

ORIGINAL RESEARCH**Benchmarking Human—AI collaboration for common evidence appraisal tools**Tim Woelfle^{a,b,c,*}, Julian Hirt^{a,d,e}, Perrine Janiaud^{a,d}, Ludwig Kappos^a, John P.A. Ioannidis^{f,g},
Lars G. Hemkens^{a,d,f,h}^a*Pragmatic Evidence Lab, Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), Basel, Switzerland*^b*Department of Neurology, University Hospital Basel, Basel, Switzerland*^c*Translational Imaging in Neurology (ThInk), Department of Biomedical Engineering, University Hospital and University of Basel, Basel, Switzerland*^d*Department of Clinical Research, University Hospital Basel and University of Basel, Basel, Switzerland*^e*Institute of Nursing Science, Department of Health, Eastern Switzerland University of Applied Sciences, St. Gallen, Basel, Switzerland*^f*Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA*^g*Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, Stanford University, Stanford, CA, USA*^h*Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany*

Accepted 9 September 2024; Published online 12 September 2024

Abstract

Background and Objective: It is unknown whether large language models (LLMs) may facilitate time- and resource-intensive text-related processes in evidence appraisal. The objective was to quantify the agreement of LLMs with human consensus in appraisal of scientific reporting (Preferred Reporting Items for Systematic reviews and Meta-Analyses [PRISMA]) and methodological rigor (A MeaSurement Tool to Assess systematic Reviews [AMSTAR]) of systematic reviews and design of clinical trials (PRagmatic Explanatory Continuum Indicator Summary 2 [PRECIS-2]) and to identify areas where collaboration between humans and artificial intelligence (AI) would outperform the traditional consensus process of human raters in efficiency.

Study Design and Setting: Five LLMs (Claude-3-Opus, Claude-2, GPT-4, GPT-3.5, Mixtral-8x22B) assessed 112 systematic reviews applying the PRISMA and AMSTAR criteria and 56 randomized controlled trials applying PRECIS-2. We quantified the agreement between human consensus and (1) individual human raters; (2) individual LLMs; (3) combined LLMs approach; (4) human—AI collaboration. Ratings were marked as deferred (undecided) in case of inconsistency between combined LLMs or between the human rater and the LLM.

Results: Individual human rater accuracy was 89% for PRISMA and AMSTAR, and 75% for PRECIS-2. Individual LLM accuracy was ranging from 63% (GPT-3.5) to 70% (Claude-3-Opus) for PRISMA, 53% (GPT-3.5) to 74% (Claude-3-Opus) for AMSTAR, and 38% (GPT-4) to 55% (GPT-3.5) for PRECIS-2. Combined LLM ratings led to accuracies of 75%–88% for PRISMA (4%–74% deferred), 74%–89% for AMSTAR (6%–84% deferred), and 64%–79% for PRECIS-2 (29%–88% deferred). Human—AI collaboration resulted in the best accuracies from 89% to 96% for PRISMA (25/35% deferred), 91%–95% for AMSTAR (27/30% deferred), and 80%–86% for PRECIS-2 (76/71% deferred).

Conclusion: Current LLMs alone appraised evidence worse than humans. Human—AI collaboration may reduce workload for the second human rater for the assessment of reporting (PRISMA) and methodological rigor (AMSTAR) but not for complex tasks such as PRECIS-2. © 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Evidence appraisal; Reporting guidelines; Systematic reviews; Clinical trials; Artificial intelligence; Large language models

Funding: The research activities of RC2NB (Research Center for Clinical Neuroimmunology and Neuroscience Basel) are supported by the University Hospital and the University of Basel and the Foundation Clinical Neuroimmunology and Neuroscience Basel, including grants from Novartis and Roche.

* Corresponding author. RC2NB, Spitalstrasse 2, Basel CH-4031, Switzerland.

E-mail address: tim.woelfle@usb.ch (T. Woelfle).

1. Introduction

The assessment of reporting, methodological rigor, and design features of biomedical research is essential for evidence-based medicine. However, these evaluations require extensive resources.

What is new?

Key findings

- Evidence appraisal tools (e.g., PRISMA, AMSTAR, PRECIS-2) are often very resource intensive. LLMs may assist human raters with these tasks.

What this adds to what was known?

- Current state-of-the-art LLMs (Claude-3-Opus, Claude-2, GPT-4, GPT-3.5, Mixtral-8x22B) alone performed substantially worse than humans but their combination with a consistency approach improved performance.

What is the implication and what should change now?

- Human–AI collaboration may save time for a second human rater for PRISMA and AMSTAR.

Many evidence appraisal tools are primarily text-based and follow instructions of varying complexity, for example, reporting checklists or clinical trial tools such as “Preferred Reporting Items for Systematic reviews and Meta-Analyses” (PRISMA) [1], “A Measurement Tool to Assess systematic Reviews” (AMSTAR) [2], and “PRagmatic Explanatory Continuum Indicator Summary 2” (PRECIS-2) [3].

Traditional machine learning and natural language processing methods have been explored for years to extract for example “population, intervention, control, outcome” information or key elements used for “risk of bias” assessments from study reports [4,5]. More advanced deep learning applications have been applied for automatically assessing reporting on a large scale [6].

Large language models (LLMs) such as OpenAI’s ChatGPT have gained great attention due to their advanced language-processing capabilities and presumed reasoning, beating several artificial intelligence (AI) benchmarks [7–9]. While their potential to facilitate systematic reviews is generally acknowledged, there is an increasing discussion of their limitations and of the need for caution in using these tools [10]. The efficacy of LLMs for screening and data extraction for systematic reviews has been highly mixed, but there appear to be scenarios where they can help with these tasks [11,12].

Our objectives were to quantify the agreement of 5 individual LLMs with human consensus in the assessment of evidence appraisal tools of different levels of complexity: reporting (PRISMA) and methodological rigor (AMSTAR) of systematic reviews, and degree of pragmatism of clinical trials (PRECIS-2). We aimed to assess how much complexity of assessment can be handled by current LLMs and whether their combination increases accuracy. Finally,

we aimed to combine individual human raters with LLMs and to evaluate whether such human–AI collaboration might outperform a traditional consensus process of multiple human raters in efficiency.

2. Methods

2.1. Selecting datasets and evidence appraisal tools

We selected datasets for which independent ratings from 2 human raters and their consensus were available. The independent assessment by at least 2 human raters is the standard in systematic reviews (eg, for Cochrane reviews [13]), and is also commonly used to assess reporting and methodological rigor (eg, using PRISMA [14] and AMSTAR [14,15]), or study designs (eg, using PRECIS-2 [16]).

For reporting and methodological rigor, we used human assessments of PRISMA and AMSTAR for 112 systematic reviews and meta-analyses in the field of pediatric surgery (data kindly shared by Cullis and colleagues) [14]. PRISMA contains 27 items and AMSTAR 11 items (Supplement), rated as no/yes/not applicable (NA) [1,2]. The 2 raters were content experts (British pediatric surgeons).

