



Medevi-RoB: Benchmarking Large Language Models for Automated Risk-of-Bias (RoB 2) Assessment

David Nguyen, Medevi Team

T-CAIREM 2025

Why automate risk of bias assessments?

Risk-of-bias (RoB) appraisals are essential for trustworthy evidence synthesis, yet remain slow, variable, and costly.

Risk of Bias 2 (RoB 2) assesses how study design and reporting affect the trustworthiness of trial results.

Reasons for automation:

- **Resource burden:** SRs are costly and time-consuming to complete.
- **Human variance:** Inter-rater reliability reflects high variance (IRR = 0.16 even among expert reviewers).
- **Increasing adoption:** Automation tools (e.g., Elicit, RobotReviewer) increasingly adopted but limited capability or lack benchmarks.

Need: Robust, transparent benchmarks to evaluate and deploy LLM-assisted RoB 2 assessments efficiently and interpretably.

**\$50,000 to
\$140,000 USD**

Average SR cost

6-18 months

Average time to SR
completion

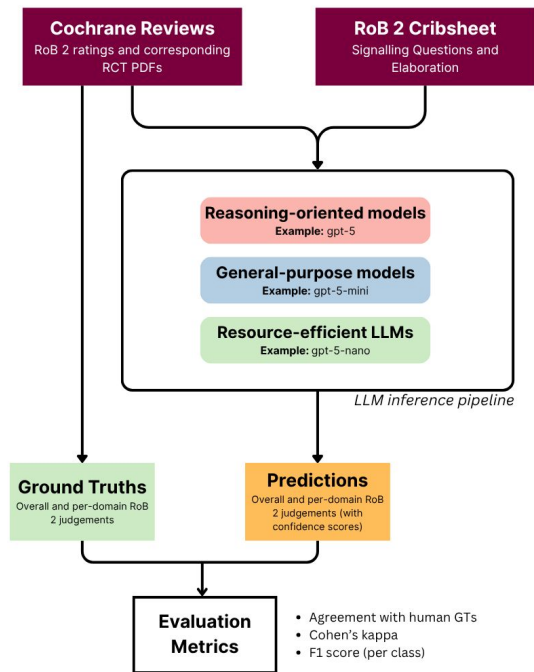
6-12 weeks

Average time for risk
of bias analysis

358 minutes

Average staff
resources allocated
to RoB 2 per study
evaluated

Medevi-RoB dataset for evaluation of LLM performance on RoB 2



Source	Field	Population	Intervention	≥2 Independent Reviewers and Consensus	Total number of RCTs
Cochrane Database of Systematic Reviews	Dermatology	Eczema	Topical anti-inflammatory treatments	Yes	291
Cochrane Database of Systematic Reviews	Dermatology	Chronic plaque psoriasis	Systemic pharmacological treatments	Yes	204
Cochrane Database of Systematic Reviews	Neurology	Parkinson's disease	Physical exercise	Yes	154
Cochrane Database of Systematic Reviews	Anesthesia	Patients undergoing neuraxial anesthesia	Ultrasound guidance vs. anatomical landmarks	Yes	65

