Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large Language Models and Generative Artificial Intelligence for Biomedical Research Appraisal and Peer Review: A Scoping Review

Ben D. Macfarlane
2619211m@student.gla.ac.uk
University of Glasgow
Yanna Hui
University of Glasgow
Latifa Yakubu
University of Glasgow
Jesse Dawson
University of Glasgow

Systematic Review

Keywords: large language models, generative artificial intelligence, research appraisal, risk of bias, peer review, scoping review

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10373594/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. BDM declares that he is Director of Strata Collective Ltd., which develops technology for the review and appraisal of biomedical research. YH, LY, and JD declare that they have no competing interests.

Large Language Models and Generative Artificial Intelligence for Biomedical Research Appraisal and Peer Review: A Scoping Review

Ben D. Macfarlane BSc[1], Yanna Hui[1], Latifa Yakubu[1], Jesse Dawson MD[2]

Affiliations

1. University of Glasgow, School of Medicine, Dentistry and Nursing, Glasgow, United Kingdom
2. University of Glasgow, School of Cardiovascular and Metabolic Health, Glasgow, United Kingdom

Corresponding Author Details

Ben D. Macfarlane BSc (Hons)

School of Medicine, Dentistry and Nursing

University of Glasgow, Glasgow, United Kingdom

Email: 2619211m@student.gla.ac.uk

Number of tables: 3

Number of figures: 1

Word count: 4,642

Abstract

Background

Large language models (LLMs) and generative artificial intelligence (AI) are increasingly proposed to support the appraisal of biomedical research, including risk-of-bias assessment, methodological quality assessment and peer review. However, evidence regarding their capabilities, limitations, and appropriate roles remains fragmented. This scoping review asked how large language models (LLMs) and other generative AI systems have been evaluated for appraising biomedical research reports, and what these evaluations have found about their performance relative to human appraisers.

Methods

This JBI-methodology scoping review was conducted in accordance with PRISMA-ScR reporting guidance and pre-registered on the Open Science Framework. Five databases were searched from 1 January 2020 to 11 June 2026, supplemented by screening of included studies' reference lists. Eligible studies evaluated LLMs or generative AI systems performing research appraisal tasks (risk-of-bias, methodological quality or reporting guideline assessment, or peer review) on biomedical research reports. Two reviewers independently screened records and verified extracted data, with findings synthesised narratively.

Results

Of 3,521 records identified, 43 studies were included. GPT-family models were the most frequently evaluated (88%), followed by Gemini (37%) and Claude (26%). Studies evaluated risk-of-bias assessment (n=16), peer review (n=12), methodological quality assessment (n=8) and multiple appraisal tasks (n=7). Most studies included human comparators (93%). Among 39 studies comparing AI with human appraisal, authors concluded that AI performance was inferior in 62%, comparable in 36% and superior in one study. Agreement was generally strongest for structured checklist-based assessments and weakest for subjective interpretive judgements.

Conclusion

Current evidence suggests LLMs may provide useful assistance for structured, rule-based appraisal tasks but remain unsuitable substitutes for expert human judgement in complex evaluative decisions. Future research should prioritise prospective, blinded validation studies using standardised benchmarks, transparent reporting, and assessment of real-world impact on research quality and peer review processes.

Keywords: large language models; generative artificial intelligence; research appraisal; risk of bias; peer review; scoping review

List of abbreviations:

- AI: artificial intelligence
- AMSTAR-2: A Measurement Tool to Assess Systematic Reviews 2
- AUROC: area under the receiver operating characteristic curve
- COPE: Committee on Publication Ethics
- ICC: intraclass correlation coefficient
- ICMJE: International Committee of Medical Journal Editors
- JBI: Joanna Briggs Institute
- LLM: large language model
- PCC: Population-Concept-Context
- PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
- RoB: risk of bias
- TRIPOD-LLM: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis using Large Language Models

Introduction

Research appraisal, including peer review, is fundamental to evidence-based medicine. Here, research appraisal means evaluating the validity, quality or reporting of a written research report, whether a published article, a preprint or a manuscript under review, rather than a research project as a process.[1, 2] This process underpins evidence synthesis, clinical guideline development, and editorial decision-making. However, the rapid expansion of biomedical literature has increased demands on reviewers, while growing methodological complexity, such as adaptive trial design, requires increasingly specialised expertise.[3-6] These pressures have raised concerns about the efficiency and sustainability of current appraisal processes, including limited evidence that editorial peer review improves report quality.[7, 8] Reviewer agreement is also only moderate for many appraisal decisions.[9-11]

Large language models (LLMs) and other generative artificial intelligence (AI) systems have emerged as promising tools to augment research appraisal. Recent studies suggest that LLMs can assist with research appraisal tasks, including risk-of-bias assessment, methodological quality assessment, reporting guideline assessment, and peer review.[12, 13] Large-scale evaluations have shown that LLM-generated feedback demonstrates substantial overlap with human reviewer comments and may improve the actionability of reviewer feedback.[12] Consequently, interest in integrating LLMs into editorial and evidence synthesis workflows has increased rapidly.

We examined these tasks together because each asks an assessor to judge the validity, quality or reporting of a written research report, and an LLM applied to any of them is performing the same underlying operation. The tasks differ along two dimensions. They range from structured instruments with defined items, such as risk-of-bias tools, to open-ended judgements, such as editorial peer review; and they are used in two settings: evidence synthesis, where reviewers appraise published studies, and editorial review, where editors and peer reviewers assess submitted manuscripts. Mapping the evidence across tasks allows performance to be compared along the structured-to-interpretive dimension, which a review of a single task could not do.

Despite rapid growth, the evidence base remains fragmented and methodologically heterogeneous. Studies evaluate diverse appraisal tasks using different LLMs, prompting strategies, reference standards, and performance metrics, making direct comparison between studies difficult. To our knowledge, no review has systematically mapped the application of LLMs and generative AI to the appraisal of biomedical research reports.

This scoping review addressed the question: how have large language models and other generative AI systems been evaluated for appraising biomedical research reports, and what have these evaluations found about their performance relative to human appraisers? A scoping review methodology was selected because this is a rapidly evolving field characterised by heterogeneous study designs, diverse AI models, varied appraisal tasks, and inconsistent outcome measures, making formal evidence synthesis inappropriate at this stage. Specifically, we asked: (1) which appraisal tasks and instruments have been evaluated; (2) which AI systems were evaluated, and how they were accessed, prompted and reported; (3) how evaluations were designed, including manuscript source, sample, reference standard, human comparator, blinding and metrics; (4) what studies concluded about AI performance compared with human appraisal, and whether this differed by task or item type; (5) what efficiency and cost outcomes were reported; and (6) what gaps exist in the evidence and its reporting.

