BOND UNIVERSITY | INSTITUTE FOR Evidence-Based Healthcare

Generative Artificial Intelligence to conduct systematic reviews

Justin Clark (presenter), Belinda Barton, Loai Albarqouni, Oyungerel Byambasuren, Tanisha Jowsey, Justin Keogh, Tian Liang, Christian Moro, Hayley O'Neill, Mark Jones

Bond University

Brief intro

Work at the Institute for Evidence-Based Healthcare (IEBH) at Bond University, we specialise in systematic reviews

A Cochrane Information Specialist and Cochrane author, was on Cochrane Info Specialists Executive (6 years)

Founding member of the International Collaboration for the Automation of Systematic Reviews (ICASR)

Lead of the automation program at IEBH, designing/testing/evaluating the Systematic Review Accelerator (SRA)

Co-designer of Two-Week Systematic Reviews (2weekSR)

Gen AI evaluations

Quote from a paper* evaluating Gen AI for systematic reviews

Of the 1287 studies provided by ChatGPT, only 7 (0.5%) studies were perfectly eligible and 18 (1.4%) studies could be considered suitable under the assumption that they were real studies if only the title, author, journal, and publication year matched.

Among these, only 1 study was perfectly consistent with studies finally included in Lee et al

*https://medinform.jmir.org/2024/1/e51187

Systematic review - eligibility

Comparative studies of standard tasks to conduct part or all of an evidence synthesis (e.g. a full systematic review, or the screening task of a systematic review)

Interventions involving processes utilising Gen AI or large language models (LLMs), (e.g., GPT-3, Claude2, BioBERT)

Must have been compared to humans

Must report accuracy, sensitivity, specificity, error rates, or time

Included studies conducted in all research disciplines (e.g., medicine, business)

Must be published and peer reviewed

Systematic review – search etc.

Run in PubMed, Embase, Web of Science, Scopus, and Business Source Ultimate on 15th May 2024

Backwards and forwards citation search done on 18 June 2024

Screening, extraction and risk of bias all done by two people independently

Systematic review – outcomes

Measures of accuracy, error rate, sensitivity (recall), and specificity (precision) of the GenAI tool against humans were calculated using the following formulas:

Accuracy = (TP + TN)/(TP +TN +FP +FN)

Error rate: 1 – Accuracy

Sensitivity/Recall: (TP/(TP + FN)

Specificity/Precision: TP/(TP +FP)

Number needed to read: 1/Precision

Where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative, and number needed to read is the number of publications needed to screen to include 1 additional relevant study.

Systematic review – RoB

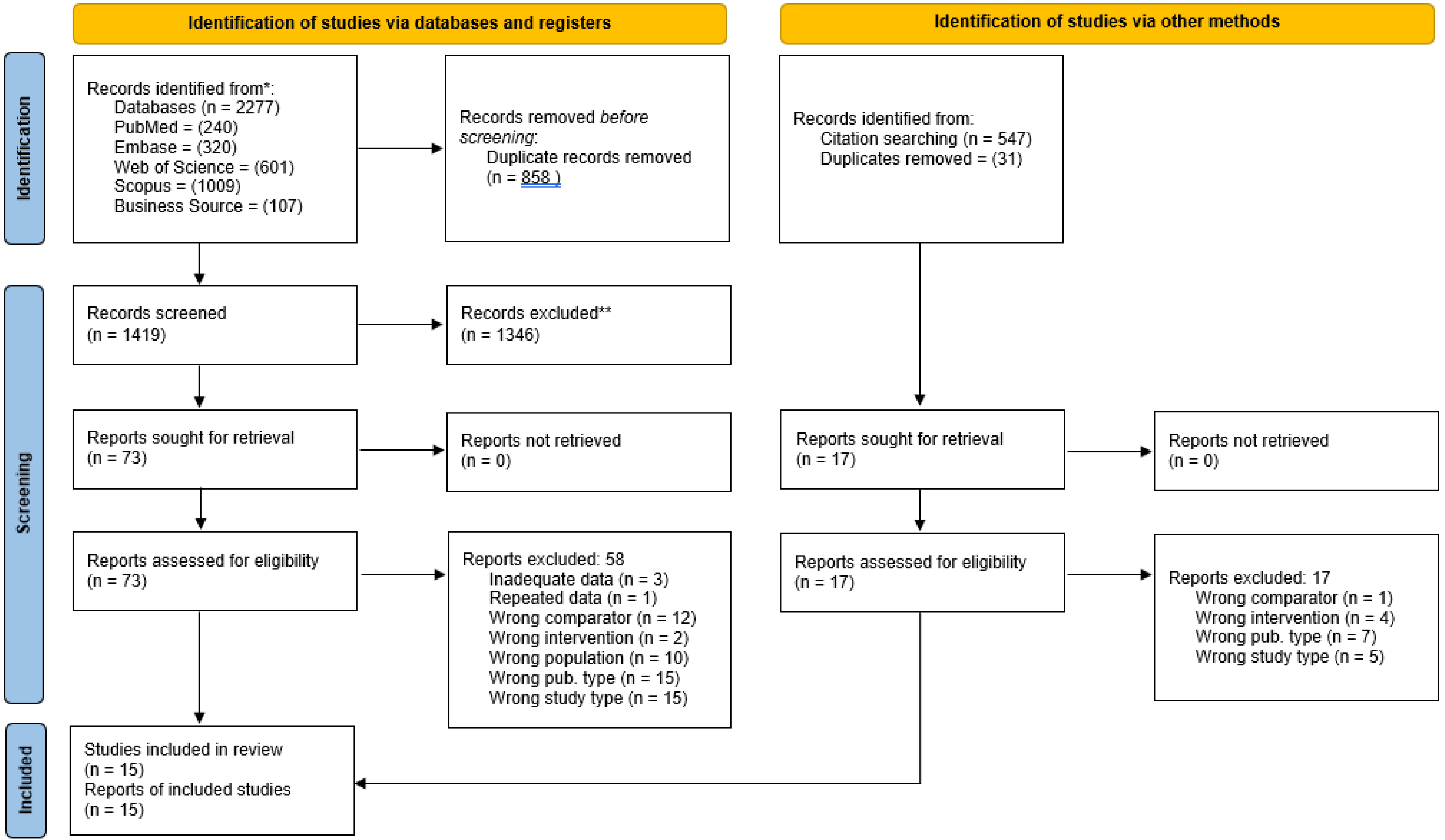
Modified QUADAS-2

Major things to note

1. Were reviews/tasked used in the study randomly selected?
2. Were prompts used pre-specified or developed iteratively
3. Was the human comparison done to an adequate level
4. Were the Gen AI tasks and human tasks done on the same topic
5. How applicable is the evaluation reviews (e.g. multiple reviews types/topics etc.)

Systematic review – results

PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources



Identification of studies via databases and registers

Identification of studies via other methods

Identification

Records identified from*:
Databases (n = 2277)
PubMed = (240)
Embase = (320)
Web of Science = (601)
Scopus = (1009)
Business Source = (107)

Records removed *before screening*:
Duplicate records removed (n = 858)

Records identified from:
Citation searching (n = 547)
Duplicates removed = (31)

Screening

Records screened
(n = 1419)

Records excluded**
(n = 1346)

Reports sought for retrieval
(n = 73)

Reports not retrieved
(n = 0)

Reports sought for retrieval
(n = 17)

Reports not retrieved
(n = 0)

Reports assessed for eligibility
(n = 73)

Reports excluded: 58
Inadequate data (n = 3)
Repeated data (n = 1)
Wrong comparator (n = 12)
Wrong intervention (n = 2)
Wrong population (n = 10)
Wrong pub. type (n = 15)
Wrong study type (n = 15)

Reports assessed for eligibility
(n = 17)

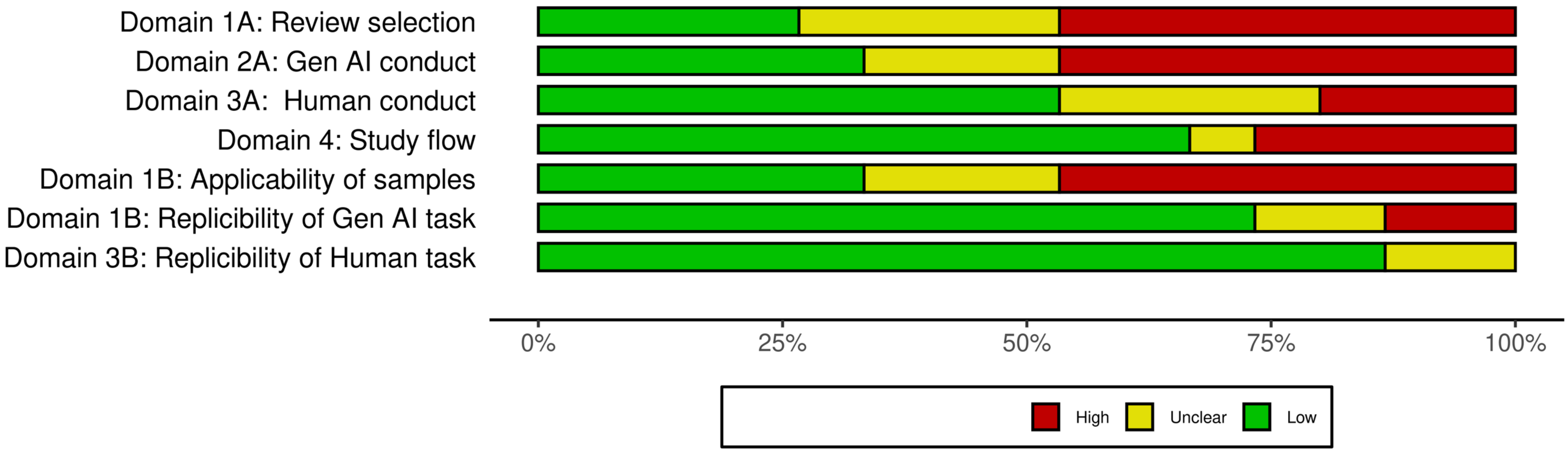
Reports excluded: 17
Wrong comparator (n = 1)
Wrong intervention (n = 4)
Wrong pub. type (n = 7)
Wrong study type (n = 5)