For pragmatism in clinical trial design, we used human ratings of PRECIS-2 for 56 randomized controlled trials within the PragMeta database [16,17]. PRECIS-2 contains nine domains (Supplement), rated ordinally from 1 (very explanatory) to 5 (very pragmatic) or NA [3]. Human rater 1 was an experienced systematic-reviewer and metaresearcher. Human rater 2 was either 1 of 2 experienced postgraduate MSc students in epidemiology with special training in PRECIS-2 assessment or a senior clinical epidemiologist and expert in pragmatic trial design. In case human raters 1 and 2 rated differently, they discussed the item and agreed on a consensus on equal terms.

2.2. LLMs

We used 4 proprietary LLMs (Anthropic’s Claude-3-Opus and Claude-2 [18,19] and OpenAI’s GPT-4 and GPT-3.5 [20,21]) and 1 open-source LLM (Mistral AI’s Mixtral-8x22B [22]), Table 1. As of June 2024, Claude-3-Opus and GPT-4 are among the best-performing LLMs according to the popular Chatbot Arena [23]. To estimate the performance gap compared to the last generation of commercial “flagship” LLMs, we also included the older, smaller, and cheaper models Claude-2 and GPT-3.5. Mistral AI’s Mixtral-8x22B is fully open source, meaning the model can be downloaded and run locally, provided the right hardware is available.

Claude-3-Opus was the only multimodal model we employed. To fully take advantage of the multimodality, we converted PDF files to plain images (one PNG file per page) for this model. This way, it processed the normal page layout, including all figures and tables alongside the full

Table 1. LLMs used: name, context length, timeframe of queries and API used, median cost and response time per publication.

Company: LLM name (exact version identifier)	Context length, n tokens (words)	Time of queries and API used	Median cost (IQR) and response time per publication
Anthropic: Claude-3-Opus (claude-3-opus-20240229)	200k (~150k)	April 2024 via Anthropic	\$ 0.40 (0.14-0.78) ~1-2 mins
Anthropic: Claude-2 (claude-2.0)	100k (~75k)	September 2023 via OpenRouter	\$ 0.19 (0.09-0.81) ~1 minute
OpenAI: GPT-4 (gpt-4-32k-0613)	32k (~24k)	September 2023 via OpenRouter	\$ 1.15 (0.65-1.98) ~2 mins
OpenAI: GPT-3.5 (gpt-3.5-turbo-16k-0613)	16k (~12k)	August 7, 2023 via OpenAI	\$ 0.06 (0.03-0.10) ~10 s
Mistral AI: Mixtral-8x22B (Mixtral-8x22b-instruct-v0.1)	64k (~48k)	April 2024 via OpenRouter	\$ 0.012 (0.010-0.017) ~ 1-2 mins

API, application programming interface; LLM, large language model.

text, which it extracted itself by implicitly performing optical character recognition. The 4 other models could only process full text (extracted by us) and no images. While later versions of GPT-4 were also multimodal, the version we used (gpt-4-32k-0613) had no vision capabilities yet.

Details on LLM queries, prompt engineering, and the extraction of ratings and quotes can be found the [Supplementary Methods](#).

2.3. Analyses and outcomes

2.3.1. Agreement with human consensus

Our main outcome was agreement with human consensus measured by accuracy (agreement fraction, i.e., the proportion of identical ratings between rater and human consensus) and Cohen's kappa. For the ordinal PRECIS-2 ratings, responses 1 and 2 ("very" and "mostly explanatory") were pooled to "1/2" and responses 4 and 5 ("very" and "mostly pragmatic") to "4/5" and a weighted version of Cohen's kappa was used ([Supplementary Table 1](#)).

Bootstrapping with 1000 resamples on the publication-level was performed to derive 95% CIs.

We performed 4 analyses for each of the 3 evidence appraisal tools (PRISMA, AMSTAR, PRECIS-2) ([Fig 1](#)) while also quantifying resources used for LLMs (costs and time effort).

- (1) Human consensus vs individual human raters
- (2) Human consensus vs individual LLMs (Claude-3-Opus, Claude-2, GPT-4, GPT-3.5, Mixtral-8x22B)
- (3) Human consensus vs combined LLMs: consistency approach

Combining multiple assessments from LLMs in a "consistency" approach improves performance in biomedical and general contexts beyond using only a single assessment [24,25]. We combined a total of nine LLM assessments: 2 × Claude-3-Opus, 2 × Claude-2, 1 × GPT-4 (due to high costs), 2 × GPT-3.5, 2 × Mixtral-8x22B [26].

Inconsistent responses without alignment among 5 to nine out of the nine LLM assessments (the

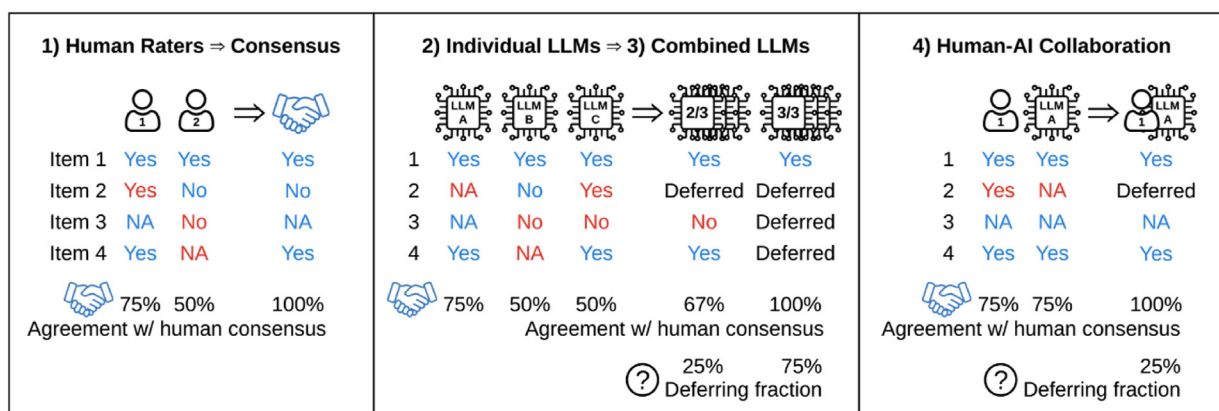


Figure 1. Schematic overview of the main outcome accuracy (agreement with human consensus, in blue) for a hypothetical tool with 4 items in a single publication. (1) Two human raters independently rate each item and agree on a consensus. (2) Three individual LLMs independently rate each item and are compared to human consensus. (3) They are combined using a consistency approach, using either all responses consistent in at least 2 out of 3 LLMs or all 3/3 LLMs. The proportion of inconsistent ratings is the deferring fraction and agreement is only calculated on the consistent ones. (4) A human rater is combined with an LLM: only items consistent between human and LLM are compared to human consensus and the remaining ones are deferred. LLM, large language model.