Methods

This scoping review was conducted in accordance with the JBI Manual for Evidence Synthesis guidance for scoping reviews.[14] Reporting follows the PRISMA-ScR checklist (Additional File 1).[15] The protocol was pre-registered on the Open Science Framework prior to database searching.[16]

Eligibility Criteria

Eligibility criteria were defined a priori using the Population-Concept-Context (PCC) framework (Supplementary Table 1) and were applied consistently during title and abstract screening and full-text review.

Population

The population of interest comprised biomedical and clinical research manuscripts, including published articles, preprints, and manuscripts submitted for peer review. Studies evaluating AI systems using non-biomedical literature were excluded. Studies limited to literature searching, information retrieval, or summarisation without a research appraisal or quality assessment component were also excluded. No restrictions were placed on the study design of the manuscripts being appraised. Only reports published in English were eligible, because the review team did not have resources for translation; two full-text reports were excluded on this basis (Supplementary Table 4). Reports published from 1 January 2020 were eligible. This start date reflects the release of GPT-3 in 2020, before which generative language models of the kind evaluated here were not available.

Concept

The concept of interest was the empirical evaluation of LLMs, generative AI, or foundation models for one or more of the following tasks: risk-of-bias assessment, methodological quality assessment, reporting guideline assessment or peer review. The registered protocol used the terms ‘critical appraisal’ and ‘research quality assessment’; in this report both are covered by ‘research appraisal’ and the defined tasks above. Eligibility criteria were unchanged. Tasks were defined a priori (Table 1) and each was assigned to its usual setting, evidence synthesis or editorial review. Studies were required to report empirical performance data; descriptive papers, editorials, perspectives, and proposals without evaluation data were excluded. Studies evaluating traditional machine learning methods without a generative or language-modelling approach were also excluded.

Context

Studies conducted within biomedical, clinical, or health sciences research were eligible, including medicine, nursing, allied health, public health, and health services research. Studies from adjacent scientific disciplines were included where the manuscripts being appraised were biomedical in nature. Conference abstracts and proceedings papers were included only where sufficient methodological detail was available to permit data extraction.

Search Strategy

MEDLINE (via PubMed), Embase, Scopus, Web of Science and IEEE Xplore were searched on 11th June 2026 for records published from 1 January 2020. IEEE Xplore was included to capture evaluations published in computer science venues. Free-text terms for generative AI systems ("large language model", "generative AI", "foundation model", "ChatGPT", "GPT-4", "Claude", "Gemini", "Llama" and "Mistral") were combined with terms for appraisal tasks ("peer review", "critical appraisal", "risk of bias", "methodological quality", "quality assessment" and "reporting guideline") in titles and abstracts (abstracts only in Embase and IEEE Xplore; topic field in Web of Science). Full strategies are given in Supplementary Table 2.

To test sensitivity, the MEDLINE search was re-run on 20 October 2026 over the same period with additional terms ("LLM", "LLMs", and model names released during the search period: GPT-5, DeepSeek, Qwen and Copilot); 48 additional records not captured by the original search were screened and none met the inclusion criteria. Grey literature and preprint servers were not searched because the review required evaluations reported in enough methodological detail to chart study design, comparators and outcomes. Reference lists of included studies were also screened for additional eligible records.

To check for existing reviews, PubMed was searched on 20th October 2026 for reviews with title terms for LLMs or generative AI ("large language model*", LLM*, ChatGPT, "generative AI", "artificial intelligence") and for appraisal ("risk of bias", "critical appraisal", "peer review", "quality assessment", "methodological quality", appraisal), limited to the review publication type (n=49 records); none mapped LLM or generative AI evaluations for research appraisal.

Study Selection

Calibration

Before title and abstract screening, both reviewers (YH and LY) independently screened a calibration sample of 20 records (10 anticipated inclusions, 10 anticipated exclusions) selected by the lead reviewer (BDM) from pilot search results. Following independent assessment, reviewers compared

decisions and resolved disagreements through discussion to ensure consistent interpretation of the eligibility criteria. Agreement exceeded the pre-specified threshold of 80%, so calibration was not repeated. Calibration was applied to title and abstract screening only. The calibration records were retained in the main screening dataset.

Title and abstract screening

Following deduplication in EndNote, records were imported into Rayyan (Rayyan Systems, Inc.) for screening. Two reviewers independently screened all titles and abstracts using four sequential eligibility questions (Supplementary Table 3). Records were excluded if the answer to any screening question was "No". Where eligibility could not be determined from the title and abstract alone, studies progressed to full-text review. Disagreements were resolved by discussion, with a third reviewer (BDM) adjudicating unresolved cases. Percentage agreement was reported descriptively.

Full-text review

Full texts of potentially eligible studies were independently assessed by two reviewers against the predefined PCC eligibility criteria. Reasons for exclusion were recorded using eight categories specified in the protocol (Supplementary Table 4, categories 1 to 8), plus two added during the review: full text not available (9) and not eligible on final check (10). Disagreements were resolved through discussion and, where necessary, adjudicated by a third reviewer (BDM). At this stage, inter-reviewer agreement was quantified as percentage agreement and Cohen's κ with 95% confidence intervals.

Data Extraction

A standardised data extraction form was developed by the review team in Microsoft Excel and refined following pilot extraction of five included studies. Two reviewers independently extracted data from a random 20% sample of included studies to ensure consistency; discrepancies were resolved through discussion, with the lead reviewer (BDM) adjudicating any that remained. The remaining studies were extracted by one reviewer and verified by a second reviewer. The charted dataset, with page

references and supporting quotations, is available on the Open Science Framework (https://osf.io/p73ry).

Extracted variables covered study characteristics; AI system characteristics (model, version, evaluation date, access method, prompting); appraisal task characteristics (task, framework, manuscript source, publication years, sample size); comparator characteristics (human comparator, blinding); outcomes (agreement measures, time, cost and verification time); and methodological quality indicators. The full data dictionary is deposited on the Open Science Framework (https://osf.io/p73ry).