Included

Studies included in review
(n = 15)
Reports of included studies
(n = 15)

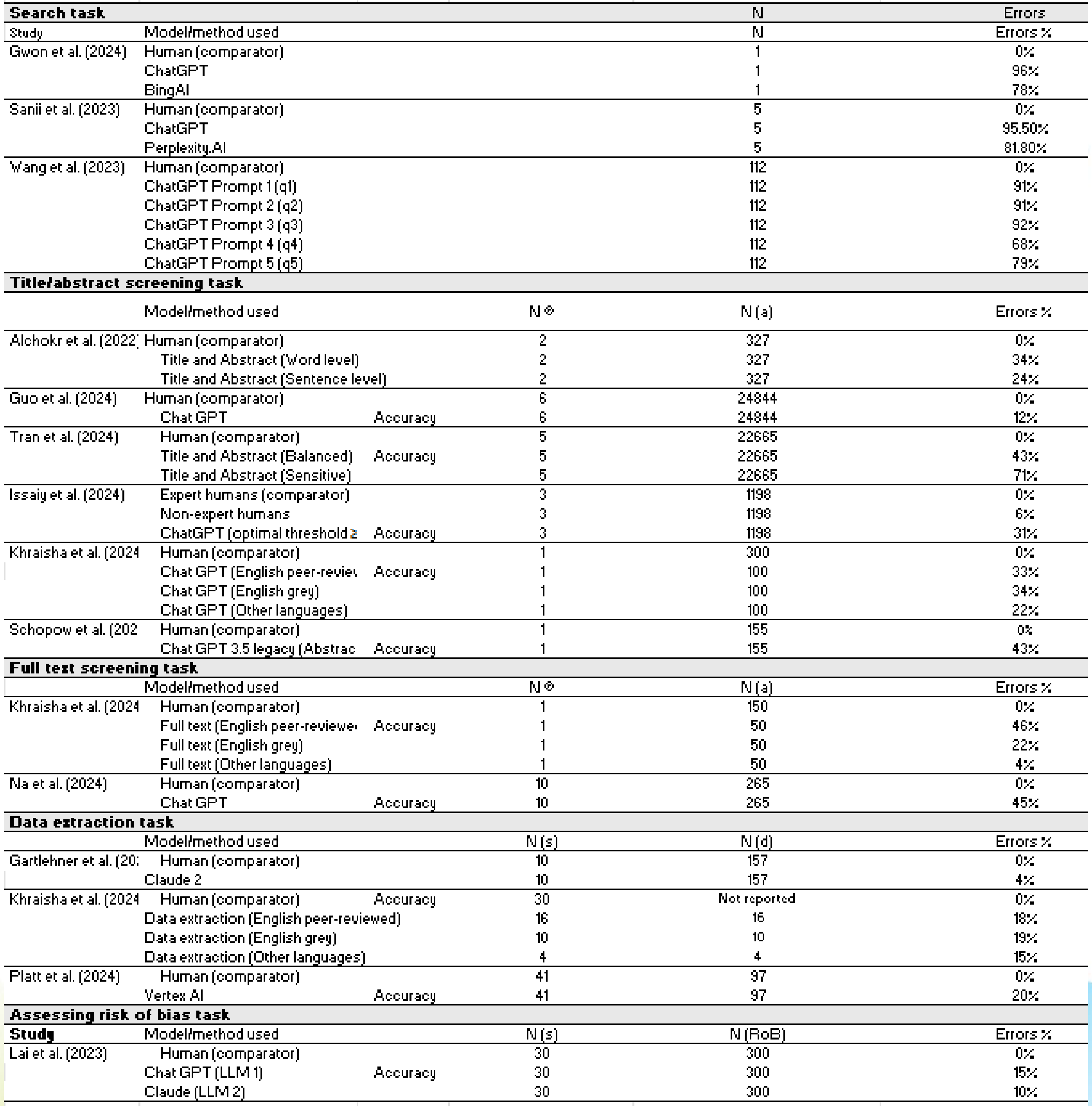
Risk of Bias

Benefits of automation tools



Results

All results

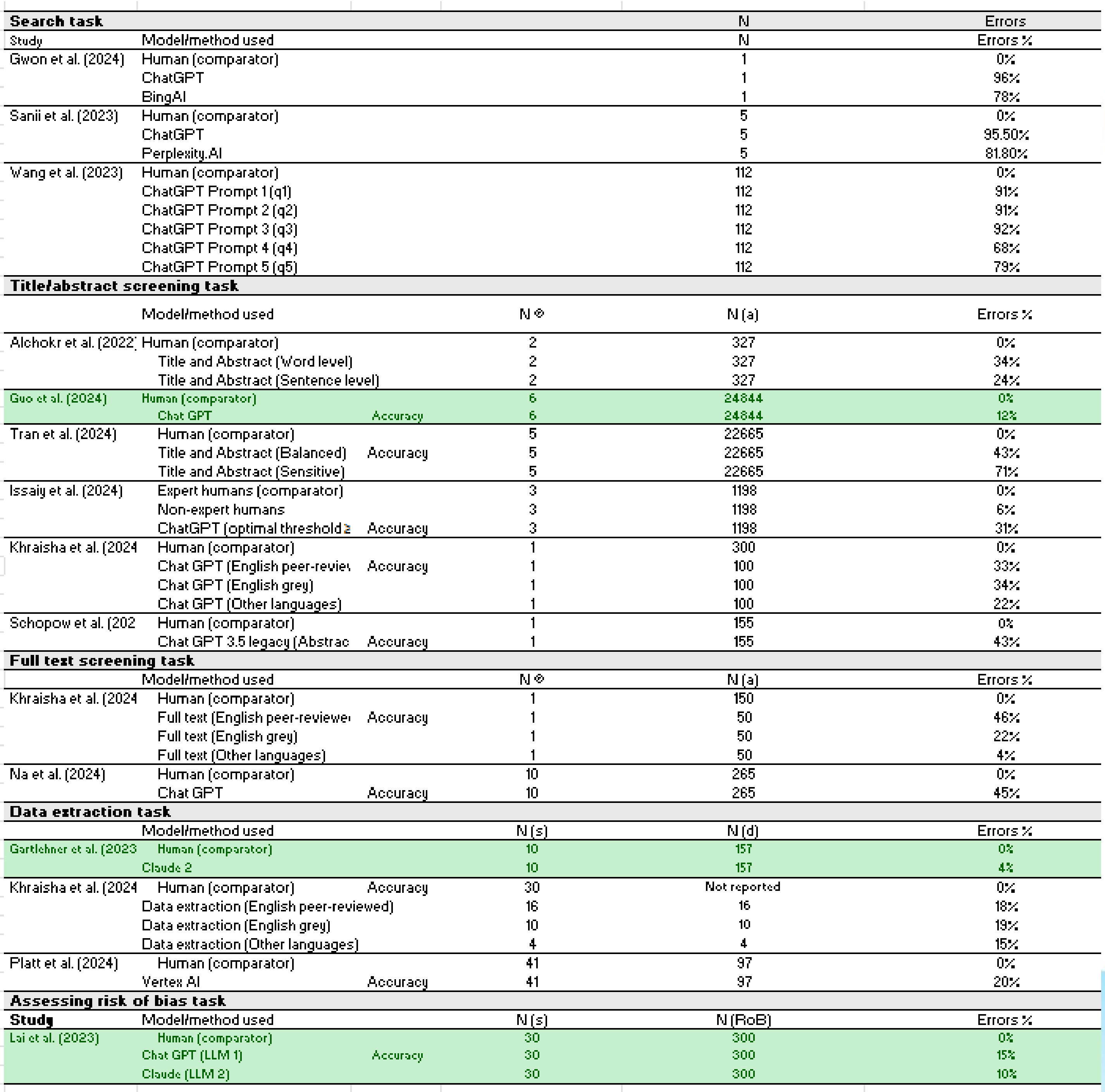


Search task				N	Errors
Study	Model/method used			N	Errors %
Gwon et al. (2024)	Human (comparator)			1	0%
	ChatGPT			1	96%
	BingAI			1	78%
Sanii et al. (2023)	Human (comparator)			5	0%
	ChatGPT			5	95.50%
	Perplexity.AI			5	81.80%
Wang et al. (2023)	Human (comparator)			112	0%
	ChatGPT Prompt 1 (q1)			112	91%
	ChatGPT Prompt 2 (q2)			112	91%
	ChatGPT Prompt 3 (q3)			112	92%
	ChatGPT Prompt 4 (q4)			112	68%
	ChatGPT Prompt 5 (q5)			112	79%
Title/abstract screening task					
	Model/method used		N ®	N (a)	Errors %
Alchokr et al. (2022)	Human (comparator)		2	327	0%
	Title and Abstract (Word level)		2	327	34%
	Title and Abstract (Sentence level)		2	327	24%
Guo et al. (2024)	Human (comparator)		6	24844	0%
	Chat GPT	Accuracy	6	24844	12%
Tran et al. (2024)	Human (comparator)		5	22665	0%
	Title and Abstract (Balanced)	Accuracy	5	22665	43%
	Title and Abstract (Sensitive)		5	22665	71%
Issaiy et al. (2024)	Expert humans (comparator)		3	1198	0%
	Non-expert humans		3	1198	6%
	ChatGPT (optimal threshold ≥	Accuracy	3	1198	31%
Khraisha et al. (2024	Human (comparator)		1	300	0%
	Chat GPT (English peer-reviev	Accuracy	1	100	33%
	Chat GPT (English grey)		1	100	34%
	Chat GPT (Other languages)		1	100	22%
Schopow et al. (202	Human (comparator)		1	155	0%
	Chat GPT 3.5 legacy (Abstrac	Accuracy	1	155	43%
Full text screening task					
	Model/method used		N ®	N (a)	Errors %
Khraisha et al. (2024	Human (comparator)		1	150	0%
	Full text (English peer-reviewe	Accuracy	1	50	46%
	Full text (English grey)		1	50	22%
	Full text (Other languages)		1	50	4%
Na et al. (2024)	Human (comparator)		10	265	0%
	Chat GPT	Accuracy	10	265	45%
Data extraction task					
	Model/method used		N (s)	N (d)	Errors %
Gartlehner et al. (20:	Human (comparator)		10	157	0%
	Claude 2		10	157	4%
Khraisha et al. (2024	Human (comparator)	Accuracy	30	Not reported	0%
	Data extraction (English peer-reviewed)		16	16	18%
	Data extraction (English grey)		10	10	19%
	Data extraction (Other languages)		4	4	15%
Platt et al. (2024)	Human (comparator)		41	97	0%
	Vertex AI	Accuracy	41	97	20%
Assessing risk of bias task					
Study	Model/method used		N (s)	N (RoB)	Errors %