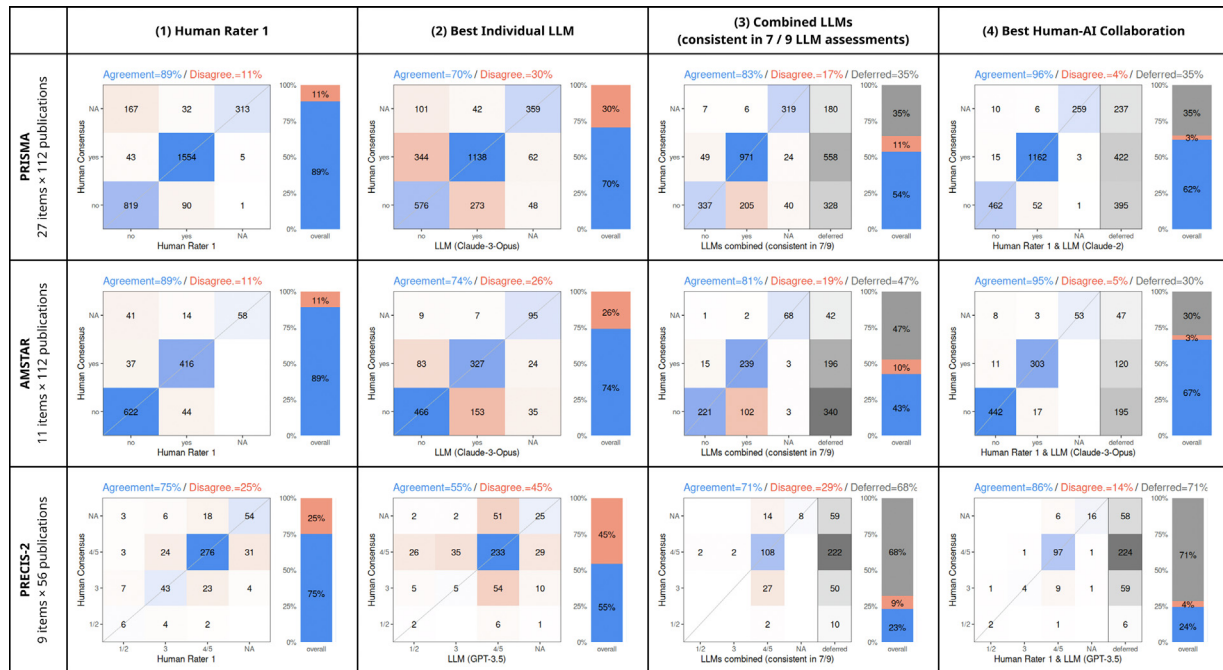


Figure 2. Confusion matrices for “agreement” (blue) and “disagreement” (red) with human consensus for the three tools (in rows). For 3) and 4), “deferred” (gray) refers to the proportion of responses inconsistent between LLMs or between human and LLM, which are thus considered unclear and deferred to a (second) human rater. Accuracy is only calculated for consistent responses, which thus represent a hypothetical workload reduction. AMSTAR, A MeaSurement Tool to Assess systematic Reviews; PRECIS-2, PRagmatic-Explanatory Continuum Indicator Summary-2; LLM, large language model; PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses.

“consistency-threshold”) were considered uncertain and thus “deferred to human raters”. In a practical application of this approach, human raters would then have to decide through their normal consensus process.

We assumed that agreement with human consensus of the nondeferred responses would

increase with increasing consistency thresholds at the cost of higher deferring fractions, that is, higher proportions of deferred responses among all responses (as seen by others, e.g., Figure 3 in [24]).

(4) Human consensus vs human–AI collaboration

We combined ratings of individual human raters

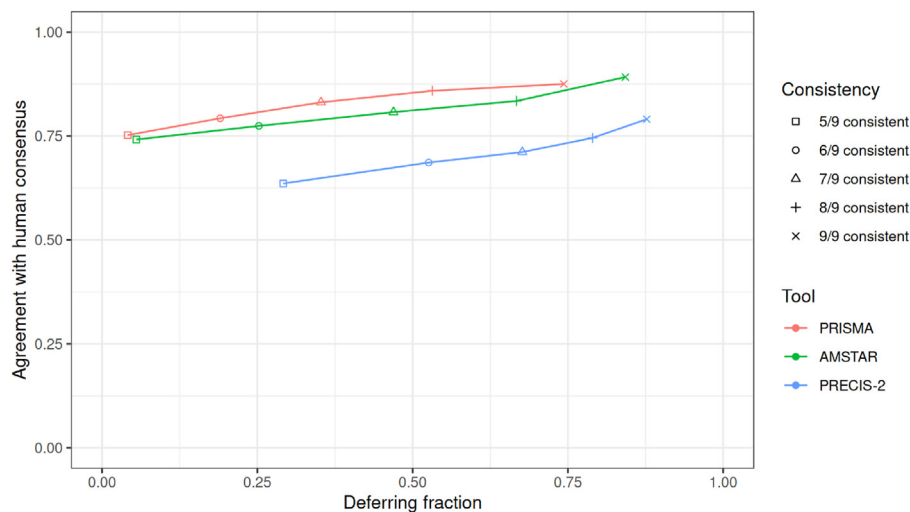


Figure 3. Combined LLMs through consistency approach. Combining all 12 LLM ratings per tool and using only the responses that are consistent among 5–9 out of nine ratings led to increasing performances for all 3 tools measured by accuracy. This comes at the cost of an increasing “deferring fraction”, comprising the inconsistent responses. Combining LLMs thus allows an estimation of uncertainty.

with individual LLM ratings for each of the three tools. Items where the LLM aligned with the human rater were compared to human consensus. Inconsistent items were considered uncertain and thus “deferred to a second human rater”. In a practical application of this approach, the first and second human rater would then have to go through their normal consensus process.

For both (3) and (4), accuracy and kappa have to be interpreted in combination with the deferring fraction.

2.3.2. Reliability

We performed LLM prompts twice (ie, in duplicate for Claude-3-Opus, Claude-2, GPT-4 [only 25% of publications, due to high cost], GPT-3.5, Mixtral-8x22B) and compared the ratings of each of the 2 runs. Due to the nature of LLMs, these duplicate runs are not independent, which is why we consider their agreement “intrarater reliability”. All application programming interface (API) queries were performed with minimal randomness (“temperature” 0) to allow the highest possible intrarater

reliability. We compared LLM intrarater reliability with human inter-rater reliability.

2.3.3. Code and data availability

API querying, extraction of ratings, fixing minor formatting issues, and quantification of quote accuracy were performed in Python 3.11.4 using the parasail and rapidfuzz libraries [27,28]. Statistical analyses and visualizations were performed in R 4.3.

Codes and data are openly available on GitHub [29]. All prompt templates are in the Supplement. We provide a web dashboard which allows interactive exploration of results [26].

