



Medevi: Large Language Model-Assisted Critical Appraisal for Medical Education Research

David Nguyen, Andrew Ganea, Minh Hoang Ton, Matthew Sibbald, Medevi Team



ICAM 2026



Why automate critical appraisal assessments?

Risk-of-bias (RoB) assessments are essential for trustworthy evidence synthesis in **evidence-based medical education (EBME)**, yet remain slow, variable, and costly.

In medical education, the **Medical Education Research Quality Index (MERSQI)** and **Newcastle Ottawa Scale-Education (NOS-E)** scales are frequently used to assess trustworthiness of studies.

The need for rigorous, validated AI-assisted critical appraisal:

- **Resource burden:** SRs are costly and time-consuming to complete but are the **gold standard** for evidence in EBME.
- **Limitations of existing tools:** Automation tools (e.g., Elicit, RobotReviewer) are increasingly adopted in evidence synthesis but are (1) limited capability or (2) lack benchmarks for critical appraisal tasks.
- **Consumer tools:** Consumer tools such as ChatGPT are widely used in knowledge work; however, are not validated for research tasks, are unreliable and can hallucinate critical information.

Need: Robust, transparent benchmarks to evaluate and deploy LLM-assisted MERSQI & NOS-E 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

Comparison of MERSQI and NOS-E

Tool	Domains included	Inter-rater reliability (between human raters)	Usage - When each Scale is Used	Strengths / Limitations
MERSQI (Medical Education Research Study Quality Instrument)	6 domains, 10 items: Study Design; Sampling (institutions, response rate); Type of Data; Validity Evidence (Content, Internal Structure, Relationships with Other Variables); Data Analysis (appropriateness, sophistication); Outcomes	Per-item reliability ranges from poor to almost perfect ; with ICCs ranging from 0.38-0.95 , with pooled MERSQI overall reliability around 0.79 .	Use when appraising quantitative experimental, quasi-experimental, or observational medical education studies	Strengths: More specific and objective scoring criteria; tailored specifically to medical education Limitations: Does not directly assess blinding or comparability of cohorts/groups, requires subjective judgement, particularly in validity domains
NOS-E (Newcastle-Ottawa Scale-Education)	5 domains: Representativeness of Intervention Group; Selection of Comparison Group; Comparability of Comparison Group; Study Retention; Blinding of Assessment	Per-domain reliability ranges also from poor to good agreement ; with ICCs ranging from 0.44-0.75 , with pooled NOS-E overall reliability around 0.82 .	Use for quantitative, non-randomized comparative studies that have a separate comparison group or one-group pre/post comparison	Strengths: Better captures issues related to group comparability, retention, and blinding, Widely recognized due to derivation from a general epidemiological tool Limitations: Less broadly applicable to medical education literature; Omits some features covered by MERSQI which can lead to oversimplifying methodological quality.