Lai et al. (2023)	Human (comparator)		30	300	0%
	Chat GPT (LLM 1)	Accuracy	30	300	15%
	Claude (LLM 2)		30	300	10%

Results

Good results

Tempted to try it out



Search task				N	Errors
Study	Model/method used			N	Errors %
Gwon et al. (2024)	Human (comparator)			1	0%
	ChatGPT			1	96%
	BingAI			1	78%
Sanii et al. (2023)	Human (comparator)			5	0%
	ChatGPT			5	95.50%
	Perplexity.AI			5	81.80%
Wang et al. (2023)	Human (comparator)			112	0%
	ChatGPT Prompt 1 (q1)			112	91%
	ChatGPT Prompt 2 (q2)			112	91%
	ChatGPT Prompt 3 (q3)			112	92%
	ChatGPT Prompt 4 (q4)			112	68%
	ChatGPT Prompt 5 (q5)			112	79%
Title/abstract screening task					
	Model/method used		N ®	N (a)	Errors %
Alchokr et al. (2022)	Human (comparator)		2	327	0%
	Title and Abstract (Word level)		2	327	34%
	Title and Abstract (Sentence level)		2	327	24%
Guo et al. (2024)	Human (comparator)		6	24844	0%
	Chat GPT	Accuracy	6	24844	12%
Tran et al. (2024)	Human (comparator)		5	22665	0%
	Title and Abstract (Balanced)	Accuracy	5	22665	43%
	Title and Abstract (Sensitive)		5	22665	71%
Issaiy et al. (2024)	Expert humans (comparator)		3	1198	0%
	Non-expert humans		3	1198	6%
	ChatGPT (optimal threshold ≥	Accuracy	3	1198	31%
Khraisha et al. (2024	Human (comparator)		1	300	0%
	Chat GPT (English peer-reviev	Accuracy	1	100	33%
	Chat GPT (English grey)		1	100	34%
	Chat GPT (Other languages)		1	100	22%
Schopow et al. (202	Human (comparator)		1	155	0%
	Chat GPT 3.5 legacy (Abstrac	Accuracy	1	155	43%
Full text screening task					
	Model/method used		N ®	N (a)	Errors %
Khraisha et al. (2024	Human (comparator)		1	150	0%
	Full text (English peer-reviewe	Accuracy	1	50	46%
	Full text (English grey)		1	50	22%
	Full text (Other languages)		1	50	4%
Na et al. (2024)	Human (comparator)		10	265	0%
	Chat GPT	Accuracy	10	265	45%
Data extraction task					
	Model/method used		N (s)	N (d)	Errors %
Gartlehner et al. (2023	Human (comparator)		10	157	0%
	Claude 2		10	157	4%
Khraisha et al. (2024	Human (comparator)	Accuracy	30	Not reported	0%
	Data extraction (English peer-reviewed)		16	16	18%
	Data extraction (English grey)		10	10	19%
	Data extraction (Other languages)		4	4	15%
Platt et al. (2024)	Human (comparator)		41	97	0%
	Vertex AI	Accuracy	41	97	20%
Assessing risk of bias task					
Study	Model/method used		N (s)	N (RoB)	Errors %
Lai et al. (2023)	Human (comparator)		30	300	0%
	Chat GPT (LLM 1)	Accuracy	30	300	15%
	Claude (LLM 2)		30	300	10%

