MSR29

AI-Assisted Risk Of Bias Assessment In Observational Studies: Isolating Judgment Using A Passage-Based Validation Approach

Eitan Agai[1], Karen A. Robinson[2], Alon Agai[1]

1 PICO Portal, St. Petersburg, FL, USA, 2 Johns Hopkins University, Baltimore, MD, USA

Introduction

Risk of bias (RoB) assessment is a required step in systematic reviews (SRs), yet it is time-intensive, resource-heavy, and prone to inconsistency across reviewers.

Observational studies present unique methodological challenges, including selection bias and confounding, and are the norm in environmental health SRs.

Several tools assess RoB in non-randomized studies, including ROBINS-I, ROBINS-E, and the Navigation Guide (NavGuide), developed specifically for environmental exposure research.

Most prior LLM evaluations have focused on randomized controlled trials (RCTs); performance in observational studies remains unclear and inconsistent.

Objective

To evaluate how well two LLMs assess RoB in observational studies of environmental exposures and health outcomes across eight NavGuide domains, using human-identified text passages as input.

Results

Overall Agreement

For exact matches: GPT-5 achieved **64.5%** agreement and Gemini **50.7%** with human consensus.

For partial matches (direction agreement): GPT-5 achieved **91.5%** and Gemini **95.6%**.

Weighted Kappa for exact matches was low for both (GPT-5: κ = 0.073; Gemini: κ = 0.042); partial match Kappa improved but remained modest (GPT-5: κ = 0.241; Gemini: κ = 0.269).

Human-to-consensus exact agreement was substantially higher: 87.5%–91.3% (κ = 0.326–0.486).

Table 2. Overall results for each LLM and humans

	Exact Match			Partial Match		
	N	Agreement, %	Weighted Kappa	N	Agreement, %	Weighted Kappa
GPT-5-Human Consensus	1007	64.5	0.0733	1007	91.5	0.241
Gemini-Human Consensus	1007	50,7	0.0422	1007	95.6	0.269
Human 1-Human Consensus	1007	87.5	0.326	1007	97.9	0.478
Human 2-Human Consensus	1007	91.3	0.486	1007	98.5	0.643
Human 1-Human 2	1007	81.1	0.108	1007	97.4	0.181

Notes: *Color does not convey meaning but distinguishes exact match and partial match results. Partial match = agreement in direction but necessarilly in magnitude (e.g. low and probably low).*

Performance by Domain

- GPT-5 exact match ranged from **9%** (Domain 7: selective outcome reporting) to **91%** (Domain 8: conflict of interest).
- Gemini exact match ranged from **27%** (Domain 1: selection bias) to **90%** (Domain 4: outcome assessment).
- Both LLMs showed consistently strong performance in **Domain 8** and **Domain 4**; worst performance in **Domain 1** and **Domain 7.**

Conservativeness

- Both LLMs were systematically more conservative than human reviewers — frequently assigning probably low when humans assigned low, or high when humans assigned probably high.
- GPT-5 was more conservative in **32.6%** of assessments; Gemini in **46.3%**. Both models were less conservative **3%** of the time.

Reproducibility

- Across five runs: GPT-5 showed **62% exact** and **90% partial** consistency; Gemini showed **59% exact** and **95% partial** consistency.
- At least one run matched the human consensus in **78%** (GPT-5) and **75%** (Gemini) of citations for exact match; **94%** and **97%** for partial match.

Post-Adjudication (GPT-5)

- Directional disagreements occurred in **8%** of total records (range: 1% in Domain 3 to 24% in Domain 1).
- After expert adjudication incorporating LLM thinking traces: agreed with humans **47%**, GPT-5 **31%**, and between both **22%** of the time.
- LLM input prompted a change in the final consensus rating in approximately **4.25% of cases.**

Methods

- We evaluated two LLMs — **GPT-5-mini** (version 2025-08-07) and **Google Gemini-3** (January 2026) — on **128 open-access observational studies** from a published SR on per- and polyfluoroalkyl substances (PFAS) and health outcomes ([1] National Academies of Sciences, 2022).
- For each article-domain pair, the LLM received: (1) the NavGuide RoB question and guidance [2], and (2) the relevant text passages identified by human reviewers — but not the full-text article or human rationales.
- LLMs returned a structured RoB rating using the NavGuide's five-level scale: high, probably high, probably low, low, not applicable.
- Both LLMs were run in a **zero-shot configuration** with default settings; each article-domain pair was run **five times** to assess reproducibility.
- Ratings were compared to human-adjudicated consensus across **1,007 article-domain pairs** spanning 8 of 9 NavGuide domains.
- Agreement was assessed as: **Exact Match** (identical labels), **Partial Match** (same direction, e.g., low vs. probably low), and **Disagreement**.
- Weighted Cohen's Kappa was calculated for both exact and partial matches.
- For GPT-5, human reviewers qualitatively evaluated LLM thinking traces in cases of directional disagreement; a senior reviewer adjudicated these cases.
- All work was conducted within the **PICO Portal** web-based evidence synthesis platform; protocol was prospectively registered via OSF.

Conclusions

Although LLMs did not reach the accuracy of human reviewers, they show potential utility as a **second reviewer** with human oversight, particularly for flagging cases where humans may be making insufficiently conservative assessments.

Both LLMs tended to treat **missing information as a risk of bias**, while human reviewers more often assumed acceptable methodology in the absence of explicit reporting — a key source of discrepancy.

LLM thinking traces provided meaningful transparency, and incorporating them into adjudication workflows may **improve consistency** in human RoB assessment.

Performance was moderate and varied across domains and runs; domains involving nuanced judgement (e.g., selection bias, selective outcome reporting) remain challenging.

Future work will evaluate whether LLMs can independently identify relevant passages, removing the need for human passage extraction, and will integrate LLM-assisted RoB assessment directly into PICO Portal.

ACKNOWLEDGMENTS: We would like to acknowledge the National Academies of Sciences, Engineering and Mathematics for allowing us to use their systematic review as an example.

This research was funded by the National Institute of Environmental Health Sciences under NIH Grant 1R43ES037589-0

REFERENCES [1] National Academies of Sciences E, Medicine, Health, et al. The National Academies Collection: Reports funded by National Institutes of Health. Guidance on PFAS Exposure, Testing, and Clinical Follow-Up. National Academies Press (US) [2] Woodruff TJ, Sutton P. The Navigation Guide systematic review methodology: a rigorous and transparent method for translating environmental health science into better health outcomes. Environ Health Perspect. Oct 2014;122(10):1007-14. doi:10.1289/ehp.1307175