MERSQI Operational Definitions

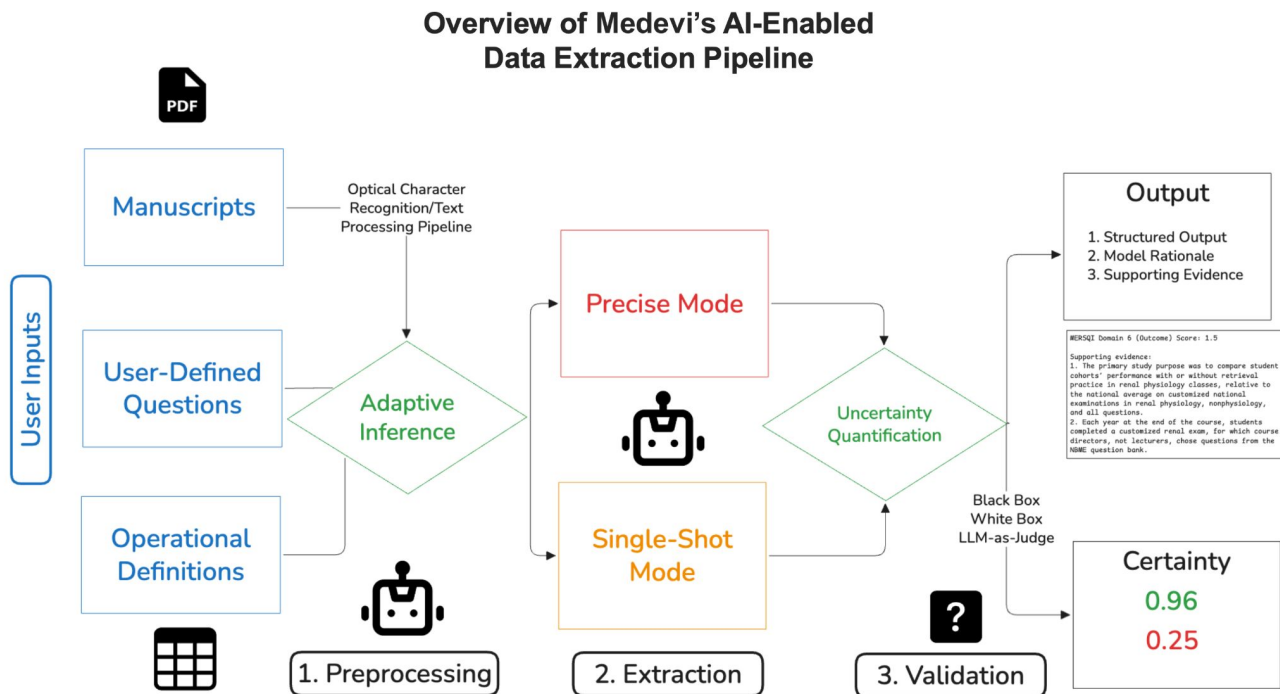
Domain: Item	Response Options: Scores	Operational Definitions
Study design	Single-group cross-sectional or single-group posttest only: 1, Single-group pretest and posttest: 1.5, Nonrandomized, 2 group: 2, Randomized controlled trial: 3	- Survey studies are cross-sectional - Case-control and cohort studies (2 or more defined cohorts) are considered 2-group nonrandomized.
Sampling	1 institution: 0.5, 2 institutions: 1, 3 or more institutions: 1.5 < 50% or not reported: 0.5, 50%–74%: 1, ≥ 75%: 1.5	- Number of institutions refers to origin of study participants - Response rate is the proportion of those eligible who completed the posttest or survey
Type of Data	Assessment by study participant: 1, Objective: 3	Observer ratings are considered objective.
Validity Evidence	Content: 1, Internal structure: 1, Relationships to other variables: 1	Relevant evidence includes using theory, guidelines, experts, and existing instruments to identify or refine the instrument, all reliability and factor analysis, and concurrent or predictive correlation with other variables.
Data Analysis	Descriptive analysis only: 1, Beyond descriptive analysis: 2, Data analysis appropriate for study design and type of data: 1	- Any test of statistical inference is considered “beyond descriptive.” - Data Analysis: Considered “no” if there is a statistical error or if authors failed to analyze data at all.
Outcome	Satisfaction, attitudes, perceptions, opinions, general facts: 1, Knowledge, skills: 1.5, Behaviors: 2, Patient/health care outcome: 3	Patient/health care outcomes are actual effects on real patients, programs, or society.

NOS-E Operational Definitions

Domain: Item	Response Options: Scores	Operational Definitions
Representativeness of Intervention Group	- Not representative: 0 - Very or somewhat representative of average learner in community: 1	“Very” representative indicates that all or a random sample of eligible learners enrolled. “Somewhat” representative indicates that 75%–99% of eligible learners enrolled
Selection of Comparison Group	- No separate comparison group (e.g., single-group pretest–posttest): 0 - Drawn from a different community: 0 - Drawn from the same community: 1	“Same” community indicates that there is no obvious reason to suspect systematic difference between intervention and comparison group.
Comparability of Comparison Group	- No separate comparison group: 0 - Randomized study: - Allocation not concealed: 1 - Allocation concealed: 2 - Nonrandomized study: - Controlled for 1 subject characteristic: 1 - Controlled for 2 or more subject characteristics: 2	- Allocation is considered concealed if enrollment, consent, or baseline assessment preceded randomization. - Controlling for subject characteristics requires statistical covariate analysis
Study Retention	- Poor retention could introduce bias: 0 - Retention unlikely to introduce bias: 1	High if $\geq 75\%$ of those enrolled provided outcome data, or if authors described those lost to follow-up.
Blinding of Assessment	- Outcome assessment not blinded: 0 - Outcome assessment blinded: 1	Blinded if the assessor cannot be influenced by group assignment.