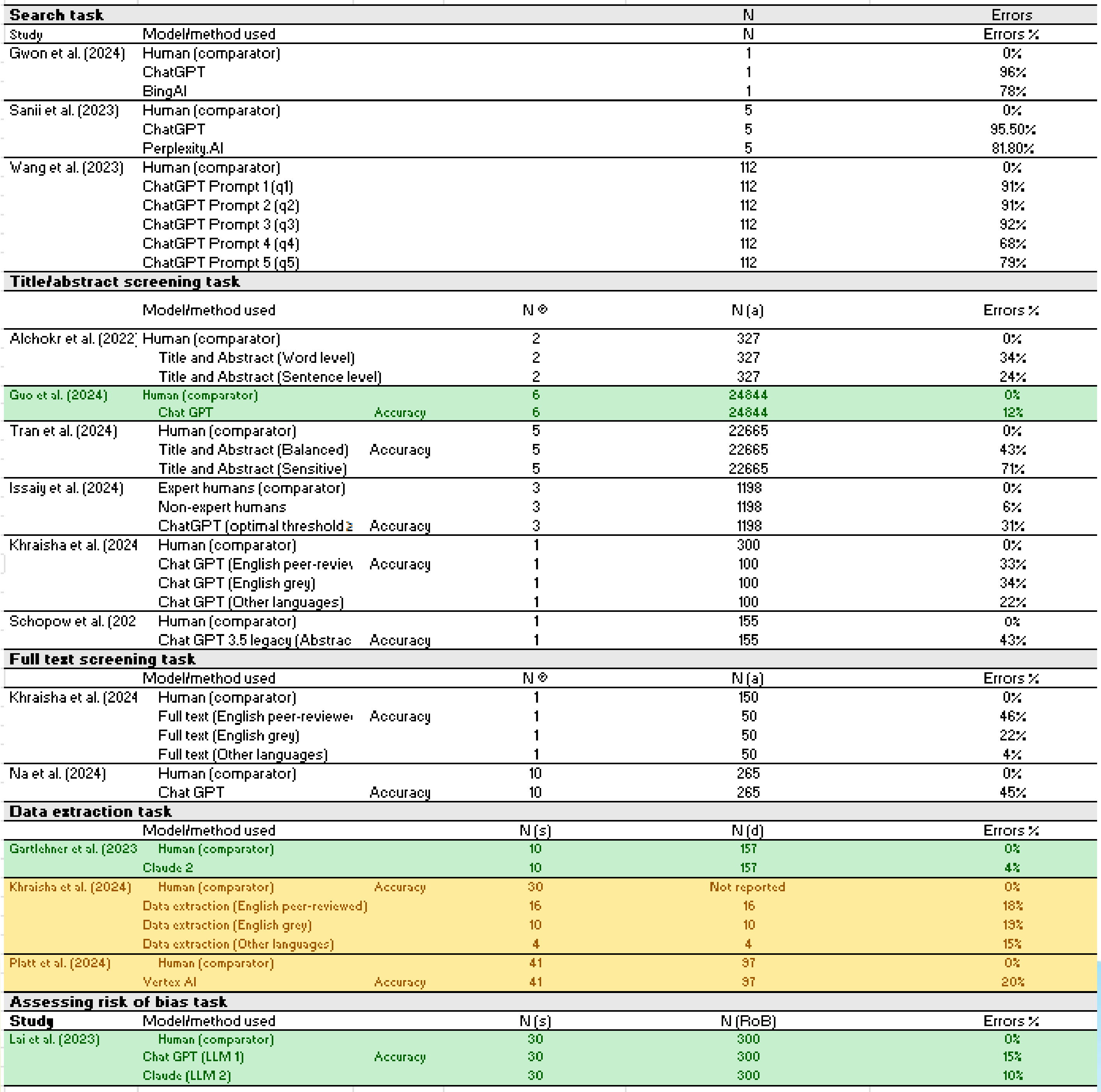
INSTITUTE FOR
Evidence-Based Healthcare

iebh.bond.edu.au

Results

Not so good results

Possibly use it paired/checked with a human expert



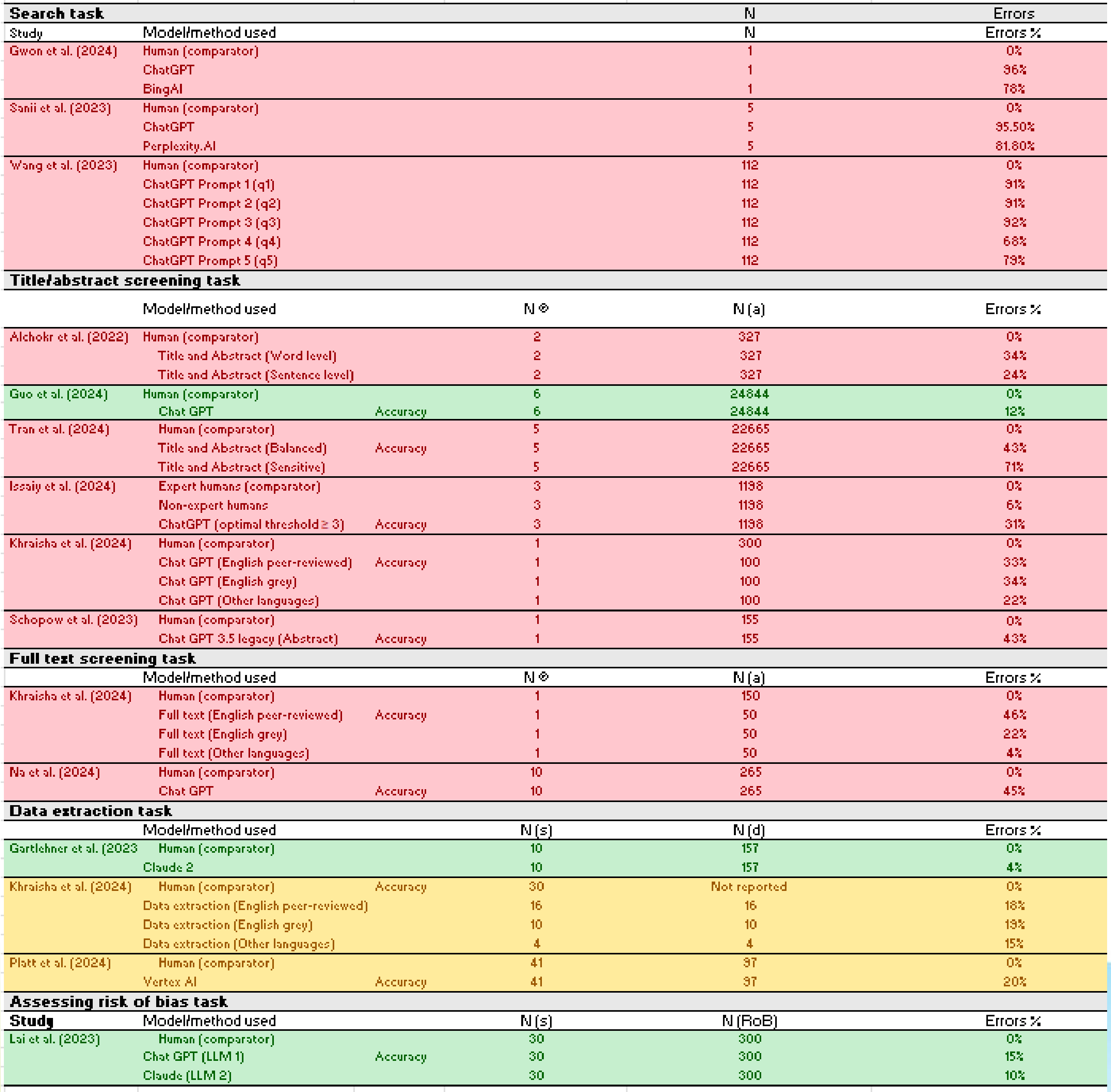
Search task				N	Errors
Study	Model/method used			N	Errors %
Gwon et al. (2024)	Human (comparator)			1	0%
	ChatGPT			1	96%
	BingAI			1	78%
Sanii et al. (2023)	Human (comparator)			5	0%
	ChatGPT			5	95.50%
	Perplexity.AI			5	81.80%
Wang et al. (2023)	Human (comparator)			112	0%
	ChatGPT Prompt 1 (q1)			112	91%
	ChatGPT Prompt 2 (q2)			112	91%
	ChatGPT Prompt 3 (q3)			112	92%
	ChatGPT Prompt 4 (q4)			112	68%
	ChatGPT Prompt 5 (q5)			112	79%
Title/abstract screening task					
	Model/method used		N ®	N (a)	Errors %
Alchokr et al. (2022)	Human (comparator)		2	327	0%
	Title and Abstract (Word level)		2	327	34%
	Title and Abstract (Sentence level)		2	327	24%
Guo et al. (2024)	Human (comparator)		6	24844	0%
	Chat GPT	Accuracy	6	24844	12%
Tran et al. (2024)	Human (comparator)		5	22665	0%
	Title and Abstract (Balanced)	Accuracy	5	22665	43%
	Title and Abstract (Sensitive)		5	22665	71%
Issaiy et al. (2024)	Expert humans (comparator)		3	1198	0%
	Non-expert humans		3	1198	6%
	ChatGPT (optimal threshold ≥	Accuracy	3	1198	31%
Khraisha et al. (2024	Human (comparator)		1	300	0%
	Chat GPT (English peer-reviev	Accuracy	1	100	33%
	Chat GPT (English grey)		1	100	34%
	Chat GPT (Other languages)		1	100	22%
Schopow et al. (202	Human (comparator)		1	155	0%
	Chat GPT 3.5 legacy (Abstrac	Accuracy	1	155	43%
Full text screening task					
	Model/method used		N ®	N (a)	Errors %
Khraisha et al. (2024	Human (comparator)		1	150	0%
	Full text (English peer-reviewe	Accuracy	1	50	46%
	Full text (English grey)		1	50	22%
	Full text (Other languages)		1	50	4%
Na et al. (2024)	Human (comparator)		10	265	0%
	Chat GPT	Accuracy	10	265	45%
Data extraction task					
	Model/method used		N (s)	N (d)	Errors %
Gartlehner et al. (2023	Human (comparator)		10	157	0%
	Claude 2		10	157	4%
Khraisha et al. (2024)	Human (comparator)	Accuracy	30	Not reported	0%
	Data extraction (English peer-reviewed)		16	16	18%
	Data extraction (English grey)		10	10	13%
	Data extraction (Other languages)		4	4	15%
Platt et al. (2024)	Human (comparator)		41	37	0%
	Vertex AI	Accuracy	41	37	20%
Assessing risk of bias task					
Study	Model/method used		N (s)	N (RoB)	Errors %
Lai et al. (2023)	Human (comparator)		30	300	0%
	Chat GPT (LLM 1)	Accuracy	30	300	15%
	Claude (LLM 2)		30	300	10%

