Evaluating the Application of GPT-4o and Retrieval-Augmented Generation (RAG) for Assessing Risk of Bias and Study Quality in Systematic Literature Reviews (SLRs): Preliminary Findings from a Comparative Study

MSR22

Ellen Kasireddy, Cuthbert Chow, Mir-Masoud Pourrahmat, Jean-Paul Collet, Jun Collet, Mir Sohail Fazeli
Evidinno Outcomes Research Inc., Vancouver, British Columbia, Canada

Background

- Assessing risk of bias is critical in systematic literature reviews (SLRs) to ensure study validity, as it helps determine the quality of included studies.
- Traditional risk of bias assessment methods rely on expert judgment, making the process time-consuming and resource-intensive.
- Recent advancements in artificial intelligence (AI), particularly large language models (LLMs) integrated with retrieval-augmented generation (RAG), offer promise to automate risk of bias assessment.[1]

Objective

- This study aimed to assess the performance of a custom AI model integrating GPT-4o and RAG in conducting risk of bias assessment for SLRs.

Methods

MODEL FRAMEWORK

- A custom AI model was developed to automate risk of bias assessment by integrating GPT-4o with RAG via the OpenAI Assistants API.[1]
- The model systematically retrieved relevant study content and generated structured evaluations based on predefined checklists. **Figure 1** outlines the model workflow.

RAG PROCESS

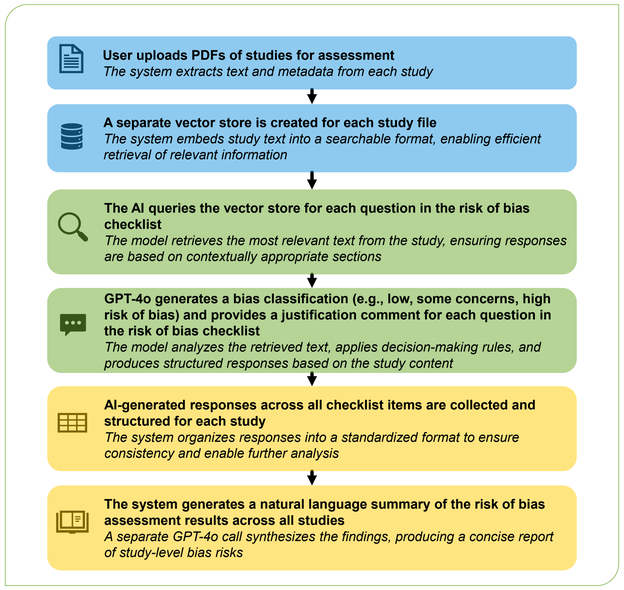
1. **User Upload & Vector Store Creation:**
   - Users upload PDF study files, which are processed individually.
   - A vector store is created for each file, enabling efficient retrieval of relevant study content.
2. **Question-Guideline Pairing & Query Execution:**
   - Each risk of bias checklist contains specific assessment questions.
   - Each question is queried against the vector store for every study file.
   - GPT-4o processes the retrieved text and generates an answer with explanatory comments.
3. **Summarization:**
   - AI-generated responses are collected across all studies.
   - A natural language summary of the assessment results is generated using a separate GPT-4o call.
   - The final output includes a structured risk of bias classification alongside a synthesized summary.

RISK OF BIAS TOOLS & STUDY SELECTION

- The model was tested using 30 randomly selected studies (10 per tool) across three risk of bias tools assessing risk of bias regarding study design, selection of participants, assessment of outcomes, statistical analysis, and reporting of results, requiring context-aware evaluation:
  - **Cochrane Risk of Bias Version 2 (ROB2)** for randomized controlled trials (RCTs)[2]
  - **JBI Critical Appraisal Checklist** for cross-sectional studies[3]
  - **Newcastle-Ottawa Scale (NOS)** for cohort studies[4]

Methods (continued)

Figure 1: Workflow for Automated Risk of Bias Assessment



MODEL EVALUATION

- To assess the model's accuracy, AI-generated risk of bias assessments were compared to human expert assessments.
- Risk of bias classifications were defined as follows:
  - **True Positives (TP):** AI and human agree on a "satisfactory" classification (low risk of bias).
  - **True Negatives (TN):** AI and human agree on an "unsatisfactory" classification (some concerns or high risk of bias).
  - **False Positives (FP):** AI incorrectly marks an item as "satisfactory" when the human does not.
  - **False Negatives (FN):** AI incorrectly marks an item as "unsatisfactory" when the human does not.

PERFORMANCE METRICS

- Key performance metrics were used to evaluate different aspects of the model's performance:
  - **Accuracy:** Overall correctness, calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

  - **Sensitivity (Recall):** The proportion of low-risk studies identified by AI out of all low-risk studies, calculated as

$$Sensitivity = \frac{TP}{TP+FN}.$$

  - **Specificity:** The proportion of high-risk studies identified by AI out of all high-risk studies, calculated as

$$Specificity = \frac{TN}{TN+FP}.$$

  - **Positive Predictive Value (PPV):** Likelihood of a "satisfactory" classification being truly "satisfactory", calculated as

$$PPV = \frac{TP}{TP+FP}.$$

  - **Negative Predictive Value (NPV):** Likelihood of an "unsatisfactory" classification being truly "unsatisfactory", calculated as

$$NPV = \frac{TN}{TN+FN}.$$

Results

PERFORMANCE METRICS

- The model demonstrated high specificity across all tools (73.1-87.5%), effectively identifying high-risk-of-bias items.
- The sensitivity was lower (33.3-67.0%), particularly for the NOS tool, resulting in a higher rate of false negatives for low-risk-of-bias items.
- Overall accuracy ranged from 56.7% to 72.5% (**Figures 2-4**).
- The rationale accompanying the model's bias classifications was logically sound and consistent with the assigned judgments.