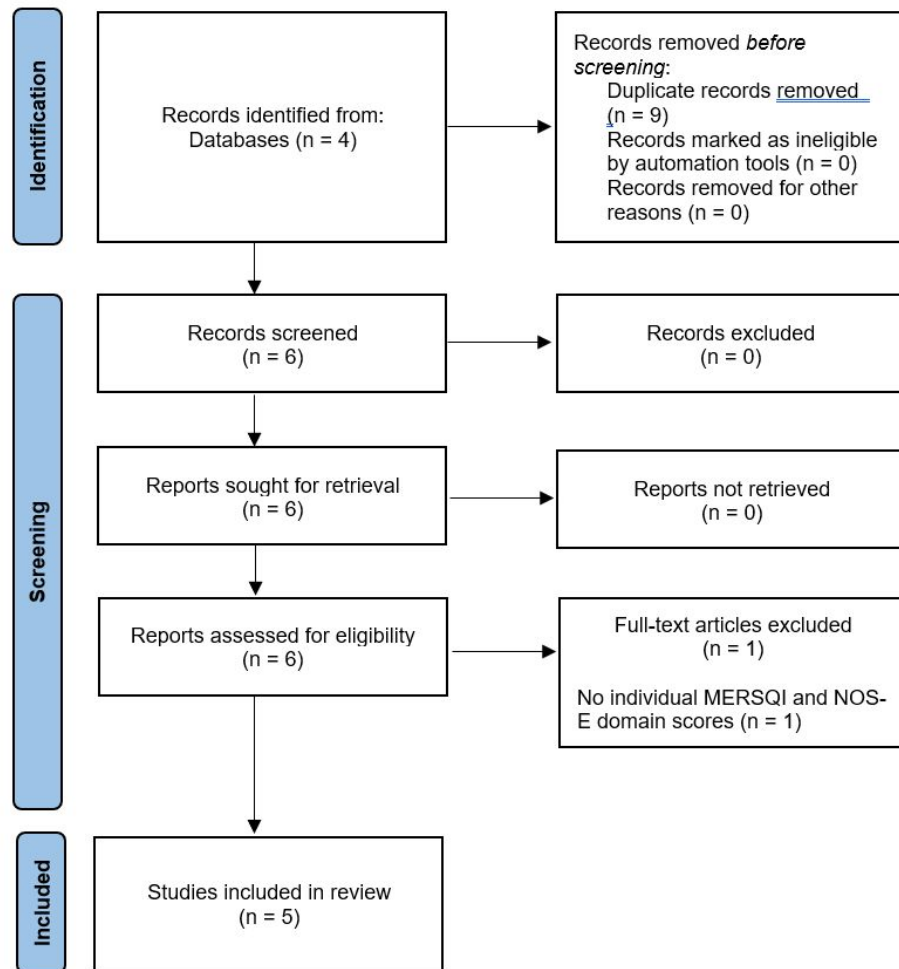
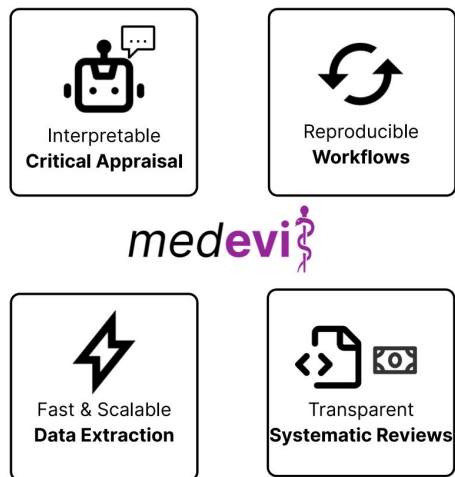
Medevi platform enables interpretable LLM evaluations

Overview of Medevi's AI-Enabled Data Extraction Pipeline for MERSQI/NOS-E Assessment



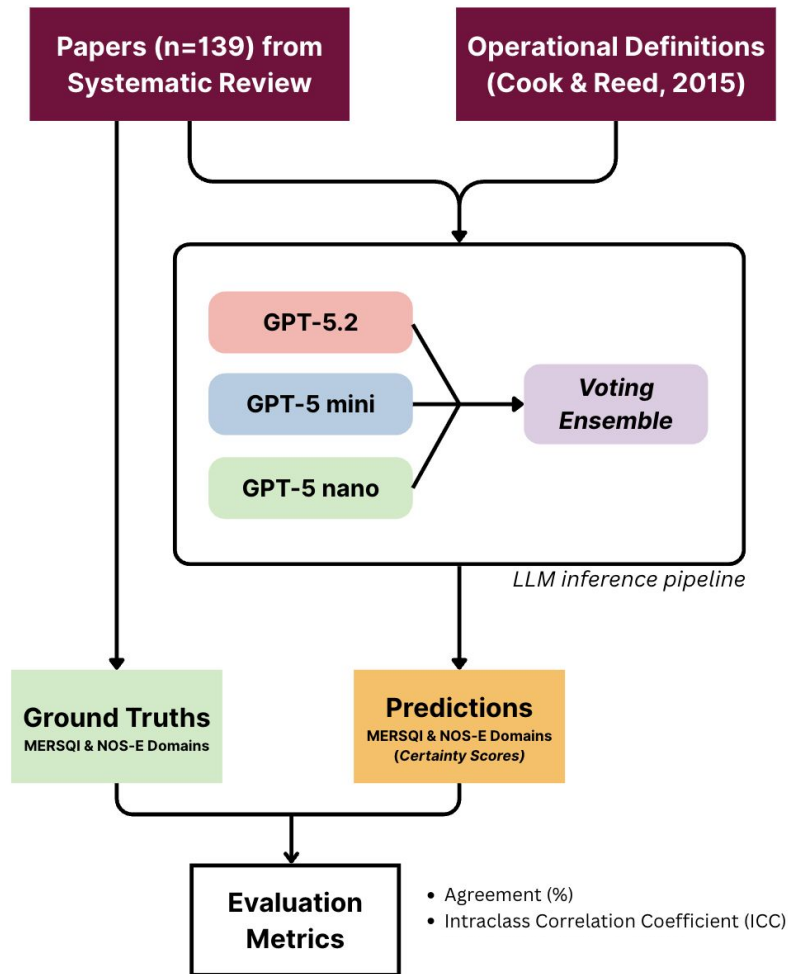
Methods

- We focused on evaluating Medevi's performance on appraising medical education studies
- Medevi offers substantial flexibility and improvements over existing options and manual data extraction