3. Results

3.1. Agreement with human consensus

3.1.1. Individual human raters

Accuracies of human raters 1 and 2 were 89% and 90% for PRISMA, 89% and 89% for AMSTAR, and 75% and

Table 2. Comparison of agreement between human consensus and (1) individual human rater 1; and (2) individual LLMs (Claude-3-Opus, Claude-2, GPT-4, GPT-3.5, Mixtral-8x22B) and their reliability

Rater	Agreement with human consensus		Reliability	
	Cohen's kappa (95% CI)	Agreement (%; 95% CI)	Cohen's kappa (95% CI)	Agreement (%; 95% CI)
PRISMA (27 items × 112 publications = up to 3024 ratings)				
(1) Human Rater 1	0.81 (0.78–0.83)	2686/3024 (89%, 87%–90%)	0.84 (0.80–0.88) ^a	2384/2619 (91%, 89%–93%) ^a
(2) Claude-3-Opus	0.51 (0.47–0.55) ^b	2073/2943 (70%, 68%–72%) ^b	0.68 (0.64–0.71) ^b	2318/2889 (80%, 78%–82%) ^b
(2) Claude-2	0.51 (0.47–0.55)	2116/3024 (70%, 68%–72%)	0.83 (0.80–0.85)	2695/3024 (89%, 87%–91%)
(2) GPT-4	0.45 (0.42–0.48) ^b	2018/2943 (69%, 66%–71%) ^b	0.80 (0.72–0.86) ^c	677/756 (90%, 85%–93%) ^c
(2) GPT-3.5	0.40 (0.37–0.44) ^b	1868/2943 (63%, 62%–66%) ^b	0.90 (0.87–0.93) ^b	2762/2943 (94%, 92%–96%) ^b
(2) Mixtral-8x22B	0.42 (0.37–0.47) ^b	1922/2997 (64%, 61%–67%) ^b	0.46 (0.40–0.52) ^b	1997/2997 (67%, 63%–70%) ^b
AMSTAR (11 items × 112 publications = up to 1232 ratings)				
(1) Human Rater 1	0.80 (0.76–0.83)	1096/1232 (89%, 87%–91%)	0.77 (0.72–0.82) ^a	936/1067 (88%, 85%–90%) ^a
(2) Claude-3-Opus	0.56 (0.52–0.60) ^b	888/1199 (74%, 72%–76%) ^b	0.68 (0.63–0.72) ^b	947/1177 (80%, 78%–83%) ^b
(2) Claude-2	0.39 (0.34–0.43)	777/1232 (63%, 60%–66%)	0.91 (0.87–0.94)	1163/1232 (94%, 92%–96%)
(2) GPT-4	0.45 (0.40–0.51) ^b	836/1199 (70%, 67%–73%) ^b	0.79 (0.68–0.87) ^c	274/308 (89%, 83%–94%) ^c
(2) GPT-3.5	0.27 (0.22–0.32)	652/1232 (53%, 50%–56%)	0.79 (0.73–0.83)	1067/1232 (87%, 83%–90%)
(2) Mixtral-8x22B	0.34 (0.29–0.39) ^b	726/1221 (59%, 56%–63%) ^b	0.44 (0.38–0.50) ^b	784/1221 (64%, 60%–68%) ^b
PRECIS-2 (9 items × 56 publications = up to 504 ratings)				
(1) Human Rater 1	0.57 (0.48–0.65)	379/504 (75%, 70%–80%)	0.29 (0.20–0.37)	288/504 (57%, 51%–63%)
(2) Claude-3-Opus	0.12 (0.06–0.19)	225/504 (45%, 40%–49%)	0.65 (0.57–0.71)	364/504 (72%, 68%–77%)
(2) Claude-2	0.01 (–0.08 to 0.08)	222/504 (44%, 39%–49%)	0.71 (0.63–0.77)	384/504 (76%, 71%–81%)
(2) GPT-4	0.02 (–0.04 to 0.08)	193/504 (38%, 33%–43%)	0.68 (0.52–0.80) ^c	80/126 (63%, 52%–75%) ^c
(2) GPT-3.5	0.04 (–0.05 to 0.12) ^b	265/486 (55%, 50%–58%) ^b	0.62 (0.52–0.71) ^b	370/477 (78%, 71%–84%) ^b
(2) Mixtral-8x22B	0.05 (–0.03 to 0.13)	244/504 (48%, 43%–54%)	0.40 (0.29–0.50)	305/504 (61%, 54%–67%)

AMSTAR, A MeaSurement Tool to Assess systematic Reviews; LLM, large language model; PRECIS-2, PRagmatic Explanatory Continuum Indicator Summary 2; PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses.

^a For human rater 2 ratings for 15 publications missing for PRISMA and AMSTAR.

^b Responses for 1–3 publications unsuccessful and thus missing.

^c GPT-4 intrarater reliability based on 25% of publications (28 for PRISMA & AMSTAR, 14 for PRECIS-2).

Box 1 Prompt and example response for GPT-4 / PRECIS-2

System prompt

You are an expert in clinical trial design and are tasked to assess the PRECIS-2 scores of a trial based on the full text of its publication. You have the briefing in the attached "Loudon 2015 shortened.txt" and "PRECIS Toolkit page 4.txt".
<Loudon 2015 shortened.txt> <PRECIS Toolkit page 4.txt>

User prompt

This is the full text to be assessed: <722.txt>

For each of the nine domains of PRECIS-2, perform the following steps.

1. Extract 1-3 relevant quotes from the full text.
2. Explain your reasoning in 1 paragraph.
3. Give a score X from 1 (very explanatory) to 5 (very pragmatic) in square brackets like this: "Score: [X]". If the full text does not contain enough information to assess a specific domain, report "Score: [NA]".

Do not provide a final summary paragraph in the end.

Response (3 of 9 domains)

1. Eligibility criteria

- Quote: "Participants had established CVD or an estimated 5-year CVD risk of 15%, with indications for antiplatelet, statin and 2 blood pressure lowering drugs ('combination treatment'). Participants randomized to the 'polypill-based strategy' received a polypill containing aspirin 75 mg, simvastatin 40 mg, lisinopril 10 mg and either atenolol 50 mg or hydrochlorothiazide 12.5 mg."
- Reasoning: The trial includes participants with established cardiovascular disease or a high risk of cardiovascular disease, which is a broad and inclusive group. However, the trial does exclude individuals who do not have indications for the combination treatment, which is a somewhat restrictive criterion.
- Score [4].

2. Recruitment

- Quote: "A total of 623 participants were randomized from 33 centers (12 Aboriginal Medical Services) between January 2010 and May 2012."
- Reasoning: The trial recruited participants from a variety of centers, including Aboriginal Medical Services, which suggests a pragmatic approach to recruitment. However, the text does not provide detailed information about the recruitment process, such as whether participants were recruited through usual appointments or other methods.
- Score: [NA]

3. Setting

- Quote: "Participants attended the primary health-care centers for trial assessments at randomization and 12 month intervals thereafter."
- Reasoning: The trial was conducted in primary health care centers, which is a typical setting for the delivery of the intervention in usual care. This suggests a pragmatic approach to the setting.
- Score [5].

