

High Average Accuracy Coexists with Substantial Between-Setting Variability: A Systematic Review and Meta-analysis of Large Language Models for Data Extraction in Evidence Synthesis

Carlos Darcy Alves Bersot

carlosbersot@gmail.com

Federal University of Sao Paulo <https://orcid.org/0000-0002-0464-7990>

Systematic Review

Keywords: Large language models, Data extraction, Systematic review, Meta-analysis, Evidence synthesis, Artificial intelligence, Automated data extraction, Human verification

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11332938/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

High Average Accuracy Coexists with Substantial Between-Setting Variability: A Systematic Review and Meta-analysis of Large Language Models for Data Extraction in Evidence Synthesis

Carlos Darcy Alves Bersot^{1*}; Lucas Ferreira Gomes Pereira²; Vitor Alves Felipe¹; Guilherme Queiroz Bersot³; Rebeca Martinelli Galvão³; Gustavo Fernandes Nunes¹; Bruno Martins Santiago⁴; José Eduardo Guimarães Pereira⁵; Luiz Fernando dos Reis Falcão⁶

Affiliations

1. Postgraduate Program in Translational Medicine, Paulista School of Medicine, Federal University of São Paulo (EPM-UNIFESP), São Paulo, Brazil.
2. Discipline of Anesthesiology, Hospital das Clínicas, Faculty of Medicine, University of São Paulo (FMUSP), São Paulo, Brazil.
3. Faculty of Medicine of Campos, Benedito Pereira Nunes Foundation, Campos dos Goytacazes, Rio de Janeiro, Brazil.
4. Department of Anesthesiology, Rio de Janeiro State University (UERJ), Rio de Janeiro, Brazil.
5. Department of Anesthesiology, Hospital Unimed Volta Redonda, Volta Redonda, Rio de Janeiro, Brazil.
6. Department of Anesthesiology, Pain and Critical Care Medicine, Paulista School of Medicine, Federal University of São Paulo (EPM-UNIFESP), São Paulo, Brazil.

*Corresponding author: Carlos Darcy Alves Bersot

Email: carlos.bersot@unifesp.br

Abstract

Large language models (LLMs) are increasingly used for data extraction in systematic reviews, but reported performance varies across extraction tasks, models, domains, reference standards, and definitions of correctness.

We systematically reviewed empirical studies of automated or AI-assisted data extraction for evidence synthesis and performed an estimand-driven meta-analysis of uncorrected LLM extraction against human or otherwise validated reference standards. Forty-six reports representing 45 independent study units were included in the systematic review. Eight independent studies contributed 10 model-level estimates and 29,643 item-level evaluations to the broad primary quantitative synthesis. Proportions were analyzed on the logit scale using a multilevel random-effects model with model-level estimates nested within study and cluster-robust inference. Pooled item-level accuracy was 86.4% (cluster-robust 95% CI, 62.6%–96.0%), with generalized I^2 of 99.6% and a model-based 95% prediction interval of 19.2%–99.4%, indicating marked between-setting variability. In a construct-restricted sensitivity analysis excluding an open-ended, recall-intensive extraction setting, seven studies contributed nine model-level estimates and 29,343 item-level evaluations; pooled accuracy was 91.2% (95% CI, 83.1%–95.6%) and the prediction interval narrowed to 66.7%–98.2%. Leave-one-study-out analyses showed that omission of individual studies other than this methodologically distinct setting had comparatively limited influence on the pooled mean, whereas heterogeneity remained substantial. Across the broader evidence base, performance varied according to extraction task, data type, workflow design, and definition of correctness. These findings indicate that average LLM extraction accuracy can be high, but performance cannot be assumed to generalize across evidence-synthesis settings. Task-specific validation and human verification remain important, particularly for ambiguous, missing, derived, numerical, or otherwise consequential information.

Keywords: Large language models; Data extraction; Systematic review; Meta-analysis; Evidence synthesis; Artificial intelligence; Automated data extraction; Human verification.

1. Background

Systematic reviews depend on accurate extraction of study characteristics, intervention and comparator information, and outcome data. Independent duplicate extraction, or extraction by one reviewer followed by independent verification, is recommended to reduce transcription, interpretation, and omission errors, particularly for information used in quantitative synthesis [1,2]. This safeguard is methodologically important because extraction errors may propagate into effect estimates, assessments of methodological limitations, certainty of evidence, and ultimately the conclusions of a review. However, duplicate extraction is resource-intensive and can account for a substantial proportion of reviewer time.

Efforts to automate data extraction in evidence synthesis predate generative artificial intelligence. Earlier approaches relied on rule-based text processing, conventional natural language processing, and task-specific machine-learning systems to identify and extract structured information from scientific reports [3,4]. More recently, large language models (LLMs) have expanded the possibilities for automation by enabling the extraction of heterogeneous information without requiring task-specific model training. Empirical studies have reported encouraging performance for LLM-based or LLM-assisted extraction, including high accuracy for selected structured fields and reductions in reviewer workload [5–8]. Nevertheless, performance has varied considerably across models, extraction tasks, document formats, scientific domains, prompting strategies, reference standards, and definitions of correctness.

The growing evidence base has prompted systematic reviews of LLM applications in evidence synthesis. Shankar et al. reviewed LLM-based data extraction and identified substantial methodological and performance heterogeneity, including lower accuracy for some numerical variables, omission as an important failure mode, and potential reductions in extraction time [9]. A broader systematic review by Laignelot et al. evaluated LLM performance across multiple stages of systematic-review production and similarly found greater variability for complex tasks, including detailed data extraction and methodological assessment, than for simpler screening tasks [10]. These reviews highlight both the potential of LLM-assisted evidence synthesis and the challenges associated with heterogeneous performance measures and evaluation methods. However, a narrower quantitative question remains: when studies are restricted to a sufficiently comparable item-level correctness construct, what average extraction accuracy is observed, and how much does that accuracy vary across extraction settings?

This distinction is important because a pooled accuracy estimate describes average performance among contributing studies but does not establish that comparable performance should be expected in another setting. Heterogeneity statistics characterize variability beyond sampling error, whereas prediction intervals provide a model-based description of the range of underlying effects that may occur across comparable settings [11,12]. High pooled accuracy may therefore coexist with substantial between-setting variability and a wide prediction interval, limiting the usefulness of the pooled mean as a general performance benchmark for automated extraction.

We therefore conducted a systematic review and an estimand-driven quantitative synthesis designed to complement previous descriptive reviews. The primary objective was to estimate uncorrected item-level LLM extraction accuracy relative to a human, adjudicated, verified, or otherwise validated reference standard, using directly reported or arithmetically reconstructable counts of correctly extracted and total evaluated items. Performance measures such as agreement, precision, recall, sensitivity, specificity, F1 score, semantic similarity, concordance, and accuracy after human correction were treated as distinct constructs rather than interchangeable measures of extraction correctness. The primary meta-analysis included eligible item-level correctness estimates, while sensitivity analyses examined the influence of extraction-task comparability on pooled accuracy and between-setting variability.

Secondary objectives were to characterize quantitative performance outside the primary estimand, particularly numerical-data extraction; describe omission, fabrication, and other reported error patterns; examine reproducibility, processing time, reviewer workload, and human–AI workflows; and explore methodological characteristics potentially contributing to variation across extraction settings. The overarching objective was not to establish a universal accuracy benchmark for LLMs, but to determine whether high average extraction accuracy is accompanied by sufficient consistency across settings to support generalization across evidence-synthesis workflows.

2. Methods

2.1. Study Design, Eligibility Criteria, and Search Strategy

This systematic review and meta-analysis evaluated the performance of automated and artificial intelligence-assisted data extraction from scientific reports within evidence-synthesis workflows. Reporting followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [13], with methodological procedures informed by the Cochrane Handbook for Systematic Reviews of Interventions [1,2]. The review was not prospectively registered. A retrospective registration was subsequently completed through the Open Science Framework (OSF) on October 6, 2026 (DOI: 10.17605/OSF.IO/5RGCN), after the systematic review process and part of the quantitative analyses had already been conducted.

Eligible studies were empirical investigations evaluating automated or artificial intelligence-assisted extraction of information from scientific reports in systematic reviews, meta-analyses, scoping reviews, living reviews, health technology assessments, guideline development, or closely related evidence-synthesis activities. Eligible technologies included artificial intelligence, machine learning, natural language processing, text mining, generative artificial intelligence, and large language models (LLMs).

Studies were required to evaluate extracted information against a human-generated, consensus-based, adjudicated, verified, curated, or otherwise validated reference standard and to report at least one relevant performance or implementation outcome. Eligible outcomes included extraction accuracy or correctness, agreement, precision, recall or sensitivity, specificity, F1 score, extraction errors, omissions, hallucinations or fabrications, unsupported inferences, reproducibility, processing time, reviewer workload, and cost.

Studies exclusively addressing literature screening without data extraction were excluded, as were conceptual or methodological articles without empirical extraction results, protocols, editorials, commentaries, narrative reviews, and applications of artificial intelligence unrelated to evidence-synthesis data extraction.

The systematic literature search was conducted in four bibliographic databases: PubMed/MEDLINE, Embase, Scopus, and the Cochrane Library. Search strategies combined terminology related to systematic reviews and evidence synthesis, artificial intelligence and large language models, and automated data extraction. Controlled vocabulary, where available, was combined with free-text terms, and search syntax was adapted to each database. No minimum extraction-performance threshold was applied.

The archived database exports contained 10,638 records, comprising 1,535 from PubMed/MEDLINE, 2,054 from Embase, 1,644 from Scopus, and 5,405 from the Cochrane Library. Reconciliation and deduplication of the archived records identified 1,978 duplicate records, leaving 8,660 unique records for title and abstract screening. Of these, 8,612 were excluded, leaving 48 database-derived reports for retrieval and full-text eligibility assessment. All 48 reports were retrieved.

One additional eligible report, Felizardo et al. [22], was identified through a supplementary manual search of Google Scholar.

Preprints identified during the literature search were considered eligible when they satisfied the review criteria. No independent search of dedicated preprint repositories was conducted. When a preprint and a subsequent peer-reviewed publication represented the same underlying investigation, the final published version was preferentially retained to avoid duplicate representation of the evidence.

Overall, 49 reports underwent full-text eligibility assessment. Three reports were excluded because they did not meet the eligibility criteria, leaving 46 reports included in the systematic review. The study-identification and selection process is presented in the PRISMA 2020 flow diagram (Figure 1).

2.2. Study Selection, Report Mapping, and Data Extraction

Study selection was conducted in two stages. Four reviewers participated in the initial screening of titles and abstracts retrieved through the bibliographic database searches. Reports considered potentially eligible subsequently underwent full-text assessment by two reviewers. Eligibility disagreements were resolved through discussion and consensus.

Eligibility was assessed at both the report level and the underlying study level to avoid duplicate representation of evidence. When multiple publications described the same investigation, including preprint and peer-reviewed versions, these reports were linked rather than treated as independent studies. Reports evaluating substantially overlapping source-document sets were also assessed for evidentiary dependence. Following report mapping, the 46 included reports represented 45 independent study units.

Eligible reports were classified according to their contribution to the present review as primary quantitative, secondary quantitative, or narrative/descriptive evidence. These categories reflected their role in the synthesis rather than the original methodological design of each investigation.

Data were extracted independently by two investigators using a structured extraction framework. Disagreements were resolved through discussion and consensus. Extracted information included publication year, scientific domain, evidence-synthesis setting, number and characteristics of source documents, artificial intelligence technology, exact model and version, access method, document input and preprocessing procedures, prompting strategy, extraction task, and variable type.

Additional variables included workflow configuration, reference-standard characteristics, construction and independence of the reference standard, unit of evaluation, number of evaluated extraction items, reported performance measures, error categories, reproducibility, processing and verification time, reviewer workload, cost, and verification or adjudication procedures.

Performance values directly reported by the original investigators were distinguished from values reconstructed from published counts, percentages, tables, figures, or supplementary materials. Reconstructed values were accepted only when the relevant numerator and denominator could be established through reproducible arithmetic without converting one performance construct into another.

When multiple LLMs were evaluated using the same source-document set, eligible model-specific estimates were retained separately while preserving their common study identifier. This approach allowed the quantitative analysis to account for dependence among estimates originating from the same investigation.

2.3. Outcomes and Eligibility for Quantitative Synthesis

The primary outcome was uncorrected item-level extraction accuracy, defined as the proportion of evaluated extraction items correctly generated by an LLM before human correction or adjudication, relative to a human, adjudicated, verified, curated, or otherwise validated reference standard.

An extraction item was operationally defined as the smallest study-defined unit of information explicitly evaluated against the reference standard. Such units included individual fields, cells, variables, and outcome data. Because the granularity of extraction items varied across investigations, no artificial uniform item definition was imposed beyond the requirement for explicit item-level evaluation.

Eligibility for the primary meta-analysis required evaluation of uncorrected LLM output, use of an appropriate reference standard, and directly reported or arithmetically reconstructable counts of correctly extracted and total evaluated items. The extraction task also had to provide an interpretable measure of item-level correctness.

Agreement, precision, recall or sensitivity, specificity, F1 score, semantic similarity, correlation, concordance, and accuracy measured after human correction or adjudication were not converted into extraction accuracy because these measures represent distinct performance constructs.

Studies evaluating extraction restricted to specific data classes, particularly numerical outcomes, were retained as secondary quantitative evidence when inclusion in the primary synthesis would alter the target estimand. Similarly, collaborative multi-LLM systems, agentic workflows, multistage pipelines, and human-verified extraction procedures were not assumed to estimate the same construct as uncorrected standalone LLM extraction.

Potential overlap in source-document sets was considered when establishing evidentiary independence, preventing substantially overlapping evaluations from being treated as independent study clusters.

The principal construct-restricted sensitivity analysis examined the influence of an open-ended, recall-intensive extraction setting that differed materially from the more structured item-level extraction tasks represented elsewhere in the primary pool. Additional sensitivity analyses examined studies with methodological concerns related to iterative prompt optimization and alternative inclusion decisions.

Secondary outcomes included numerical-data extraction performance, omission and missed-data errors, hallucination or fabrication, unsupported inference, consequential extraction errors, reproducibility, processing and verification time, reviewer workload, cost, and the performance of human–AI workflows. These outcomes were synthesized descriptively when differences in

definitions, evaluation procedures, or workflow configurations precluded meaningful quantitative pooling.