Dataset

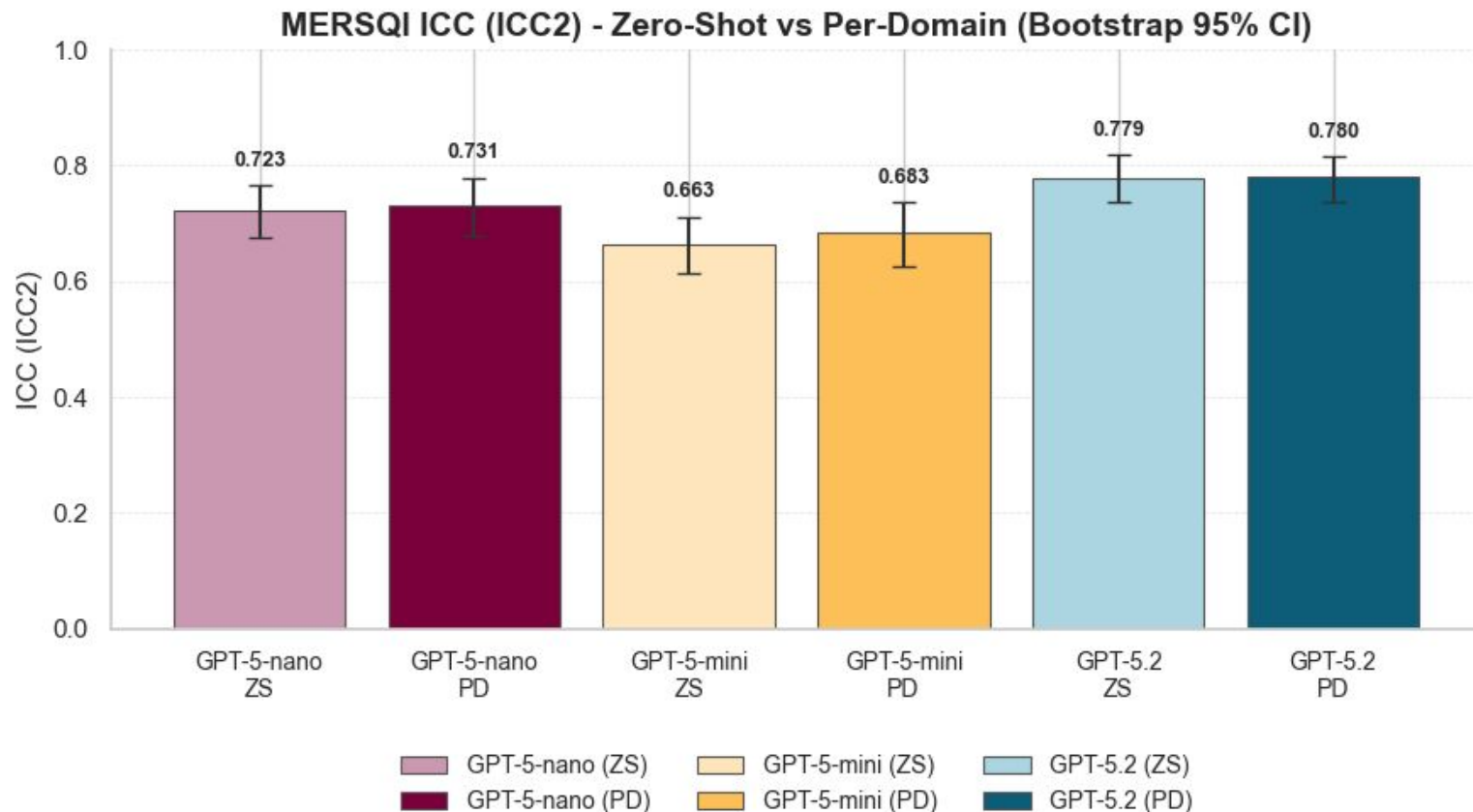
- 139 peer-reviewed articles with corresponding **MERSQI** and **NOS-E** scores, sourced from **5 distinct systematic reviews**
- Field representation was diverse, including retrieval practices, surgical skills, pain management education & virtual reality simulation



Reviews included in benchmark dataset

Title	Author	Year of Publication	Number of Articles	Journal	Field of Systematic Review
Systematic review of distributed practice and retrieval practice in health professions education	Emma Trumble	2024	58	Advances in Health Sciences Education	Retrieval Practice
Crowdsourced assessment of surgical skills: A systematic review	Rikke G Olsen	2022	32	The American Journal of Surgery	Surgical Skills
A Systematic Review of Pain Management Education in Graduate Medical Education	Zayir Malik	2022	25	Journal of Graduate Medical Education	Pain Management Education
Effectiveness of virtual reality interactive simulation practice in prosthodontic education: A systematic review and meta-analysis	Hang-Nga Mai	2025	15	European Journal of Dental Education	Virtual Reality Simulation
Virtual reality simulation in robot-assisted surgery: Meta-analysis of skill transfer and predictability of skill	Mona Wanda Schmidt	2021	8	BJS Open	Virtual Reality Simulation

MERSQI: LLMs achieve parity in terms of inter-rater reliability compared to human evaluators. Zero-shot CoT evaluations with the latest model context sizes are non-inferior to per-domain CoT evaluations.



