at 61.4% (95% CI 48.1%-74.7%) compared to 53.4% (95% CI 41.7%-65.0%) in the second, suggesting slightly reduced effectiveness in TP identification by the second assessment (RD 8.0%, 95% CI –7.7% to 23.7%). Specificity remained high in both assessments, increasing from 75.8% (95% CI 65.3%-86.3%) in the first to 81.1% (95% CI 67.1%-95.0%) in the second, demonstrating strong performance in identifying low-risk judgments. F_1-scores were similar between assessments, with the first one slightly lower at 49.5% (95% CI 43.9%-55.1%) compared to 49.7% (95% CI 42.5%-56.9%) in the second.

Domain-Specific Accuracy

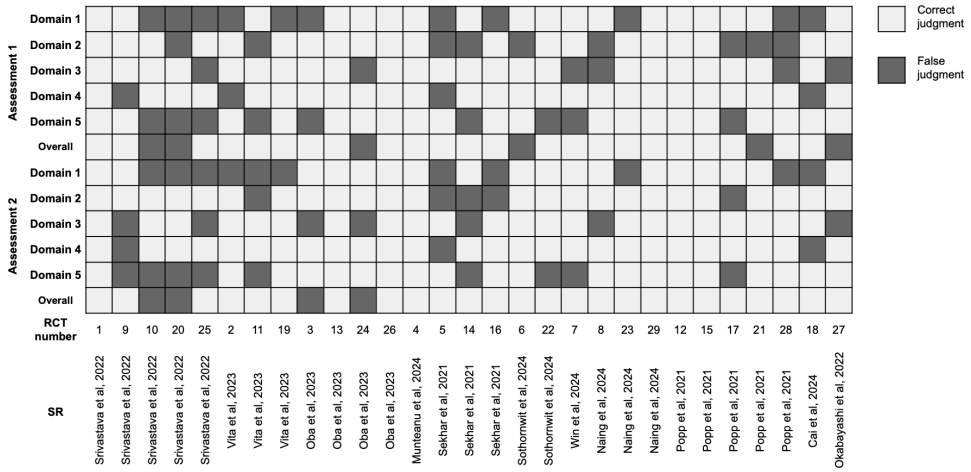
Across the 5 ROB 2 domains, the mean accuracy was 74.5% (95% CI 66.0%-83.0%) based on pooled results from 2 assessments. The lowest accuracy was observed in domain 1 (randomization process) at 62.1%, whereas the highest was observed in domain 4 (outcome measurement) at 87.9%. A total of 74 discrepancies were identified in ROB assessments, with 55 (74.3%) resulting from differences in judgments between the LLM and the reference standard and 19 (25.7%) arising from inconsistencies in data extraction. Domain 1 (randomization process) exhibited the highest number of total discrepancies (22/74, 29.7%), with 68.2% (15/22) related to judgment and 31.8% (7/22) related to data extraction. Domain 5 (selection of the reported results) accounted for 24.3% (18/74) of discrepancies, including 66.7% (12/18) judgment-related and 33.3% (6/18) extraction-related differences. Domains 2 and 3 showed intermediate discrepancy rates (14/74, 18.9% and 13/74, 17.6%, respectively), whereas domain 4 exhibited the fewest discrepancies (7/74, 9.5%), predominantly due to judgment differences.

Sensitivity ranged from 42.9% in domain 2 to 71.4% in domain 1, highlighting variability in identifying high-risk judgments. Specificity was consistently high (range 59.1%-94%), indicating reliable identification of low-risk judgments. The F_1-score was highest in domain 4 (53.3%) and lowest in domain 2 (46.2%), highlighting the influence of both sensitivity and precision on performance in nuanced domains.

Trial-Specific Accuracy

Across 58 assessments for the 29 trials (Table S4 in Multimedia Appendix 3), 7 (24.1%) trials achieved full accuracy (100% correct), whereas 7 (24.1%) trials attained an average accuracy between 80% and 90%. In total, 51.7% (15/29) of the trials had an accuracy of 70% or lower, with the lowest at 40% (Figure 3 [42-52]).

Figure 3. Heat map of accuracy across trial- and domain-specific contexts [42-52]. RCT: randomized controlled trial; SR: systematic review.