2.4. Methodological Quality, Risk of Bias, and Certainty of Evidence

Methodological limitations and risk of bias were assessed using a framework informed by QUADAS-2 and adapted to artificial intelligence-based data-extraction validation research [14]. Because QUADAS-2 was developed for diagnostic-accuracy studies, its use represented a structured methodological adaptation rather than application of a formally validated instrument specifically designed for LLM extraction research.

The assessment considered four domains: selection and representativeness of source documents; specification and conduct of the artificial intelligence extraction procedure; independence and quality of the reference standard; and flow and analysis.

Assessment of the extraction procedure considered model identification and version, prompting strategy, document preprocessing, reproducibility, and potential model access to reference information. Assessment of the reference standard addressed its construction, independence, verification, adjudication procedures, and potential modification after comparison with model output.

Flow and analysis were evaluated with particular attention to completeness of source-document and item-level assessment, post-extraction exclusions, treatment of missing information, operational definitions of correctness, repeated evaluations, and statistical dependence among observations.

Additional methodological considerations included availability of prompts, reporting of generation settings, stability of proprietary model versions, separation between development and evaluation data, and correspondence between experimental extraction tasks and practical evidence-synthesis workflows.

Domain-level judgments were categorized as low risk, some concerns or unclear risk, and high risk. Overall interpretation was based on the pattern of domain assessments rather than a numerical quality score. Disagreements were resolved through consensus.

Confidence in the primary quantitative conclusion was additionally considered using a GRADE-informed framework incorporating the conceptual domains of risk of bias, inconsistency, indirectness, imprecision, and publication bias [15]. This framework was used to support structured interpretation of certainty without implying formal validation of conventional GRADE methodology for LLM-based data-extraction accuracy.

2.5. Statistical Analysis

The primary meta-analysis included eight independent studies contributing 10 model-level estimates and 29,643 item-level evaluations.

For each eligible estimate, extraction accuracy was calculated as the number of correctly extracted items divided by the total number of evaluated items. Proportions were transformed to the logit scale for statistical analysis and subsequently back-transformed to the proportion scale for presentation.

A multilevel random-effects meta-analysis was fitted using restricted maximum likelihood estimation. Model-level estimates were nested within their corresponding studies to accommodate dependence among multiple estimates originating from the same investigation. Separate variance components were estimated for between-study heterogeneity and additional model-level heterogeneity within studies.

Cluster-robust variance estimation was used to account for dependence among estimates originating from the same study. Inference used the CR2 small-sample correction with Satterthwaite-adjusted degrees of freedom and t-based confidence intervals [16,18].

The independent study, rather than the individual extraction item, was treated as the primary evidentiary unit. Large item-level denominators contributed to the precision of study-specific estimates but were not interpreted as independent study replications.

Heterogeneity was summarized using a generalized I^2 statistic representing the proportion of observed variability attributable to heterogeneity beyond sampling error. A model-based 95% prediction interval was calculated on the logit scale using the fitted heterogeneity structure and subsequently back-transformed to the proportion scale [11,12]. Prediction intervals were interpreted as model-dependent descriptions of variability in underlying extraction accuracy across settings rather than precise forecasts for individual models or future evidence-synthesis workflows.

Sensitivity analyses examined the robustness of the primary findings to alternative decisions regarding extraction-task comparability and study inclusion. The principal construct-restricted analysis excluded the open-ended, recall-intensive extraction setting while retaining the remaining eligible studies. Additional analyses excluded the study in which iterative prompt optimization on the evaluation dataset raised concern regarding overfitting, excluded both methodologically distinctive studies simultaneously, and replaced the final optimized-prompt estimate with the corresponding initial-prompt result.

Sequential leave-one-study-out analyses were performed across all eight independent study clusters. Changes in pooled accuracy, variance components, generalized I^2 , and prediction intervals were examined to assess the influence of individual studies.

Meta-regression was not performed because the limited number of independent studies was considered insufficient for reliable moderator analysis. Funnel-plot asymmetry and regression-based small-study-effect tests were likewise not performed because their interpretation would be unreliable with only eight independent study clusters.

All statistical analyses were conducted using R version 4.6.1 [19] within the Posit/RStudio environment. Multilevel random-effects models were fitted using the **metafor** package (version 5.2-1) [17], and cluster-robust inference was implemented using **clubSandwich** (version 0.7.0).

The analytical code, frozen study-level inputs, quantitative eligibility decisions, transformations, model specifications, and sensitivity analyses were retained in the reproducibility archive.

3. Results

3.1. Study Selection and Characteristics of the Evidence Base

The 46 included reports covered diverse evidence-synthesis settings, including systematic reviews, meta-analyses, living evidence syntheses, guideline-development workflows, scoping reviews, and related review contexts. Evaluated technologies ranged from conventional machine-learning and natural language processing approaches to generative LLMs, including GPT-family models, Claude, Gemini, Moonshot, purpose-built extraction systems, and multimodel or agentic workflows. Considerable methodological variation was observed in document preprocessing, prompting strategies, model-access methods, extraction schemas, degree of human involvement, and intended applications of automation.

Reference standards and performance measures also varied across the included studies. Automated extractions were compared with single-reviewer extraction, independently verified human extraction, duplicate human extraction, consensus or adjudicated reference datasets, previously curated systematic-review datasets, and other validated reference standards. Reported outcomes included item-level accuracy, agreement, precision, recall or sensitivity, specificity, F1 score, semantic equivalence, extraction errors, reproducibility, processing time, reviewer workload, and performance following human verification. These differences supported distinguishing eligibility for the systematic review from eligibility for the primary quantitative synthesis.

The study-selection process is presented in the PRISMA 2020 flow diagram (Figure 1). Characteristics of the 46 included reports, including the evaluated AI systems, extraction settings, principal

performance measures, synthesis classifications, and relationships among overlapping reports, are summarized in Table 1.

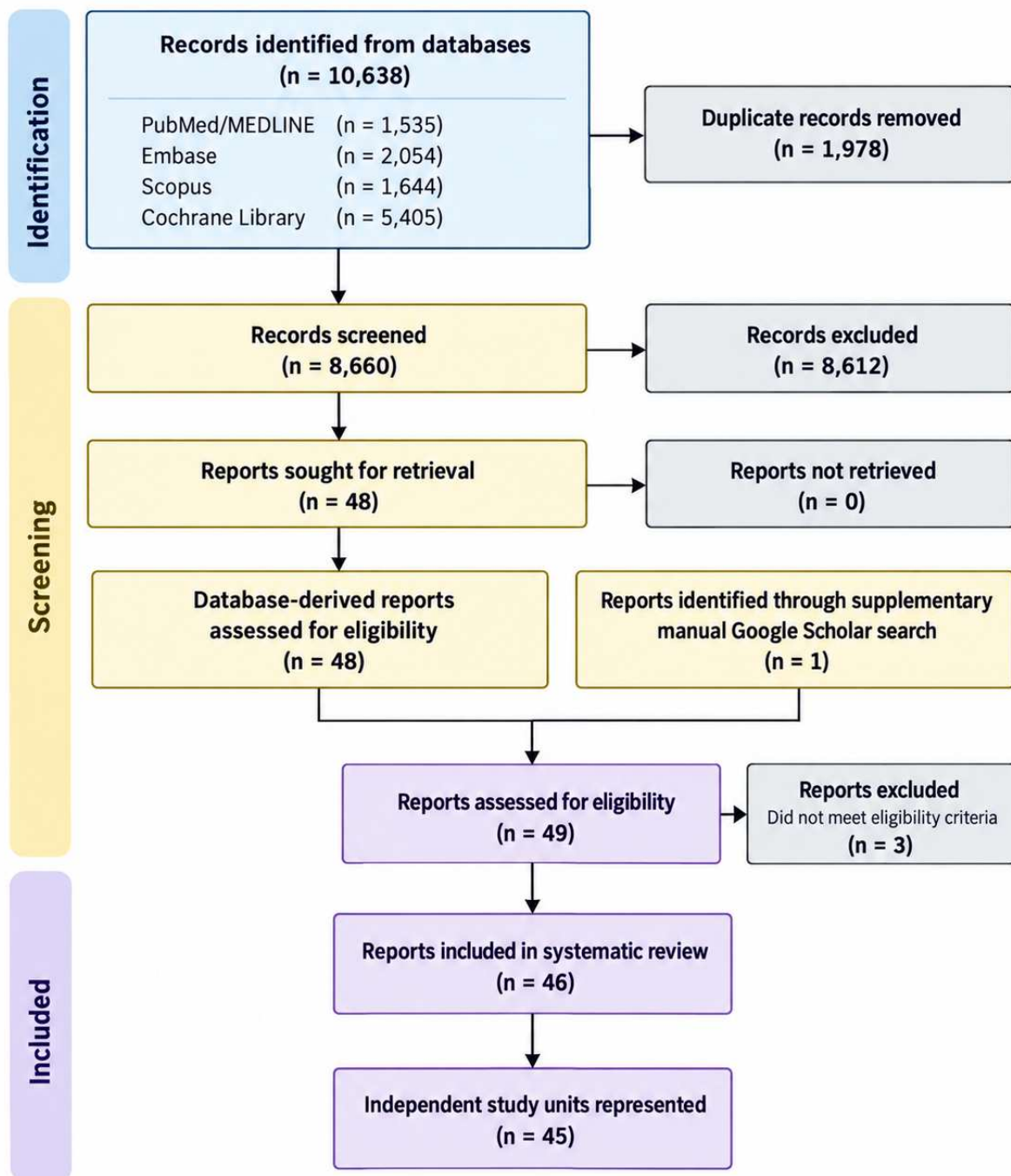


Figure 1. PRISMA 2020 flow diagram of study identification, deduplication, screening, full-text assessment, and inclusion in the systematic review.

Table 1. Characteristics and synthesis classification of the 46 reports included in the systematic review.

Study/report	AI system/technology	Main extraction setting	Main performance construct	Synthesis classification
Gartlehner et al., 2024 [5]	Claude 2	Evidence-synthesis data extraction	Item-level accuracy	Primary quantitative
Liu et al., 2025 [20]	Claude 2	RCT data extraction	Item-level accuracy	Primary quantitative
Motzfeldt Jensen et al., 2025 [6]	ChatGPT-4o	Second-rater extraction	Item-level accuracy	Primary quantitative
Vivekanantha et al., 2026 [21]	GPT-5.2; Gemini 3 Pro	Evidence extraction	Item-level accuracy	Primary quantitative
Lai et al., 2025 [8]	Claude 3.5 Sonnet; Moonshot-v1-128k	LLM-only / LLM-assisted extraction	Accuracy	Primary quantitative

Felizardo et al., 2024 [22]	ChatGPT-4o	Systematic-mapping data extraction	Accuracy	Primary quantitative
Mahmoudi et al., 2025 [23]	GPT-4 / ChatGPT	Systematic-review data extraction	Item-level accuracy	Primary quantitative
Stewart-Evans et al., 2025 [24]	Claude 3.5 Sonnet	Open-ended evidence extraction	Item-level correctness / recall-intensive extraction	Primary quantitative
Caponio et al., 2026 [25]	Claude 3.5 Sonnet; DeepSeek; ChatGPT-4o; Gemini	Numerical data extraction	Numerical accuracy	Secondary quantitative
Peng et al., 2025 [26]	Claude 3.5; ChatGPT-4o	Numerical RCT data extraction	Numerical accuracy	Secondary quantitative
Kataoka et al., 2026 [27]	GPT-4o; o3	Multi-dataset systematic-review extraction	Accuracy by dataset / variable class	Secondary quantitative
Oami et al., 2026 [28]	ChatGPT-4o; Claude 3 Sonnet; Gemini 1.5 Pro	Guideline/SR data extraction	Cell-level no-error proportion	Secondary quantitative
Zhang et al., 2026 [29]	Multiple GPT-based systems	Meta-analysis data extraction	Correct extraction / error categories	Secondary quantitative

Wang et al. (TrialMind), 2025 [30]	TrialMind multi-stage pipeline	LLM	Systematic-review workflow	AI-assisted extraction workflow performance	Secondary / quantitative
Rokhshad et al., 2026 [31]	LLM-based extraction		Evidence-synthesis extraction	Study-specific extraction-quality score	Narrative/descriptive
Daraqel et al., 2025 [32]	AI-assisted extraction		Evidence-synthesis data extraction	Agreement	Narrative/descriptive
Kartchner et al., 2023 [33]	Generative language-model workflow		Meta-analysis data extraction	Accuracy; ROUGE; fuzzy matching	Narrative/descriptive
Jansen et al., 2025 [34]	Eight generative AI/LLM systems		Systematic-review databases	Multiple accuracy measures	Narrative/descriptive
Schroeder et al., 2025 [35]	Multiple LLMs		Systematic-review data extraction	Agreement; match-based measures	Narrative/descriptive
Susnjak et al., 2025 [36]	Fine-tuned / LLM workflow	LLM	Evidence-synthesis automation	Fidelity; workflow performance	Narrative/descriptive
Barros et al., 2026 [37]	LLM-assisted workflow		Systematic mapping	Agreement; workflow/time	Narrative/descriptive

Gartlehner et al., 2025 [7]	Claude 2.1/3.0/3.5 + human verification	Prospective study-within-review workflow	Accuracy; sensitivity; PPV; time	Narrative/descriptive
Bianchi et al., 2025 [38]	Elicit	Evidence extraction	Extraction equivalence	Narrative/descriptive
Helms Andersen et al., 2025 [39]	Elicit; ChatGPT	AI as second reviewer	Recall; performance; confabulation	Narrative/descriptive
Murton et al., 2025 [40]	LLM with prompt engineering	Study and outcome extraction	Precision; recall; F1	Narrative/descriptive
Fan et al., 2026 [41]	Claude 3.5 Sonnet	Large-scale RCT extraction	Extraction error rate	Narrative/descriptive
Kim et al., 2025 [42]	GPT-4	AI-assisted systematic-review workflow	Concordance	Narrative/descriptive
Lamanna et al., 2025 [43]	ChatGPT-4	Meta-analysis data extraction	Accuracy	Narrative/descriptive
Kasireddy et al., 2026 [44]	Claude 3.7 Sonnet	Structured data extraction	Precision; recall; F1	Narrative/descriptive
Yerramilli et al., 2026 [45]	Six LLMs	Semiautomated extraction	Concordance	Narrative/descriptive