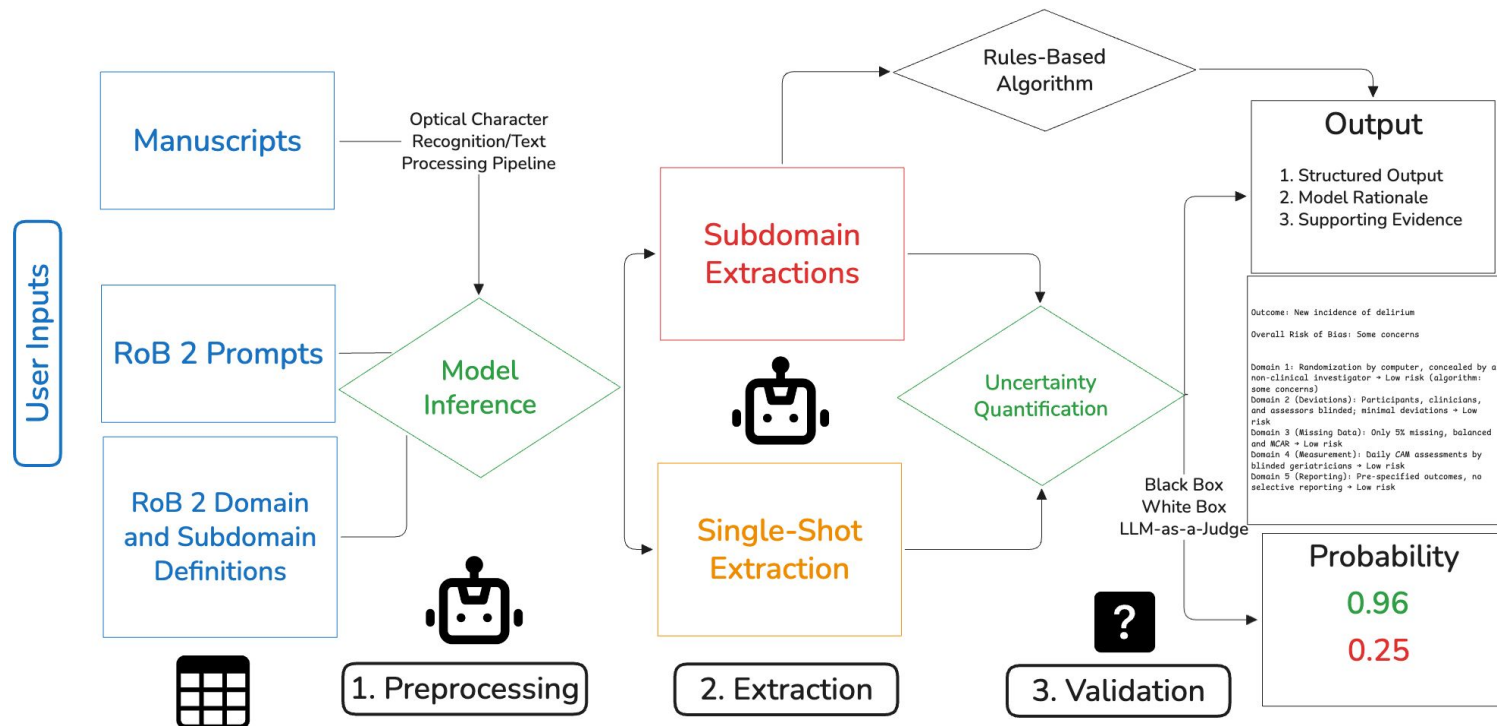
Baseline Characteristics of Evaluated Outcome-Specific RoB 2 Judgements

Baseline Characteristics of Evaluated Outcome-Specific RoB 2 Judgements (n=273)

Domain	Low risk of bias	Some concerns	High risk of bias
Domain 1: Randomization process	194 (71.1%)	72 (26.4%)	7 (2.6%)
Domain 2: Deviations from attended intervention	209 (76.6%)	46 (16.8%)	18 (6.6%)
Domain 3: Missing outcome data	238 (87.2%)	12 (4.4%)	23 (8.4%)
Domain 4: Measurement of the outcome	215 (78.8%)	30 (11.0%)	28 (10.3%)
Domain 5: Selection of reported results	164 (60.1%)	105 (38.5%)	4 (1.5%)

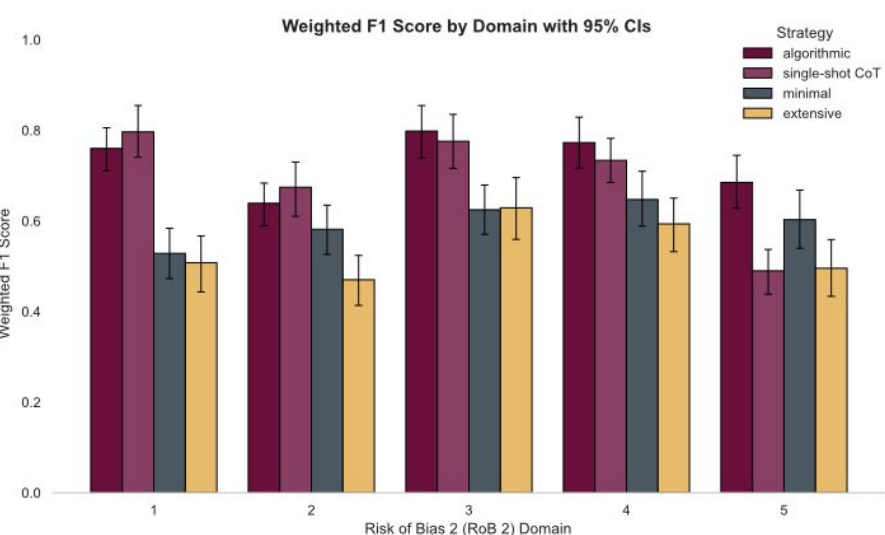
Medevi platform enables reproducible LLM evaluations

Overview of Medevi's AI-Enabled Data Extraction Pipeline for Risk of Bias 2 (RoB 2) Assessment



Modern open-source reasoning models and structured (algorithmic) prompting approaches substantially outperform earlier GPT-4 baselines across all RoB domains.