Performance varied across medical disciplines, although these findings were limited by small subgroup sample sizes. The reported counts represent aggregated domain-level assessments across 5 ROB 2 domains and 2 assessment cycles per trial. For bone health (1/29, 3.4%), the single trial was correctly judged in both assessments. For other disciplines, we observed the following aggregated domain-level accuracies: in gastroenterology, 39 out of 50 domains were correctly assessed (across 5/29, 17.2% of the trials); in infectious diseases, 48 out of 60 domains were correctly assessed (across 6/29, 20.7% of the trials); in pulmonology, 35 out of 40 domains were correctly assessed (across 4/29, 13.8% of the trials); and in obstetrics and gynecology, 17 out of 20 domains were correctly assessed (across 2/29, 6.9% of the trials). Cardiology and psychology exhibited the lowest accuracy, with 31 out of 50 domains (across 5/29, 17.2% of the trials) and 36 out of 60 domains (across 6/29, 20.7% of the trials) correctly assessed, respectively, with no domains in either discipline achieving full accuracy in the first or second assessment.

Sensitivity Analysis

Notably, a typographical error in the initial assessments by Toouli et al [20] was identified in domains 2, 3, and 4. After attempting to contact the study authors without receiving a response, the research team conducted a sensitivity analysis to evaluate the impact of manual adjudications on the reference standard. Correction of the typographical error led to improvements in domain-specific performance. In the first assessment, F_1-scores increased from 40.0% to 50.0% for domain 2 and from 57.1% to 61.5% for domain 3, whereas accuracy increased from 69.0% to 72.4% for domain 2 and from 79.3% to 82.8% for domain 3. In the second assessment, while the F_1-score for domain 3 improved from 36.4% to 40.0% and its accuracy rose from 75.9% to 79.3%, the F_1-score for domain 2 paradoxically decreased from 54.6% to 50.0%, and accuracy decreased from 82.8% to 79.3%. This decrease occurred because the LLM's original judgment in assessment 2 happened to align with the uncorrected Cochrane reference. After manual correction of the reference standard to reflect the true methodological assessment, the LLM's response was reclassified as discrepant. Additionally, we performed a sensitivity check by excluding the subsequently retracted trial by Maher et al [10]. Removing this single trial resulted in a slight decrease in the aggregate domain accuracy, shifting from 73.1% to 72.9% in the first assessment and from 75.9% to 75% in the second. These changes were of less than 1 percentage point and did not materially alter the overall trends or conclusions regarding the LLM's performance. Following exclusion of this trial, obstetrics and gynecology was represented by 1 remaining trial in which 8 of 10 domains across the 2 assessments were correctly classified; this result was therefore reported descriptively.

Consistency

The assessment rate (observed agreement; P_o) demonstrated high consistency across domains, with a mean of 89.0% (SD 7.5%), ranging from 79.3% to 96.6%. Cohen κ values indicated near-perfect agreement in domains 1, 4, and 5 (κ=0.86, 0.84, and 0.85, respectively); moderate agreement in domain 3 (κ=0.56); and fair agreement in domain 2 (κ=0.39). PABA κ values followed similar trends (Table S5 in Multimedia Appendix 3).

At the trial level, of all 29 RCT assessments, 17 (58.6%) achieved full consistency (proportion of agreement; P_o=1.00) across all domains, whereas 9 (31%) attained a proportion of agreement of exactly 0.80.

The Cohen κ was reported as "not available" due to zero variance in 24.1% (7/29) of the trials. The mean agreement across all studies was 0.89 (SD 0.16), reflecting high consistency (Table S6 in Multimedia Appendix 3).

Discussion

Principal Findings

In this study, we used GPT-o3-mini-high, a compute-intensive and reasoning-focused LLM to conduct structured ROB assessments for RCTs using the ROB 2. The LLM's performance reflected a hybrid cognitive task: first performing semantic reasoning to extract and interpret complex clinical narratives for answering signaling questions and subsequently strictly following instructions by applying the

structured logic of the ROB 2 decision tree. The judgments by review authors from published Cochrane systematic reviews served as a widely recognized reference standard for assessing the generalizability and representativeness of the LLM's evaluations.

The LLM demonstrated a moderate mean accuracy of 74.5%. At the domain level, the averaged sensitivity (57.4%) and specificity (78.4%), calculated from 2 independent assessments, indicated a lower likelihood of identifying high-risk domain-level judgments while reliably detecting low-risk judgments. However, this pattern was not reflected in the overall trial-level assessment. When domain-level judgments were integrated into an overall ROB classification, the model showed an asymmetric pattern in the opposite direction, tending to classify trials as high risk rather than overlook truly high-risk trials. Specifically, across the 2 assessments, the overall judgments resulted in 9 FPs and only 1 FN. This tendency may be advantageous in screening applications by increasing the likelihood that studies with potential bias will be prioritized for subsequent expert review.