Data Synthesis

Given the anticipated heterogeneity in AI systems, appraisal tasks, study designs, and outcome measures, findings were synthesised descriptively in accordance with JBI guidance for scoping reviews. Results were organised by the six review questions; within questions 1 and 4, findings are reported by appraisal task (Table 1). Where multiple studies reported comparable outcome measures, descriptive summaries, including ranges of reported values, were presented rather than formal quantitative synthesis. Formal critical appraisal of individual included studies was not performed; instead, the quality-relevant indicators listed above were charted descriptively as part of data extraction.

For each study that compared AI with human appraisal, the review team classified the primary authors' stated conclusion as inferior, comparable or superior AI performance relative to the human comparator and recorded the supporting text. Studies whose authors drew no comparative conclusion were classified as such. We did not apply a common performance threshold because studies used metrics without a shared scale, including κ, ICC, accuracy, F1-score and Likert ratings.

Deviations from the Protocol

Two changes were made to the protocol. First, the four registered task categories (peer review, critical appraisal, risk-of-bias assessment and reporting guideline assessment) were replaced by risk-of-bias assessment, methodological quality assessment, peer review and multiple-task studies; reporting

guideline assessment remained eligible. Second, the four registered research questions were reorganised into one review question with six sub-questions, which cover the registered questions and add evaluation design and efficiency outcomes.

Results

Study Selection

The database search yielded 3,521 records. After removal of duplicates (n=1,782), 1,739 records were screened at title and abstract level. Following screening, 123 full-text articles were assessed for eligibility, of which 43 met the inclusion criteria and were included in the final synthesis (Figure 1).

Reviewers agreed on 95.2% of decisions at title/abstract screening, and on 95.1% of decisions at full-text review (Cohen's κ=0.90, 95% CI 0.83–0.98; almost perfect agreement).

Characteristics of Included Studies

A total of 43 studies were included, published between 2024 and 2026 (2024, n=7; 2025, n=21; 2026, n=15). Full references for all 43 studies are given in Supplementary Table 5 and in the Supplementary Reference List. Two studies by Mese and Kocak (2025b and 2026) evaluated the same 48 radiomics articles using different AI approaches, and two studies (Wang 2025; Kim and Yoon 2025) were published as conference proceedings papers. Included studies were conducted across 23 countries. The most common countries were the United States (n=11) and China (n=6); studies with authors from more than one country were counted in each country.

Study designs were heterogeneous: 31 studies (72%) used a validation design, 4 (9%) were experimental, and 8 (19%) were observational, longitudinal or retrospective diagnostic accuracy designs. Funding was reported in 14 (33%) studies, most commonly a mix of academic and government sources; no study reported industry funding. A conflict of interest was declared in 29 (67%) studies, none was declared in 11 (26%) and the declaration was not reported in 3 (7%). Only 3 (7%) studies were prospectively pre-registered.

1. Appraisal tasks and instruments

Included studies evaluated LLM performance across four main appraisal-related domains: risk of bias assessment (n=16), peer review (n=12), methodological quality assessment (n=8), and studies combining multiple tasks, such as screening, extraction, and appraisal together (n=7) (Table 1; Supplementary Table 6). Reporting guideline assessment was eligible but was not the primary task of any included study.

Risk of bias assessment studies most commonly applied the Cochrane RoB tool (RoB1 or RoB2; used in 12 of 16 studies in this domain), with a smaller number using JBI's Critical Appraisal Checklist, a bespoke instrument (ROBUST-RCT), or no validated framework. Sixteen studies overall (37%) used no named appraisal framework, instead applying custom rubrics or unstructured peer-review-style prompts. The manuscripts appraised were most commonly randomised controlled trials (n=13) or of mixed or unspecified design (n=22), with systematic reviews (n=3), observational studies (n=4), and scientific letters (n=1) appraised less frequently.

2. AI systems and how they were deployed

Across included studies, a wide range of LLMs and generative AI systems were evaluated. GPT-family models (including ChatGPT-4o, GPT-4, GPT-3.5, GPT-4-turbo, and GPT-5) were the most frequently assessed (88% of studies), followed by Gemini (versions 1.0-2.5, Flash and Pro variants) family models (n=16, 37%) and Claude (versions 2, 3, 3.5 Sonnet, 3 Opus, and Sonnet-4.5) (n=11, 26%) (Supplementary Table 6). Open-source foundation models, including Llama, Mistral, DeepSeek, and Qwen variants, were evaluated in 17 studies (40%), most often as comparators to proprietary models rather than as the primary system under test.

Reporting of the information needed to reproduce an evaluation was limited. A named version was given for every model in 31 studies (72%), but a dated snapshot or build identifier in only 5 (12%). A query date or windowwas reported in 18 studies (42%), and 12 studies (28%) gave both a named version and a date. The publication window of the appraised articles was reported in 19 studies (44%).

Study-level evaluation dates, publication windows and contamination safeguards are given in Supplementary Table 7.

Most studies (n=25) used web-based chat interfaces to access models (one with PDF upload), with a smaller number using API-based access (n=10), local deployment (n=2), or a third-party plug-in (n=1); access was mixed in 2 studies and not reported in 3. Prompting strategies varied substantially, with structured prompting (including checklist-based and role-based prompts) most common (n=21), followed by zero-shot prompting (n=5), chain-of-thought prompting (n=3) and zero-shot with few-shot examples (n=2); the other 12 studies used mixed or study-specific approaches. The majority of studies (n=32, 74%) compared two or more different AI systems head-to-head. Reporting of model parameters, including temperature and token limits, was inconsistent and frequently not stated.

3. Designs of evaluations

Included studies evaluated a median of 60 manuscripts each (range 3–7,334; IQR 28–160), reflecting substantial variation in study scale. Just over half of studies (n=23, 53%) were general or multi-specialty in scope; the remainder (n=20, 47%) focused on a single clinical specialty. Most studies evaluated AI performance against already-published manuscripts (n=34, 79%), with a smaller number using genuinely unpublished or in-review submissions (n=5, 12%) or a mixed/preprint source (n=4, 9%).

A human comparator or human-derived reference standard was included in 40 (93%) studies: expert reviewers (n=14), published expert ratings (n=13), peer-reviewers or journal editors (n=7), expert and junior reviewers (n=5) and trained assessors (n=1). Of these, 39 reported a comparative conclusion; Nyrhi (2026) used published expert ratings but drew no comparative conclusion. Three studies had no human comparator (Table 2). Blinding of human assessors to AI-generated outputs (or vice versa) was rarely reported: blinding of assessors was not reported in 22 studies (51%), and not applicable in 10 (23%), mostly where the reference standard was published ratings; blinding was described in 7 studies (16%) and partly in 4 (9%).