Konet et al., 2024 [46]	Claude 2; GPT-4		Evidence extraction	Accuracy		Narrative/descriptive ; related report to Gartlehner 2024
Khraisha et al., 2024 [47]	GPT-4		Multilingual screening and extraction	Extraction performance		Narrative/descriptive
Achterberg et al., 2026 [48]	OpenExtract / LLM pipeline		Automated extraction	Precision; recall		Narrative/descriptive
Muthu et al., 2026 [49]	Agentic framework	AI	Systematic-review workflow	Agreement / κ		Narrative/descriptive
Ye et al., 2024 [50]	Hybrid semiautomated workflow		Systematic-review automation	Workflow extraction performance	/	Narrative/descriptive
Sanlier et al., 2026 [51]	Multi-LLM consensus		Complexity-stratified extraction	Semantic-equivalence accuracy		Narrative/descriptive
Mitchell et al., 2025 [52]	AI-LES		Living evidence synthesis	Accuracy; reliability		Narrative/descriptive
Ishikane et al., 2026 [53]	LLM-assisted workflow		Randomized crossover extraction	Accuracy; time		Narrative/descriptive
Jansen et al., 2026 [54]	Gemini 2.5 Pro; GPT-4.1; o3		Meta-analysis extraction	ICC; agreement		Narrative/descriptive

Gates et al., 2021 [3]	Machine learning / text mining	Semi-automated extraction	Accuracy	Narrative/descriptive
Yisha et al., 2026 [55]	ChatGPT-4; Claude 3 Opus	Numerical data extraction	Accuracy; reproducibility	Narrative/descriptive
Panayi et al., 2023 [4]	BERT + CRF	Machine-learning extraction	Precision; recall; F1	Narrative/descriptive
Li et al., 2026 [56]	Gemini-2.0-flash; Grok-3; GPT-4o-mini	Meta-analysis extraction benchmark	Precision; recall; F1	Narrative/descriptive
Meliante et al., 2025 [57]	Multiple AI-assisted review tools	Systematic-review workflow	Content/extraction performance	Narrative/descriptive
Wang et al. (SciDaSynth), 2025 [58]	SciDaSynth GPT-based extraction	+ Interactive structured data extraction	Ordinal extraction-quality score; time	Narrative/descriptive
Khan et al., 2025 [59]	GPT-4-turbo; Claude-3-Opus	Collaborative multi-LLM extraction	Concordance; accuracy	Narrative/descriptive

Note: Synthesis classification reflects the role of each report in the present review rather than whether the original investigation itself used quantitative or qualitative methods. Primary quantitative studies contributed to the broad item-level accuracy meta-analysis; secondary quantitative studies reported quantitative outcomes that were not pooled because of differences in extraction construct, task specificity, workflow architecture, overlap, or reconstructibility.

Report dependence: Konet et al. [46] evaluated the same underlying set of 10 randomized trials used in the related Gartlehner et al. 2024 proof-of-concept study [5] and was therefore retained as a related report rather than treated as an additional independent quantitative study. The published JAMIA version of Khan et al. [59] represents that study; its earlier preprint was not counted separately.

3.2. Descriptive Synthesis of Extraction Performance

The descriptive synthesis preserved the performance constructs reported by individual studies according to their original definitions. No attempt was made to transform or combine non-equivalent measures into a common metric. Accordingly, accuracy, agreement, precision, recall or sensitivity, specificity, F1 score, semantic equivalence, error rates, reproducibility, and workflow-level performance after human verification were interpreted as distinct outcomes.

Across the 46 eligible reports, extraction performance varied substantially according to task complexity, data type, AI system, workflow configuration, and evaluation methodology. Several studies reported high performance for explicitly stated and structured variables, whereas performance was less consistent when extraction required interpretation, integration of information distributed across text, tables, or figures, identification of missing or ambiguous information, or recovery of numerical outcome data [5,20,21,25–28,40,41,55].

Differences in workflow design further influenced the interpretation of reported performance. Some studies evaluated uncorrected LLM-generated outputs, whereas others assessed AI-assisted workflows involving subsequent human verification, correction, adjudication, or collaborative and multi-agent systems [7,8,30,39,49,53,59]. Consequently, apparently similar performance estimates did not necessarily represent equivalent extraction procedures. Results obtained after human verification were therefore distinguished from estimates of autonomous LLM extraction performance.

Substantial methodological variation was also observed in the construction of reference standards and operational definitions of correctness. Reference datasets included previously curated systematic-review tables, independently verified human extractions, and consensus-based or adjudicated reference standards [5–8,20,21,28,38,58]. Some studies applied strict criteria requiring fully correct extraction, whereas others accepted partial correctness, semantic equivalence, or graded study-specific assessments [21,28,38,51,58]. These differences were preserved rather than retrospectively harmonized because combining them under a single definition of accuracy would have obscured meaningful differences in the underlying performance constructs.

The unit of evaluation also varied across studies. Assessments were conducted at the level of individual fields, table cells, variables, outcome data, complete study records, or integrated workflow outputs. Several investigations additionally evaluated multiple LLMs, alternative prompting strategies, or repeated extraction sessions using overlapping source documents [5,8,20,21,25–29,40,55]. Consequently, the number of evaluated extraction items did not necessarily correspond to the number of independent studies, source documents, or experimental conditions represented in a given report.

Across studies reporting extraction errors, recurring problems included omitted information, incorrect numerical values, misallocation of information to inappropriate fields or treatment groups, unsupported inferences, and fabricated content [5,6,20,21,25–28,40,41,55]. Reproducibility across repeated model executions also varied, particularly when the requested information was absent, ambiguous, or required interpretation [6,21,28,52,55]. These findings indicated that aggregate performance measures could conceal clinically or methodologically consequential extraction errors.

Studies examining human–AI collaboration provided additional evidence regarding the practical implementation of automated extraction. Several investigations reported improvements in processing efficiency or comparable final extraction performance when LLM-generated information was subsequently verified by human reviewers [7,8,30,39,53]. However, these findings addressed the performance of assisted workflows rather than autonomous LLM extraction and were therefore interpreted separately from the primary item-level accuracy estimand. The principal reported performance outcomes across the included reports are summarized in Table 2.

Table 2. Principal reported performance outcomes across the 46 reports included in the systematic review.

Study/report	AI system / workflow	Principal reported performance outcome	Key finding relevant to synthesis	Synthesis classification
Gartlehner et al., 2024 [5]	Claude 2	154/160 correct; 96.3% accuracy repeat runs and 95.0%	High item-level accuracy; omissions represented an important error type.	Primary quantitative

Liu et al., 2025 [20]	Claude 2	1,775/1,873 correct; accuracy 94.8%	High overall accuracy, although performance varied substantially by extraction variable and participant-number items accounted for many errors.	Primary quantitative
Motzfeldt Jensen et al., 2025 [6]	ChatGPT-4o as second extractor	447/484 correct or satisfactory; accuracy 92.4%; reproducibility 94.1%	Performance was high overall but poorer when requested information was absent or ambiguously reported.	Primary quantitative
Vivekanantha et al., 2026 [21]	GPT-5.2; Gemini 3 Pro	GPT-5.2: 328/384 correct (85.4%); Gemini 3 Pro: 326/384 (84.9%)	Performance varied across extraction domains; dual-model use increased the probability that at least one output was correct.	Primary quantitative
Lai et al., 2025 [8]	Claude 3.5 Sonnet; Moonshot-v1-128k	Claude: 12,324/12,814 (96.2%); Moonshot: 12,189/12,814 (95.1%)	LLM-only extraction achieved high accuracy; subsequent human refinement further improved performance.	Primary quantitative

Felizardo et al., 2024 [22]	ChatGPT-4o		325/370 correct; accuracy 87.8%	Bibliometric variables were extracted much more accurately than research-question-specific information.	Primary quantitative
Mahmoudi et al., 2025 [23]	GPT-4 / ChatGPT		Final iteration: 43/60 correct (71.7%); initial iteration: 26/60	Performance improved markedly after iterative prompt engineering, raising concern about evaluation-set overfitting.	Primary quantitative
Stewart-Evans et al., 2025 [24]	Claude Sonnet	3.5	Simple-protocol micro-accuracy 51/300 (17.0%); much higher performance for simple citation items than complex extraction	Open-ended, recall-intensive excerpt extraction produced substantially lower performance than structured field extraction.	Primary quantitative
Caponio et al., 2026 [25]	Claude Sonnet; DeepSeek-v3 R1; ChatGPT-4o; Gemini 2.0-flash	3.5	Numerical accuracy: 99.0%, 99.0%, 95.8%, and 89.6%, respectively	Numerical extraction could be highly accurate under a narrowly structured task, but performance differed by model.	Secondary quantitative
Peng et al., 2025 [26]	Claude ChatGPT-4o	3.5;	4,006/5,714 correct with Claude (70.1%); 3,619/5,714 with ChatGPT-4o (63.3%)	Numerical-data performance was substantially lower than headline accuracy reported in	Secondary quantitative

				several broader extraction studies.	
Kataoka et al., 2026 [27]	GPT-4o; o3	String-variable accuracy approximately 74.0%–81.7%; numerical-variable accuracy approximately 60.4%–74.2% across evaluated settings	Performance differed by data type and extraction method; the external validation dataset partially overlapped with Gartlehner 2024.	Secondary quantitative	
Oami et al., 2026 [28]	ChatGPT-4o; Claude 3 Sonnet; Gemini 1.5 Pro	Background no-error proportion 81.6%–92.4%; outcome no-error proportion 27.8%–80.7%	Background characteristics were extracted substantially more reliably than clinical outcomes; repeated-session consistency was also lower for outcomes.	Secondary quantitative	
Zhang et al., 2026 [29]	ChatPDF, Bing Chat, GPT-4o	Study 1 mean accuracy approximately 0.54, 0.73, 0.79, and 0.80, respectively; DCR workflow increased GPT-4o accuracy to approximately 0.95	Model choice and multi-pass evidence-grounded processing materially improved performance.	Secondary quantitative	

Wang et al. (TrialMind), 2025 [30]	TrialMind multi-stage LLM pipeline	AI-assisted extraction increased accuracy by approximately 23.5% and reduced extraction time by approximately 63.4% relative to human-only extraction in the user study	Human–AI collaboration improved workflow efficiency, but the result does not represent uncorrected standalone LLM accuracy.	Secondary quantitative
Rokhshad et al., 2026 [31]	LLM-based extraction	Extraction quality evaluated using a study-specific 0–5 scoring framework	Performance varied across extracted domains, but the ordinal scoring system was not directly comparable with binary item-level correctness.	Narrative/descriptive
Daraqel et al., 2025 [32]	AI-assisted extraction	Agreement ranged approximately 0.16–1.00 across categorical variables	Straightforward variables showed stronger agreement than more complex methodological variables.	Narrative/descriptive
Kartchner et al., 2023 [33]	ChatGPT and foundation-model extraction	Field-specific accuracy, fuzzy Jaccard similarity, and ROUGE-L	Performance varied markedly across extraction fields; ChatGPT generally outperformed GPT-JT, particularly for dosage and summarization tasks.	Narrative/descriptive

Jansen et al., 2025 [34]	Eight generative AI/LLM systems	Overall performance highly variable across variables; approximately 20% of variables unacceptable across all accuracy metrics	Extraction difficulty was strongly variable-specific; effect-size extraction was particularly unreliable.	Narrative/descriptive
Schroeder et al., 2025 [35]	Multiple LLMs	Agreement and κ varied substantially across models and variable types; reported κ values included values as low as approximately 0.19	Explicitly stated information was extracted more reliably than derived or interpretive variables.	Narrative/descriptive
Susnjak et al., 2025 [36]	Fine-tuned LLM / RAG variants	NEFTune: 89.2% supported responses; LoRA: 87.7%; NEFTune + RAG substantially lower	Fine-tuning markedly improved factual fidelity, whereas combining some fine-tuned models with RAG could reduce consistency.	Narrative/descriptive
Barros et al., 2026 [37]	ChatGPT-4-supported systematic mapping workflow	Comparative manual-versus-LLM workflow performance and execution time	LLM assistance reduced effort across review stages, but the end-to-end design did not yield a directly comparable primary extraction estimand.	Narrative/descriptive

Gartlehner et al., 2025 [7]	Claude-based extraction + human verification	AI-assisted accuracy 91.0% vs 89.0% human-only; sensitivity 89.4%; PPV 98.9%; median 41 min saved/study	Prospective workflow evidence suggested comparable or improved final performance with reduced reviewer time.	Narrative/descriptive
Bianchi et al., 2025 [38]	Elicit	29.3% “more,” 20.7% “equal,” 45.7% “partially equal,” 4.3% “deviating”	Elicit frequently retrieved relevant information but often produced only partial equivalence with human extraction.	Narrative/descriptive
Helms Andersen et al., 2025 [39]	Elicit; ChatGPT	Recall 92.2% for Elicit and 88.8% for ChatGPT; confabulation in 4/90 and 3/90 tasks	Both tools showed potential as second reviewers, but errors and confabulations remained.	Narrative/descriptive
Murton et al., 2025 [40]	LLM with alternative prompting strategies	F1 >0.85 for study/baseline characteristics; approximately 0.22–0.50 for complex efficacy/adverse-event data	Prompt engineering improved performance, but complex outcome extraction remained substantially more difficult.	Narrative/descriptive
Fan et al., 2026 [41]	Claude Sonnet	3.5 Overall extraction error rate 6.6% (95% CI 5.4%–8.2%)	Misallocation and missed or omitted information were prominent failure modes, with error	Narrative/descriptive

						rates varying by data class.	
Kim et al., 2025 [42]	GPT-4			Approximately 88.6% concordance with the published systematic review	GPT-4 reproduced most study-characteristic data and generated broadly comparable review outputs.	Narrative/descriptive	
Lamanna et al., 2025 [43]	ChatGPT-4			Global extraction accuracy approximately 75%; labeling accuracy 67.3%	Item-level discrepancies remained even when aggregate review findings were relatively similar.	Narrative/descriptive	
Kasireddy et al., 2026 [44]	Claude Sonnet	3.7		Precision 98.2%; recall 96.6%; F1 97.4%	High overall performance was accompanied by meaningful variation across extraction domains.	Narrative/descriptive	
Yerramilli et al., 2026 [45]	Six LLMs			Mean concordance approximately 0.57–0.75 across evaluated models	Performance differed by model and by numerical, textual, and categorical data type.	Narrative/descriptive	
Konet et al., 2024 [46]	Claude 2; GPT-4 + PDF plugin			Claude 96.3%; GPT-4 68.8%; selected-text analysis 98.7% and 100%, respectively	PDF parsing accounted for much of GPT-4's poor performance. Source documents overlapped with Gartlehner 2024.	Narrative/descriptive; related report	