In evaluating performance across trial- and domain-specific contexts, we observed substantial variation in accuracy across medical disciplines, suggesting that the model's alignment with expert judgments may depend on the characteristics of the underlying literature. This variability may partly reflect known challenges in ROB assessment, such as complex study structures, nonintuitive reporting terminology, and inconsistencies even among experienced reviewers, as noted in previous research [53,54]. Our findings provide a preliminary overview of model performance across medical disciplines. While high accuracy was observed in structured, objective end points such as bone health, these results should be interpreted with caution due to the limited number of trials in these subgroups. Accuracy diminished to approximately 60% in disciplines such as psychology [55] and cardiology [56], suggesting that while the model excels at processing structured data, its reliability diminishes when critical details such as cointerventions or adherence are implicit or conveyed through nuanced narratives. Future implementation should prioritize automated screening for trials with objective outcomes while maintaining intensive expert oversight for behavioral research.

At the domain-specific level, the model demonstrated the highest accuracy in domain 4 (measurement of the outcome), which presented the fewest discrepancies. Meanwhile, domains with well-defined criteria and clearly stated descriptions, such as domain 1 (randomization process), exhibited the highest κ value (0.86) despite a relatively higher frequency of judgment discrepancies compared with the reference standard. In contrast, domain 2 (deviations from intended interventions) exhibited the lowest consistency, with a κ value of 0.39 and the lowest observed agreement. This discrepancy likely reflects that the fixed structure of the prompt struggles with the inherent complexity of this domain. Following the ROB 2, our prompts required the LLM to initially categorize the trial as evaluating either an intention-to-treat or per-protocol effect. When trial reports use obscure or nonstandard terminology without further elaboration, the model's initial classification is prone to error, resulting in cascading inconsistencies in the subsequent assessment. This aligns with empirical evidence [57] suggesting that human raters also demonstrate the lowest interrater reliability in domain 2, often due to the nuanced interpretive requirements of intention-to-treat and per-protocol effects. These issues present challenges not only for LLMs but also for human reviewers, contributing to discrepancies in judgment and highlighting the need for more precise information extraction, robust reasoning, and contextual understanding in automated ROB assessments.

A total of 74 discrepancies were observed between LLM assessments and the reference standard, with 55 (74.3%) related to differences in final risk judgments rather than data extraction errors. Most judgment-related discrepancies occurred in domains 1 and 5. Analysis showed that the explanations generated by LLMs often highlighted the presence or absence of specific keywords (eg, "concealment") rather than synthesizing indirect cues or context. For instance, in domain 1, the absence of explicit terms such as "allocation concealment" in the model's reasons was associated with a "some concerns" rating, whereas human reviewers inferred low risk based on indirect phrases such as "centralized randomization." A similar discrepancy was observed in domain 5, where vague reporting often led to inconsistent bias judgments. This suggests that the textual outputs observed in the model's generated justifications remain highly sensitive to surface-level terminology.

These discrepancies highlight a fundamental distinction between human and LLM-based ROB assessments: humans integrate contextual understanding and inferential logic, whereas LLMs' output may appear influenced by surface features. Nevertheless, the model's strong internal consistency and performance in well-structured domains suggest its potential as a tool in systematic review workflows.

In addition, our study used GPT-o3-mini-high, an LLM optimized for reasoning and instruction-following capabilities that was explicitly selected for its ability to perform inference tasks to conduct a rule-guided assessment task [7]. Compared to the models used in the previous study [5] (ChatGPT and Claude), which were based on earlier-generation architectures, the GPT-o3-mini model has been shown to generate more accurate and clearer answers by using deeper internal reasoning processes. The GPT-o3-mini model has demonstrated a 39% reduction in major errors on complex questions and was preferred over GPT-o1-mini in 56% of expert tester comparisons [7]. Furthermore, its performance on high-level scientific reasoning benchmarks such as GPQA Diamond (79.7%) significantly outperforms contemporary nonreasoning models such as Claude 3.5 Sonnet (65.0%) [58]. This demonstrates its superior capacity for tasks requiring nuanced judgment, such as ROB evaluation.