Across studies, agreement between AI systems and human reviewers or reference standards was most commonly assessed using Cohen's kappa (n=19, 44%), followed by intraclass correlation coefficients (ICC; n=8, 19%), percentage or observed agreement (n=6, 14%), Krippendorff's alpha (n=3, 7%), Kendall's tau or W (n=3, 7%) and Gwet's AC2 (n=1, 2%); several studies reported more than one agreement statistic. A smaller number of studies reported diagnostic accuracy metrics (sensitivity, specificity, AUROC, and F1-score) against a reference standard.

4. Reported performance relative to human appraisal

Where reported, agreement values varied substantially and were generally poor to moderate: kappa values for risk-of-bias domains commonly fell in the 0.1-0.5 range, although higher agreement was reported for structured, item-level checklists such as AMSTAR-2 (kappa up to 0.60) and for fine-tuned models applied to custom methodological-quality tools (kappa up to 0.90). Of the 39 studies that compared AI with human appraisal, authors concluded that AI performance was inferior in 24 (62%), comparable in 14 (36%) and superior in 1 (3%) (Table 2).

A recurring finding was that newer or larger models did not consistently outperform smaller or older ones, and that agreement was highest for structured, rule-based checklist items and lowest for domains requiring contextual or interpretive judgement.

5. Efficiency and cost

Fourteen studies (33%) reported a time or cost outcome alongside their primary performance measure (study-level figures in Supplementary Table 8). Where reported, findings consistently favoured AI systems: reported completion times ranged from seconds to about 11 minutes per manuscript, compared with roughly 3 to 45 minutes reported for human reviewers performing the equivalent task (several human times were cited from earlier work or estimated rather than measured). Several studies also reported per-manuscript API costs, generally under $1–2. No study measured the human time needed to verify AI-generated outputs; one noted only that verification effort was required (Lim 2025).

Gaps in the evidence and its reporting

Commonly cited limitations included variability in model versions and prompting strategies across studies, absence of external validation datasets, small or single-specialty samples, and reliance on previously published human judgements (rather than live, blinded comparators) as reference standards (Table 3).

Discussion

Principal Findings

This scoping review identified 43 studies evaluating generative AI for the appraisal of biomedical research reports. Most compared AI output with human appraisal, and most of those authors concluded that AI performance fell short of human performance, with the gap smallest for structured, checklist-based items.

Three main findings warrant further discussion.

First, AI performance appeared to depend more on task structure than on model scale. Across risk of bias assessment, peer review, and methodological quality assessment alike, LLMs performed comparatively well on items with an explicit, verifiable answer and poorly on items requiring integrative or contextual judgement. This pattern was reported for both frontier proprietary and smaller open-source models, and fine-tuning narrowed but did not eliminate the gap for interpretive items.

Second, the finding that authors of 62% of comparative studies concluded AI performance was inferior requires careful interpretation. Human inter-rater reliability for the same instruments is itself frequently poor. Indeed, a cross-sectional evaluation of RoB 2 by four independent raters across 70 randomised trials reported only slight agreement for the overall risk-of-bias judgement (Fleiss' kappa 0.16, 95% CI 0.08-0.24), with fair-to-moderate agreement confined to more concrete domains such as the randomisation process.[17] Viewed against this benchmark, several of the AI-human kappa values identified in the present review are not obviously worse than the ceiling set by human-human

agreement on the same instrument. This does not establish that AI performance is adequate for independent use, but it reframes the comparator: the appropriate benchmark for AI-assisted appraisal is realistic human reliability, not a hypothetical perfectly consistent human reviewer, and few included studies made this comparison explicit.

Third, newer or larger models did not consistently outperform smaller or older ones, whereas task-specific fine-tuning and structured, checklist-anchored prompting produced more consistent gains than model choice alone. This suggests that near-term improvements in this domain are more likely to come from better task specification, training data, and structured output formats than from general-purpose model scaling.

Interpretation by Appraisal Domain

For risk of bias assessment, the largest single task category (n=16), LLMs applying Cochrane RoB tools showed consistently modest agreement with expert consensus, with several studies reporting near-chance performance on interpretive domains such as selective reporting and blinding, while structured, objectively verifiable domains showed comparatively higher agreement, mirroring the domain-level reliability pattern previously reported for human reviewers applying the same instrument.[17]

For peer review (n=12), AI-generated reviews were, in several studies, more generic, less critical, and less likely to recommend rejection than human reviewers, though several studies noted that AI feedback was faster and more actionable on specific dimensions. This pattern is consistent with a system that reliably detects checklist-style issues but struggles with more judgement-dependent problems, such as generalisability and novelty. Sycophancy, the tendency of language models trained on human preferences to agree with the position they are given, may also make AI-generated reviews less likely to recommend rejection.[18] None of the included studies tested this explanation.

For methodological quality assessment (n=8), performance was more variable and depended heavily on whether the framework was a simple, well-specified checklist or a broader, more subjective instrument; fine-tuning consistently improved performance over zero-shot prompting where this was

tested. Studies combining multiple tasks (e.g., screening, extraction, and appraisal in sequence) suggested that current LLMs may be more reliably deployed for full-text screening and structured data extraction than for judgement-heavy appraisal steps, a distinction with direct implications for where in the editorial or evidence-synthesis pipeline these tools are best positioned.

Comparison with Existing Literature

This review extends, rather than begins, a longer methodological tradition of automating research appraisal. Machine learning tools predating the current generation of LLMs, most notably RobotReviewer, achieved approximately 71% overall accuracy against published Cochrane risk-of-bias judgements, compared with 78.3% accuracy between two human reviewers, using supervised classifiers trained specifically for this task.[19] Roughly a decade later, general-purpose LLMs evaluated in the studies captured by this review do not appear to have decisively surpassed that earlier, narrower, purpose-built baseline for risk-of-bias classification, despite substantially greater general language capability. These findings suggest that the primary limitation may lie in the subjectivity of risk-of-bias judgements rather than model capability alone, an interpretation consistent with the human-human reliability data discussed above.[17]

Within general scientific publishing and computer science venues, a parallel and more mature evidence base has emerged. Liang et al. found that GPT-4-generated feedback overlapped with human reviewer comments by 31% for Nature-family journals and 39% for abstract submissions to the International Conference on Learning Representations (ICLR), closely comparable to human-human overlap in the same venues (28.58% and 35.25% respectively).[12] A field experiment comparing AI and human quality assessments of 305 conference abstracts similarly found only modest human-human and human-AI agreement, concluding that AI may be best suited to a prescreening rather than an adjudicating role.[20] Most directly relevant to editorial workflow design, a randomised controlled deployment of an LLM feedback system across more than 20,000 ICLR 2025 reviews found that 27% of reviewers who received AI-generated suggestions revised their reviews, producing feedback that was rated as more specific and actionable, without replacing the reviewers' own judgements.[13] Taken together, this literature and the present findings converge on a consistent picture across both computer

science and biomedical contexts: current LLMs are more reliably positioned as a feedback-augmentation or prescreening layer than as an independent adjudicator of research quality.