73% for PRECIS-2, respectively (Table 2, Fig 2, Supplementary Figs 1-6).

3.1.2. Individual LLMs

For PRISMA, individual LLM accuracies ranged from 63% (GPT-3.5) to 70% (Claude-3-Opus), for AMSTAR from 53% (GPT-3.5) to 74% (Claude-3-Opus), and for PRECIS-2 from 38% (GPT-4) to 55% (GPT-3.5) (Table 2,

Fig 2, Supplementary Figs 7-21). The Box 1 shows an exemplary GPT-4 prompt and response for PRECIS-2.

While differences in the prompts used complicate comparisons between individual LLMs, averaging their performance nevertheless gives a broad overview: The best average accuracy was achieved by Claude-3-Opus (63%), followed by Claude-2 (59%), GPT-4 (59%), Mixtral-8x22B (57%), and GPT-3.5 (57%). This ranking is also

reflected by their average kappas: 0.40, 0.30, 0.31, 0.27, and 0.24, respectively.

3.1.3. Combined LLMs: consistency approach

All nine LLM assessments were combined using only ratings consistent in a majority of LLM assessments. Responses without such a consistent majority would be deferred to human raters. This approach led to substantial improvements with accuracies ranging from 75% to 88% for PRISMA (while deferring 4%–74% of ratings), from 74% to 89% for AMSTAR (while deferring 6%–84% of ratings), and from 64% to 79% for PRECIS-2 (with deferring fractions from 29% to 88%) (Fig 2, Supplementary Figs 22–36). By consecutively increasing the consistency threshold

from $\geq 5/9$ to finally 9/9, accuracies and deferring fractions increased gradually (Fig 3). The performance of combined LLMs at high-consistency thresholds (9/9 for PRISMA and AMSTAR and $\geq 8/9$ for PRECIS-2) was statistically indifferent from human performance, evidenced by overlapping CIs for both accuracy and kappa, Table 3.

3.1.4. Human–AI collaboration

For PRISMA, combining human rater 1 with individual LLMs by only considering ratings consistent between human and LLM led to accuracies ranging from 89% (GPT-4) to 96% (Claude-2) while deferring 25% and 35% of responses to the second human rater, respectively. Conversely, this would spare the second human rater 65%

Table 3. Comparison of agreement between human consensus and (3) combined LLMs (consistent in 5 to 9 of 9 LLM ratings), (4) human-AI collaboration for each of the 3 tools. The confidence intervals of bold cells are not overlapping with those of the human rater alone (compare Table 1)

Rater	Cohen's kappa (95% CI)	Agreement (%; 95% CI)	Deferring fraction (%; 95% CI)
PRISMA (27 items $\times$ 112 publications = up to 3024 ratings)			
(3) Combined LLMs (consistent in 9/9)	0.77 (0.72–0.81)	679/776 (88%, 85%–90%)	2248/3024 (74%, 72%–77%)
(3) Combined LLMs (consistent in 8/9)	0.76 (0.72–0.79)	1216/1416 (86%, 84%–88%)	1608/3024 (53%, 50%–56%)
(3) Combined LLMs (consistent in 7/9)	0.71 (0.68–0.74)	1627/1958 (83%, 81%–85%)	1066/3024 (35%, 33%–38%)
(3) Combined LLMs (consistent in 6/9)	0.65 (0.62–0.68)	1941/2449 (79%, 78%–81%)	575/3024 (19%, 17%–21%)
(3) Combined LLMs (consistent in 5/9)	0.59 (0.55–0.62)	2180/2899 (75%, 73%–77%)	125/3024 (4%, 3%–5%)
(4) Human Rater 1 & Claude-3-Opus	0.90 (0.88–0.93)^a	1878/1989 (94%, 93%–96%)^a	954/2943 (32%, 30%–34%)^a
(4) Human Rater 1 & Claude-2	0.92 (0.90–0.94)	1883/1970 (96%, 94%–97%)	1054/3024 (35%, 33%–37%)
(4) Human Rater 1 & GPT-4	0.81 (0.78–0.83) ^a	1955/2194 (89%, 87%–91%) ^a	749/2943 (25%, 23%–28%) ^a
(4) Human Rater 1 & GPT-3.5	0.91 (0.88–0.94)^a	1653/1740 (95%, 93%–97%)^a	1203/2943 (41%, 39%–43%)^a
(4) Human Rater 1 & Mixtral-8x22B	0.92 (0.89–0.94) ^a	1715/1801 (95%, 94%–97%) ^a	1196/2997 (40%, 37%–44%) ^a
AMSTAR (11 items $\times$ 112 publications = up to 1232 ratings)			
(3) Combined LLMs (consistent in 9/9)	0.82 (0.75–0.89)	173/194 (89%, 85%–94%)	1038/1232 (84%, 82%–87%)
(3) Combined LLMs (consistent in 8/9)	0.73 (0.66–0.79)	342/410 (83%, 79%–87%)	822/1232 (67%, 64%–70%)
(3) Combined LLMs (consistent in 7/9)	0.68 (0.63–0.73)	528/654 (81%, 78%–84%)	578/1232 (47%, 44%–50%)
(3) Combined LLMs (consistent in 6/9)	0.62 (0.58–0.66)	713/921 (77%, 75%–80%)	311/1232 (25%, 23%–28%)
(3) Combined LLMs (consistent in 5/9)	0.56 (0.52–0.60)	863/1164 (74%, 72%–76%)	68/1232 (6%, 4%–7%)
(4) Human Rater 1 & Claude-3-Opus	0.92 (0.89–0.94)^a	798/837 (95%, 94%–97%)^a	362/1199 (30%, 28%–33%)^a
(4) Human Rater 1 & Claude-2	0.88 (0.85–0.91)	703/753 (93%, 92%–95%)	479/1232 (39%, 36%–42%)
(4) Human Rater 1 & GPT-4	0.83 (0.80–0.87) ^a	793/872 (91%, 89%–93%) ^a	327/1199 (27%, 25%–30%) ^a
(4) Human Rater 1 & GPT-3.5	0.93 (0.90–0.96)	563/587 (96%, 94%–97%)	645/1232 (52%, 49%–56%)
(4) Human Rater 1 & Mixtral-8x22B	0.91 (0.87–0.94)^a	643/679 (95%, 93%–96%)^a	542/1221 (44%, 41%–48%)^a
PRECIS-2 (9 items $\times$ 56 publications = up to 504 ratings)			
(3) Combined LLMs (consistent in 9/9)	0.49 (0.20–0.73)	49/62 (79%, 67%–89%)	442/504 (88%, 84%–91%)
(3) Combined LLMs (consistent in 8/9)	0.32 (0.04–0.57)	79/106 (75%, 65%–83%)	398/504 (79%, 74%–84%)
(3) Combined LLMs (consistent in 7/9)	0.22 (0.05–0.40)	116/163 (71%, 64%–78%)	341/504 (68%, 62%–73%)
(3) Combined LLMs (consistent in 6/9)	0.17 (0.04–0.32)	164/239 (69%, 63%–74%)	265/504 (53%, 47%–58%)
(3) Combined LLMs (consistent in 5/9)	0.11 (0.01–0.22)	266/415 (64%, 58%–69%)	89/504 (18%, 14%–21%)
(4) Human Rater 1 & Claude-3-Opus	0.73 (0.62–0.83)	106/127 (83%, 77%–90%)	377/504 (75%, 72%–78%)
(4) Human Rater 1 & Claude-2	0.56 (0.35–0.76)	101/126 (80%, 72%–88%)	378/504 (75%, 71%–79%)
(4) Human Rater 1 & GPT-4	0.63 (0.49–0.76)	98/123 (80%, 72%–87%)	381/504 (76%, 72%–80%)
(4) Human Rater 1 & GPT-3.5	0.68 (0.49–0.83) ^a	119/139 (86%, 79%–91%) ^a	347/486 (71%, 68%–75%) ^a
(4) Human Rater 1 & Mixtral-8x22B	0.61 (0.45–0.77)	123/149 (83%, 76%–89%)	355/504 (70%, 66%–75%)