Finally, the methodological tools differed between studies. Lai et al [5] assessed bias using a modified version of the original Cochrane ROB tool developed by the CLARITY group [59]. In contrast, our study applied the updated version of the tool, which is currently recommended by the Cochrane

Collaboration for evaluating randomized trials [3]. Despite its conceptual improvements, prior research has demonstrated that the ROB 2 exhibits relatively poor interrater reliability [54,60]. It involves more steps and requires more logical reasoning to apply consistently, thereby presenting inherent challenges for consistent evaluation. This complexity makes it particularly suitable for testing LLMs with higher reasoning capacity.

Limitations

This study has several limitations that should be noted. First, while Cochrane systematic reviews represent a high standard of methodological rigor [12,13], their assessments are not definitive or immune to error. Our comparisons reflect alignment with expert assessments as a reference standard rather than assuming those assessments to be correct. Moreover, the ROB 2 is known to have relatively poor interrater reliability and produce inconsistent judgments even among human raters [53,54,60]. This inherent variability likely contributed to the discrepancies between the LLM-generated and expert assessments. Second, we evaluated a single LLM under fixed prompting conditions without exploring performance across different models, prompt styles, or fine-tuning approaches. Although we iteratively refined prompts using a small pilot set of RCTs based on a previous study [5] and achieved consistent alignment with experts, the approximately 75% accuracy observed in the validation dataset suggests potential limitations in generalizing prompts optimized on a small pilot test to a broader validation dataset. Third, our dataset, limited to 29 RCTs from 11 Cochrane systematic reviews, may not fully represent the diversity across all medical disciplines. Fourth, a limitation involves the potential for pretraining data contamination. Given the nature of LLMs, it is difficult to definitively distinguish between the retrieval of memorized patterns and de novo reasoning. Although mitigation efforts were made, the possibility of prior exposure to public medical datasets remains a fundamental constraint when evaluating LLM outputs. Finally, this study using the ChatGPT web interface introduced certain methodological fragility. As the precise mechanism of PDF parsing within the interface remains uncertain, the risk of data omission or formatting errors cannot be entirely ruled out, highlighting the need for future validation studies to use version-controlled APIs.

Conclusions

In this exploratory feasibility study using one of the most advanced LLMs available, we found that the accuracy of ROB assessments for randomized trials was largely comparable to that of Cochrane review authors when using the ROB 2. These preliminary findings suggest the potential of LLMs in methodological evaluations, although domains with less structured reporting demand careful consideration. With further development, LLMs could significantly enhance the efficiency and capacity for large-scale application of systematic review processes in biomedical research.

Acknowledgments

During this work, the authors used the GPT-o3-mini-high model (OpenAI) as part of a formal research design, with a clear description provided in the Methods section. After using this tool, the authors reviewed and edited the outputs as needed and take full responsibility for the content of the publication.

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and analyzed during this study are included in this published article and its supplementary materials.

Authors' Contributions

YJL contributed to conceptualization, methodology, data curation, investigation, formal analysis, validation, statistical analysis, writing—original draft, and visualization. SHL contributed to writing—original draft, writing—review and editing, and critical revision for important intellectual content. JWL contributed to conceptualization, methodology, supervision, writing—review and editing, validation, project administration, and resources.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompt for the large language model to assess risk of bias (ROB) in randomized controlled trials using version 2 of the Cochrane ROB tool for randomized trials.
[PDF File (Adobe File), 206 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Response from ChatGPT.
[PDF File (Adobe File), 1689 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Supplementary tables for reference standard, interpretation criteria, and assessment results.
[PDF File (Adobe File), 1594 KB-Multimedia Appendix 3]