Reporting Standards and Governance Implications

This review identified substantial inconsistency in how included studies reported the AI systems, prompts, and parameters used. This gap is increasingly addressed by AI-specific reporting frameworks developed since 2024: TRIPOD-LLM, an extension of TRIPOD+AI, provides a 19-item checklist specifically designed for studies developing, tuning, or evaluating LLMs in healthcare, covering model versioning, prompt transparency, and human-oversight reporting.[21] TRIPOD-LLM was published in January 2025, so the evaluations in 12 of the 43 included studies (28%) predate it: 7 were published in 2024 and 5 reported evaluation dates before 2025. Adoption of this framework by future primary studies in this space is important to improve reporting quality and cross-study comparability.

A further threat to validity is training-data contamination. Of the 43 studies, 34 appraised already published articles, and only 5 reported steps to avoid material the models may have seen in training, such as unpublished reference standards, manuscripts published after a model's knowledge cut-off, or prompts that did not name the tool being applied; one more acknowledged the risk only as a limitation. Because many evaluations used published Cochrane reviews as the reference standard, reported agreement may overestimate performance on genuinely new submissions. Evaluation dates were reported in only 18 studies (42%) and dated snapshots in 5 (12%) (Supplementary Table 7), which limits the reproducibility of results for models that are updated without notice. This reporting gap sits alongside a separate governance question. Editorial decisions to adopt AI-assisted appraisal tools should be guided by evidence of validated performance before routine implementation. Journals and editorial bodies are increasingly defining policies on the use of AI in peer review itself. Examples include the International Committee of Medical Journal Editors' 2024 guidance restricting confidential manuscript content from third-party AI tools, and the Committee on Publication Ethics's 2023 position statement on AI and authorship.[22, 23]

Implications for Practice and Future Research

Priorities differ by task. For risk-of-bias assessment, the evidence supports using an LLM as a structured second reviewer for domains with an explicit, verifiable answer, with a human confirming every judgement, but not for unsupervised overall ratings; studies should report agreement by domain alongside human-human reliability for the same instrument. For peer review, the evidence supports LLM feedback as an adjunct for checking completeness and clarity, but not for making or recommending editorial decisions, because AI-generated reviews were more generic and less likely to recommend rejection; journals should evaluate any tool prospectively against editorial outcomes and apply their confidentiality policies. For methodological quality assessment, well-specified checklists such as AMSTAR-2 are the most promising target, and checklist-anchored or fine-tuned approaches should be compared with zero-shot prompting in the same sample. For workflows that combine several steps, screening and structured data extraction appear more tractable than appraisal, and each step should be validated separately, including the human time needed to verify the output. For all tasks, studies should record the model version, access date, prompts and parameters.

The evidence gaps in Table 3 set the priorities for future primary research: comparisons of validated frameworks with bespoke rubrics; studies of how model, prompt and parameter choices affect reproducibility; prospective evaluations with blinded, independently recruited human comparators, because most included studies appraised already-published manuscripts and few described blinding; and reporting by appraisal domain and item type rather than aggregate agreement alone.

Strengths

This review followed a pre-registered, JBI-consistent methodology with dual independent screening, page-referenced charting of every included study, and to our knowledge is the first to map evidence across risk-of-bias assessment, methodological quality assessment and peer review.

Limitations

This review has several limitations. First, the pace of the field means that newer models and studies will inevitably postdate the final search. Second, outcome measures are heterogeneous, precluding

formal quantitative synthesis. Third, many studies rely on previously published human judgements as reference standards rather than live, blinded comparators, which may not fully capture contemporary human performance. Fourth, the search used free-text terms in titles and abstracts and did not include preprint servers or grey literature, where evaluations of LLMs often appear first; a supplementary MEDLINE search adding recent model names identified no additional eligible studies. Finally, only English-language reports were included, and two full texts were excluded on this basis.

Conclusion

Current evidence supports LLMs as assistive tools for selected appraisal tasks, but does not support their independent use for research appraisal, including peer review. Performance is strongest for structured, rule-based appraisal items and weakest for tasks requiring contextual or interpretive judgement, a pattern that recurs across risk-of-bias assessment, peer review, and methodological quality assessments alike and is consistent with earlier machine-learning approaches to the same problem. Human reviewers themselves show only modest reliability on many of these same instruments; AI performance should therefore be benchmarked against realistic human agreement rather than an idealised standard. Reporting standards across the evidence base remain inconsistent, and prospective, adequately powered, blinded validation studies are needed before AI-assisted appraisal tools can be recommended for independent use in biomedical research.

Declarations

Ethical Approval and Consent to Participate

Not applicable. This study was a scoping review of published literature and did not involve human participants, human tissue, or identifiable personal data.

Consent for Publication

Not applicable.

Availability of Data and Materials

The protocol is registered on the Open Science Framework (https://osf.io/9muas). The charted dataset for all 43 included studies, the records from the sensitivity search, and the reference list are available on the Open Science Framework (https://osf.io/p73ry).

Competing Interests

BDM declares that he is Director of Strata Collective Ltd., which develops technology for the review and appraisal of biomedical research. YH, LY, and JD declare that they have no competing interests.

Funding

None to declare.