Khraisha et al., 2024 [47]	GPT-4	Study-specific extraction and review-performance measures	GPT-4 showed useful extraction capability but performance varied across tasks and language/context requirements.	Narrative/descriptive
Achterberg et al., 2026 [48]	OpenExtract / LLM pipeline	Precision and recall >0.80 in evaluated extraction tasks	Automated extraction was promising but evaluated with precision/recall rather than the primary item-level correctness estimand.	Narrative/descriptive
Muthu et al., 2026 [49]	Locally executed agentic AI framework	Structured data-extraction agreement $\kappa \approx 0.87$	The agentic workflow achieved substantial agreement with human reviewers within a broader review pipeline.	Narrative/descriptive
Ye et al., 2024 [50]	Hybrid semiautomated workflow	Study-specific workflow and extraction-performance measures	Hybrid automation reduced manual effort but retained human involvement across critical review steps.	Narrative/descriptive
Sanlier et al., 2026 [51]	Multi-LLM consensus workflow	Semantic-equivalence accuracy approximately 80.5%–83.4%; unanimous 4/4	Consensus did not guarantee correctness, and performance declined with	Narrative/descriptive

		consensus in 45.6% of items	increasing extraction complexity.	
Mitchell et al., 2025 [52]	AI-LES	100% accuracy and reliability when target statistics were clearly reported	Performance deteriorated when reporting was incomplete or ambiguous.	Narrative/descriptive
Ishikane et al., 2026 [53]	LLM-assisted randomized crossover workflow	Extraction accuracy 100% in both assisted and control conditions; 27.5 vs 34.5 min/article	Accuracy was preserved while completion time numerically favored LLM assistance, although the time effect was imprecise.	Narrative/descriptive
Jansen et al., 2026 [54]	Gemini 2.5 Pro; GPT-4.1; o3	ICC approximately 0.96–0.97; percentage agreement approximately 77%–81%	LLM-derived estimates approached human-expert performance under the study-specific adjudication framework.	Narrative/descriptive
Gates et al., 2021 [3]	Machine learning / text mining	Median 91% accuracy for identifying reported elements; median 48% of complete solutions fully correct	Semi-automation generated modest efficiency gains, but complete extraction remained considerably less reliable than field detection.	Narrative/descriptive

Yisha et al., 2026 [55]	ChatGPT-4; Claude 3 Opus	Binary group size approximately 91%–94%; event counts 57%–71%; continuous means/SDs 24%–56%	Numerical extraction accuracy depended strongly on variable type.	Narrative/descriptive
Panayi et al., 2023 [4]	BERT + CRF / ML pipeline	Entity-recognition F1 approximately 70%–73%; relation-extraction F1 >90% in well-represented classes	Performance depended strongly on entity type and availability of training data.	Narrative/descriptive
Li et al., 2026 [56]	Gemini-2.0-flash; Grok-3; GPT-4o-mini	High precision but more variable recall; customized prompting increased recall by approximately 14.8 percentage points	Omission rather than false extraction was the dominant limitation; task-specific prompting improved completeness.	Narrative/descriptive
Meliante et al., 2025 [57]	Elicit; ChatPDF; other AI tools	Extraction performance approximately 51.4% for Elicit and 60.3% for ChatPDF in the reported comparison	No single tool performed best across all review tasks; extraction from tables and figures remained challenging.	Narrative/descriptive

Wang et al. (SciDaSynth), 2025 [58]	SciDaSynth; manual baseline; GPT baseline	Dataset I quality score: 17.58 vs 16.08 vs 13.25; Dataset II: 17.79 vs 16.56 vs 14.50	SciDaSynth improved extraction quality and efficiency relative to both manual and GPT-only baselines, but used an ordinal expert-scored outcome rather than binary correctness.	Narrative/descriptive
Khan et al., 2025 [59]	Collaborative GPT-4-turbo / Claude-3-Opus workflow	Test-set concordance approximately 87%; accuracy among concordant outputs approximately 0.94	Cross-critique resolved a substantial proportion of discordant outputs and improved reliability, but the collaborative pipeline represented a distinct workflow construct.	Narrative/descriptive

Abbreviations: AI, artificial intelligence; F1, harmonic mean of precision and recall; ICC, intraclass correlation coefficient; LLM, large language model; PPV, positive predictive value; RAG, retrieval-augmented generation; RCT, randomized controlled trial. Performance measures are reported according to the definitions used by the original investigators and should not be interpreted as directly comparable across rows. Classification refers to the role of each report in the present review. Konet et al. [46] was retained as a related report because its source-document set substantially overlapped with Gartlehner et al. [5].

3.3. Eligibility for the Primary Quantitative Synthesis

The primary meta-analysis evaluated uncorrected item-level extraction accuracy, defined as the proportion of extraction items correctly generated by an LLM before human correction or adjudication, relative to a human, adjudicated, verified, or otherwise validated reference standard.

Application of the quantitative eligibility criteria identified eight independent studies from the broader evidence base of 46 eligible reports. These studies contributed 10 model-level estimates to the primary quantitative synthesis [5,6,8,20–24]. Eligible estimates were required to represent uncorrected

LLM output, use an appropriate reference standard, and provide directly reported or arithmetically reconstructable counts of correctly extracted and total evaluated items.

The remaining reports were not included in the primary meta-analysis because their performance measures, extraction tasks, or workflow characteristics did not adequately correspond to the target estimand. Reasons for non-pooling included the use of non-equivalent performance constructs, such as agreement, precision, recall, sensitivity, specificity, F1 score, semantic equivalence, ordinal scoring, or outcomes measured after human verification. Other reports lacked reconstructable item-level numerators and denominators or evaluated collaborative, agentic, or multistage workflows that could not be interpreted as uncorrected LLM extraction.

Importantly, quantitative eligibility was determined by the comparability of the underlying performance construct rather than the magnitude or direction of the reported results. The availability of a numerical accuracy estimate alone was insufficient to justify pooling when the evaluation represented a substantively different extraction task. Studies restricted to numerical-data extraction, for example, were retained as secondary quantitative evidence because their results represented a narrower task-specific estimand. Similarly, accuracy estimates averaged across repeated model executions, prompting conditions, variable classes, or multistage workflows were not converted into a single pooled estimate when doing so would require assumptions beyond the defined primary estimand.

Potential dependence among reports was also considered when determining eligibility for quantitative synthesis. Konet et al. [46] evaluated substantially overlapping source documents with the Gartlehner et al. proof-of-concept study [5] and was therefore retained as a related report rather than treated as an additional independent quantitative study. Similarly, although Kataoka et al. [27] represented an independent investigation overall, its external-validation Dataset 3 reused the source-document set evaluated by Gartlehner et al. [5] and was not treated as an additional independent contribution for that component.

Thus, restriction of the primary quantitative synthesis was based on compatibility with the defined estimand, extraction construct, and evidentiary independence rather than observed performance. The synthesis classifications and principal reported performance outcomes of the included reports are summarized in Tables 1 and 2.

3.4. Primary Meta-analysis of Item-Level Extraction Accuracy

Eight independent studies contributed 10 model-level estimates comprising 29,643 item-level evaluations to the primary meta-analysis [5,6,8,20–24]. Two studies evaluated more than one LLM,

and their model-specific estimates were retained as dependent observations within the corresponding study clusters. Study- and model-level data contributing to the primary meta-analysis are summarized in Table 3, and the corresponding forest plot is presented in Figure 2.

The multilevel random-effects model yielded a pooled item-level extraction accuracy of 86.4% (cluster-robust 95% confidence interval [CI], 62.6%–96.0%). Substantial heterogeneity was observed across the included studies. The estimated between-study variance component was 2.4723, compared with an additional model-within-study variance component of 0.0194. This difference indicated that most of the estimated heterogeneity occurred between independent studies rather than between model-specific estimates within the same study. The generalized I^2 was 99.6%.

The model-based 95% prediction interval ranged from 19.2% to 99.4%, indicating considerable variability in underlying extraction accuracy across evaluation settings. Although the pooled estimate suggested relatively high average performance, the wide prediction interval indicated that comparable accuracy could not be assumed across different evidence-synthesis workflows.

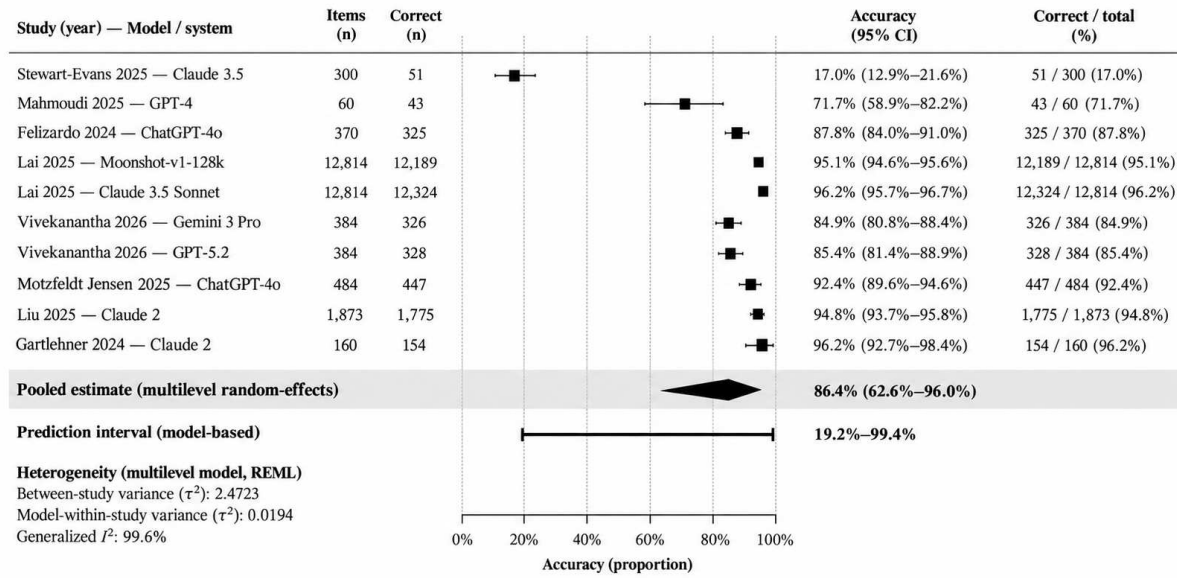
The predominance of between-study heterogeneity was consistent with the methodological diversity of the included investigations, which differed in extraction tasks, source-document characteristics, reference standards, prompting strategies, and workflow configurations. These characteristics represent plausible contributors to between-setting variability. However, their individual contributions could not be formally established because the limited number of independent studies precluded reliable moderator analyses.

The substantial difference between the pooled confidence interval and the prediction interval highlighted the distinction between uncertainty surrounding average extraction accuracy and variability in underlying performance across settings. Consequently, the pooled estimate should not be interpreted as a universally applicable performance benchmark for automated data extraction.

Because only eight independent study clusters contributed to the primary synthesis, estimates of heterogeneity and the prediction interval remain subject to considerable uncertainty. The prediction interval should therefore be interpreted as a model-dependent description of between-setting variability rather than a precise forecast of accuracy for a particular LLM or future evidence-synthesis workflow.

Forest plot of item-level extraction accuracy in the broad primary meta-analysis

Eight independent studies, 10 model-level estimates, 29,643 item-level evaluations



Note: Squares represent individual model-level estimates with 95% confidence intervals; square sizes are proportional to the inverse of the variance under the multilevel model. Multiple estimates from the same study were modeled as dependent observations within the corresponding study cluster.

Figure 2. Forest plot of item-level extraction accuracy in the broad primary meta-analysis. Individual model-level estimates are shown with 95% confidence intervals. Multiple estimates originating from the same study were modeled as dependent observations within the corresponding study cluster. The pooled estimate was obtained from a multilevel random-effects model fitted on the logit scale, with cluster-robust CR2 inference at the study level. The diamond represents the pooled accuracy and its 95% confidence interval, whereas the separate horizontal line represents the model-based 95% prediction interval for underlying extraction accuracy across comparable settings. Square sizes reflect the precision of individual estimates and should not be interpreted as conventional inverse-variance study weights, which are not directly interpretable as single weights in a multilevel model accounting for dependent estimates.

Table 3. Study- and model-level data contributing to the broad primary meta-analysis of item-level extraction accuracy

Study	Model	Correct items	Total items	Accuracy (%)	Analytical treatment
-------	-------	---------------	-------------	--------------	----------------------

Gartlehner et al., 2024 [5]	Claude 2	154	160	96.3	Independent study estimate
Liu et al., 2025 [20]	Claude 2	1,775	1,873	94.8	Independent study estimate
Motzfeldt Jensen et al., 2025 [6]	ChatGPT-4o	447	484	92.4	Independent study estimate
Vivekanantha et al., 2026 [21]	GPT-5.2	328	384	85.4	Model-level estimate; clustered within study
Vivekanantha et al., 2026 [21]	Gemini 3 Pro	326	384	84.9	Model-level estimate; clustered within study
Lai et al., 2025 [8]	Claude Sonnet 3.5	12,324	12,814	96.2	Model-level estimate; clustered within study
Lai et al., 2025 [8]	Moonshot-v1-128k	12,189	12,814	95.1	Model-level estimate; clustered within study
Felizardo et al., 2024 [22]	ChatGPT-4o	325	370	87.8	Independent study estimate
Mahmoudi et al., 2025 [23]	GPT-4 ChatGPT	/ 43	60	71.7	Independent study estimate; final optimized-prompt iteration used in the primary analysis

Stewart-Evans et al., 2025 [24]	Claude Sonnet simple protocol	3.5	51	300	17.0	Independent study estimate; open-ended, recall-intensive task examined in construct-restricted sensitivity analysis
---------------------------------	-------------------------------	-----	----	-----	------	---

Note: Model-specific estimates originating from the same study were retained separately and modeled as dependent observations within the corresponding study cluster. The Mahmoudi et al. primary estimate represents the final optimized-prompt iteration; the initial-prompt estimate was examined in sensitivity analysis. Stewart-Evans et al. was retained in the broad primary meta-analysis but excluded from the construct-restricted sensitivity analysis because its open-ended, recall-intensive extraction task differed materially from the more structured extraction settings represented elsewhere in the primary pool.