References

1. Fanaroff AC, Califf RM, Lopes RD. High-quality evidence to inform clinical practice. Lancet. Aug 24, 2019;394(10199):633-634. [doi: 10.1016/S0140-6736(19)31256-5] [Medline: 31448730]
2. Guyatt G, Agoritsas T, Brignardello-Petersen R, et al. Core GRADE 1: overview of the Core GRADE approach. BMJ. Apr 22, 2025;389:e081903. [doi: 10.1136/bmj-2024-081903] [Medline: 40262844]
3. Sterne JA, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. Aug 28, 2019;366:l4898. [doi: 10.1136/bmj.l4898] [Medline: 31462531]
4. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
5. Lai H, Ge L, Sun M, et al. Assessing the risk of bias in randomized clinical trials with large language models. JAMA Netw Open. May 1, 2024;7(5):e2412687. [doi: 10.1001/jamanetworkopen.2024.12687] [Medline: 38776081]
6. Pitt SC, Schwartz TA, Chu D. AAPOR reporting guidelines for survey studies. JAMA Surg. Aug 1, 2021;156(8):785-786. [doi: 10.1001/jamasurg.2021.0543] [Medline: 33825811]
7. OpenAI o3-mini. OpenAI. URL: https://openai.com/index/openai-o3-mini/ [Accessed 2025-04-26]
8. Scope of human research projects exempt from institutional review board review [Article in Chinese]. The Executive Yuan Gazette. 2012. URL: https://gazette.nat.gov.tw/EG_FileManager/eguploadpub/eg018127/ch08/type1/gov70/num35/Eg.htm [Accessed 2025-05-25]
9. RoB 2: a revised Cochrane risk-of-bias tool for randomized trials. Cochrane Methods Bias. URL: https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials [Accessed 2025-04-26]
10. Maher MA, Sayyed TM, Elkhouly NI. Different routes and forms of uterotonics for treatment of retained placenta: a randomized clinical trial. J Matern Fetal Neonatal Med. Sep 2017;30(18):2179-2184. Retracted in: J Matern Fetal Neonatal Med. 2025 Dec;38(1):2509344. [doi: 10.1080/14767058.2025.2509344] [Medline: 40533268]
11. Statement of retraction: Different routes and forms of uterotonics for treatment of retained placenta: a randomized clinical trial. J Matern Fetal Neonatal Med. Dec 2025;38(1):2509344. [doi: 10.1080/14767058.2025.2509344] [Medline: 40533268]
12. Dosenovic S, Jelicic Kadic A, Vucic K, Markovina N, Pieper D, Puljak L. Comparison of methodological quality rating of systematic reviews on neuropathic pain using AMSTAR and R-AMSTAR. BMC Med Res Methodol. May 8, 2018;18:37. [doi: 10.1186/s12874-018-0493-y]
13. Gao Y, Cai Y, Yang K, et al. Methodological and reporting quality in non-Cochrane systematic review updates could be improved: a comparative study. J Clin Epidemiol. Mar 2020;119:36-46. [doi: 10.1016/j.jclinepi.2019.11.012] [Medline: 31759063]
14. Marenzi G, Muratori M, Cosentino ER, et al. Continuous ultrafiltration for congestive heart failure: the CUORE trial. J Card Fail. Jan 2014;20(1):9-17. [doi: 10.1016/j.cardfail.2013.11.004] [Medline: 24269855]
15. Liu P, Li P, Li Q, et al. Effect of pretreatment of S-ketamine on postoperative depression for breast cancer patients. J Invest Surg. Aug 2021;34(8):883-888. [doi: 10.1080/08941939.2019.1710626] [Medline: 31948296]
16. Katial RK, Bernstein D, Prazma CM, Lincourt WR, Stempel DA. Long-term treatment with fluticasone propionate/salmeterol via Diskus improves asthma control versus fluticasone propionate alone. Allergy Asthma Proc. 2011;32(2):127-136. [doi: 10.2500/aap.2011.32.3426] [Medline: 21189151]
17. Munteanu SE, Landorf KB, McClelland JA, et al. Shoe-stiffening inserts for first metatarsophalangeal joint osteoarthritis: a randomised trial. Osteoarthritis Cartilage. Apr 2021;29(4):480-490. [doi: 10.1016/j.joca.2021.02.002] [Medline: 33588086]
18. Lebares CC, Hershberger AO, Guvva EV, et al. Feasibility of formal mindfulness-based stress-resilience training among surgery interns: a randomized clinical trial. JAMA Surg. Oct 1, 2018;153(10):e182734. [doi: 10.1001/jamasurg.2018.2734] [Medline: 30167655]
19. Daher EF, Nogueira CB. Evaluation of penicillin therapy in patients with leptospirosis and acute renal failure. Rev Inst Med Trop Sao Paulo. 2000;42(6):327-332. [doi: 10.1590/s0036-46652000000600005] [Medline: 11136519]
20. Toouli J, Roberts-Thomson IC, Kellow J, et al. Manometry based randomised trial of endoscopic sphincterotomy for sphincter of Oddi dysfunction. Gut. Jan 2000;46(1):98-102. [doi: 10.1136/gut.46.1.98] [Medline: 10601063]
21. Hanna MA, Tang WH, Teo BW, et al. Extracorporeal ultrafiltration vs. conventional diuretic therapy in advanced decompensated heart failure. Congest Heart Fail. 2012;18(1):54-63. [doi: 10.1111/j.1751-7133.2011.00231.x] [Medline: 22277179]