Authors' Contributions

BDM conceived the study, developed the protocol, performed data extraction, analysed the findings, and drafted the manuscript. YH and LY performed screening and data extraction. JD provided senior methodological supervision and contributed to interpretation of findings and manuscript revision. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. Bmj. 1996;312(7023):71-2.
2. Smith R. Peer review: a flawed process at the heart of science and journals. J R Soc Med. 2006;99(4):178-82.
3. Bastian H, Glasziou P, Chalmers I. Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLOS Medicine. 2010;7(9):e1000326.
4. Aczel B, Szaszi B, Holcombe AO. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. Res Integr Peer Rev. 2021;6(1):14.
5. Ware M MM. The STM Report: An overview of scientific and scholarly journal publishing2015 [cited 2026. Available from: http://digitalcommons.unl.edu/scholcom/9.
6. Bhatt DL, Mehta C. Adaptive Designs for Clinical Trials. N Engl J Med. 2016;375(1):65-74.
7. Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. Cochrane Database Syst Rev. 2007;2007(2):Mr000016.
8. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet. 2014;383(9913):267-76.
9. Rothwell PM, Martyn CN. Reproducibility of peer review in clinical neuroscience. Is agreement between reviewers any greater than would be expected by chance alone? Brain. 2000;123 (Pt 9):1964-9.
10. Armijo-Olivo S, Ospina M, da Costa BR, Egger M, Saltaji H, Fuentes J, et al. Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials. PLOS ONE. 2014;9(5):e96920.
11. Gates M, Gates A, Duarte G, Cary M, Becker M, Prediger B, et al. Quality and risk of bias appraisals of systematic reviews are inconsistent across reviewers and centers. Journal of Clinical Epidemiology. 2020;125:9-15.
12. Liang W, Zhang Y, Cao H, Wang B, Ding Daisy Y, Yang X, et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. NEJM AI. 2024;1(8):AIoa2400196.
13. Thakkar N, Yuksekgonul M, Silberg J, Garg A, Peng N, Sha F, et al. A large-scale randomized study of large language model feedback in peer review. Nature Machine Intelligence. 2026;8(3):326-36.
14. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18(10):2119-26.
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467-73.
16. Macfarlane BD. Large Language Models and Generative Artificial Intelligence for Critical Appraisal, Peer Review, and Research Quality Assessment in Biomedical Research Open Science Framework2026 [Available from: https://doi.org/10.17605/OSF.IO/9MUAS.
17. Minozzi S, Cinquini M, Gianola S, Gonzalez-Lorenzo M, Banzi R. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. J Clin Epidemiol. 2020;126:37-44.
18. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, et al. Towards understanding sycophancy in language models. International Conference on Learning Representations (ICLR)2024.
19. 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(1):193-201.
20. Shcherbiak A, Habibnia H, Böhm R, Fiedler S. Evaluating science: A comparison of human and AI reviewers. Judgment and Decision Making. 2024;19:e21.
21. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-9.

22. Editors ICoMJ. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals: Artificial Intelligence (AI)-Assisted Technology 2024 [cited 2026 9th July]. Available from: https://www.icmje.org.
23. (COPE) CoPE. Authorship and AI tools: COPE position statement 2023 [cited 2026 10th July]. Available from: https://publicationethics.org.

Table 1. Summary of appraisal tasks, frameworks, and study scale by domain.

Appraisal domain	Definition	Setting	N studies	Predominant framework	Median manuscripts appraised (range)
Risk of bias assessment	Judging whether the design, conduct or analysis of a study is likely to have biased its results, using a structured tool or domain-based criteria.	Evidence synthesis	16	Cochrane RoB n=12; JBI n=1; ROBUST-RCT n=1; no framework n=2	68 (9-7334)
Peer review	Evaluating a submitted manuscript to advise an editorial decision, producing comments, ratings or a recommendation.	Editorial review	12	Predominantly bespoke/journal-specific instruments (n=7); no single framework predominates	97 (8-763)
Methodological quality assessment	Rating how well a study or review was designed and reported against a checklist or rubric, which may extend beyond risk of bias.	Evidence synthesis	8	METRICS n=2; NOS n=1; AMSTAR-2 n=1; CASP n=1; no framework n=3	50 (3-200)
Multiple tasks (e.g., screening, extraction, appraisal)	Evaluations in which one AI workflow combined two or more steps, such as full-text screening,	Evidence synthesis	7	Cochrane RoB n=2; no framework n=2; ROBIS, AMSTAR-2, QUADAS-2 and PICO n=1 each (one	60 (40-112)

	data extraction and appraisal.			study applied 2 frameworks)	

Note: Research appraisal is used as an umbrella term for evaluating the validity, quality or reporting of a written research report. Reporting guideline assessment (checking a report against a standard such as CONSORT or PRISMA) was eligible but was not the primary task of any included study. Overall, 16 of 43 studies (37%) applied no named appraisal framework, instead using custom rubrics or unstructured prompts (see Results, Appraisal tasks).

Table 2. Primary outcome measures, human comparator and authors' conclusion in each included study, grouped by appraisal task.

Study	Primary outcome measure(s)	Human comparator	Authors' conclusion: AI vs human	N manuscripts appraised
Risk of bias assessment (n = 16)				
Di Pumpo et al. 2025	Cohen's kappa (IRA); mixed-effect ordinal logistic regression (OR)	Expert reviewers	Inferior to human	36
Vidor et al. 2026	Gwet's AC2 inter-rater reliability coefficient (ordinal weights, 4 levels, 95% confidence; probabilistic benchmarking)	Expert reviewers	Comparable	9
Xiong, Wang, Sun et al. 2026	Accuracy (3-category confusion matrix); weighted Cohen's kappa; Krippendorff's alpha	Published expert ratings	Comparable	158 RCTs
Li, Yang, Wu & Li 2026	Accuracy (3-class classification, standalone LLM); overall accuracy and intervention rate (Human-AI collaboration modes)	Published expert ratings	Comparable	2,100 domain-level samples (300 RCTs per domain)
Nyrhi, Ponkilainen, Laaksonen et al. 2026	Interobserver reliability (Cohen κ); intraobserver agreement; diagnostic accuracy (sensitivity, specificity, PPV, NPV, F1)	Published expert ratings	No comparative conclusion	200 RCTs
Wang et al. 2025	Macro-F1 (main task); Aspect Recall Ratio @ Optimal (SSR); Accuracy (SJS)	Expert reviewers	Inferior to human	700 papers
Xia et al. 2025	Correct assessment rate; Cohen's kappa; PABAK; sensitivity/specificity/F1	Published expert ratings	Comparable	30
Leucuța et al. 2025	Accuracy (correct answer + correct reasoning on signaling questions)	Published expert ratings	Inferior to human	10