Cochrane Risk of Bias Version 2 Tool for Randomized Controlled Trials

- The model performed moderately well in identifying low-risk-of-bias items (sensitivity: 67.0%), minimizing the misclassification of these items as high risk (**Figure 2**).
- However, the relatively low NPV (32.8%) indicates that the model tended to have a higher false-negative rate, which suggests the model struggled to correctly classify items as unsatisfactory.

PERFORMANCE METRICS (CONTINUED)

JBI Tool for Cross-Sectional Studies

- The model demonstrated high specificity (87.5%), effectively identifying high-risk-of-bias items, but showed lower sensitivity (62.5%), reflecting its moderate ability in detecting low-risk-of-bias items (**Figure 3**).
- The model performed well in correctly classifying items as low risk of bias (PPV: 88.2%), whereas the ability to correctly classify items as high risk of bias was slightly lower (NPV: 60.9%).

Newcastle-Ottawa Scale for Cohort Studies

- The model showed a strong specificity (83.8%) and moderate PPV (69.6%), indicating it was able to successfully identify high-risk items and correctly classified most low-risk-of-bias items (**Figure 4**).
- The sensitivity was low (33.3%) indicating that the model had a significant false-negative rate, highlighting its difficulty in accurately classifying low-risk-of-bias items.

The model demonstrated high specificity across all tools (73.1–87.5%), indicating strong performance in identifying high-risk-of-bias items.

Figure 2: AI Model Performance on the Cochrane Risk of Bias Version 2 Tool

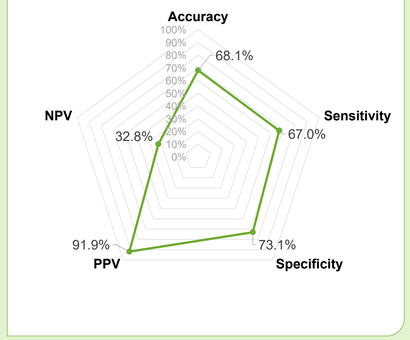


Figure 3: AI Model Performance on the JBI Tool for Cross-Sectional Studies

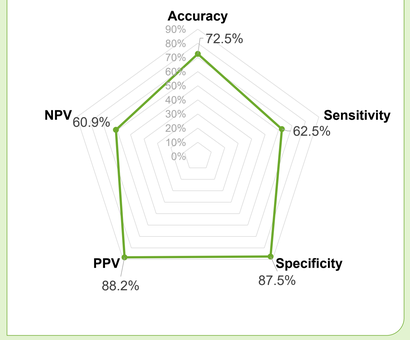
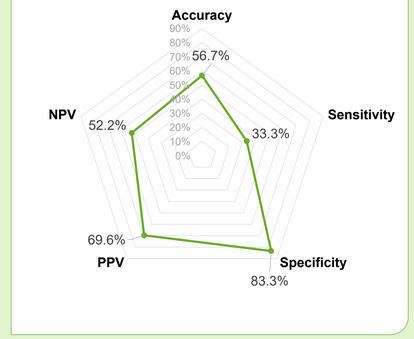


Figure 4: AI Model Performance on the Newcastle-Ottawa Scale for Cohort Studies



Conclusions

- The current version of the model demonstrated high specificity across all tools, effectively identifying high-risk-of-bias items and minimizing false positives.
- However, its limited sensitivity, particularly with NOS, and low NPV with ROB2, indicate a high false-negative rate, risking misclassification of low-risk-of-bias studies.
- Risk of bias assessment is inherently complex: the nature and severity of individual biases must be considered.
- For evaluation, risk of bias responses were dichotomized (low vs. high/some concerns), reducing granularity and potentially contributing to lower sensitivity as the model's conservative judgments (e.g., classifying borderline studies as "some concerns") were counted as false negatives.
- While the current model shows promise for supporting risk of bias assessments alongside human reviewers, further optimization is needed to reduce false negatives and enhance sensitivity.
- Future work should focus on refinement and validation in a larger, more diverse set of studies to ensure generalizability and practical utility.

References

1. OpenAI GPT-4. 2023. https://openai.com/index/gpt-4-research
2. Sterne JAC, et al. *BMJ*. 2019;366:l4898.
3. Moola S, et al. *JBI Manual for Evidence Synthesis*. 2020. https://synthesismanual.jbi.global
4. Wells GA, et al. 2013. https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Acknowledgments

This study was conducted by Evidinno Outcomes Research Inc. EK, CC, MP, JPC, JC, and MSF report employment with Evidinno Outcomes Research Inc. Authors report no other conflict of interest.

EVIDINNO
EVIDENCE NAVIGATION & SYNTHESIS

Presented at ISPOR 2025; May 13-16; Montreal, Canada

Email: jun@evidinno.com