3.5. Robustness and Sensitivity Analyses

Sensitivity analyses were conducted to examine the robustness of the primary findings and to assess the influence of studies with potentially distinct extraction constructs or methodological features. The principal construct-restricted sensitivity analysis excluded Stewart-Evans et al. [24] because its open-ended, recall-intensive extraction task differed materially from the more structured item-level extraction settings represented in the remainder of the primary pool. The corresponding forest plot is shown in Figure 3.

After exclusion of Stewart-Evans et al. [24], seven independent studies contributed nine model-level estimates comprising 29,343 item-level evaluations. Under this construct-restricted model, pooled item-level extraction accuracy increased from 86.4% in the broad primary analysis to 91.2% (cluster-robust 95% CI, 83.1%–95.6%). Heterogeneity remained substantial, but its magnitude was attenuated: the between-study variance component decreased from 2.4723 to 0.5897, the model-within-study variance component remained 0.0194, and generalized I^2 decreased from 99.6% to 98.6%. The model-based 95% prediction interval also narrowed markedly, from 19.2%–99.4% in the broad primary analysis to 66.7%–98.2%. These findings indicate reduced estimated between-setting variability after exclusion of the constructually distinctive study, although substantial residual heterogeneity remained.

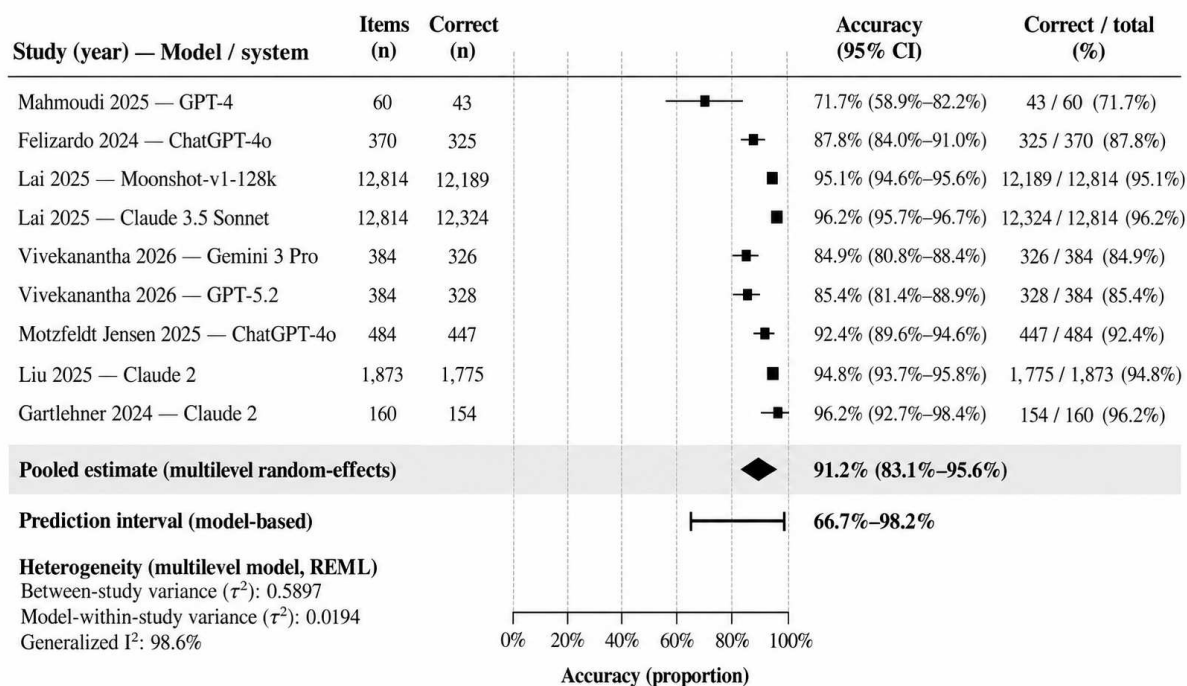


Figure 3. Forest plot of item-level extraction accuracy in the construct-restricted sensitivity analysis.

After exclusion of Stewart-Evans et al. [24], seven independent studies contributed nine model-level estimates and 29,343 item-level evaluations. Squares represent individual model-level accuracy estimates with 95% confidence intervals. Multiple estimates originating from the same study were modeled as dependent observations within the corresponding study cluster. The diamond represents the pooled accuracy of 91.2% (cluster-robust 95% CI, 83.1%–95.6%), estimated using a multilevel random-effects model. The horizontal line represents the model-based 95% prediction interval (66.7%–98.2%). Between-study variance was 0.5897, model-within-study variance was 0.0194, and generalized I² was 98.6%.

Additional sensitivity analyses further examined the influence of methodological differences. Exclusion of Mahmoudi et al. [23], a study raising concerns regarding iterative prompt optimization on the evaluation set, yielded a pooled accuracy of 87.8% (95% CI, 60.6%–97.1%), with a between-study variance of 2.7343, generalized I² of 99.7%, and a prediction interval of 18.2%–99.6%. Thus, exclusion of Mahmoudi et al. alone had relatively little effect on the overall result. In contrast, simultaneous exclusion of Stewart-Evans et al. [24] and Mahmoudi et al. [23] yielded a synthesis of six independent studies and eight model-level estimates, with a pooled accuracy of 92.7% (95% CI, 87.0%–96.0%), between-study variance of 0.3238, generalized I² of 97.8%, and a narrower prediction interval of 78.5%–97.8%.

Because Mahmoudi et al. [23] used iterative prompt refinement, an additional analysis replaced the final optimized-prompt estimate with the study's initial-prompt result (26/60) to examine the sensitivity of the findings to this methodological choice. Under this scenario, pooled accuracy was 84.5% (95% CI, 56.1%–95.9%), with a between-study variance of 2.9532, generalized I² of 99.7%, and a prediction interval of 13.1%–99.5%. Replacing the optimized-prompt estimate modestly reduced pooled accuracy but did not materially alter the finding of substantial between-setting heterogeneity.

Sequential leave-one-study-out analyses showed that the pooled estimate was directionally stable, although its magnitude and dispersion were influenced by individual studies. Omitting any study other than Stewart-Evans or Mahmoudi yielded pooled accuracies ranging from 84.0% to 86.5%, with persistently wide confidence intervals and prediction intervals. In contrast, omitting Stewart-Evans et al. [24] increased pooled accuracy to 91.2% (95% CI, 83.1%–95.6%) and substantially narrowed the prediction interval, whereas omitting Mahmoudi et al. [23] increased pooled accuracy more modestly to 87.8% (95% CI, 60.6%–97.1%). Stewart-Evans et al. therefore exerted the greatest observed influence on the pooled estimate and estimated between-setting dispersion in these leave-one-study-out analyses. The complete results are summarized in Table 4.

Overall, the sensitivity analyses support two complementary conclusions. First, relatively high average item-level extraction accuracy was observed across alternative analytical specifications, although pooled estimates varied according to study inclusion and prompt-selection decisions. Second, the magnitude of heterogeneity was sensitive to the comparability of extraction constructs. Exclusion of the most methodologically distinctive setting increased pooled accuracy and substantially narrowed the prediction interval, consistent with extraction-task differences contributing to between-setting variability. Nevertheless, substantial residual heterogeneity, wide uncertainty intervals in several sensitivity analyses, and the limited number of independent studies constrain the precision and generalizability of these findings.

Table 4. Leave-one-study-out sensitivity analyses of the broad primary item-level accuracy meta-analysis

Study excluded	Pooled accuracy	Robust 95% CI	Generalized I ²	Model-based 95% PI
Gartlehner et al., 2024 [5]	84.0%	54.2%–95.9%	99.7%	15.5%–99.3%
Liu et al., 2025 [20]	84.5%	54.0%–96.2%	99.7%	14.8%–99.4%
Motzfeldt Jensen et al., 2025 [6]	85.2%	54.6%–96.5%	99.7%	14.5%–99.5%

Vivekanantha et al., 2026 [21]	86.5%	56.6%–96.9%	99.7%	15.2%–99.6%
Lai et al., 2025 [8]	84.1%	53.8%–96.0%	99.0%	15.1%–99.4%
Felizardo et al., 2024 [22]	86.1%	55.9%–96.8%	99.7%	14.8%–99.6%
Mahmoudi et al., 2025 [23]	87.8%	60.6%–97.1%	99.7%	18.2%–99.6%
Stewart-Evans et al., 2025 [24]	91.2%	83.1%–95.6%	98.6%	66.7%–98.2%

Note: Leave-one-study-out sensitivity analyses were performed by sequentially excluding each independent study cluster and refitting the multilevel random-effects model. When a study contributed multiple model-level estimates, all estimates originating from that study were excluded together. Confidence intervals were calculated using cluster-robust inference. Prediction intervals were derived from the fitted heterogeneity structure.

Abbreviations: CI, confidence interval; PI, prediction interval; I^2 , proportion of variability attributable to heterogeneity beyond sampling error.

3.6. Numerical-Data Extraction

The estimand for this secondary quantitative analysis was the proportion of numerical extraction items correctly reproduced relative to the study-specific reference standard. Although six studies were classified as secondary quantitative evidence overall, only two provided numerical-data strata with directly interpretable correct/total denominators suitable for this specific analysis [25,26]. Because these studies differed substantially in extraction task, data structure, source material, and evaluation design, no pooled estimate was calculated.

Across the two studies, numerical-data accuracy ranged from **63.3% to 99.0%**. Caponio et al. [25] reported **190/192** correct numerical extractions with Claude 3.5 Sonnet (**99.0%**), **190/192** with DeepSeek-v3 R1 (**99.0%**), **184/192** with ChatGPT-4o (**95.8%**), and **172/192** with Gemini 2.0-flash (**89.6%**). Peng et al. [26] reported **4,006/5,714** correct numerical extractions with Claude 3.5 (**70.1%**) and **3,619/5,714** with ChatGPT-4o (**63.3%**).

The marked difference between the two studies demonstrates that numerical extraction performance cannot be characterized adequately by a single generic accuracy estimate and may reflect differences in task structure, numerical-variable type, source material, and evaluation framework. Accordingly, these numerical-data results were retained as a separate secondary quantitative analysis rather than incorporated into the broader item-level primary meta-analysis. Study- and model-level results are summarized in **Table 5**.

Table 5. Secondary quantitative analysis of numerical-data extraction performance

Study	Model	Correct/Total	Accuracy
Caponio et al. [25]	Claude 3.5 Sonnet	190/192	99.0%
Caponio et al. [25]	DeepSeek-v3 R1	190/192	99.0%
Caponio et al. [25]	ChatGPT-4o	184/192	95.8%
Caponio et al. [25]	Gemini 2.0-flash	172/192	89.6%
Peng et al. [26]	Claude 3.5	4,006/5,714	70.1%
Peng et al. [26]	ChatGPT-4o	3,619/5,714	63.3%

The observed range should not be interpreted as evidence that numerical extraction is universally inferior to other extraction classes. Rather, it demonstrates that aggregate accuracy can conceal substantial variation according to data type, task structure, and extraction setting. Given the potential for errors in event counts, denominators, means, standard deviations, or time points to alter downstream effect estimates, numerical-data performance was retained as a distinct secondary outcome.

3.7. Human-Assisted Workflows, Time, and Reviewer Workload

The estimands for this section were study-specific measures of processing time, reviewer time, workload reduction, and workflow efficiency. Differences in interfaces, extraction scope, human verification procedures, and definitions of elapsed or reviewer time precluded meaningful quantitative pooling.

Among studies reporting efficiency outcomes, automation or human–AI collaboration was generally associated with shorter extraction or processing times. However, the magnitude of apparent time savings varied substantially according to whether studies measured model-processing time alone, complete extraction time, prompt preparation, document upload or transfer, discrepancy resolution, or subsequent human verification.

Selected directly interpretable results included approximately 88.8 seconds per randomized trial for automated extraction in Liu et al. [20]; approximately 3.5 hours versus 25–30 hours for extraction from 11 articles in the second-rater workflow evaluated by Motzfeldt Jensen et al. [6]; and a median reduction of 41 minutes per study, from 125 to 84 minutes, in the prospective AI-assisted study-within-review comparison reported by Gartlehner et al. [7]. Ishikane et al. [53] reported 27.5 versus 34.5 minutes per article, corresponding to an adjusted mean difference of approximately 7.9 minutes per article; however, the confidence interval included no difference (95% CI, –1.5 to 17.3; $P = 0.099$).

Other studies reported rapid LLM-only processing, including approximately 82–96 seconds per randomized trial in Lai et al. [8] and approximately 4.5 minutes per article in Kasireddy et al. [44]. These estimates should be interpreted cautiously because the corresponding human comparators were not always measured concurrently and because model-processing time does not necessarily capture the full reviewer workload required for preparation, verification, and correction.

Because automation may redistribute rather than eliminate human effort, model-processing speed alone was not considered equivalent to overall workflow efficiency. Studies directly comparing complete human-only and human–AI workflows were therefore considered particularly informative for implementation, especially when verification time and downstream correction were incorporated into the comparison. The principal directly interpretable time, workload, and efficiency outcomes are summarized in **Table 6**.

Table 6. Reported time, workload, and efficiency outcomes in automated or human–AI data-extraction workflows

Study	Workflow	Reported time/workload outcome	Interpretation
-------	----------	--------------------------------	----------------

Liu et al., 2025 [20]	Automated LLM extraction	88.8 s per randomized trial	Very rapid automated extraction; human comparison was not concurrently measured.
Motzfeldt Jensen et al., 2025 [6]	ChatGPT-4o as second rater	Approximately 3.5 h vs 25–30 h for 11 articles	Large estimated reduction in extraction time; workflow still required manual document and prompt handling.
Gartlehner et al., 2025 [7]	AI-assisted extraction with human verification	Median 84 vs 125 min per study; 41-min reduction	Direct prospective study-within-review comparison; AI-assisted extraction was faster in five of six reviews.
Ishikane et al., 2026 [53]	LLM-assisted randomized crossover workflow	27.5 vs 34.5 min/article; adjusted mean difference approximately 7.9 min	Direction favored LLM assistance, but the confidence interval included no difference and did not establish a definitive time advantage.
Lai et al., 2025 [8]	LLM-only and LLM-assisted extraction	LLM-only approximately 82–96 s/RCT; assisted workflow approximately 14.7 min/RCT	Automated generation was rapid, but total efficiency depended on subsequent human verification.
Kasireddy et al., 2026 [44]	Claude Sonnet extraction	3.7 Approximately 4.5 min/article	Short processing time; no concurrent human comparator was available.