Study	Primary outcome measure(s)	Human comparator	Authors' conclusion: AI vs human	N manuscripts appraised
Dias et al. 2025	AUROC; macro F1-score; precision; recall	Published expert ratings	Superior	7,334
Lai et al. 2024	Correct assessment rate (accuracy); sensitivity; specificity; F1 score; Cohen's kappa (consistency)	Expert reviewers	Comparable	30
Eisele-Metzger et al. 2025	Cohen's weighted kappa; observed percent agreement; Matthews correlation coefficient	Published expert ratings	Inferior to human	100
Kuitunen et al. 2024	Weighted Fleiss' kappa (overall and domain-specific)	Published expert ratings	Inferior to human	100
Kim et al. 2025	Weighted kappa (inter-rater); intraclass correlation coefficient (intra-rater)	Published expert ratings	Comparable	17
Šuster, Baldwin & Verspoor 2024	Agreement between LLM-predicted and published RoB judgement	Published expert ratings	Inferior to human	218 studies
Lettner et al. 2026	Inter-rater agreement and deviation from expert reviewer across 13 JBI domains	Expert + junior reviewers	Inferior to human	13 RCTs
Barsby et al. 2025	Concurrence/disagreement rate between ChatGPT and human gold-standard RoB judgement across 6 domains + overall	Expert reviewers	Inferior to human	15 papers
Peer review (n = 12)				
Tang et al. 2026	Observed agreement (Po); fuzzy Po (≤1 point difference); Cohen's kappa	Trained assessors	Inferior to human	180
Li et al. 2026	Category-level comment distribution; grant score comparison; citation prediction accuracy	Peer reviewers/editors	Inferior to human	763 preprints; 12 grant proposals; 100 published papers
Rajakumar et al. 2026	Domain scores (adapted Superchi et al. 24-item rubric, Likert 1–5); thematic analysis frequency; Shannon Diversity Index	Peer reviewers/editors	Inferior to human	119
Moshirfar, Han, Jaafar et al. 2026	Decision-frequency distribution (accept/minor/major/reject); favorability score (-2 to +2); 4-category quality score (0-3)	Expert reviewers	Inferior to human	300
Joachim et al. 2025	Weighted Cohen's kappa (AI vs human recommendation); intraclass correlation coefficient (temporal stability & intermodel)	Peer reviewers/editors	Inferior to human	22

Study	Primary outcome measure(s)	Human comparator	Authors' conclusion: AI vs human	N manuscripts appraised
Lim et al. 2025	Mean 5-point Likert scores for feedback/recommendation quality (relevance, completeness, accuracy, error-identification, constructiveness); evaluation time	Expert + junior reviewers	Comparable	8
Zhu et al. 2025	5-point Likert consistency/reasonableness rating (2 independent oncology experts); AI-detector probability (zeroGPT, GPTzero)	Expert reviewers	Inferior to human	20
Liu et al. 2026	Intraclass correlation coefficient (ICC); Bland-Altman mean difference	Expert reviewers	Comparable	160
Suleiman et al. 2024	Qualitative agreement category (complete match/partial match/no match/contradiction); inter-assessor kappa	Peer reviewers/editors	Inferior to human	20
Fernández-Cisnal et al. 2024	Quality score of AI-generated vs human review, and editor preference/ability to distinguish	Peer reviewers/editors	Inferior to human	75 letters
Sallam et al. 2024	Agreement between genAI and human editorial decision (send for review/reject)	Peer reviewers/editors	Inferior to human	34
Tomazini et al. 2026	Change rate in AI recommendation after hidden nudge injection (vs initial recommendation)	No	No comparative conclusion	180 (90 preprints; 90 published)
Methodological quality assessment (n = 8)				
Mese & Kocak 2026	Kendall's rank correlation (tau B); ICC (two-way random effects, absolute agreement)	Published expert ratings	Comparable	48
Shen et al. 2026	Exact match accuracy; loose match accuracy (within 1 quartile)	No	No comparative conclusion	200
Beber, Groff, Mange, Bram & Fabricant 2026	Intraclass correlation coefficient (ICC(2,1) and ICC(2,k)); percent agreement	Peer reviewers/editors	Inferior to human	27
Yao et al. 2025	Congruence (accuracy); sensitivity; specificity; F1 score	Expert reviewers	Comparable	114
Marques-Cruz et al. 2025	Accuracy; unweighted Cohen's kappa; F1-score; Concordance Correlation Coefficient	Expert + junior reviewers	Comparable	100 (test set)
Mese & Kocak 2025a	Intraclass correlation coefficient (ICC); Fleiss' kappa	Expert reviewers	Comparable	52

Study	Primary outcome measure(s)	Human comparator	Authors' conclusion: AI vs human	N manuscripts appraised
Mese & Kocak 2025b	Intraclass correlation coefficient (ICC); Kendall's tau-B; Fleiss'/Cohen's kappa (item-wise)	Expert reviewers	Inferior to human	48
Landerholm 2025	Krippendorff's alpha (AI interrater reliability); Cramer's V (model-pair association)	No	No comparative conclusion	3
Multiple tasks (n = 7)				
Emami & Shirani 2025	Kendall W concordance (inter-account agreement); GEE-based response correctness; weighted accuracy	Published expert ratings	Inferior to human	40
Rokhshad, Motie, Shirani, Sameie & Revilla-León 2026	Accuracy, sensitivity, specificity, precision, F1-score (screening); 0–5 ordinal scores (extraction); Cohen's kappa (RoB)	Expert + junior reviewers	Comparable	59 screened; 29 to 31 for extraction and risk of bias
Forero et al. 2025	Accuracy; Cohen's kappa; Krippendorff's alpha	Expert reviewers	Inferior to human	40 (20 ROBIS + 20 AMSTAR 2)
Rokhshad et al. 2025	Cohen's kappa (RoB); accuracy/sensitivity/specificity/F1 (screening); modified Global Quality Score (PICO/extraction)	Expert reviewers	Inferior to human	Not stated
Moussa et al. 2025	Intraclass correlation coefficient (ICC); Cohen's unweighted kappa	Expert reviewers	Comparable	61
Kim & Yoon 2025	Accuracy (RoB domain-level); recall/precision (abstract screening)	Published expert ratings	Inferior to human	81
Woelfle et al. 2024	Accuracy (agreement with human consensus); Cohen's kappa	Expert + junior reviewers	Inferior to human	112

Note: AI vs human conclusions are those stated by the study authors. Four studies (no human comparator, n = 3; Nyrhi 2026, which used published expert ratings but drew no comparison) reached no comparative conclusion. Task, framework, manuscript design and full sample-size detail for each study are in Supplementary Table 6. Full references for all 43 studies are in Supplementary Table 5 and the Supplementary Reference List; evaluation dates, publication years of the appraised articles and contamination safeguards are in Supplementary Table 7; time and cost outcomes are in Supplementary Table 8. *Mese & Kocak (2025b, 2026) evaluated the same cohort of 48 radiomics articles under distinct AI experiments and were retained as two separate studies.