Results

INSTITUTE FOR Evidence-Based Healthcare

Bad results

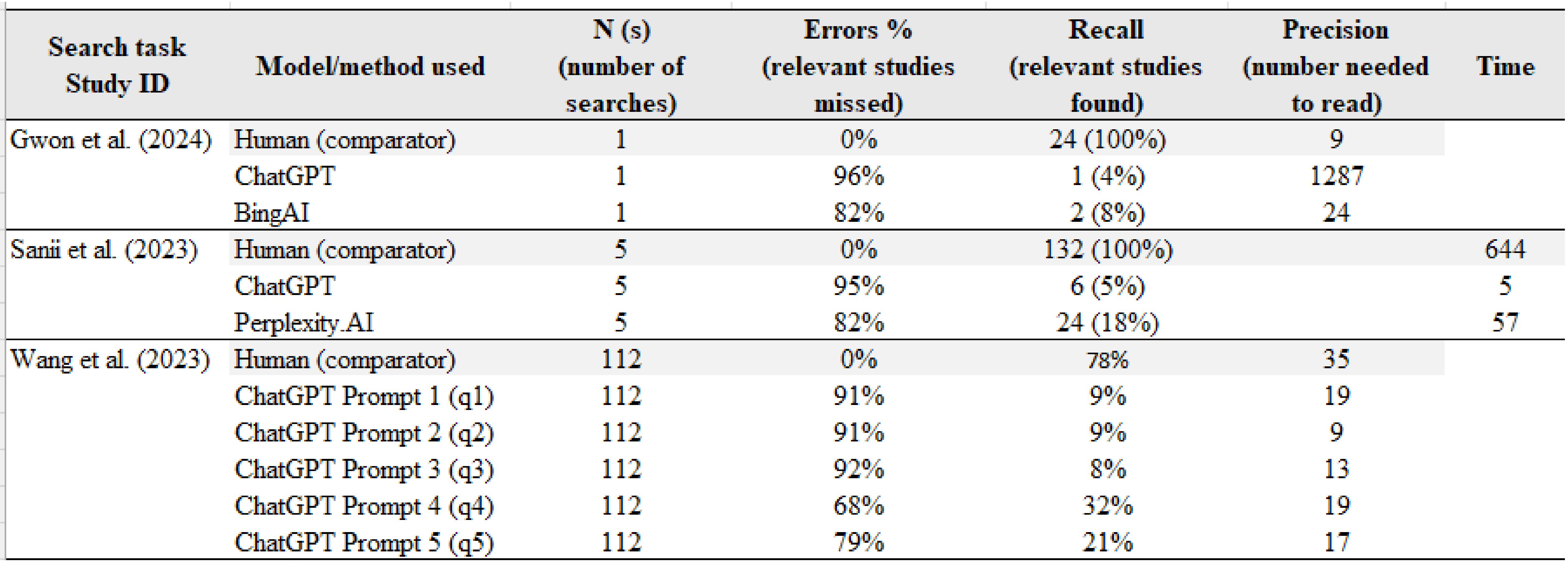
Would not use it



Search task				N	Errors
Study	Model/method used			N	Errors %
Gwon et al. (2024)	Human (comparator)			1	0%
	ChatGPT			1	96%
	BingAI			1	78%
Sanii et al. (2023)	Human (comparator)			5	0%
	ChatGPT			5	95.50%
	Perplexity.AI			5	81.80%
Wang et al. (2023)	Human (comparator)			112	0%
	ChatGPT Prompt 1 (q1)			112	91%
	ChatGPT Prompt 2 (q2)			112	91%
	ChatGPT Prompt 3 (q3)			112	92%
	ChatGPT Prompt 4 (q4)			112	68%
	ChatGPT Prompt 5 (q5)			112	79%
Title/abstract screening task					
	Model/method used		N ®	N (a)	Errors %
Alchokr et al. (2022)	Human (comparator)		2	327	0%
	Title and Abstract (Word level)		2	327	34%
	Title and Abstract (Sentence level)		2	327	24%
Guo et al. (2024)	Human (comparator)		6	24844	0%
	Chat GPT	Accuracy	6	24844	12%
Tran et al. (2024)	Human (comparator)		5	22665	0%
	Title and Abstract (Balanced)	Accuracy	5	22665	43%
	Title and Abstract (Sensitive)		5	22665	71%
Issaiy et al. (2024)	Expert humans (comparator)		3	1198	0%
	Non-expert humans		3	1198	6%
	ChatGPT (optimal threshold ≥ 3)	Accuracy	3	1198	31%
Khraisha et al. (2024)	Human (comparator)		1	300	0%
	Chat GPT (English peer-reviewed)	Accuracy	1	100	33%
	Chat GPT (English grey)		1	100	34%
	Chat GPT (Other languages)		1	100	22%
Schopow et al. (2023)	Human (comparator)		1	155	0%
	Chat GPT 3.5 legacy (Abstract)	Accuracy	1	155	43%
Full text screening task					
	Model/method used		N ®	N (a)	Errors %
Khraisha et al. (2024)	Human (comparator)		1	150	0%
	Full text (English peer-reviewed)	Accuracy	1	50	46%
	Full text (English grey)		1	50	22%
	Full text (Other languages)		1	50	4%
Na et al. (2024)	Human (comparator)		10	265	0%
	Chat GPT	Accuracy	10	265	45%
Data extraction task					
	Model/method used		N (s)	N (d)	Errors %
Gartlehner et al. (2023	Human (comparator)		10	157	0%
	Claude 2		10	157	4%
Khraisha et al. (2024)	Human (comparator)	Accuracy	30	Not reported	0%
	Data extraction (English peer-reviewed)		16	16	18%
	Data extraction (English grey)		10	10	19%
	Data extraction (Other languages)		4	4	15%
Platt et al. (2024)	Human (comparator)		41	97	0%
	Vertex AI	Accuracy	41	97	20%
Assessing risk of bias task					
Study	Model/method used		N (s)	N (RoB)	Errors %
Lai et al. (2023)	Human (comparator)		30	300	0%
	Chat GPT (LLM 1)	Accuracy	30	300	15%
	Claude (LLM 2)		30	300	10%