3.8. Methodological Limitations and Risk of Bias

Methodological limitations and risk of bias were evaluated across the 46 eligible reports representing 45 independent study units using a QUADAS-2-informed framework adapted to AI

extraction-validation research [14]. Because QUADAS-2 was originally developed for diagnostic-accuracy studies, the framework was used as a structured methodological adaptation rather than as a formally validated risk-of-bias instrument for LLM-based data-extraction studies. The assessment considered four domains: selection and representativeness of source documents; specification and conduct of the AI extraction procedure; independence and quality of the reference standard; and flow and analysis.

Methodological concerns were common across the evidence base. Recurring issues included incomplete reporting of prompts, model versions, temperature or other generation settings; limited assessment of model stability and reproducibility; prompt development or optimization using material related to the evaluation set; unclear separation between model development and final validation; imperfect independence between AI output and construction or revision of the reference standard; limited reporting of blinding during adjudication; heterogeneous definitions of correctness; post-extraction exclusion or reclassification of difficult or missing items; and failure to account for dependence among multiple extraction items obtained from the same source document.

The flow-and-analysis domain was particularly important in this literature because large numbers of item-level observations were frequently derived from relatively small numbers of source articles. Treating these items as statistically independent can substantially overstate the effective amount of independent evidence. Accordingly, studies were judged conservatively when within-document clustering, repeated model evaluations, or other forms of dependence were not explicitly addressed.

Reference standards were generally based on human extraction, adjudicated or consensus datasets, or previously curated evidence-synthesis data. However, their methodological strength varied according to whether extraction was performed independently, whether disagreements were adjudicated, whether reviewers were blinded to model output, and whether the reference dataset was modified after comparison with the AI system. Where these features were insufficiently reported, the corresponding domain was judged as having some concerns or unclear risk rather than assuming independence.

Additional concerns specific to LLM-based studies included reproducibility of proprietary or continuously updated models, incomplete reporting of prompts and preprocessing procedures, potential contamination of evaluation material through model training or prior exposure, and use of the same source material for iterative prompt refinement and final performance assessment. These issues were considered particularly relevant when interpreting very high accuracy estimates derived from highly optimized workflows.

Study-level risk-of-bias judgments across the 46 reports representing 45 independent study units are presented in the traffic-light plot (Figure 4), and the distribution of judgments across methodological domains is summarized in Figure 5. Across domains, source-document selection and representativeness and the reference-standard domain were more often judged at low risk than the AI extraction procedure and flow-and-analysis domains. By contrast, the flow-and-analysis domain accounted for most high-risk judgments, reflecting frequent concerns regarding dependence among observations, completeness of evaluated items, and post-extraction analytical decisions. Consistent with this pattern, overall judgments were predominantly high risk, driven largely by recurrent concerns in the flow-and-analysis domain. Risk-of-bias findings for the studies contributing to the

primary meta-analysis were considered separately when interpreting the pooled estimate and the certainty of the quantitative conclusion.

Risk of bias across the 46 included reports

Adapted QUADAS-2-informed framework · 46 reports representing 45 independent study unit

	D1 Source-document selection	D2 AI extraction procedure	D3 Reference standard	D4 Flow and analysis	Overall
Gartlehner 2024	+	+	+	?	?
Motzfeldt Jensen 2025	+	?	+	-	-
Liu 2025	+	?	+	-	-
Vivekanantha 2026	+	?	+	-	-
Lai 2025	+	?	+	-	-
Felizardo 2024	+	+	-	-	-
Mahmoudi 2025	?	-	-	-	-
Stewart-Evans 2025	+	+	?	-	-
Caponio 2026	+	?	+	-	-
Peng 2025	+	?	+	-	-
Kataoka 2026	+	+	+	-	-
Oami 2026	+	?	+	-	-
Zhang 2026	+	+	+	-	-
Wang (TrialMind) 2025	+	+	+	-	-
Rokhshad 2026	+	?	+	-	-
Daraqel 2025	+	?	?	-	-
Kartchner 2023	+	?	+	-	-
Jansen 2025	+	?	+	-	-
Schroeder 2025	+	+	?	-	-
Susnjak 2026	+	+	+	-	-
Barros 2026	+	?	+	-	-
Gartlehner 2025	+	+	+	?	?
Bianchi 2025	+	?	+	-	-
Helms Andersen 2025	+	?	+	-	-
Murton 2025	+	-	+	-	-
Fan 2026	+	?	+	-	-
Kim 2025	?	-	?	-	-
Lamanna 2025	+	?	+	-	-
Kasireddy 2026	?	-	?	-	-
Yerramilli 2026	+	?	+	-	-
Konet 2024	+	?	+	-	-
Khraisha 2024	+	+	?	-	-
Achterberg 2026	?	?	?	-	-
Muthu 2026	?	-	?	-	-
Ye 2024	+	?	+	-	-
Sanlier 2026	+	?	+	-	-
Mitchell 2025	+	?	+	-	-
Ishikane 2026	+	+	+	?	?
Jansen 2026	+	?	+	-	-
Gates 2021	+	+	+	?	?
Yisha 2026	+	?	+	-	-
Panayi 2023	+	+	+	?	?
Li 2026	+	?	+	-	-
Meliante 2025	+	?	?	-	-
Wang (SciDaSynth) 2025	+	+	+	?	?
Khan 2025	+	+	+	-	-

Figure 4. Traffic-light plot of methodological quality and risk-of-bias judgments across the included evidence base. Domain-level judgments were assigned using a QUADAS-2-informed framework adapted to AI extraction-validation research. The four assessed domains were: D1, source-document selection and representativeness; D2, specification and conduct of the AI extraction procedure; D3, independence and quality of the reference standard; and D4, flow and analysis, including completeness of evaluated items and handling of statistical dependence. Judgments were categorized as low risk, some concerns/unclear, or high risk. Because the framework was adapted from QUADAS-2, these judgments should be interpreted as structured methodological assessments rather than results from a formally validated tool specific to LLM-based data extraction. Konet et al. [46] and Gartlehner et al. [5] represent related reports from the same independent study unit.

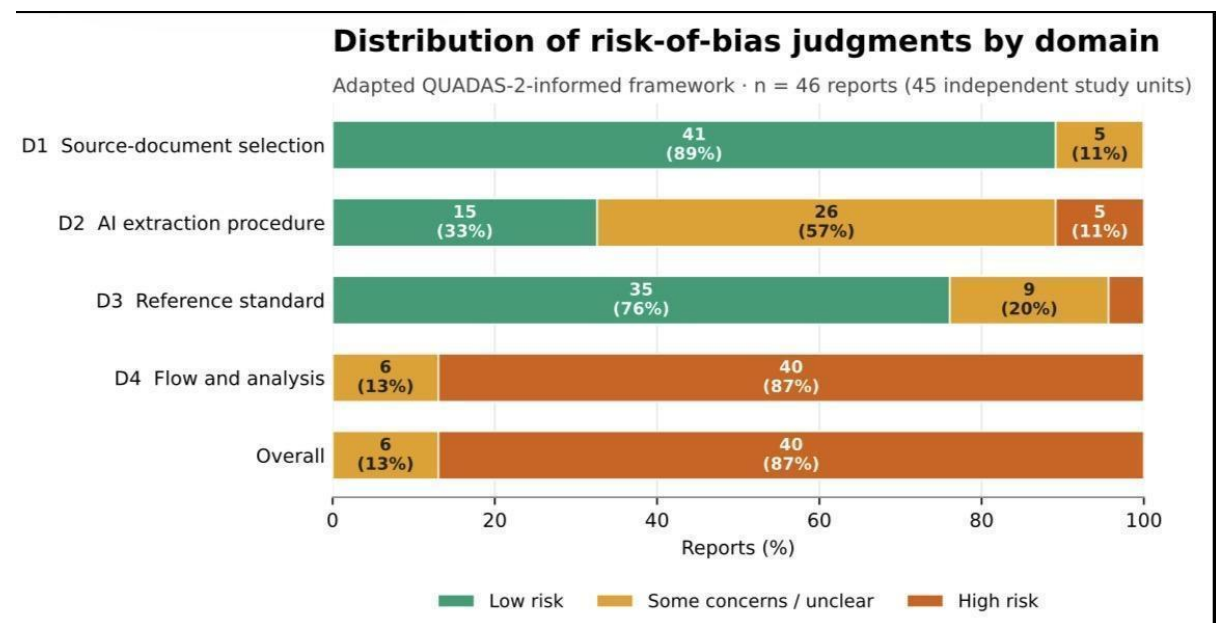


Figure 5. Summary distribution of methodological quality and risk-of-bias judgments across assessment domains. The figure shows the proportion of included evidence units rated as low risk, some concerns/unclear, or high risk within each of the four adapted domains: source-document selection and representativeness; AI extraction procedure; reference standard; and flow and analysis. Domain-level distributions are intended to highlight recurring methodological limitations across the evidence base rather than to provide a numerical quality score. The predominance of high-risk judgments in the flow-and-analysis domain and in the overall assessment reflects recurring concerns regarding dependence among observations, incomplete handling of evaluated items, and limited transparency in analytical decisions.

3.9. Certainty of the Primary Quantitative Conclusion

Confidence in the primary quantitative conclusion was considered low under the adapted GRADE-informed framework [15]. The principal reasons for reduced confidence were methodological

limitations across the contributing evidence, extreme between-setting heterogeneity, the limited number of independent study clusters, and uncertainty regarding the applicability of the pooled estimate beyond the extraction settings directly evaluated. This assessment represents a structured appraisal informed by GRADE principles rather than a formally validated application of GRADE to LLM-based data-extraction accuracy.

The broad primary meta-analysis included eight independent studies contributing 10 model-level estimates and 29,643 item-level evaluations. Although pooled item-level extraction accuracy was 86.4% (cluster-robust 95% CI, 62.6%–96.0%), generalized I^2 was 99.6%, and the model-based 95% prediction interval extended from 19.2% to 99.4%. Thus, the relatively high pooled mean coexisted with substantial variability in underlying extraction accuracy across evaluation settings.

Inconsistency was a major consideration in assessing certainty. The estimated between-study variance substantially exceeded the additional model-within-study variance, indicating that heterogeneity was concentrated predominantly between independent studies. Differences in extraction tasks, source documents, prompting strategies, model versions, reference standards, definitions of correctness, and workflow configurations were plausible contributors to this variability. However, their individual contributions could not be formally established given the limited number of independent studies.

Indirectness was also relevant because the available investigations represented a heterogeneous but limited sample of potential evidence-synthesis applications. The construct-restricted sensitivity analysis further illustrated the importance of extraction-task comparability: exclusion of the methodologically distinctive open-ended, recall-intensive setting increased pooled accuracy to 91.2% (95% CI, 83.1%–95.6%) and narrowed the prediction interval to 66.7%–98.2%. Although this finding was consistent with extraction-task differences contributing to heterogeneity, it did not establish the independent effects of specific methodological characteristics. Accordingly, the pooled accuracy should not be regarded as a universal performance benchmark for LLM-based extraction.

Imprecision remained important despite the large number of item-level evaluations. Only eight independent study clusters contributed to the primary meta-analysis, and evidentiary replication was therefore determined primarily by the number of independent studies rather than by the 29,643 individual extraction items. The wide cluster-robust confidence interval and particularly the broad prediction interval further limited confidence in the performance that might be expected in a new extraction setting.

Publication bias could not be meaningfully assessed because only eight independent studies contributed to the primary meta-analysis. Funnel-plot asymmetry and regression-based small-study-effect tests were therefore not performed, and no inference regarding the presence or absence of publication bias was made.

Taken together, the evidence indicates that LLMs can achieve high average item-level extraction accuracy under the conditions evaluated, but confidence in the generalizability of this performance across evidence-synthesis settings remains low. The principal concern is not that LLM-based extraction is uniformly inaccurate, but that average performance conceals substantial, incompletely explained variability across extraction contexts. Consequently, the pooled estimate should be interpreted as a summary of the evaluated settings rather than as a transferable accuracy guarantee.

4. Discussion

4.1. Principal Findings

The central finding of this systematic review and meta-analysis is that high average item-level extraction accuracy does not necessarily translate into consistent performance across evidence-synthesis settings. Across eight independent studies contributing 10 model-level estimates and 29,643 item-level evaluations [5,6,8,20–24], pooled accuracy was 86.4% (cluster-robust 95% CI, 62.6%–96.0%). However, the model-based 95% prediction interval extended from 19.2% to 99.4%, demonstrating that the relatively high pooled mean coexisted with substantial variability in underlying extraction performance.

The estimated heterogeneity was concentrated predominantly between independent studies. The between-study variance component was 2.4723, compared with an additional model-within-study variance component of 0.0194, and generalized I^2 was 99.6%. This pattern suggests that differences between evaluation settings were more influential than differences between model-specific estimates within the same study. Extraction-task characteristics, source documents, reference standards, prompting strategies, operational definitions of correctness, and workflow configurations are plausible contributors, although their individual effects could not be established reliably with the available number of independent studies.

Sensitivity analyses reinforced the importance of construct comparability. Exclusion of Stewart-Evans et al. [24], which evaluated a substantially different open-ended, recall-intensive extraction task, increased pooled accuracy to 91.2% (95% CI, 83.1%–95.6%) and narrowed the prediction interval to 66.7%–98.2%. Simultaneous exclusion of Stewart-Evans et al. [24] and Mahmoudi et al. [23] yielded a pooled accuracy of 92.7% (95% CI, 87.0%–96.0%) and a prediction interval of 78.5%–97.8%. These analyses illustrate how the composition of the evidence base influences both the estimated mean and the dispersion of performance.

Importantly, the influence of Stewart-Evans et al. should not be interpreted as grounds for removing a low-performing study merely because its result differed from the remainder of the evidence. Its inclusion in the broad primary analysis was retained, while the construct-restricted analysis examined the consequences of a substantive difference in extraction task. The findings therefore support a

conditional interpretation: high extraction accuracy is achievable in evaluated settings, but comparable performance cannot be presumed when the extraction construct or workflow changes.

4.2. Why the Pooled Mean Is Not Enough

A pooled accuracy estimate summarizes average performance under a specified analytical model; it does not, by itself, establish the reliability of an LLM in a new extraction setting. This distinction is especially important when studies differ in source-document characteristics, information complexity, extraction schemas, model versions, prompting procedures, and reference-standard definitions.