MERSQI: LLMs achieve strong performance across MERSQI domains, with weakness in domains requiring domain-specific knowledge.

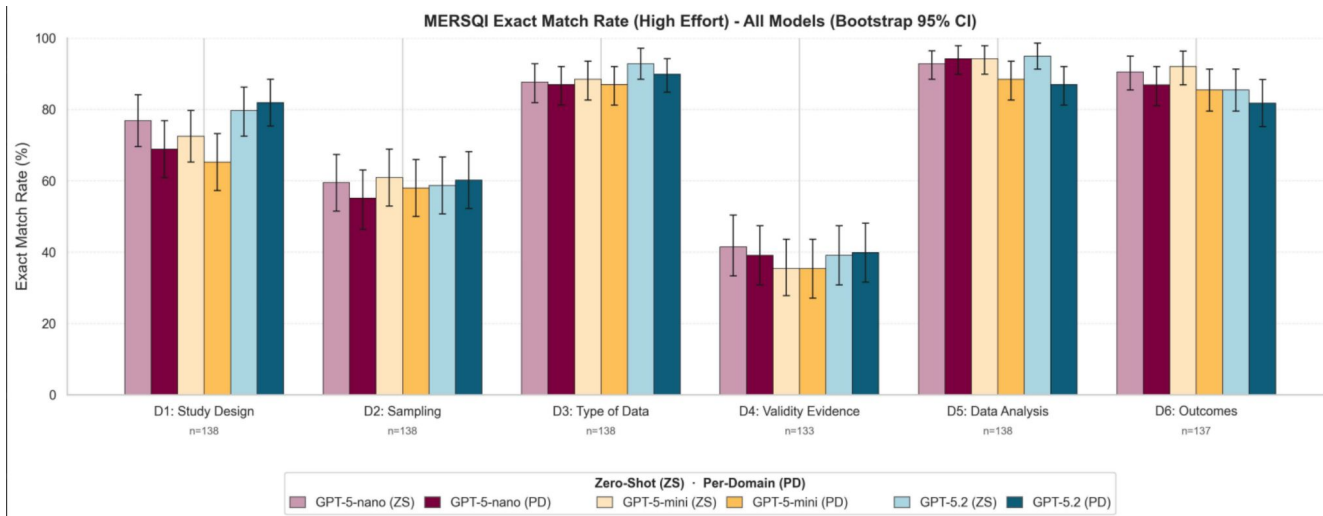
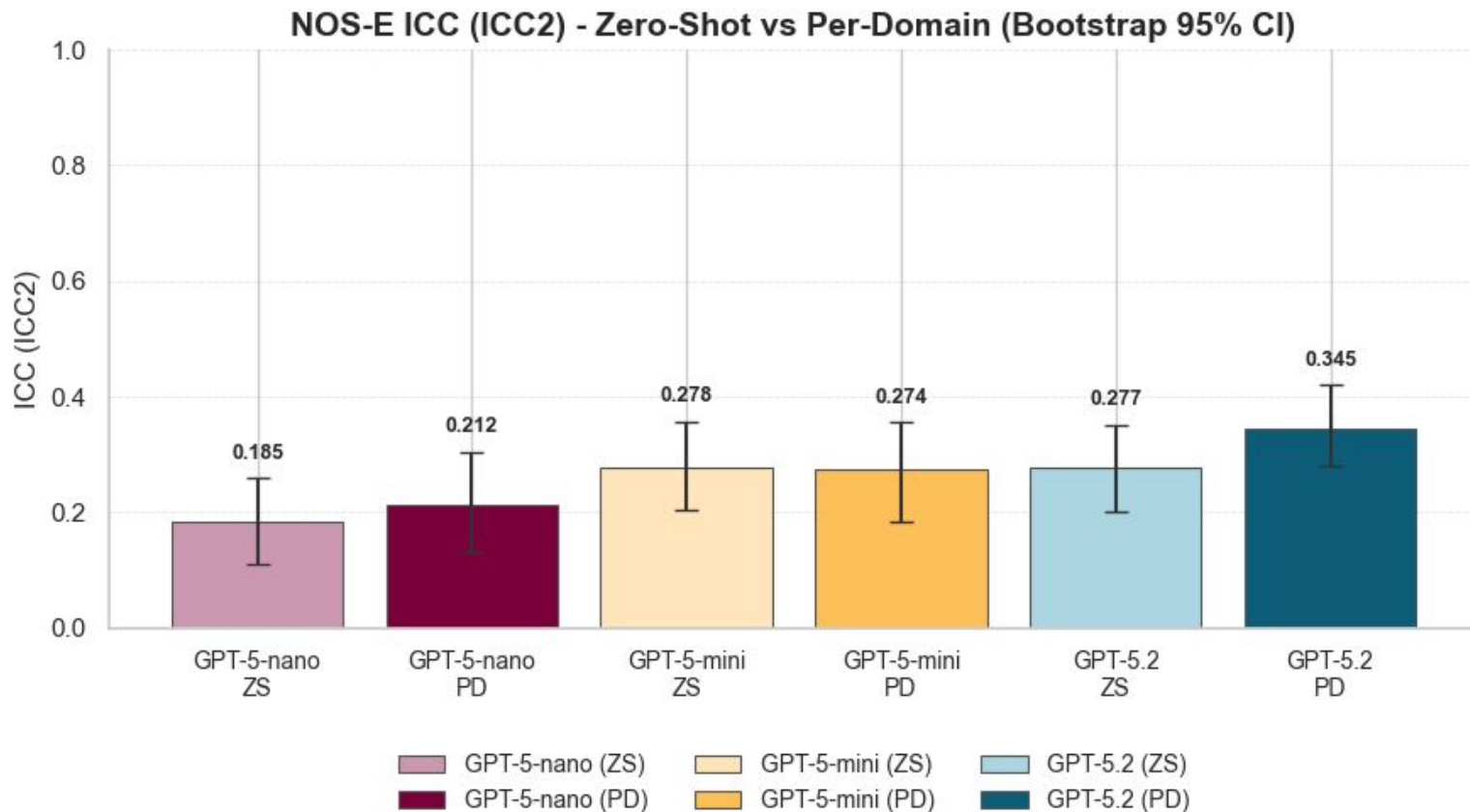