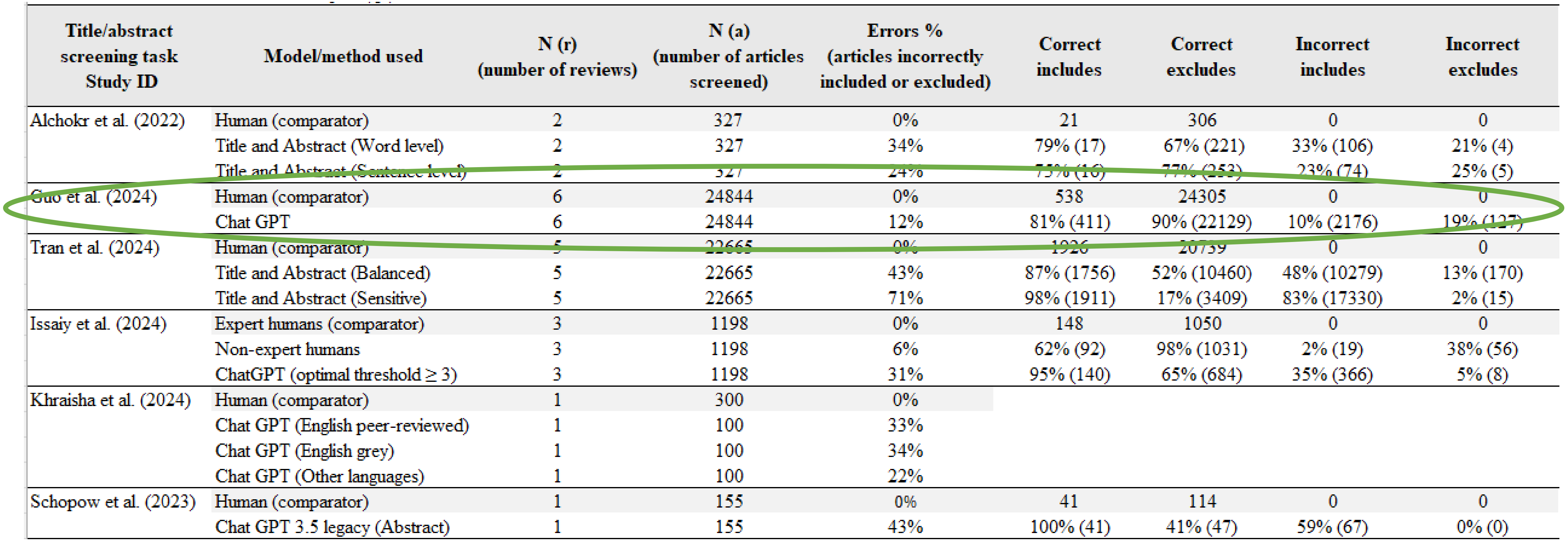
iebh.bond.edu.au

Results - searching



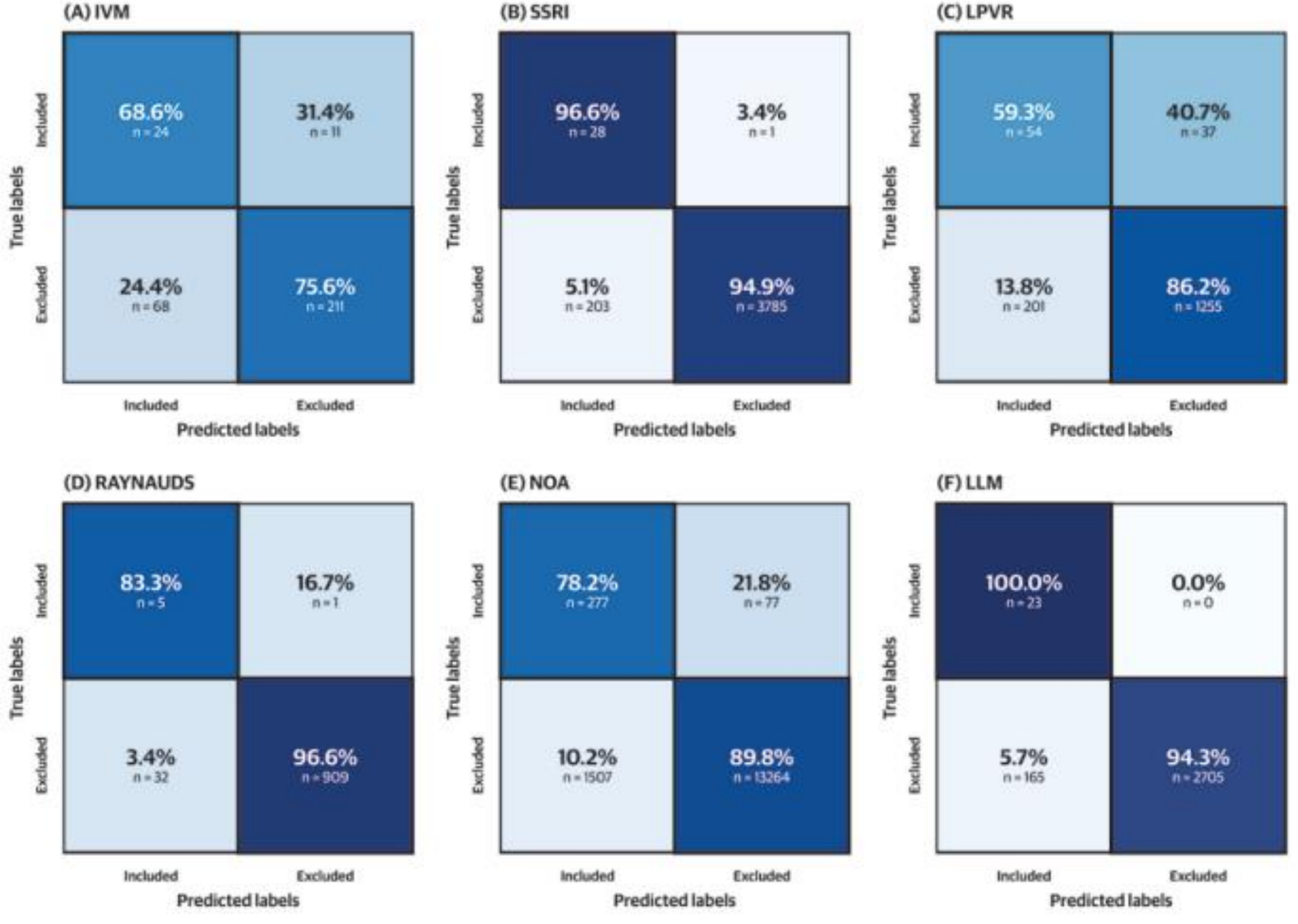
Search task Study ID	Model/method used	N (s) (number of searches)	Errors % (relevant studies missed)	Recall (relevant studies found)	Precision (number needed to read)	Time
Gwon et al. (2024)	Human (comparator)	1	0%	24 (100%)	9	
	ChatGPT	1	96%	1 (4%)	1287	
	BingAI	1	82%	2 (8%)	24	
Sanii et al. (2023)	Human (comparator)	5	0%	132 (100%)		644
	ChatGPT	5	95%	6 (5%)		5
	Perplexity.AI	5	82%	24 (18%)		57
Wang et al. (2023)	Human (comparator)	112	0%	78%	35	
	ChatGPT Prompt 1 (q1)	112	91%	9%	19	
	ChatGPT Prompt 2 (q2)	112	91%	9%	9	
	ChatGPT Prompt 3 (q3)	112	92%	8%	13	
	ChatGPT Prompt 4 (q4)	112	68%	32%	19	
	ChatGPT Prompt 5 (q5)	112	79%	21%	17	

Results – title/abstract screening



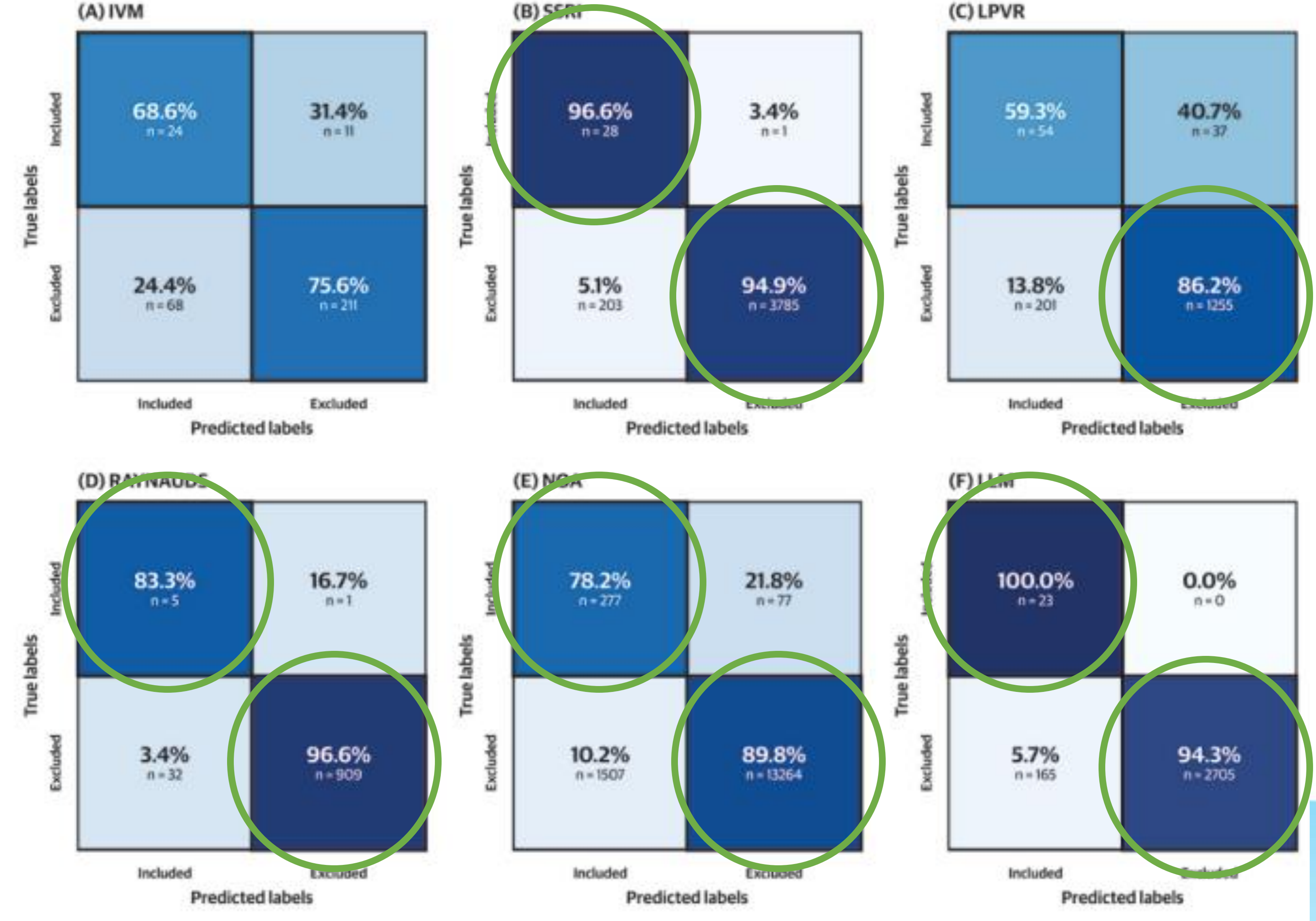
Title/abstract screening task Study ID	Model/method used	N (r) (number of reviews)	N (a) (number of articles screened)	Errors % (articles incorrectly included or excluded)	Correct includes	Correct excludes	Incorrect includes	Incorrect excludes
Alchokr et al. (2022)	Human (comparator)	2	327	0%	21	306	0	0
	Title and Abstract (Word level)	2	327	34%	79% (17)	67% (221)	33% (106)	21% (4)
	Title and Abstract (Sentence level)	2	327	24%	75% (16)	77% (253)	23% (74)	25% (5)
Guo et al. (2024)	Human (comparator)	6	24844	0%	538	24305	0	0
	Chat GPT	6	24844	12%	81% (411)	90% (22129)	10% (2176)	19% (127)
Tran et al. (2024)	Human (comparator)	5	22665	0%	1926	20739	0	0
	Title and Abstract (Balanced)	5	22665	43%	87% (1756)	52% (10460)	48% (10279)	13% (170)
	Title and Abstract (Sensitive)	5	22665	71%	98% (1911)	17% (3409)	83% (17330)	2% (15)
Issaiy et al. (2024)	Expert humans (comparator)	3	1198	0%	148	1050	0	0
	Non-expert humans	3	1198	6%	62% (92)	98% (1031)	2% (19)	38% (56)
	ChatGPT (optimal threshold ≥ 3)	3	1198	31%	95% (140)	65% (684)	35% (366)	5% (8)
Khraisha et al. (2024)	Human (comparator)	1	300	0%				
	Chat GPT (English peer-reviewed)	1	100	33%				
	Chat GPT (English grey)	1	100	34%				
	Chat GPT (Other languages)	1	100	22%				
Schopow et al. (2023)	Human (comparator)	1	155	0%	41	114	0	0
	Chat GPT 3.5 legacy (Abstract)	1	155	43%	100% (41)	41% (47)	59% (67)	0% (0)

Results – Guo et al. (2024)*



*Guo E, et al., Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024;26:e48996.