LLM Performance on Individual RoB 2 Domains
Structured vs. unstructured approaches



Note: F1 score and inter-rater reliability per-outcome on a dataset of 273 outcome-RoB 2 pairs (154 studies) from 34 systematic reviews with no fine-tuning.

Inter-rater reliability (95% CI) compared to human expert consensus agreement

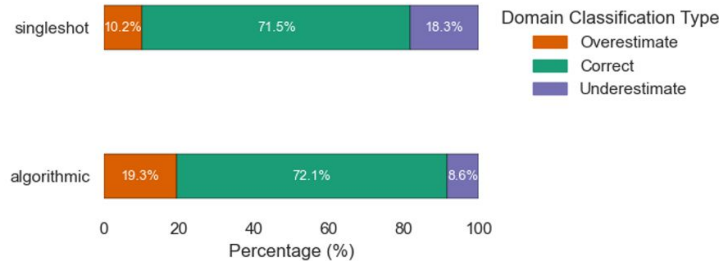
Domain	Algorithmic gpt-oss-120b	Single-shot CoT gpt-oss-120b	Minimal prompt gpt-4	Extensive prompt gpt-4
Domain 1: Risk of bias due to randomization	0.396 (0.300-0.502)	0.573 (0.454-0.678)	0.074 (-0.044-0.201)	0.050 (-0.063-0.164)
Domain 2: Risk of bias due to deviations from the intended intervention	0.137 (0.028-0.245)	0.190 (0.040-0.342)	0.127 (0.001-0.240)	0.093 (0.002-0.186)
Domain 3: Risk of bias due to missing outcome data	0.252 (0.112-0.397)	0.341 (0.205-0.469)	0.142 (0.025-0.259)	0.120 (0.020-0.218)
Domain 4: Risk of bias due to measurement of the outcome	0.525 (0.388-0.656)	0.345 (0.218-0.474)	0.041 (-0.073-0.149)	0.095 (-0.040-0.218)
Domain 5: Risk of bias due to selective reporting	0.373 (0.267-0.475)	0.055 (-0.020-0.138)	0.274 (0.155-0.379)	0.180 (0.105-0.267)

LLM-derived predictions achieve human-level agreement on all RoB domains

Domain	LLM Approach Best performing	LLM Cohen's κ	LLM Agreement	Human (Minozi et al., 2020) Fleiss κ	Human Agreement
Domain 1: Randomization process	Single-shot CoT (gpt-oss-120b)	0.573 (0.454-0.678)	Moderate	IRR 0.45, (95% CI: 0.37-0.53)	Moderate
Domain 2: Deviations from attended intervention	Single-shot CoT (gpt-oss-120b)	0.190 (0.040-0.342)	Slight	IRR 0.04, (95% CI: -0.06-0.14)	Slight
Domain 3: Missing outcome data	Single-shot CoT (gpt-oss-120b)	0.341 (0.205-0.469)	Fair	IRR 0.22 (95% CI: 0.14-0.30)	Fair
Domain 4: Measurement of the outcome	Algorithmic (gpt-oss-120b)	0.525 (0.388-0.656)	Moderate	IRR 0.27 (95% CI: 0.19-0.35)	Fair
Domain 5: Selection of reported results	Algorithmic (gpt-oss-120b)	0.373 (0.267-0.475)	Fair	IRR 0.30 (95% CI: 0.22-0.38)	Fair
Overall Risk of Bias	Per-domain assessment (algorithmic)	0.213 (0.125-0.297)	Fair	IRR 0.16 (95% CI: 0.08-0.24)	Fair

Strengths and limitations of algorithmic vs. single-shot chain-of-thought prompting approaches

RoB 2 Domain Classification Type - single-shot CoT vs. algorithmic



Comparison of single-shot versus algorithmic domain-level risk of bias classifications. The algorithmic method shows a higher tendency toward overestimation of bias (due to missing subdomain scores), whereas single-shot approach tends towards underestimation of bias. Both approaches demonstrate similar overall accuracy.