Table 1b - MERSQI Exact Match Rate % by Domain (Bootstrap 95% CI) (bold = best per domain, higher is better)

	D1 Study Design	D2 Sampling	D3 Type of Data	D4 Validity Evidence	D5 Data Analysis	D6 Outcomes
GPT-5-nano (ZS)	76.8 (69.6, 84.1)	59.4 (51.4, 67.4)	87.7 (81.9, 92.8)	41.5 (33.3, 50.4)	92.8 (88.4, 96.4)	90.5 (85.4, 94.9)
GPT-5-nano (PD)	68.8 (60.9, 76.8)	55.1 (46.4, 63.0)	87.0 (81.2, 92.0)	39.1 (30.8, 47.4)	94.2 (89.9, 97.8)	86.9 (81.0, 92.0)
GPT-5-mini (ZS)	72.5 (65.2, 79.7)	60.9 (52.9, 68.8)	88.4 (82.6, 93.5)	35.3 (27.8, 43.6)	94.2 (89.9, 97.8)	92.0 (86.9, 96.4)
GPT-5-mini (PD)	65.2 (57.2, 73.2)	58.0 (50.0, 65.9)	87.0 (81.2, 92.0)	35.3 (27.1, 43.6)	88.4 (82.6, 93.5)	85.4 (79.5, 91.2)
GPT-5.2 (ZS)	79.7 (72.5, 86.2)	58.7 (50.7, 66.7)	92.8 (88.4, 97.1)	39.1 (30.8, 47.4)	94.9 (91.3, 98.6)	85.4 (79.5, 91.2)
GPT-5.2 (PD)	81.9 (75.4, 88.4)	60.1 (52.2, 68.1)	89.9 (84.8, 94.2)	39.8 (31.6, 48.1)	87.0 (81.2, 92.0)	81.8 (75.2, 88.3)

NOS-E: Models demonstrate poor ($ICC < 0.5$) overall agreement with human ratings, with the per-domain CoT approach performing best.



NOS-E: Zero-shot models made similar errors to per-domain models across GPT 5 model sizes.

Logic-heavy domains had marginally increased performance with increased model size.