Results – Guo et al. (2024)*

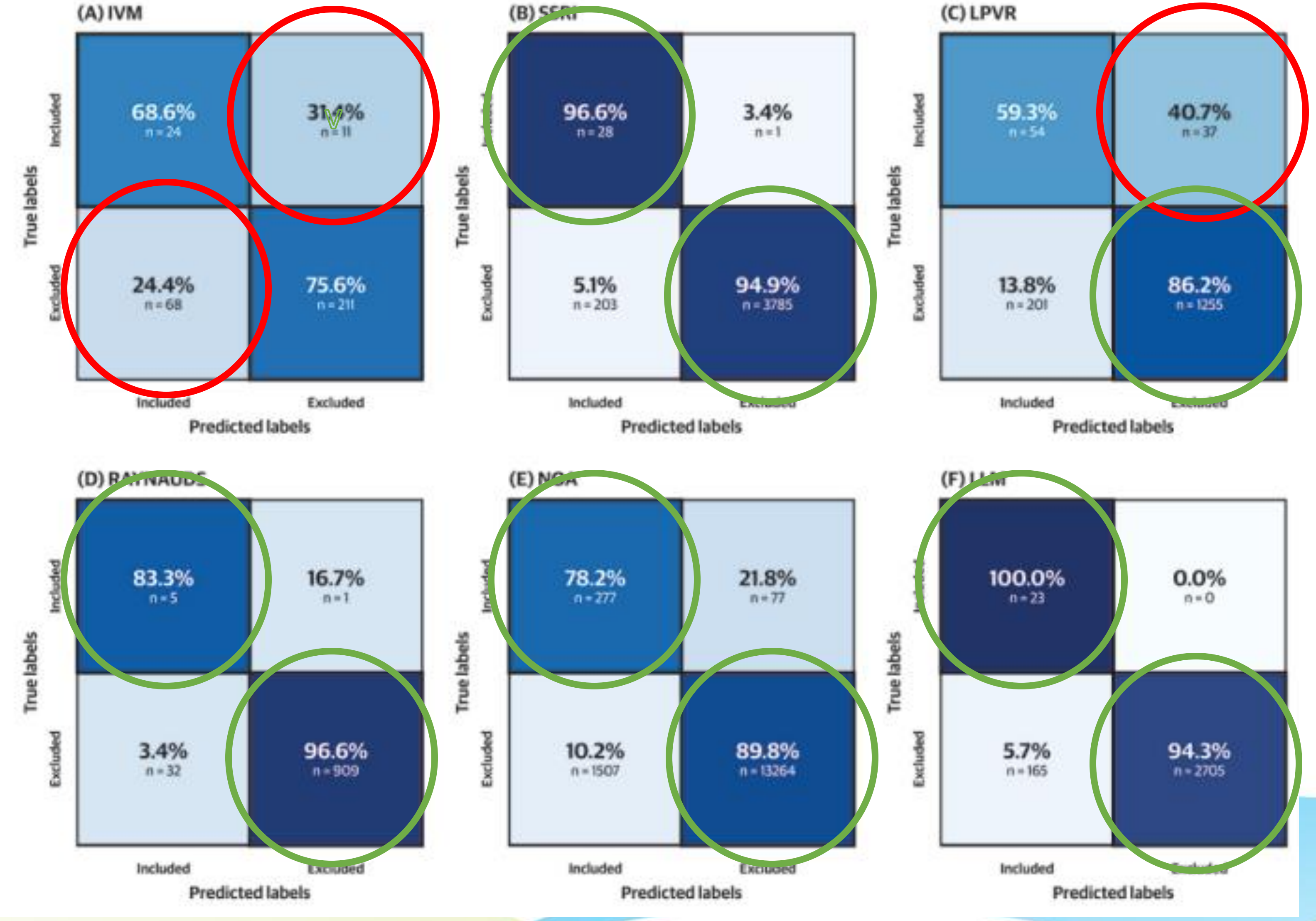


*Guo E, et al., Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024;26:e48996.

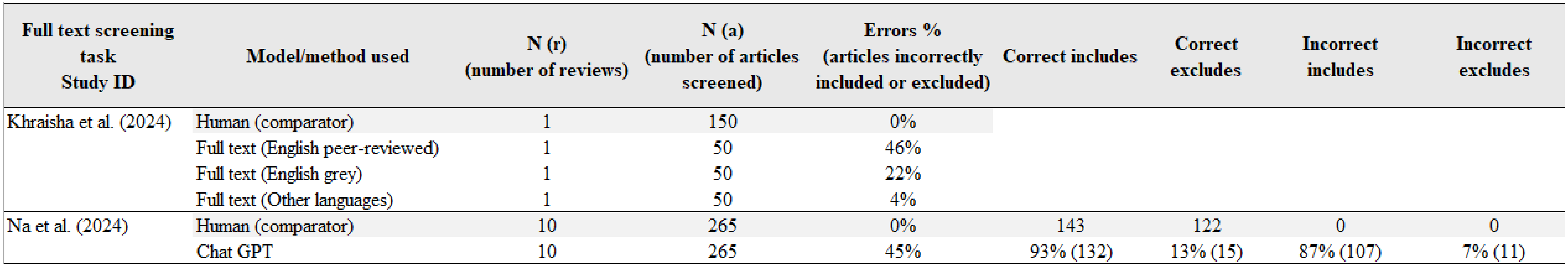
Results – Guo et al. (2024)*

BOND UNIVERSITY | INSTITUTE FOR Evidence-Based Healthcare

*Guo E, et al., Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024;26:e48996.

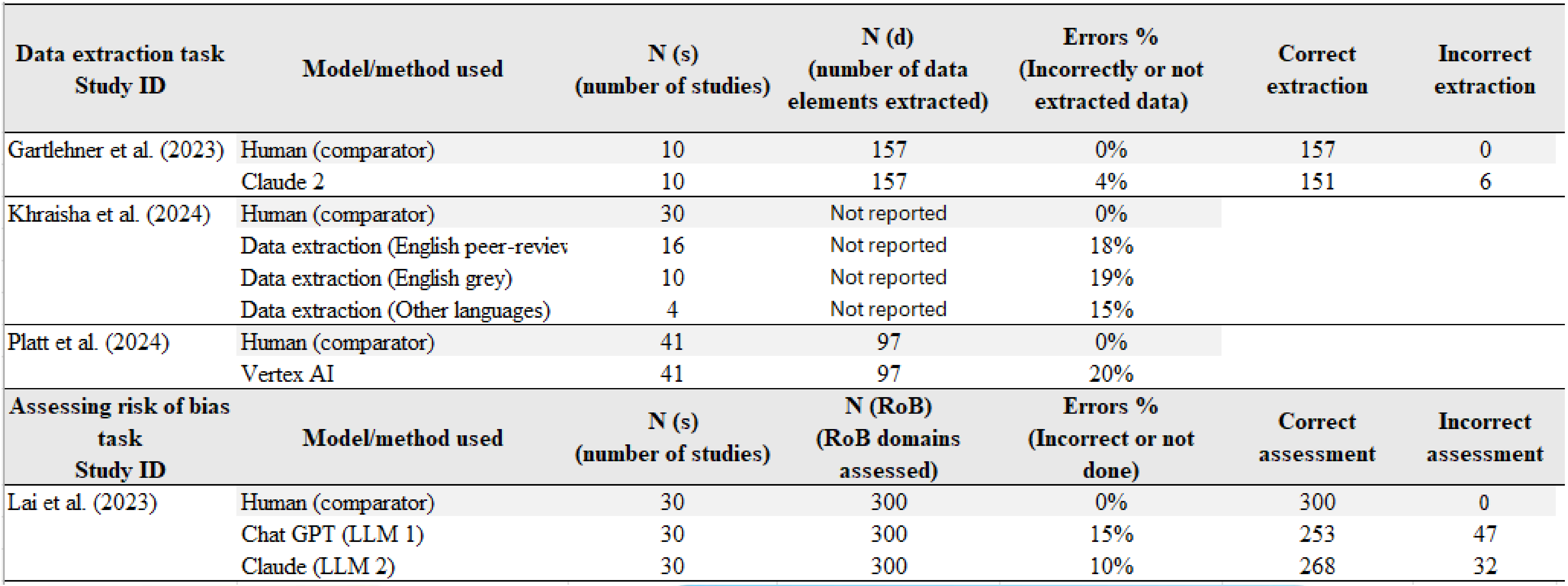


Results – full text screening



Full text screening task Study ID	Model/method used	N (r) (number of reviews)	N (a) (number of articles screened)	Errors % (articles incorrectly included or excluded)	Correct includes	Correct excludes	Incorrect includes	Incorrect excludes
Khraisha et al. (2024)	Human (comparator)	1	150	0%				
	Full text (English peer-reviewed)	1	50	46%				
	Full text (English grey)	1	50	22%				
	Full text (Other languages)	1	50	4%				
Na et al. (2024)	Human (comparator)	10	265	0%	143	122	0	0
	Chat GPT	10	265	45%	93% (132)	13% (15)	87% (107)	7% (11)

Results – Extraction & RoB



Data extraction task Study ID	Model/method used	N (s) (number of studies)	N (d) (number of data elements extracted)	Errors % (Incorrectly or not extracted data)	Correct extraction	Incorrect extraction
Gartlehner et al. (2023)	Human (comparator)	10	157	0%	157	0
	Claude 2	10	157	4%	151	6
Khraisha et al. (2024)	Human (comparator)	30	Not reported	0%		
	Data extraction (English peer-reviev	16	Not reported	18%		
	Data extraction (English grey)	10	Not reported	19%		
	Data extraction (Other languages)	4	Not reported	15%		
Platt et al. (2024)	Human (comparator)	41	97	0%		
	Vertex AI	41	97	20%		

Assessing risk of bias task Study ID	Model/method used	N (s) (number of studies)	N (RoB) (RoB domains assessed)	Errors % (Incorrect or not done)	Correct assessment	Incorrect assessment
Lai et al. (2023)	Human (comparator)	30	300	0%	300	0
	Chat GPT (LLM 1)	30	300	15%	253	47
	Claude (LLM 2)	30	300	10%	268	32