AMSTAR, A MeaSurement Tool to Assess systematic Reviews; LLM, large language model; PRECIS-2, PRagmatic Explanatory Continuum Indicator Summary 2; PRISMA, Preferred Reporting Items for Systematic reviews and Meta-Analyses.

^a Responses for 1–3 publications unsuccessful and thus missing.

of responses when accepting 96% accuracy, that is, 1 wrong response for every ~25 responses spared. Human–AI collaboration with Claude-3-Opus, Claude-2, GPT-3.5, and Mixtral-8x22B led to significantly more accurate responses than either human rater 1 or 2 alone (8 of 10 possible human–AI pairs, Table 3, Figure 2, Supplementary Table 2, Supplementary Figures 37–41).

For AMSTAR, human–AI collaboration led to accuracies ranging from 91% (GPT-4) to 95% (Claude-3-Opus) while deferring 27% and 30% of responses, respectively. Conversely, this would spare the second human rater 70% responses when accepting 95% accuracy, that is, 1 wrong response for every ~20 responses spared. Human–AI collaboration with Claude-3-Opus, Claude-2, GPT-3.5, and Mixtral-8x22B led to significantly more accurate responses than either human rater 1 or 2 alone (8 of 10 possible human–AI pairs, Table 3, Figure 2, Supplementary Table 2, Supplementary Figures 42–46).

For PRECIS-2, accuracies were ranging from 80% (GPT-4) to 86% (GPT-3.5) while deferring 76% and 71% of responses, respectively. Conversely, this would spare the second human rater 29% of responses when accepting 86% accuracy, that is, 1 wrong response for every ~7 responses spared. Only the combination of human rater 2 with GPT-3.5 led to significantly more accurate responses than human rater 2 alone (1 of 10 possible human–AI pairs, Table 3, Figure 2, Supplementary Table 2, Supplementary Figures 47–51).

In all cases, median agreements with human consensus and deferring fractions for individual publications and for individual items were very similar to the overall metrics (Supplementary Figures 1–51).

3.2. Reliability

Human inter-rater reliability measured by agreement was 91%, 88%, and 57% and by kappa 0.84, 0.77, and 0.29 for PRISMA, AMSTAR, and PRECIS-2, respectively.

For PRISMA, intrarater reliability of LLMs was similar to human inter-rater-reliability for Claude-2, GPT-4, and GPT-3.5 but worse for Claude-3-Opus and Mixtral-8x22B. For AMSTAR, it was better than humans for Claude-2, similar for GPT-4 and GPT-3.5, but worse for Claude-3-Opus and Mixtral-8x22B. For PRECIS-2, it was better than humans for Claude-3-Opus, Claude-2, GPT-4, and GPT-3.5, and similar for Mixtral-8x22B (Table 2).

3.3. Cost, effort, formatting, and quoting accuracy

Mixtral-8x22B was the most affordable model with a median of \$1.20 per 100 papers and GPT-4 the most expensive one with a median of \$115.00. Model response speeds were ranging from ~10 seconds (GPT-3.5) to 2 minutes (GPT-4) per paper (Table 1). Rarely, API rate limits required a break until the end of the day to continue calling Claude-3-Opus, Claude-2, and GPT-4.

Claude-3-Opus could persistently not process 3 of 112 (2.7%) publications for PRISMA and AMSTAR because of the output being blocked by Anthropic’s content filtering policy or being too long. Claude-2 was able to process all publications. GPT-4 could not process 3 lengthy publications (2.7%) for PRISMA and AMSTAR combined due to its context length. GPT-3.5 could persistently not process 3 of 112 publications (2.7%) for PRISMA and 2 of 56 publications (3.6%) for PRECIS-2. Mixtral-8x22B could persistently not process 1/112 publications (1%) for PRISMA and AMSTAR. Claude-3-Opus, GPT-3.5, and Mixtral-8x22B had to be reprompted several times for up to 13% of publications until success. Claude-3-Opus and Claude-2 rarely and GPT-3.5 and Mixtral-8x22B often exhibited minor automatically fixable formatting issues like forgotten squared brackets around ratings or wrong responses like “[Unclear]” (Supplementary Table 3).

All prompts required the LLMs to “extract 1–3 relevant quotes from the full text” per item. For PRISMA (27 items), a median of 14 quotes (range 5–17) were provided per publication, 0.5 quotes/item. For AMSTAR (11 items), a median of 7 quotes (range 4–8) were provided per publication, 0.6 quotes/item. For PRECIS-2 (9 domains), a median of 10 quotes (range 9–10) were provided per publication, 1 quote/domain. Claude-3-Opus, Claude-2, and Mixtral-8x22B sometimes quoted from the provided briefings instead of the full text to be assessed, which was not part of the instructions, while GPT-4 and GPT-3.5 rarely did this. The median quote similarity with the original full text was 99%, with some quotes slightly shortened (eg, removing references or brackets) or rephrased (Supplementary Table 3).

4. Discussion

This analysis of the agreement of LLMs with human consensus in evidence appraisal of different levels of complexity identified some areas where human–AI collaboration might outperform a traditional consensus process of human raters in efficiency but also indicated clear limitations.

Individual LLMs performed significantly worse than humans for all three evidence appraisal tools. They generally performed best for reporting assessment with PRISMA, followed by the assessment of systematic review methodological rigor with AMSTAR. For the assessment of clinical trial pragmatism through PRECIS-2, the accuracy was poor and not useful in research practice. This graded performance likely reflects the increasing complexity between these tasks. Reporting assessment with PRISMA is almost exclusively a task of text and language assessment, thus an optimal scenario for LLMs. AMSTAR uses (like many risk-of-bias tools) checklists and questions where specific words and phrases are highly indicative of methodological features (eg, “systematic search”, “gray literature”) and

sufficient for a superficial categorization. Contrarily, pragmatism of clinical trials (PRECIS-2) is determined by complex features, and it requires more than a search for specific signal words and phrases. This increasing complexity may also be reflected by the decreasing human inter-rater reliability with 91% for PRISMA, 88% for AMSTAR, and 57% for PRECIS-2.