Results

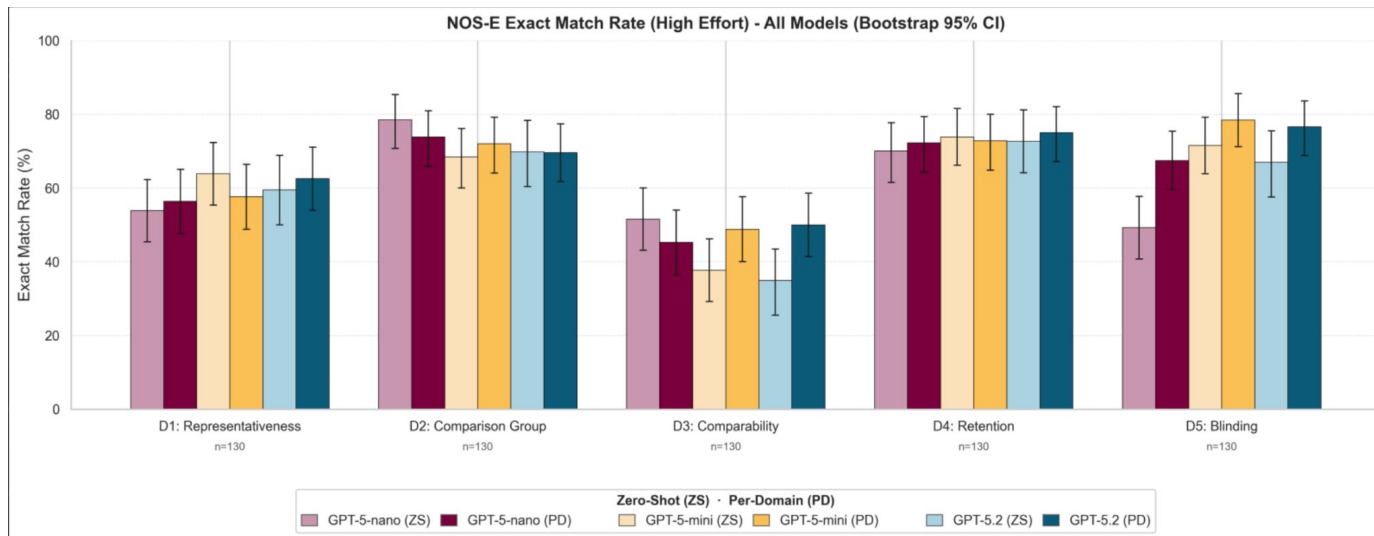


Table 2b - NOS-E Exact Match Rate % by Domain (Bootstrap 95% CI) (bold = best per domain, higher is better)

	D1 Representativeness	D2 Comparison Group	D3 Comparability	D4 Retention	D5 Blinding
GPT-5-nano (ZS)	53.8 (45.4, 62.3)	78.5 (70.8, 85.4)	51.5 (43.1, 60.0)	70.0 (61.5, 77.7)	49.2 (40.8, 57.7)
GPT-5-nano (PD)	56.3 (47.6, 65.1)	73.8 (65.9, 81.0)	45.2 (36.5, 54.0)	72.2 (64.3, 79.4)	67.5 (59.5, 75.4)
GPT-5-mini (ZS)	63.8 (55.4, 72.3)	68.5 (60.0, 76.2)	37.7 (29.2, 46.2)	73.8 (66.2, 81.5)	71.5 (63.8, 79.2)
GPT-5-mini (PD)	57.6 (48.8, 66.4)	72.0 (64.0, 79.2)	48.8 (40.0, 57.6)	72.8 (64.8, 80.0)	78.4 (71.2, 85.6)
GPT-5.2 (ZS)	59.4 (50.0, 68.9)	69.8 (60.4, 78.3)	34.9 (25.5, 43.4)	72.6 (64.2, 81.1)	67.0 (57.5, 75.5)
GPT-5.2 (PD)	62.5 (53.9, 71.1)	69.5 (61.7, 77.3)	50.0 (41.4, 58.6)	75.0 (67.2, 82.0)	76.6 (68.8, 83.6)

Increased CoT context length improves model performance for both NOS-E and MERSQI irrespective of thinking model size.

Delta: MERSQI Exact Match % (High Effort - Standard) [green = HE higher accuracy]

	D1 Study Design	D2 Sampling	D3 Type of Data	D4 Validity Evidence	D5 Data Analysis	D6 Outcomes
GPT-5-nano (ZS)	+5.1	+6.5	+2.2	+3.3	+0.7	+0.7
GPT-5-nano (PD)	-2.9	+2.9	-1.4	-1.5	+1.4	-5.1
GPT-5-mini (ZS)	+10.9	+8.0	+5.8	+0.8	0.0	+2.2
GPT-5-mini (PD)	-3.6	+1.4	-0.7	+0.8	-1.4	-2.2
GPT-5.2 (ZS)	+2.9	+0.7	+5.8	-4.5	+0.7	-4.4
GPT-5.2 (PD)	+8.7	+9.4	0.0	-5.3	-7.2	-10.9

Delta: NOS-E Exact Match % (High Effort - Standard) [green = HE higher accuracy]

	D1 Representativeness	D2 Comparison Group	D3 Comparability	D4 Retention	D5 Blinding
GPT-5-nano (ZS)	+6.2	+3.8	0.0	+8.5	-1.5
GPT-5-nano (PD)	-5.2	-0.8	+6.0	-0.1	+6.7
GPT-5-mini (ZS)	+5.4	-0.8	-6.2	+3.1	+1.5
GPT-5-mini (PD)	+0.7	+2.0	+3.4	+2.0	+8.4
GPT-5.2 (ZS)	-0.6	+1.3	-19.7	+10.3	+4.7
GPT-5.2 (PD)	+6.3	-4.3	-5.4	+8.1	+0.4

Comparison between human data extraction and LLM-assisted data extraction with Medevi