The broad primary analysis illustrates the limitation of relying on a single summary measure. An average accuracy of 86.4% may appear encouraging, but its model-based prediction interval of 19.2%–99.4% indicates substantial variation in underlying performance across settings. Even after restricting the analysis to more comparable extraction tasks, the prediction interval remained relatively wide at 66.7%–98.2%. These intervals are model-dependent estimates, subject to considerable uncertainty with only eight independent studies in the broad synthesis, rather than precise predictions for individual future applications.

Moreover, item-level accuracy does not incorporate the methodological consequences of different errors. An incorrectly extracted publication year and an incorrectly extracted event count may each contribute one error to the denominator, yet their implications for evidence synthesis differ substantially. Errors involving event counts, denominators, means, standard deviations, treatment groups, or follow-up times may alter effect estimates and propagate into subsequent meta-analytic calculations.

The secondary numerical-data analysis further illustrated the importance of task specificity. Across the two studies with directly interpretable numerical-extraction strata, accuracy ranged from 63.3% to 99.0% [25,26]. These results were not pooled because of meaningful differences in extraction tasks and evaluation designs. Their value lies instead in demonstrating that performance for consequential data classes should be evaluated directly rather than inferred from overall extraction accuracy.

Consequently, the practical interpretation of LLM performance requires consideration of at least three distinct dimensions: average correctness, variability across extraction settings, and the consequences of individual errors. A high pooled estimate addresses the first dimension but cannot substitute for evaluation of the others.

4.3. Estimand-Based Novelty and Relation to Previous Reviews

The principal methodological contribution of this review is its estimand-driven distinction between the breadth of evidence eligible for systematic review and the narrower evidence appropriate for quantitative synthesis. The review included 46 eligible reports representing 45 independent study units, but only eight independent studies [5,6,8,20–24] satisfied the criteria for the broad primary meta-analysis of uncorrected item-level extraction accuracy.

This distinction is not merely procedural. Agreement, precision, recall, sensitivity, specificity, F1 score, semantic equivalence, ordinal quality assessments, and performance after human verification do not necessarily estimate the same quantity as the proportion of uncorrected extraction items correctly generated relative to a validated reference standard. Even when studies report mathematically similar proportions, substantive differences in the extraction task, evaluation unit, or workflow may prevent their direct interpretation as equivalent measures.

Evidentiary independence was another essential consideration. Multiple model estimates generated from the same source-document set were retained where appropriate but modeled as dependent observations within their study cluster. Related reports, including Konet et al. [46] and Gartlehner et al. [5], were not treated as independent quantitative replications. These decisions prevented the breadth of reported model evaluations from being confused with the number of independent investigations.

Previous systematic reviews have already established that LLM extraction performance is heterogeneous. Shankar et al. [9] examined accuracy, reliability, and efficiency across extraction studies and used narrative synthesis because methodological differences limited meaningful meta-analysis. Laignelot et al. [10] evaluated LLM performance across several systematic-review tasks, reporting task-specific performance summaries while identifying substantial variability and limitations in quantitative comparability.

The present study addresses a complementary question. Rather than treating all reported performance measures as interchangeable or abandoning quantitative synthesis altogether, it defines a specific accuracy estimand, restricts pooling to studies meeting that definition, accounts for dependent model estimates, and evaluates both average performance and between-setting variability. Its contribution is therefore not the observation that LLM performance varies, but the quantitative demonstration that substantial variability persists despite a relatively high pooled mean and that the estimated performance distribution changes materially when extraction-task comparability is more tightly defined.

This distinction has direct methodological relevance. It separates evidence that an extraction system can perform well under particular conditions from evidence that such performance is likely to generalize. The former is supported by several existing evaluations; the latter remains substantially uncertain.

4.4. Error Architecture: Omission, Confabulation, and Reproducibility

The descriptive evidence indicates that extraction errors differ in mechanism, detectability, and consequence. Reported failures included omitted information, incorrect numerical values, inappropriate allocation of information to extraction fields or treatment groups, unsupported inferences, and fabricated content. Such errors were particularly relevant when information was absent, ambiguously reported, distributed across document components, or required interpretation beyond direct identification.

Technical aspects of document processing may also contribute to apparent model failure. Konet et al. [46] demonstrated that extraction performance differed substantially according to how source information was supplied to the model. In their evaluation, GPT-4 performed less accurately when document access depended on a PDF-processing workflow, whereas performance improved when relevant text was provided directly. This observation highlights the importance of distinguishing limitations of the language model from failures introduced by document ingestion and preprocessing.

Reproducibility adds another dimension that cannot be inferred from accuracy alone. An LLM may produce different answers across repeated executions of the same extraction task, but consistent outputs are not necessarily correct. Similarly, agreement between multiple LLMs does not establish validity when the systems share assumptions, overlook the same information, or reproduce similar errors.

These findings suggest that a sufficiently informative evaluation of automated extraction should distinguish correctness against an independent reference standard, reproducibility across executions, error type, error severity, and the likelihood that errors will be identified during verification. For quantitative evidence synthesis, downstream consequences are particularly important because a small number of consequential extraction errors may influence effect estimates or review conclusions despite high overall item-level accuracy.

4.5. When Is LLM-Assisted Second Extraction Defensible?

LLM-assisted second extraction is a plausible application for prospective evaluation, but the available evidence does not establish general equivalence between an LLM and an independent second human reviewer. Motzfeldt Jensen et al. [6] directly examined ChatGPT-4o as a second extractor, while Gartlehner et al. [7] prospectively evaluated AI-assisted extraction followed by human verification. These investigations provide relevant implementation evidence, although their workflows and outcomes should not be interpreted as equivalent to autonomous extraction.

Automation may reduce some components of manual extraction while redistributing reviewer effort toward verification, discrepancy resolution, and adjudication. Consequently, model-processing speed and final workflow efficiency must be evaluated separately. The relevant practical question is not simply whether an LLM can extract information quickly, but whether its integration reduces workload while maintaining an acceptable level of residual error.

The present meta-analysis adds an important qualification to this implementation question. Average accuracy was relatively high, but performance varied substantially across evaluation settings. A workflow that performs well for explicit and structured variables may not exhibit comparable reliability when numerical outcomes, ambiguous reporting, missing information, or derived variables are involved.

A reasonable near-term research strategy is therefore risk-adapted human-LLM collaboration rather than unverified autonomous extraction. One human reviewer could perform the primary extraction while an LLM independently generates a second extraction. Discrepancies would then undergo human adjudication, with additional verification of high-consequence variables even when both

extraction streams agree. Such safeguards are important because agreement does not exclude shared errors.

Explicit, structured, and relatively low-consequence variables may be suitable initial candidates for prospective evaluation. By contrast, event counts, denominators, measures of dispersion, intervention-group assignments, multiple time points, ambiguous findings, and derived outcomes warrant particularly careful verification.

The effectiveness of these arrangements remains an empirical question. Prospective comparisons should evaluate complete human-human and human-LLM workflows using the same source documents and extraction schemas, with outcomes including residual errors after adjudication, reviewer workload, processing time, cost, and consequences for subsequent evidence synthesis. Until such evidence is available, LLM-assisted second extraction is better regarded as a promising workflow strategy requiring validation than as an established replacement for independent human extraction.

4.6. Implications for Meta-analysis Automation

Data extraction is a critical dependency in the automation of systematic reviews and meta-analyses. Screening determines which studies enter the evidence base, whereas extraction supplies the structured information used to calculate effect estimates and conduct quantitative synthesis. Errors introduced at this stage may therefore propagate into effect-size calculations, study weights, heterogeneity estimates, subgroup analyses, certainty assessments, and ultimately the conclusions of a review.

This creates an important distinction between automating statistical computation and ensuring the validity of the information being analyzed. A system may execute an otherwise correct meta-analytic procedure using incorrectly extracted event counts or denominators. In that situation, automation increases the efficiency of downstream computation without resolving the underlying validity problem.

The findings of this review therefore favor a task-specific and risk-adapted approach to automation. Extraction tasks characterized by explicit information, reproducible outputs, readily detectable discrepancies, and relatively low error consequences may be more amenable to automation. Human verification remains particularly important where errors are difficult to detect or capable of materially altering quantitative conclusions.

Performance should also be evaluated at the level of the complete extraction workflow. Model version, document ingestion, preprocessing, prompts, extraction schema, source-document characteristics, reference-standard construction, and verification procedures may all influence the observed result.

Accordingly, the relevant unit of implementation is not a named LLM in isolation, but a defined model-task-workflow combination operating under specified verification conditions. This framing

provides a more defensible basis for evaluating whether performance demonstrated in one setting can reasonably inform deployment in another.

4.7. Implications for Future Research

Future validation studies should prioritize the intended extraction task and workflow rather than reporting model performance without sufficient operational context. Investigators should provide directly reported or reproducibly reconstructable numerators and denominators, identify the unit of evaluation, distinguish numerical from categorical and textual variables, and specify whether the assessment concerns uncorrected model output or a human-verified workflow.

Technical reporting should include the exact model and version, access method, document-processing procedures, prompts, relevant generation settings, and the timing of evaluations. Reference standards should be independently constructed and, where feasible, adjudicated without access to model outputs. Repeated-run assessments should distinguish reproducibility from correctness.

Prospective study-within-a-review evaluations are particularly valuable. Processing the same source documents through conventional dual-human extraction and human-LLM workflows would permit direct comparison of extraction accuracy, residual errors, discrepancy resolution, reviewer time, and cost under comparable conditions.

A further research priority is the measurement of downstream error propagation. Most extraction-validation studies focus on whether individual items match a reference standard. For meta-analysis, however, an equally consequential question is whether extraction errors materially change effect estimates, heterogeneity, certainty judgments, or the substantive conclusions of a review.

This distinction motivates prospective evaluation of consequence-sensitive performance measures. A risk-weighted approach, in which errors are characterized according to their potential impact on evidence synthesis, may complement conventional item-level accuracy. Such an approach would require explicit definitions, transparent weighting procedures, and independent validation before being adopted as a standardized metric.

Statistical design must also account for the structure of extraction evaluations. Multiple items obtained from the same source document are not independent experimental replications, and evaluations involving several models, prompts, extraction classes, or repeated runs introduce additional dependence. Future studies should preserve this hierarchical structure and use analytical methods consistent with the actual units of independent evidence.

Finally, external validation should be prioritized. Models and extraction procedures should be assessed using reviews, source documents, and extraction schemas independent of those used during prompt development or workflow optimization. Because LLM systems evolve rapidly, validation should be considered specific to the evaluated model version and workflow configuration rather than permanently transferable across technological updates.

4.8. Strengths and Limitations

This review has several methodological strengths. It distinguishes systematic-review eligibility from quantitative eligibility for a defined estimand; avoids combining non-equivalent performance measures; considers evidentiary overlap among reports; accounts for dependence between model-specific estimates from the same study; and reports both average accuracy and a prediction interval. Construct-restricted, methodological, and leave-one-study-out sensitivity analyses further clarify how analytical decisions and extraction-task comparability influence the findings.

The broader evidence base comprised 46 reports representing 45 independent study units, permitting the descriptive synthesis to incorporate relevant findings that could not appropriately contribute to the primary meta-analysis. This approach retained information concerning numerical extraction, errors, reproducibility, and human–AI workflow performance without forcing incompatible outcomes into a single quantitative measure.

Important limitations nevertheless constrain interpretation. Only eight independent studies contributed to the primary meta-analysis [5,6,8,20–24], limiting reliable moderator analysis and precluding conventional publication-bias assessment. Cluster-robust inference incorporated a small-sample correction, but the limited number of independent study clusters remains a source of uncertainty. Estimates of heterogeneity and prediction intervals should therefore be interpreted cautiously.

Although the primary synthesis encompassed 29,643 item-level evaluations, these observations did not constitute an equivalent number of independent experimental replications. Extraction items were nested within source documents, tasks, models, and studies, and their definitions differed across investigations. Thus, the large aggregate denominator should not be mistaken for a correspondingly large independent evidence base.

The broad primary analysis was deliberately inclusive within the specified item-level correctness estimand, but meaningful differences in extraction constructs remained. Stewart-Evans et al. [24] evaluated an open-ended, recall-intensive task that was substantively different from the more structured extraction settings represented elsewhere. Its inclusion widened the estimated performance distribution; however, excluding it solely because of its low observed accuracy would have been methodologically inappropriate. The construct-restricted sensitivity analysis instead examined the influence of the identifiable task difference while preserving the study in the broad primary synthesis.

Mahmoudi et al. [23] raised a separate concern because iterative prompt optimization was conducted using evaluation material. Replacing the final optimized-prompt estimate with the initial-prompt result reduced pooled accuracy from 86.4% to 84.5%, while substantial heterogeneity persisted. This analysis demonstrates sensitivity to prompt-selection decisions but cannot establish the absence or magnitude of optimization-related bias.

Potential dependence among source-document sets also required methodological judgment. Identifiable overlap, including that between Konet et al. [46] and Gartlehner et al. [5], was addressed

when defining independent quantitative contributions. Nevertheless, incomplete reporting of source-document provenance in some investigations means that unrecognized overlap cannot be excluded entirely.

Rapid technological evolution creates an additional challenge. Model updates, interfaces, document-processing systems, context windows, and prompting behavior may change faster than conventional systematic-review cycles. The quantitative results therefore characterize the systems and workflows evaluated in the included studies rather than a stable or universal property of LLM technology.

Finally, the QUADAS-2-informed and GRADE-informed frameworks were used as structured methodological adaptations rather than instruments formally validated for LLM extraction research. The review was not prospectively registered; retrospective registration was subsequently completed through the Open Science Framework (DOI: 10.17605/OSF.IO/5RGCN). These limitations constrain claims regarding formal prespecification, methodological validation, and certainty of evidence.

Taken together, the findings support a distinction that is central to the interpretation and implementation of LLM-based extraction: demonstrating high average accuracy under evaluated conditions is not equivalent to demonstrating reliable performance across new extraction settings. The evidence supports cautious, task-specific, human-verified implementation and further prospective validation, rather than treating pooled accuracy as a general performance guarantee.

5. Conclusions

Large language models can achieve high average item-level accuracy for data extraction in evidence synthesis, but this performance is not necessarily consistent across evaluation settings. In the primary meta-analysis of eight independent studies, comprising 10 model-level estimates and 29,643 item-level evaluations, pooled accuracy was 86.4% (cluster-robust 95% CI, 62.6%–96.0%). However, generalized I^2 was 99.6%, and the model-based 95% prediction interval ranged from 19.2% to 99.4%. These findings indicate that a high pooled mean may coexist with substantial between-setting variability and should not be interpreted as a universally transferable measure of LLM extraction performance.