22. Hu J, Wan Q, Zhang Y, et al. Efficacy and safety of early ultrafiltration in patients with acute decompensated heart failure with volume overload: a prospective, randomized, controlled clinical trial. BMC Cardiovasc Disord. Oct 14, 2020;20(1):447. [doi: 10.1186/s12872-020-01733-5] [Medline: 33054727]
23. Tavakoli Ardakani M, Mehrpooya M, Mehdizadeh M, Beiraghi N, Hajifathali A, Kazemi MH. Sertraline treatment decreased the serum levels of interleukin-6 and high-sensitivity C-reactive protein in hematopoietic stem cell transplantation patients with depression; a randomized double-blind, placebo-controlled clinical trial. Bone Marrow Transplant. Apr 2020;55(4):830-832. [doi: 10.1038/s41409-019-0623-0] [Medline: 31383995]
24. Cavalcanti AB, Zampieri FG, Rosa RG, et al. Hydroxychloroquine with or without azithromycin in mild-to-moderate Covid-19. N Engl J Med. Nov 19, 2020;383(21):2041-2052. [doi: 10.1056/NEJMoa2019014] [Medline: 32706953]
25. Peters SP, Bleecker ER, Canonica GW, et al. Serious asthma events with budesonide plus formoterol vs. budesonide alone. N Engl J Med. Sep 1, 2016;375(9):850-860. [doi: 10.1056/NEJMoa1511190] [Medline: 27579635]
26. Damião Neto A, Lucchetti AL, da Silva Ezequiel O, Lucchetti G. Effects of a required large-group mindfulness meditation course on first-year medical students' mental health and quality of life: a randomized controlled trial. J Gen Intern Med. Mar 2020;35(3):672-678. [doi: 10.1007/s11606-019-05284-0] [Medline: 31452038]
27. Omrani AS, Pathan SA, Thomas SA, et al. Randomized double-blinded placebo-controlled trial of hydroxychloroquine with or without azithromycin for virologic cure of non-severe Covid-19. EClinicalMedicine. Dec 2020;29:100645. [doi: 10.1016/j.eclinm.2020.100645] [Medline: 33251500]
28. Erogul M, Singer G, McIntyre T, Stefanov DG. Abridged mindfulness intervention to support wellness in first-year medical students. Teach Learn Med. 2014;26(4):350-356. [doi: 10.1080/10401334.2014.945025] [Medline: 25318029]
29. Sekhavati E, Jafari F, SeyedAlinaghi S, et al. Safety and effectiveness of azithromycin in patients with COVID-19: an open-label randomised trial. Int J Antimicrob Agents. Oct 2020;56(4):106143. [doi: 10.1016/j.ijantimicag.2020.106143] [Medline: 32853672]
30. Wang J, Wang Q, Dong J, et al. Total laparoscopic uncut Roux-en-Y for radical distal gastrectomy: an interim analysis of a randomized, controlled, clinical trial. Ann Surg Oncol. Jan 2021;28(1):90-96. [doi: 10.1245/s10434-020-08710-4] [Medline: 32556870]
31. van Heeringen K, Zivkov M. Pharmacological treatment of depression in cancer patients. A placebo-controlled study of mianserin. Br J Psychiatry. Oct 1996;169(4):440-443. [doi: 10.1192/bjp.169.4.440] [Medline: 8894194]