Metric	Human Data Extraction	Medevi
<i>Time per Paper</i>	65-130 minutes (estimate: 5-10 minutes per domain)	29-57 seconds per paper (scalable with no overhead)
<i>Cost per paper</i>	\$21.67-\$43.33 (assuming conservative labour cost of \$20/h per paper)	Zero-shot CoT mode: \$0.10/paper (99%+ cost reduction) Per-domain CoT mode: \$0.57/paper (97.4-98.6% cost reduction)
<i>Scalability</i>	Limited	Infinitely scalable
<i>Agreement with Human Raters</i>	MERSQI: Substantial agreement (ICC = 0.79)	MERSQI: Substantial agreement (ICC = 0.755; 95% CI [0.704, 0.801])
	NOS-E: Almost perfect agreement (ICC = 0.82)	NOS-E: Fair agreement (ICC = 0.345; 95% CI [0.268, 0.424])

Summary of Findings

- Medevi is an **open-source tool** that enables **cost-effective & transparent AI-assisted data extraction** for systematic reviews
- LLM-assisted data extraction achieved **good agreement** with human reviewers on total MERSQI score (ICC = 0.755) and **poor agreement** on total NOS-E score (ICC = 0.345)
- Domain-specific agreement with human reviewers ranged from **35.3% to 94.9%** in **MERSQI** domains, and **34.9% to 78.5%** in **NOS-E** domains

Future Directions

- Assess Medevi's capabilities in **title/abstract** and **full-text screening** tasks for medical education studies specifically
- Assess bias in LLMs and evaluate utility of explainable outputs, as rated by human users
- Evaluate the performance of Medevi on general medical systematic review critical appraisal tasks (e.g., **RoB 2, ROBINS-I, QUADAS-2, PROBAST**)

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

Acknowledgements

Medevi Team

Evaluation & Data Extraction: Andrew Ganea, Michael Nguyen, Dylan Kwan, Nathan Lai, Henry Liu, Chris Ouyang, Jessica Song, Sarah Pimenta

Software Development: Minh Hoang Ton, Michael Nguyen, Kathy Tang

Advisors:

Matthew Sibbald (MedEd), Dena Zeratkaar, Gordon Guyatt (RoB)

Funding Support:

Microsoft, Cohere Labs

Disclosures:

Not applicable.

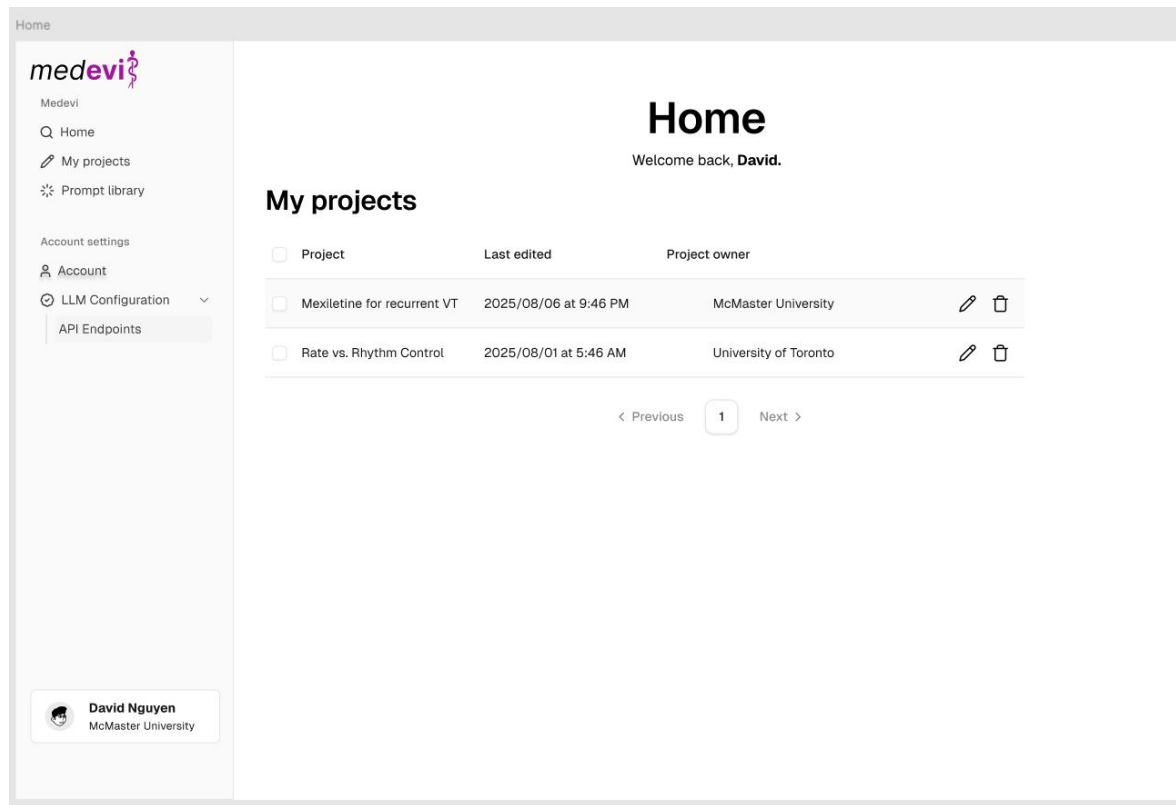
References



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