The principal contribution of this review is to quantify this distinction within an explicitly defined extraction-accuracy estimand while accounting for dependence among model-level estimates. The narrower prediction interval observed in the construct-restricted sensitivity analysis further suggests that extraction-task comparability is central to interpreting pooled performance. Nevertheless, substantial residual heterogeneity and the limited number of independent studies warrant caution regarding generalizability.

The implications are constructive rather than prohibitive. LLM-assisted data extraction merits further prospective evaluation within task-specific, risk-adapted, and human-verified workflows, including its potential use as an independent second extractor. Particular attention should be given to

numerical, ambiguous, missing, derived, and other high-consequence information, for which errors may propagate into quantitative syntheses.

Ultimately, the relevant methodological question is not simply whether LLMs can extract data accurately, but under which conditions their performance is sufficiently reliable, reproducible, and verifiable to support evidence synthesis without compromising its validity.

References

1. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.5. Cochrane; 2024.
2. Li T, Higgins JPT, Deeks JJ. Chapter 5: Collecting data. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.5. Cochrane; 2024.
3. Gates A, Gates M, Sim S, Elliott SA, Pillay J, Hartling L. Creating efficiencies in the extraction of data from randomized trials: a prospective evaluation of a machine learning and text mining tool. *BMC Med Res Methodol*. 2021;21:169. doi:10.1186/s12874-021-01354-2.
4. Panayi A, Ward K, Benhadji-Schaff A, et al. Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. *Syst Rev*. 2023;12:187. doi:10.1186/s13643-023-02351-w.
5. 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;15(4):576–589. doi:10.1002/jrsm.1710.
6. Motzfeldt Jensen M, Brix Danielsen M, Riis J, Assifuah Kristjansen K, Andersen S, Okubo Y, et al. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. *PLoS One*. 2025;20(1). doi:10.1371/journal.pone.0313401.

7. Gartlehner G, Kugley S, Crotty K, Viswanathan M, Dobrescu A, Nussbaumer-Streit B, et al. Artificial intelligence-assisted data extraction with a large language model: a study within reviews. *Ann Intern Med.* 2025;178(12):1763–1771. doi:10.7326/ANNALS-25-00739.
8. Lai H, Liu J, Bai C, et al. Language models for data extraction and risk of bias assessment in complementary medicine. *npj Digit Med.* 2025;8:74. doi:10.1038/s41746-025-01457-w.
9. Shankar R, Lim A, Qian X. Performance of large language models in data extraction for evidence synthesis: a systematic review. *J Biomed Inform.* 2026;181:105086. doi:10.1016/j.jbi.2026.105086.
10. Laignelot F, Martin GL, Ossman M, Pigeon O, Boubaker A, Picovschi E, et al. Large language models show promising performance for some systematic review tasks but call for cautious implementation: a systematic review. *J Clin Epidemiol.* 2026;194:112221. doi:10.1016/j.jclinepi.2026.112221.
11. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557–560. doi:10.1136/bmj.327.7414.557.
12. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ.* 2011;342. doi:10.1136/bmj.d549.
13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;372. doi:10.1136/bmj.n71.
14. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med.* 2011;155(8):529–536. doi:10.7326/0003-4819-155-8-201110180-00009.
15. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al.; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ.* 2008;336(7650):924–926. doi:10.1136/bmj.39489.470347.AD.
16. Hedges LV, Tipton E, Johnson MC. Robust variance estimation in meta-regression with dependent effect size estimates. *Res Synth Methods.* 2010;1(1):39–65. doi:10.1002/jrsm.5.
17. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1–48. doi:10.18637/jss.v036.i03.
18. Pustejovsky JE, Tipton E. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J Bus Econ Stat.* 2018;36(4):672–683. doi:10.1080/07350015.2016.1247004.
19. R Core Team. *R: A Language and Environment for Statistical Computing*. Version 4.6.1. Vienna, Austria: R Foundation for Statistical Computing; 2026.
20. Liu J, Lai H, Zhao W, et al. AI-driven evidence synthesis: data extraction of randomized controlled trials with large language models. *Int J Surg.* 2025;111(3):2722–2726. doi:10.1097/JS9.0000000000002215.
21. Vivekanantha P, Kahlon H, Balogun OT, et al. Automated data extraction for systematic reviews using GPT-5.2 and Google Gemini Pro 3: a dual-large language model approach in orthopaedic research. *Knee Surg Sports Traumatol Arthrosc.* 2026;34(8):3011–3029. doi:10.1002/ksa.70412.
22. Felizardo KR, Steinmacher I, Lima MS, Deizepe A, Conte TU, Barcellos MP. Data extraction for systematic mapping study using a large language model: a proof-of-concept study in software engineering. In: *Proceedings of the 18th ACM/IEEE International Symposium on*

Empirical Software Engineering and Measurement (ESEM 2024). 2024. p. 407–413.
doi:10.1145/3674805.3690743.

23. Mahmoudi H, Chang D, Lee H, Ghaffarzadegan N, Jalali MS. Critical assessment of large language models' (ChatGPT) performance in data extraction for systematic reviews: exploratory study. *JMIR AI*. 2025;4. doi:10.2196/68097.
24. Stewart-Evans J, Wilson E, Langley T, Prayle A, Hands A, Exley K, et al. Expediting data extraction using a large language model (LLM) and scoping review protocol: a methodological study within a complex scoping review. *arXiv*. 2025. doi:10.48550/arXiv.2507.06623.
25. Caponio VCA, Lorenzo-Pouso AI, Magalhaes M, et al. Accuracy of LLMs to retrieve numeric data for meta-analysis in dentistry. *J Dent*. 2026;164:106245. doi:10.1016/j.jdent.2025.106245.
26. Peng Z, Wu X, Qin Z, et al. Accuracy of large language models in data extraction from randomized controlled trials in sleep medicine: a proof-of-concept study. *Sleep Med Rev*. 2025;84:102192. doi:10.1016/j.smrv.2025.102192.
27. Kataoka Y, Takayama T, Yoshimura K, So R, Tsujimoto Y, Yamagishi Y, et al. Automating the data extraction process for systematic reviews using GPT-4o and o3. *Res Synth Methods*. 2026;17(1):42–62. doi:10.1017/rsm.2025.10030.
28. Oami T, Okada Y, Maeda K, Nakada TA. Performance of large language models and prompt engineering strategies for data extraction in systematic reviews. *Front Digit Health*. 2026;8:1799623. doi:10.3389/fdgth.2026.1799623.
29. Zhang Y, Luo N, Kim H, Li L, Gao L, Han J, et al. Large language models (LLMs) for evidence synthesis: an exploratory evaluation and a new approach for automated data extraction and validation. *PsyArXiv*. 2026. doi:10.31234/osf.io/udysp_v2.
30. Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. *npj Digit Med*. 2025;8:509. doi:10.1038/s41746-025-01840-7.
31. Rokhshad R, Motie P, Shirani M, Sameie A, Revilla-León M. Performance of large language models conducting systematic review tasks in prosthodontics. *J Prosthet Dent*. 2026;136(1):81–87. doi:10.1016/j.prosdent.2026.02.009.
32. Daraqel B, Owayda A, Khan H, Koletsi D, Mheissen S. Artificial intelligence as a tool for data extraction is not fully reliable compared to manual data extraction. *J Dent*. 2025;160:105846. doi:10.1016/j.jdent.2025.105846.
33. Kartchner D, Ramalingam S, Al-Hussaini I, Kronick O, Mitchell C. Zero-shot information extraction for clinical meta-analysis using large language models. In: *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Association for Computational Linguistics; 2023. p. 396–405. doi:10.18653/v1/2023.bionlp-1.37.
34. Jansen T, Liebenow LW, Mertens U, Schmidt FTC, Lohmann JF, Fleckenstein J, et al. Data extraction by generative artificial intelligence: assessing determinants of accuracy using human-extracted data from systematic review databases. *Psychol Bull*. 2025;151(10):1280–1306. doi:10.1037/bul0000501.
35. Schroeder NL, Jaldi CD, Zhang S. Large language models with human-in-the-loop validation for systematic review data extraction. *arXiv*. 2025. doi:10.48550/arXiv.2501.11840.
36. Susnjak T, Hwang P, Reyes NH, Barczak ALC, McIntosh T, Ranathunga S. Automating research synthesis with domain-specific large language model fine-tuning. *ACM Trans Knowl Discov Data*. 2025;19(3) 68. doi:10.1145/3715964.

37. Ferreira Barros C, Kalinowski M, Kassab M, Graciano Neto VV. On the use of a large language model to support the conduction of a systematic mapping study: a brief report from a practitioner's view. In: *Proceedings of the 2026 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE '26)*. ACM; 2026. doi:10.1145/3786149.3788298.
38. Bianchi J, Hirt J, Vogt M, Vetsch J. Data extractions using a large language model (Elicit) and human reviewers in randomized controlled trials: a systematic comparison. *Cochrane Evid Synth Methods*. 2025;3. doi:10.1002/cesm.70033.
39. Helms Andersen T, Marcussen TM, Termannsen AD, Lawaetz TWH, Nørgaard O. Using artificial intelligence tools as second reviewers for data extraction in systematic reviews: a performance comparison of two AI tools against human reviewers. *Cochrane Evid Synth Methods*. 2025;3(4). doi:10.1002/cesm.70036.
40. Murton M, Boulton E, Cross S, et al. Harnessing large-language models for efficient data extraction in systematic reviews: the role of prompt engineering. *Cochrane Evid Synth Methods*. 2025;3(6). doi:10.1002/cesm.70058.
41. Fan S, Chen M, Doi SA, et al. Evaluating data extraction error by a large language model from randomised controlled trials: a large-scale empirical study. *BMJ Evid Based Med*. 2026. doi:10.1136/bmjebm-2025-114044.
42. Kim JK, Chua ME, Li TG, Rickard M, Lorenzo AJ. Novel AI applications in systematic review: GPT-4 assisted data extraction, analysis, review of bias. *BMJ Evid Based Med*. 2025;30(5):313–322. doi:10.1136/bmjebm-2024-113066.
43. Lamanna M, Muca E, Giannone C, Bovo M, Boffo F, Romanzin A, et al. Artificial intelligence meets dairy cow research: large language model application in extracting daily time-activity budget data for a meta-analytical study. *J Dairy Sci*. 2025;108:10203–10219. doi:10.3168/jds.2025-26385.
44. Kasireddy E, Chow C, Collet J, Pourrahmat MM, Fazeli MS. Evaluating the performance of Claude 3.7 Sonnet in data extraction automation for systematic literature reviews. *Value Health Reg Issues*. 2026;53:101539. doi:10.1016/j.vhri.2025.101539.
45. Yerramilli A, Blagojevic C, Bryce A, et al. A proof-of-concept performance evaluation of six large language models for semiautomated data extraction in a systematic review on infectious diseases clinical trials. *CMI Commun*. 2026;3(3):105214. doi:10.1016/j.cmicom.2026.105214.
46. Konet A, Thomas I, Gartlehner G, Kahwati L, Hilscher R, Kugley S, et al. Performance of two large language models for data extraction in evidence synthesis. *Res Synth Methods*. 2024;15(5):818–824. doi:10.1002/jrsm.1732.
47. 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;15(4):616–626. doi:10.1002/jrsm.1715.
48. Achterberg J, van Dijk B, Meng J, et al. OpenExtract: automated data extraction for systematic reviews in health. *Stud Health Technol Inform*. 2026;336:934–938. doi:10.3233/SHTI260316.
49. Muthu S, Kolarpatti Ponnusamy DV, Viswanathan VK, Rajappan Chandra SK, Sharun K. A locally executed agentic artificial intelligence framework for deduplication, screening, and structured data extraction in spine surgery systematic reviews. *JBJS Open Access*. 2026. doi:10.2106/JBJS.OA.26.00102.

50. Ye A, Maiti A, Schmidt M, Pedersen SJ. A hybrid semi-automated workflow for systematic and literature review processes with large language model analysis. *Future Internet*. 2024;16(5):167. doi:10.3390/fi16050167.
51. Sanlier N, Sulaimanov U, Moniri A, et al. How often do large language models agree with each other—and with the truth? A consensus- and complexity-stratified analysis of data extraction for neuroimaging AI. *J Clin Med*. 2026;15:6005. doi:10.3390/jcm15156005.
52. Mitchell E, Are EB, Colijn C, Earn DJD. Using artificial intelligence tools to automate data extraction for living evidence syntheses. *PLoS One*. 2025;20(4). doi:10.1371/journal.pone.0320151.
53. Ishikane M, Kataoka Y, Tsujimoto Y, et al. Efficacy of a large language model data extraction system in evidence reviews for emerging infectious diseases: a randomized crossover trial. *Open Forum Infect Dis*. 2026;13(7). doi:10.1093/ofid/ofag401.
54. Jansen T, Liebenow LW, Schaller NJ, Hattie J, Möller J. Automated data extraction by large language models: assessing accuracy in comparison to human experts using the example of Visible Learning. *Educ Psychol Rev*. 2026;38:38. doi:10.1007/s10648-026-10136-5.
55. Yisha Z, Zou P, Li S, Zhang L, Guo L, Gu A, et al. Assessing data extraction in randomized clinical trials with large language models. *BMC Med Res Methodol*. 2026;26:33. doi:10.1186/s12874-025-02729-5.
56. Li L, Mathrani A, Susnjak T. What level of automation is “good enough”? A benchmark of large language models for meta-analysis data extraction. *Res Synth Methods*. 2026;17(4):671–692. doi:10.1017/rsm.2025.10066.
57. Meliante LA, Coco G, Rabiolo A, De Cillà S, Manni G. Evaluation of AI tools versus the PRISMA method for literature search, data extraction, and study composition in glaucoma systematic reviews: content analysis. *JMIR AI*. 2025;4. doi:10.2196/68592.
58. Wang X, Huey SL, Sheng R, Mehta S, Wang F. SciDaSynth: interactive structured data extraction from scientific literature with large language model. *Campbell Syst Rev*. 2025;21(4). doi:10.1002/cl2.70073.
59. Khan MA, Ayub U, Naqvi SAA, et al. Collaborative large language models for automated data extraction in living systematic reviews. *J Am Med Inform Assoc*. 2025;32(4):638–647. doi:10.1093/jamia/ocae325.