32. Bart BA, Boyle A, Bank AJ, et al. Ultrafiltration versus usual care for hospitalized patients with heart failure: the Relief for Acutely Fluid-Overloaded Patients With Decompensated Congestive Heart Failure (RAPID-CHF) trial. J Am Coll Cardiol. Dec 6, 2005;46(11):2043-2046. [doi: 10.1016/j.jacc.2005.05.098] [Medline: 16325039]
33. RECOVERY Collaborative Group. Azithromycin in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. Lancet. Feb 2021;397(10274):605-612. [doi: 10.1016/S0140-6736(21)00149-5] [Medline: 33545096]
34. van Stralen G, Veenhof M, Holleboom C, van Roosmalen J. No reduction of manual removal after misoprostol for retained placenta: a double-blind, randomized trial. Acta Obstet Gynecol Scand. Apr 2013;92(4):398-403. [doi: 10.1111/aogs.12065] [Medline: 23231499]
35. Takezawa M, Kida Y, Kida M, Saigenji K. Influence of endoscopic papillary balloon dilation and endoscopic sphincterotomy on sphincter of Oddi function: a randomized controlled trial. Endoscopy. Jul 2004;36(7):631-637. [doi: 10.1055/s-2004-814538] [Medline: 15243887]
36. Woodcock A, Lötvall J, Busse WW, et al. Efficacy and safety of fluticasone furoate 100 µg and 200 µg once daily in the treatment of moderate-severe asthma in adults and adolescents: a 24-week randomised study. BMC Pulm Med. Jul 9, 2014;14:113. [doi: 10.1186/1471-2466-14-113] [Medline: 25007865]
37. Şeker A, Kayataş M, Hüzmeli C, Candan F, Yılmaz MB. Comparison of ultrafiltration and intravenous diuretic therapies in patients hospitalized for acute decompensated biventricular heart failure. Turk J Nephrol. Jan 2019;25(1):79-87. URL: https://www.turkjnephrol.org/index.php/pub/article/view/1036 [Accessed 2026-08-28]
38. Lee LA, Bailes Z, Barnes N, et al. Efficacy and safety of once-daily single-inhaler triple therapy (FF/UMEC/VI) versus FF/VI in patients with inadequately controlled asthma (CAPTAIN): a double-blind, randomised, phase 3A trial. Lancet Respir Med. Jan 2021;9(1):69-84. [doi: 10.1016/S2213-2600(20)30389-1] [Medline: 32918892]
39. Schreiber S, Khaliq-Kareemi M, Lawrance IC, et al. Maintenance therapy with certolizumab pegol for Crohn's disease. N Engl J Med. Jul 19, 2007;357(3):239-250. [doi: 10.1056/NEJMoa062897] [Medline: 17634459]
40. Hinks TS, Barber VS, Black J, et al. A multi-centre open-label two-arm randomised superiority clinical trial of azithromycin versus usual care in ambulatory COVID-19: study protocol for the ATOMIC2 trial. Trials. Aug 17, 2020;21(1):718. [doi: 10.1186/s13063-020-04593-8] [Medline: 32807209]
41. Cotton PB, Durkalski V, Romagnuolo J, et al. Effect of endoscopic sphincterotomy for suspected sphincter of Oddi dysfunction on pain-related disability following cholecystectomy: the EPISOD randomized clinical trial. JAMA. May 2014;311(20):2101-2109. [doi: 10.1001/jama.2014.5220] [Medline: 24867013]