Claude-3-Opus performed best for PRISMA and AMSTAR, followed by Claude-2 and GPT-4, while the smallest and cheapest models GPT-3.5 and Mixtral-8x22B performed worst. Interestingly however, this ranking was somewhat inverted for PRECIS-2: Claude-3-Opus, Claude-2, and GPT-4 performed worse than GPT-3.5 and Mixtral-8x22B. This highlights that newer and larger LLMs do not always perform better. We believe this counterintuitive paradox can be explained by differences in prompts, model “personality,” and class imbalance: The dataset used to assess PRECIS-2 contained mostly pragmatic trials and few explanatory ones. The smaller, simpler models might somehow be biased toward more pragmatic scores, compared to the otherwise better performing larger models Claude-3-Opus, Claude-2, and GPT-4, which might be trained to be more balanced. Of note, the fully open source Mixtral-8x22B on average performed better than the proprietary GPT-3.5.

Combining all LLM assessments led to increasing accuracies and deferring fractions with increasing consistency thresholds. At the highest consistency-thresholds, performance reached that of human raters, albeit only for few responses due to very high deferring fractions. Selecting only consistent majority ratings and filtering out inconsistent ones can also be regarded as a measure of uncertainty, which is very important for AI algorithms [30]. The increasing deferring fractions may also reflect the increasing complexity between the 3 tools.

Overall, these levels of accuracy suggest that individual and combined LLMs with our prompts are not yet good enough to be used alone in these evidence appraisals, and their performance is worse in more complex situations.

The best performance and currently most promising approach was human–AI collaboration: ratings consistent between a human rater and LLMs showed significantly better accuracies than human raters alone for 8 of 10 possible collaboration-pairs for PRISMA and AMSTAR (up to 96% accuracy for PRISMA and 95% for AMSTAR compared to 89% accuracy of human rater 1 alone). Approximately one-third of the items would still need to be assessed by a second human rater, but sparing them two-thirds of the items might mean substantial workload reduction. Most models could thus be used to identify high-certainty ratings by the first human rater and filter out low-certainty ones to be deferred and double-checked. The second human rater would only have to rate these deferred items, leading to potential workload reduction. For PRECIS-2, there was still substantial improvement in accuracy but statistically significant in only 1 of 10 collaboration pairs and a large

majority of items would still need to be assessed by the second human rater.

Today’s LLMs may thus already provide a viable research partner for human raters for PRISMA and AMSTAR, but the exact reduction in workload is difficult to estimate. We cannot exclude that the reduction may not be that substantial, for example, if the context of the nondeferred items needs to be read and understood anyway before the deferred items are also appraised.

4.1. Limitations

First, there is a general concern about “train/test contamination” or “data leakage” with all LLM benchmarks because of the extensive web scraping used for their training. This refers to the risk of machine learning models having seen and potentially memorized test-outcomes from training-data. However, this typically affects popular data found multiple times on the internet, as it has been shown that “duplication encourages memorization” [31,32]. While the human consensus datasets used as comparators in this work are openly available on the internet, they can only be found as tabular data (Excel/CSV). Tabular data are unlikely to be part of an LLM training corpus, as they are not useful in self-supervised training to predict the next word of a sentence. Even if they were part of an LLM training corpus, memorization would be highly unlikely because of very low numbers of duplications. Nevertheless, only prospective replication of these results with new human consensus datasets would eliminate the risk of train/test contamination.

Second, as discussed above, the dataset used in this study to assess PRECIS-2 contains mostly pragmatic trials and few explanatory ones. A more balanced dataset could complement the current findings and challenge the surprising superiority of GPT-3.5 and Mixtral-8x22B compared to larger models.

Third, 2 of the 3 tools used have seen updates in recent years: PRISMA 2020 [33] and AMSTAR-2 [34], meaning the current results might only partly apply to them. However, differences between the versions used and the updated versions are not major.

Fourth, except Claude-3-Opus, the LLMs used could only process text and were blind to images. Figures potentially relevant for PRISMA, AMSTAR, and PRECIS-2 include flowcharts and sometimes tables are also presented as images. Newer versions of GPT-4 released after our assessment also allow multimodal prompts [20]. However, the difference in average accuracy between Claude-3-Opus (multimodal, 63%) and Claude-2 (text-only, 59%) was not major, and only a part of this difference is due to multimodality of the input data—the presumably bigger part is due to LLM-properties.

Fifth, the presented datasets contain ratings from only two human experts, whereas a higher number would make the human consensus more robust and eventually

approximate a “ground truth”—a strong term which cannot be claimed for the presented datasets. However, we believe this imperfect benchmark does not bias our results in a consistent way. Furthermore, 2 human raters present the most common setting and thus reflect real research practice.

Finally, the presented datasets do not contain information on how long it took the human raters to assess each publication and item individually and how long the consensus process took (“time on task”). This means the efficiency and potential time savings cannot be quantified. It is possible that a human has to read and process an article in the same way looking at the whole paper, even if the LLM has already arbitrated on some items. If so, there would be no or little time savings for the human.

4.2. Outlook

To validate the utility of LLMs for assessing PRISMA and AMSTAR in the human–AI collaboration approach that we propose, they should be evaluated prospectively against a traditional consensus process within the workflow of a new review with a preregistered protocol in a confirmatory fashion. This future, prospective work would benefit from having many human raters with the highest expertise to maximize their accuracy. Time on task should be recorded precisely to allow quantification of efficiency and potential time savings.

Moreover, we provide an open and reproducible benchmark that can be built upon by other groups. Testing more diverse prompt engineering techniques may further improve performance, ranging from minor variations in wording to more sophisticated multiprompt techniques like tree of thoughts and graph of thoughts [35,36]. Both industry and academia have introduced multimodal LLMs fine-tuned specifically for biomedical contexts, demonstrating encouraging performance [37,38]. However, these models are not generally available. We believe focusing on open source models instead of proprietary ones will allow more transparency and control in the future, aligning better with scientific principles [39]. Meta’s flagship open source model Llama-3 will be an exciting addition to this benchmark, once longer context windows are made available [40]. Fine-tuning such open source models toward a single evidence appraisal tool may yield the biggest benefits.

Finally, the presented framework can be extended to other evidence appraisal tools like “CONsolidated Standards Of Reporting Trials,” for which a short report evaluating ChatGPT was published [41], or “Risk of Bias,” for which a protocol intending similar evaluations was published [42].

5. Conclusion

Current LLMs alone appraised evidence substantially worse than humans. Pairing a first human rater with an

LLM as human–AI collaboration may reduce workload for the second human rater for the assessment of reporting (PRISMA) and methodological rigor (AMSTAR) while maintaining very high accuracy but not for more complex tasks such as assessing pragmatism of clinical trials (PRECIS-2).