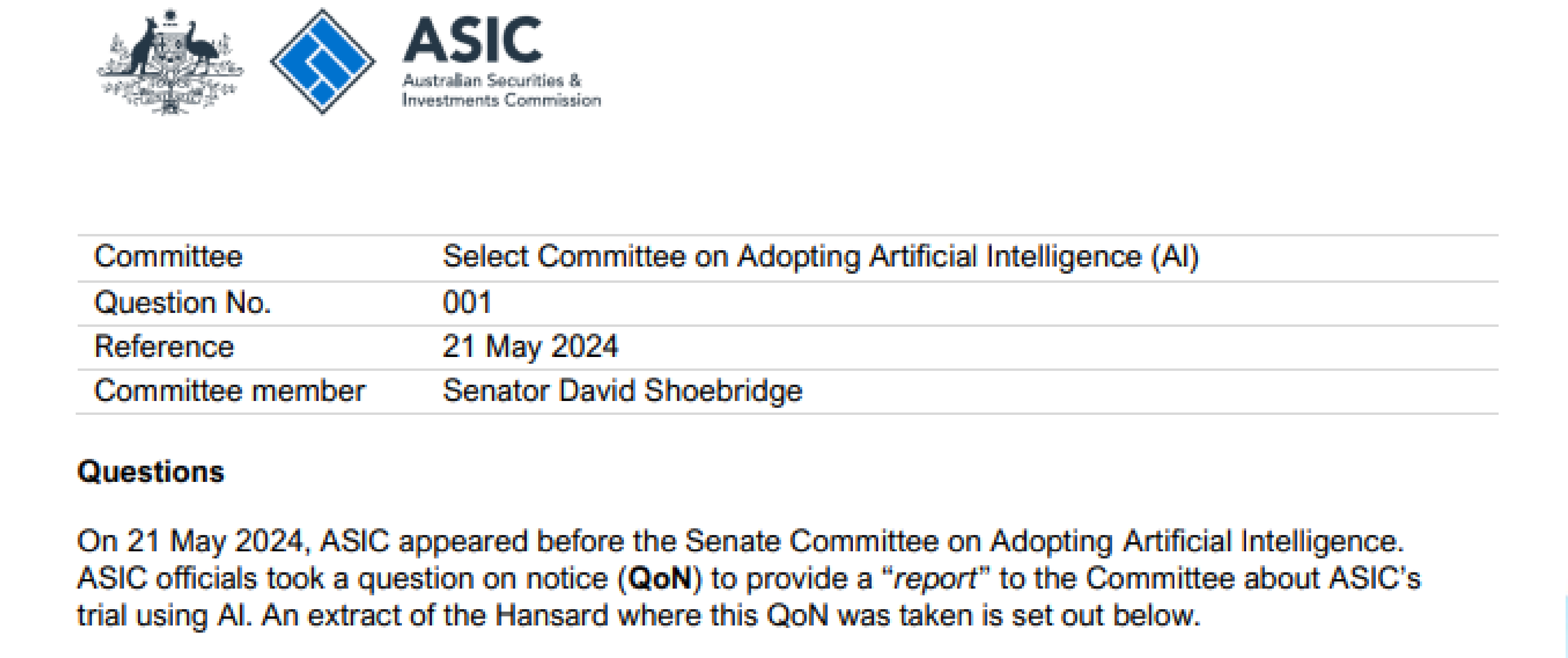
Additional points

Time outcome: reported processing time only, left out set up time, e.g., designing prompts, preparing abstracts/full texts for processing etc. One study only reported it ~2 days needed for prompt design etc.

Reporting seemed overly favourable to Gen AI, a lot of “shows promise/potential”, and emphasising positive results isolated from negative results, e.g., we correctly included 90% of studies without saying they incorrectly included 70% of results

Additional example

Provided by: Tim Repke



ASIC
Australian Securities &
Investments Commission

Committee	Select Committee on Adopting Artificial Intelligence (AI)
Question No.	001
Reference	21 May 2024
Committee member	Senator David Shoebridge

Questions

On 21 May 2024, ASIC appeared before the Senate Committee on Adopting Artificial Intelligence. ASIC officials took a question on notice (**QoN**) to provide a "*report*" to the Committee about ASIC's trial using AI. An extract of the Hansard where this QoN was taken is set out below.

Additional example

https://www.crikey.com.au/2024/09/03/ai-worse-summarising-information-humans-government-trial/

AI worse than humans in every way at summarising information, government trial finds

- A test of AI for Australia's corporate regulator found that the technology might actually make more work for people, not less.

Additional example

AI model Llama2-70B summarized submissions into audit and consultancy firms

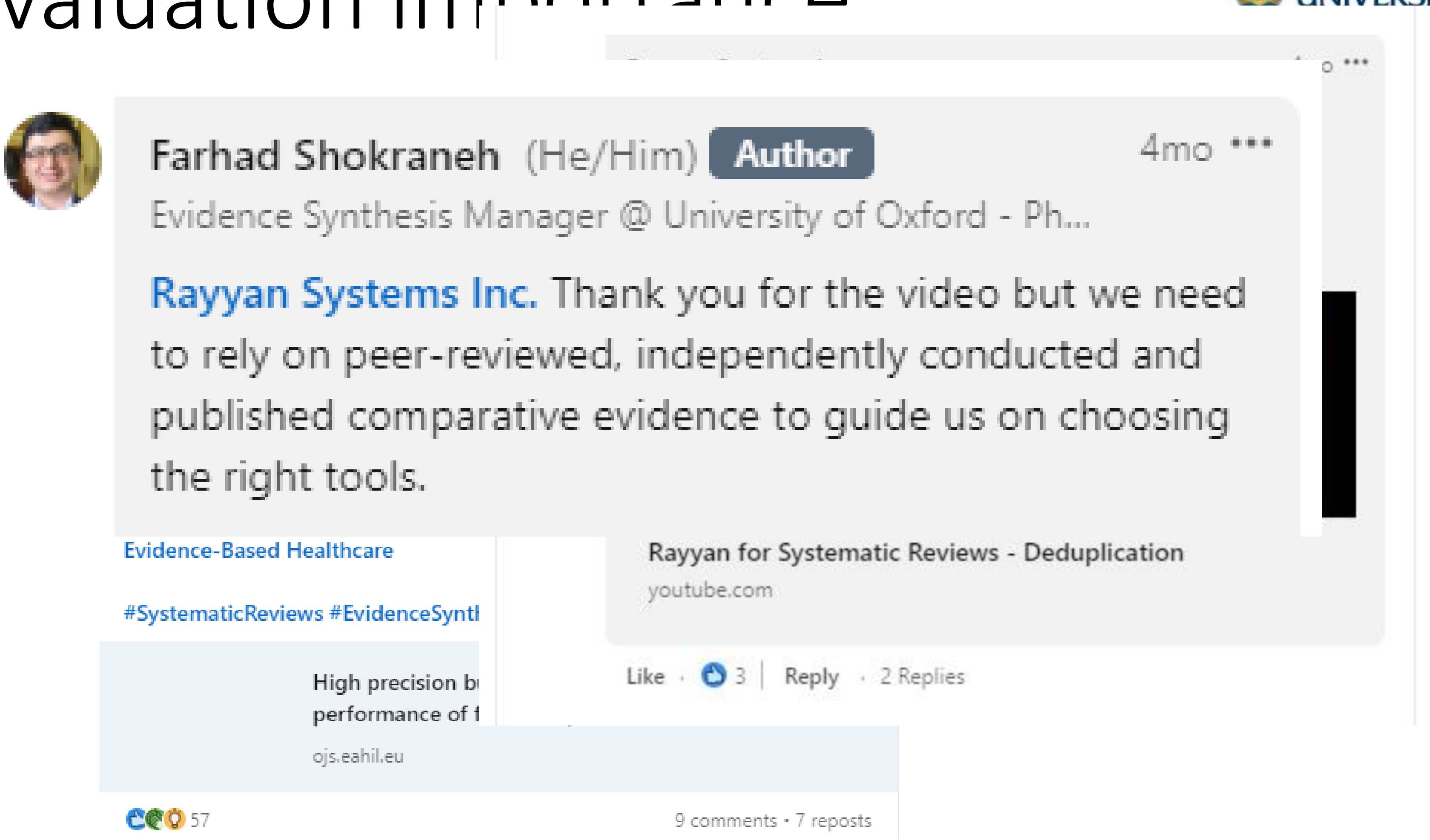
Ten human staff given the same task

Reviewers blindly assessed the summaries, unaware that this exercise involved AI

Reviewers overwhelmingly found human summaries beat AI summaries, humans = 81%, AI = 47%

Reviewers' feedback was AI summaries may be counterproductive and create further work because of the need to fact-check and refer to original submissions which communicated the message better and more concisely

Evaluation importance



Farhad Shokraneh (He/Him) Author 4mo

Evidence Synthesis Manager @ University of Oxford - Ph...

Rayyan Systems Inc. Thank you for the video but we need to rely on peer-reviewed, independently conducted and published comparative evidence to guide us on choosing the right tools.

Rayyan for Systematic Reviews - Deduplication

youtube.com

Like · 3 | Reply · 2 Replies

Evidence-Based Healthcare

#SystematicReviews #EvidenceSynt

High precision b
performance of

ojs.eahil.eu

57

9 comments • 7 reposts

Questions?