Prompting Strategy	Algorithmic (per-domain) gpt-oss-120b (effort: high)	Single-shot CoT gpt-oss-120b (effort: high)
Role of LLM	Generate judgements at a subdomain level for input into a deterministic algorithm.	Generate judgements at a domain level.
Average input tokens	75000	15000
Average output tokens (including reasoning)	12500	7500
Average cost (gpt-oss-120b)	Input: \$0.01125/paper Output: \$0.0075/paper Total: \$0.01875/paper	Input: \$0.00225/paper Output: \$0.001125/paper Total: \$0.003375/paper
Strengths	Highly sensitive for high risk of bias papers.	Stronger interpretive performance with missing data.
Limitations	May overestimate risk due to missing information.	Tends to underestimate risk compared to algorithmic approaches.

Implications and Next Steps

Key takeaways:

- **Medevi-RoB** establishes a reusable benchmark for evaluating LLM performance in structured critical appraisal.
- **Medevi** provides a transparent platform for LLM-assisted risk-of-bias assessment; enabling safe, auditable integration of LLMs into evidence synthesis pipelines.
- **Structured LLM** prompting can improve performance over baseline measures.
- **Bias detection and interpretability:** Analyses can help identify when LLMs systematically over- or under-estimate risk, supporting trustworthy automation.

Future directions:

- **Fine-tuning & hybrid NLP:** Compare performance of open, fine-tuned LLMs to save costs.
- **Dataset expansion:** Complete evaluation on ~1,000 randomized trials in dataset, spanning multiple specialties.
- **Extension to other critical appraisal tasks:** Apply framework and prompting strategies to other critical appraisal tools (ROBINS-I, QUADAS-C, AMSTAR).

With robust benchmarks, we can move from experimenting with AI to robust use and evaluation in evidence synthesis.

Acknowledgements

Medevi Team

Evaluation: Andrew Ganea, Dylan Kwan, Nathan Lai, Henry Liu, Chris Ouyang, Jessica Song, Sarah Pimenta

Software Development: Minh Hoang Ton, Michael Nguyen

Advisors:

Dena Zeraatkar, Matthew Sibbald

Funding Support:

Cohere Labs, Microsoft

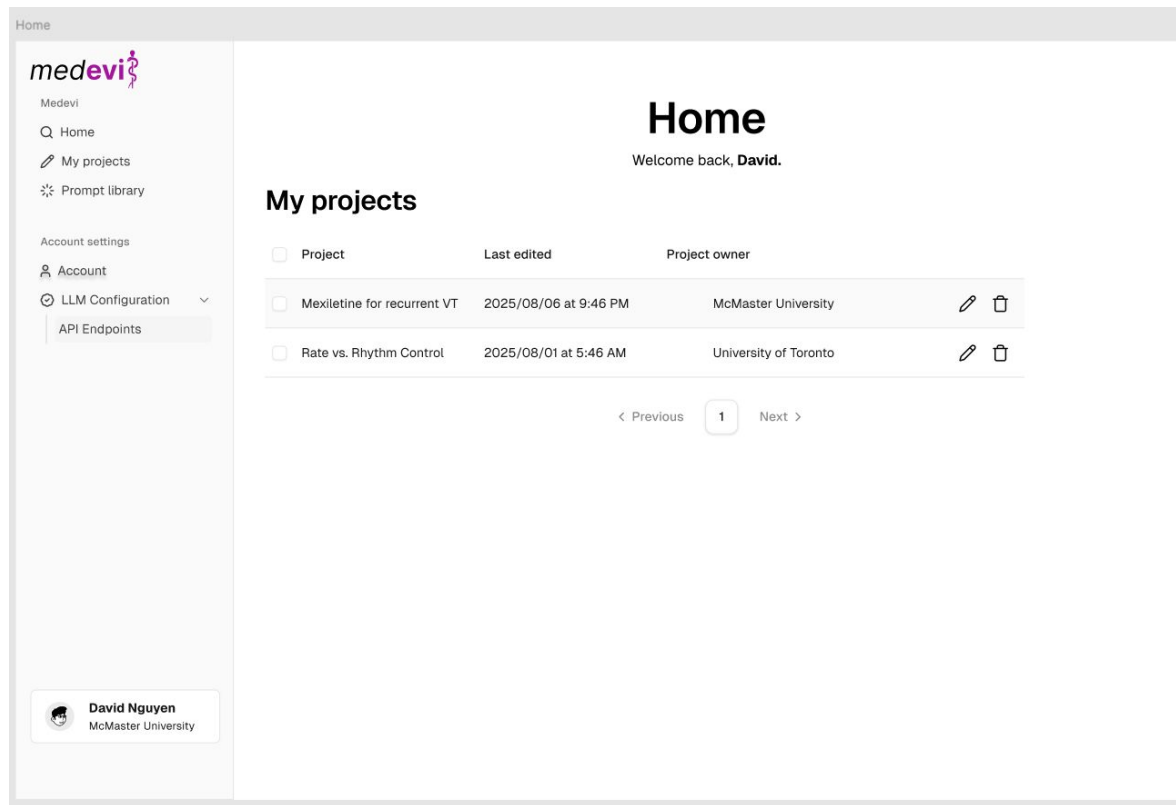
Disclosures:

Not applicable.

Contact:

David Nguyen
davidle.nguyen@medportal.ca

User Interface



User Interface

Project Home


Medevi

🔍 Home

✎ Projects

Current project

🏠 Project Home

☰ Screening

📄 Data Extraction

📊 Results

Projects > Unnamed Project

Project Scope

Project title

Title of your project.

Description of Review

Description of your project.

Population

Describe your target population.

Intervention/Exposure

Describe the intervention/exposure.

Comparator

Describe the comparator of interest.

Outcomes

Describe the outcomes of interest.

Numbered Inclusion/Exclusion Criteria

Description of your project.

Validate Criteria

Project Configuration

Endpoint URL

Endpoint URL for LLM

Model Name

Select a model

API Key

API Key

API Version

API Key

Save project details



David Nguyen
McMaster University

➔ Log out

User Interface

Screening



Medevi

🔍 Home

✎ Projects

Current project

🏠 Project Home

≡ Screening

📄 Data Extraction

🔍 Results

Projects > Project Title

Screening

Screening for your current project.

1. Title/Abstract Screening

Source name

Source of references (e.g., PubMed, Ovid MEDLINE)
Must be .bib, .ris, or PubMed .txt format.

Choose File No file chosen

Import references

Screen

2. Full-Text Screening

Upload PDF files of the manuscripts that have passed title/abstract screening. Our algorithm will automatically match the abstracts to your PDF files. For most accurate matching, rename your PDF files to "AuthorYear.pdf"

Choose PDF Files No files chosen

Match to screened abstracts

Screen

Title/Abstract Screening Results

- Total imported:** 300 abstracts (200 unique)
 - Source 1: 100 abstracts (+100 unique)
 - Source 2: 100 abstracts (+75 unique)
 - Source 3: 100 abstracts (+25 unique)
- Abstracts excluded:** 185 abstracts
- Abstracts included:** 15 abstracts

Full-Text Screening Results

- Total screened:** 15 full-text abstracts
- Abstracts excluded:** 3 abstracts
- Abstracts included:** 12 abstracts

Import History

☐ Event	Timestamp	Event type	
☐ Uploaded 15 PDF files (a.pdf, ...	2025/08/07 at 9:46 PM	Upload full-text	✎ 🗑
☐ Imported "PubMed" (100)	2025/08/07 at 9:44 PM	Import references	✎ 🗑



David Nguyen
McMaster University

➔ Log out

User Interface

Data Extraction


Medevi

Q Home

P Projects

Current project

P Project Home

≡ Screening

□ Data Extraction

○ Results

 **David Nguyen**
McMaster University

[Log out](#)

Projects > Project Title

Data Extraction

Define the **data extraction** table for your current project.

Configure Data Extraction Table

You can import your configuration parameters from a JSON file. This will allow you to load your settings and questions for data extraction

Drag and Drop file here No file chosen Limit 200MB per File * JSON

Import

Download

Save to library

Data Editor

Please fill out the following form to define all the fields and questions you would like to extract from each study.

☐ Variable	Question Prompt	Return Type	High Precision?
☐ Article Title	What is the title of the author?	Text	☐
☐ Risk of Bias	What is the risk of bias?	Multi-select	☐

Run

Extracting study 5/10 (Lee et al., 2021. doi:10.00/JOURNAL42).

Output format

Select an output format

Download output