CRedit authorship contribution statement

Tim Woelfle: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Project administration. **Julian Hirt:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Perrine Janiaud:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Ludwig Kappos:** Writing – review & editing, Supervision, Conceptualization, Funding acquisition. **John P.A. Ioannidis:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Lars G. Hemkens:** Writing – review & editing, Supervision, Investigation, Data curation, Conceptualization, Funding acquisition, Methodology, Validation.

Data availability

All data are openly available: <https://github.com/timwoelfle/Evidence-Appraisal-AI>.

Declaration of competing interest

There are no competing interests for any author.

Acknowledgments

The authors sincerely thank Paul Stephen Cullis and colleagues for providing individual human ratings for PRISMA and AMSTAR for their systematic review in addition to the already openly shared human consensus data [14].

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2024.111533>.

References

- [1] Moher D. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264.
- [2] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;7:10.

- [3] Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;350:h2147.
- [4] Wallace BC, Kuiper J, Sharma A, Zhu M, Marshall IJ. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *J Mach Learn Res* 2016;17:4572–96.
- [5] Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016;23:193–201.
- [6] Kilicoglu H, Jiang L, Hoang L, Mayo-Wilson E, Vinkers CH, Otte WM. Methodology reporting improved over time in 176,469 randomized controlled trials. *J Clin Epidemiol* 2023;162:19–28.
- [7] Katz DM, Bommarito MJ, Gao S, Arredondo P. GPT-4 Passes the bar exam. *SSRN Electron J* 2023;382:20230254.
- [8] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2303.13375>.
- [9] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183:589–96.
- [10] Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models “the answer” to bringing us closer to systematic review automation? *Syst Rev* 2023;12:72.
- [11] Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. Can large language models replace humans in systematic reviews? Evaluating GPT-4’s efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods* 2024;5:616–26.
- [12] Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods* 2024;5:576–89.
- [13] Li T, Higgins JP, Deeks JJ. Collecting data. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. *Cochrane Handb. Syst. Rev. Interv.* 1st ed. Hoboken, New Jersey: Wiley; 2019:109–41.
- [14] Cullis PS, Gudlaugsdottir K, Andrews J. A systematic review of the quality of conduct and reporting of systematic reviews and meta-analyses in paediatric surgery. *PLoS One* 2017;12:e0175213.
- [15] Pollock M, Fernandes RM, Hartling L. Evaluation of AMSTAR to assess the methodological quality of systematic reviews in overviews of reviews of healthcare interventions. *BMC Med Res Methodol* 2017;17:48.
- [16] Hirt J, Janiaud P, Dublin P, Hemkens LG. Meta-research on pragmatism of randomized trials: rationale and design of the PragMeta database. *Trials* 2023;24:437.
- [17] Hirt J, Janiaud P, Dublin P, Nicoletti GJ, Dembowska K, Nguyen TVT, et al. Use of pragmatic randomized trials in multiple sclerosis: a systematic overview. *Mult Scler* 2024;30:463–78.
- [18] Anthropic. The Claude 3 model family: opus, sonnet, haiku. 2024. Available at: <https://www.anthropic.com/claude-3-model-card>. Accessed May 14, 2024.
- [19] Anthropic. Model card and evaluations for Claude models. 2023. Available at: <https://www.anthropic.com/claude-2-model-card>. Accessed May 14, 2024.
- [20] OpenAI. GPT-4 technical report. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2303.08774>.
- [21] OpenAI. GPT-3.5 - models - OpenAI API 2023. Available at: <https://platform.openai.com/docs/models/gpt-3-5>. Accessed August 14, 2023.
- [22] Mistral AI. Mixtral 8x22B - cheaper, better, faster, Stronger. 2024. Available at: <https://mistral.ai/news/mixtral-8x22b/>. Accessed May 14, 2024.
- [23] Chiang WL, Zheng L, Sheng Y, Angelopoulos AN, Li T, Li D, et al. Chatbot Arena: an open platform for evaluating LLMs by human preference. *arXiv* 2024. <https://doi.org/10.48550/arXiv.2403.04132>.
- [24] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172–80.
- [25] Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain of thought reasoning in language models. *arXiv* 2023. <https://doi.org/10.48550/arXiv.2203.11171>.
- [26] Woelfle T. Evidence-appraisal-AI web dashboard. Available at: <https://timwoelfle.github.io/Evidence-Appraisal-AI/>. Accessed May 14, 2024.
- [27] Bachmann M. Levenshtein - RapidFuzz 3.2.0 documentation. 2023. Available at: <https://maxbachmann.github.io/RapidFuzz/Usage/distance/Levenshtein.html#normalized-similarity>. Accessed August 14, 2023.
- [28] Daily J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinf* 2016;17:81.
- [29] Woelfle T. Evidence-appraisal-AI repository. Available at: <https://github.com/timwoelfle/Evidence-Appraisal-AI>. Accessed May 14, 2024.
- [30] Granziera C, Woelfle T, Kappos L. Development and implementation of new diagnostic technologies in neurology. *Nat Rev Neurol* 2022;8:445–6.
- [31] Chang KK, Cramer M, Soni S, Bamman D. Speak, memory: an archaeology of books known to ChatGPT/GPT-4. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2305.00118>.
- [32] Zhou Z, Xiang J, Chen C, Su S. Quantifying and analyzing entity-level memorization in large language models. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2308.15727>.
- [33] Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* 2021;372:n160.
- [34] Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, et al. Amstar 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* 2017;358:j4008.
- [35] Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, et al. Tree of thoughts: deliberate problem solving with large language models. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2305.10601>.
- [36] Besta M, Blach N, Kubicek A, Gerstenberger R, Gianinazzi L, Gajda J, et al. Graph of thoughts: solving elaborate problems with large language models. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2308.09687>.
- [37] Tu T, Azizi S, Driess D, Schaeckermann M, Amin M, Chang P-C, et al. Towards generalist biomedical AI. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2307.14334>.
- [38] Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-flamingo: a multimodal medical few-shot learner. *arXiv* 2023. <https://doi.org/10.48550/ARXIV.2307.15189>.
- [39] van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023;614:224–6.
- [40] Meta AI. Introducing Meta Llama 3: the most capable openly available LLM to date 2024. Available at: <https://ai.meta.com/blog/meta-llama-3/>. Accessed May 14, 2024.
- [41] Roberts RH, Ali SR, Hutchings HA, Dobbs TD, Whitaker IS. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. *BMJ Health Care Inform* 2023;30:e100830.
- [42] Rose CJ, Ringsten M, Bidonde J, Glanville J, Berg RC, Cooper C, et al. Using a large language model (ChatGPT) to assess risk of bias in randomized controlled trials of medical interventions: protocol for a pilot study of interrater agreement with human reviewers. *Res Square* 2023. <https://doi.org/10.21203/rs.3.rs-3288515/v1>.