Table 3. Evidence gaps and implications for future research.

Evidence gap	Observation from included studies	Implication for future research
Preregistration and reproducibility	Only 3/43 studies (7%) were prospectively pre-registered. A dated model snapshot was reported in 5/43 (12%), an evaluation date in 18/43 (42%), and both a named version and an evaluation date in 11/43 (26%); prompt text and temperature were frequently unreported.	Future studies should be prospectively pre-registered and report full prompts, model versions, and parameter settings to support reproducibility.
Appraisal frameworks	16/43 studies (37%) used no named appraisal framework, relying on custom or unstructured rubrics not comparable across studies.	Anchoring future evaluations to established frameworks (e.g., Cochrane RoB2, AMSTAR-2) would improve comparability across the evidence base.
Reference standard design	Most studies compared AI output to previously published human judgements rather than a live, blinded, concurrent human comparator. Blinding of assessors was reported in 7/43 studies (16%) and partly in a further 4 (9%).	Prospective, blinded, and concurrently generated human comparators would strengthen the validity of future AI-human comparisons.
Sample size and specialty breadth	Across 42 studies with extractable sample-size data, the median evaluated was 60 manuscripts each (IQR 28-160; range 3-7,334), and roughly half were limited to a single clinical specialty.	Larger, a priori powered samples spanning multiple specialties and manuscript types would improve generalisability of future findings.
Domain-specific performance gaps	AI-human agreement was consistently higher for structured, rule-based checklist items and lower for interpretive domains (e.g., blinding, selective reporting, novelty).	Reporting outcomes stratified by item type (rule-based vs interpretive), rather than a single pooled agreement statistic, would better characterise where AI performance is and is not adequate.
Model comparison practices	Newer or larger models did not consistently outperform smaller/older ones; findings are highly prompt- and version-dependent.	Comparisons across multiple models and versions under standardised, transparently reported prompting conditions would help clarify whether performance differences

		reflect genuine model capability or prompting artefacts.
Manuscript provenance and training-data exposure	79% of studies tested AI against already-published manuscripts, which may already be present in a model's training data; only 5/43 studies (12%) reported steps to avoid such exposure.	Using genuinely unpublished, blinded submissions in future evaluations would help avoid potential memorisation or exposure confounds.
Efficiency and verification	14/43 studies (33%) reported a time or cost outcome; none timed the human effort needed to verify AI output.	Future studies should report AI time, human time and verification time for the same task, so that claimed time savings can be checked.

Figure 1. PRISMA-ScR flow diagram.

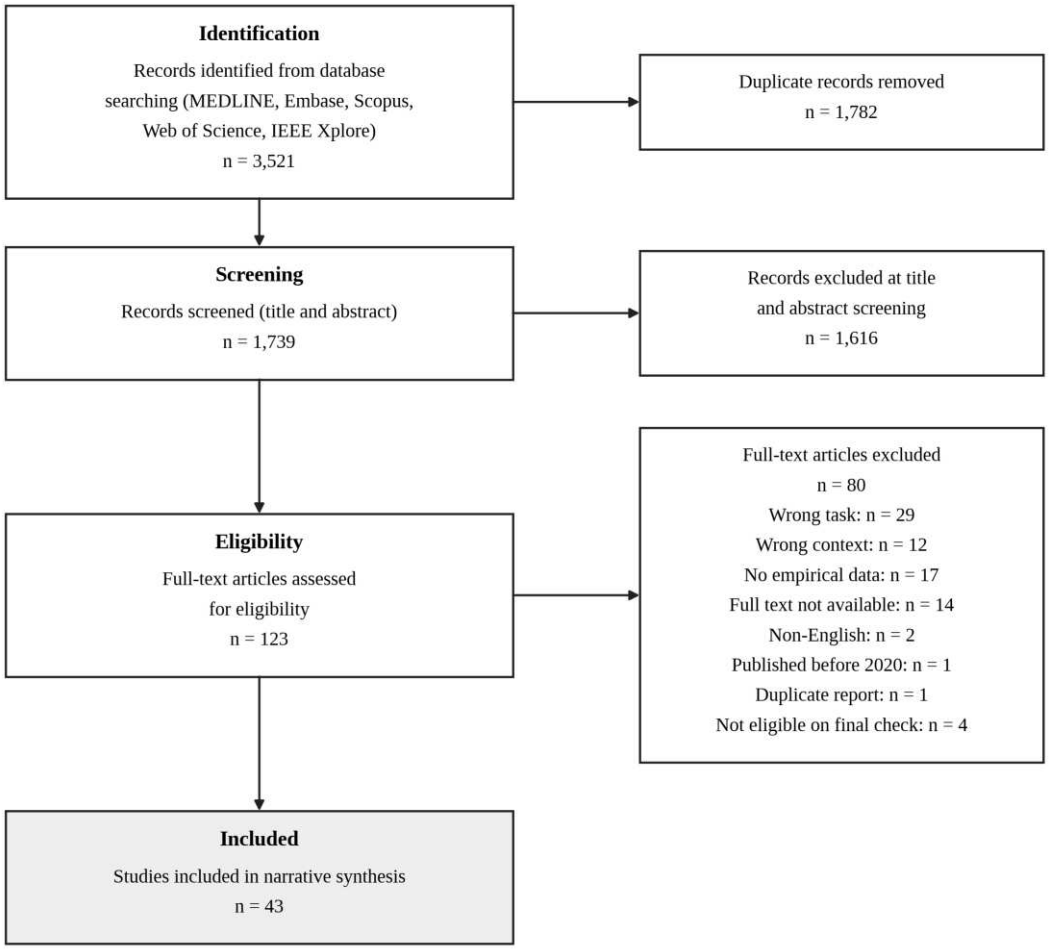


Figure 1. PRISMA-ScR flow diagram of study selection. MEDLINE, Embase, Scopus, Web of Science and IEEE Xplore were searched on 11 June 2026 for records published from 1 January 2020, yielding 3,521 records. After removal of 1,782 duplicates, 1,739 records were screened by title and abstract and 1,616 were excluded. Of 123 full-text articles assessed for eligibility, 80 were excluded for the reasons shown, leaving 43 studies included in the narrative synthesis.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial.docx