42. Srivastava M, Harrison N, Caetano AF, Tan AR, Law M. Ultrafiltration for acute heart failure. Cochrane Database Syst Rev. Jan 21, 2022;1(1):CD013593. [doi: 10.1002/14651858.CD013593.pub2] [Medline: 35061249]
43. Vita G, Compri B, Matcham F, Barbui C, Ostuzzi G. Antidepressants for the treatment of depression in people with cancer. Cochrane Database Syst Rev. Mar 31, 2023;3(3):CD011006. [doi: 10.1002/14651858.CD011006.pub4] [Medline: 36999619]
44. Oba Y, Anwer S, Patel T, Maduke T, Dias S. Addition of long-acting beta2 agonists or long-acting muscarinic antagonists versus doubling the dose of inhaled corticosteroids (ICS) in adolescents and adults with uncontrolled asthma with medium dose ICS: a systematic review and network meta-analysis. Cochrane Database Syst Rev. Aug 21, 2023;8(8):CD013797. [doi: 10.1002/14651858.CD013797.pub2] [Medline: 37602534]
45. Munteanu SE, Buldt A, Lithgow MJ, Cotchett M, Landorf KB, Menz HB. Non-surgical interventions for treating osteoarthritis of the big toe joint. Cochrane Database Syst Rev. Jun 17, 2024;6(6):CD007809. [doi: 10.1002/14651858.CD007809.pub3] [Medline: 38884172]
46. Sekhar P, Tee QX, Ashraf G, et al. Mindfulness-based psychological interventions for improving mental well-being in medical students and junior doctors. Cochrane Database Syst Rev. Dec 10, 2021;12(12):CD013740. [doi: 10.1002/14651858.CD013740.pub2] [Medline: 34890044]
47. Sothornwit J, Ngamjarus C, Pattanittum P, et al. Uterotonics for management of retained placenta. Cochrane Database Syst Rev. Oct 28, 2024;10(10):CD016147. [doi: 10.1002/14651858.CD016147] [Medline: 39465684]
48. Win TZ, Han SM, Edwards T, et al. Antibiotics for treatment of leptospirosis. Cochrane Database Syst Rev. Mar 14, 2024;3(3):CD014960. [doi: 10.1002/14651858.CD014960.pub2] [Medline: 38483092]
49. Naing C, Ni H, Aung HH, Pavlov CS. Endoscopic sphincterotomy for adults with biliary sphincter of Oddi dysfunction. Cochrane Database Syst Rev. Mar 22, 2024;3(3):CD014944. [doi: 10.1002/14651858.CD014944.pub2] [Medline: 38517086]
50. Popp M, Stegemann M, Riemer M, et al. Antibiotics for the treatment of COVID-19. Cochrane Database Syst Rev. Oct 22, 2021;10(10):CD015025. [doi: 10.1002/14651858.CD015025] [Medline: 34679203]
51. Cai Z, Mu M, Ma Q, et al. Uncut Roux-en-Y reconstruction after distal gastrectomy for gastric cancer. Cochrane Database Syst Rev. Feb 29, 2024;2(2):CD015014. [doi: 10.1002/14651858.CD015014.pub2] [Medline: 38421211]
52. Okabayashi S, Yamazaki H, Yamamoto R, et al. Certolizumab pegol for maintenance of medically induced remission in Crohn's disease. Cochrane Database Syst Rev. Jun 30, 2022;6(6):CD013747. [doi: 10.1002/14651858.CD013747.pub2] [Medline: 35771590]
53. Jordan VM, Lensen SF, Farquhar CM. There were large discrepancies in risk of bias tool judgments when a randomized controlled trial appeared in more than one systematic review. J Clin Epidemiol. Jan 2017;81:72-76. [doi: 10.1016/j.jclinepi.2016.08.012] [Medline: 27622779]
54. 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. Oct 2020;126:37-44. [doi: 10.1016/j.jclinepi.2020.06.015] [Medline: 32562833]
55. Button KS, Munafò MR. Addressing risk of bias in trials of cognitive behavioral therapy. Shanghai Arch Psychiatry. Jun 25, 2015;27(3):144-148. [doi: 10.11919/j.issn.1002-0829.215042] [Medline: 26300596]
56. Baasan O, Freihat O, Nagy DU, Lohner S. Methodological quality and risk of bias assessment of cardiovascular disease research: analysis of randomized controlled trials published in 2017. Front Cardiovasc Med. 2022;9:830070. [doi: 10.3389/fcvm.2022.830070] [Medline: 35369336]
57. Pitre T, Jassal T, Talukdar JR, Shahab M, Ling M, Zeraatkar D. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: a methods study. medRxiv. Preprint posted online on Nov 22, 2023. [doi: 10.1101/2023.11.19.23298727]
58. Claude 3.7 Sonnet and Claude Code. Anthropic. 2025. URL: https://www.anthropic.com/news/claude-3-7-sonnet [Accessed 2025-05-23]
59. Tool to assess risk of bias in randomized controlled trials. DistillerSR. URL: https://www.distillersr.com/resources/methodological-resources/tool-to-assess-risk-of-bias-in-randomized-controlled-trials-distillersr [Accessed 2025-05-23]
60. Eisele-Metzger A, Lieberum JL, 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. May 2025;16(3):491-508. [doi: 10.1017/rsm.2025.12] [Medline: 41626932]

Abbreviations

FN: false negative
FP: false positive
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
LLM: large language model

PABA: prevalence-adjusted, bias-adjusted
RCT: randomized controlled trial
RD: relative difference
ROB: risk of bias
ROB 2: version 2 of the Cochrane risk-of-bias tool for randomized trials
TN: true negative
TP: true positive

Edited by Andrew Coristine; peer-reviewed by Eon Ting, Eriberto Franchi; submitted 29.Sep.2025; final revised version received 11.Aug.2026; accepted 12.Aug.2026; published 25.Sep.2026

Please cite as:
Lai YJ, Lin SH, Liu JW
Evaluating the Accuracy of Large Language Models in Risk-of-Bias Assessment Using Version 2 of the Cochrane Risk-of-Bias Tool for Randomized Trials: Exploratory Feasibility Study
J Med Internet Res 2026;28:e84915
URL: https://www.jmir.org/2026/1/e84915
doi: 10.2196/84915