Agent-Assisted Evidence Synthesis

A Survey

Bili Dong[1]

Draft 2026-08-09 · landing page · survey record

Abstract

AI systems now target every stage of evidence synthesis. This exploratory systematic map organizes 776 included works from a 1,975-row catalog covering 2020–2026, with structured evidence notes for 31 selected works (25 full-text, 5 abstract-only, 1 secondary-only). A four-dimensional, single-pass abstract coding places screening at the centre of the retained map and appraisal and reporting at its thin end; these counts describe the map, not population prevalence or evidence quality. Across the selected studies, reported performance is heterogeneous and common metrics can obscure failures under class imbalance. Guidance repeatedly addresses tool identity, task, human role, configuration, and verification, while four unvalidated instruments cover different subsets across reporting and reproducible storage. Multi-model studies document ensemble and deferral configurations, but no selected evidence record identifies a definition of independent agent reviewers or isolates the source of ensemble gains with a matched design. The survey contributes a consistent terminology, a taxonomy, a scoped synthesis of performance and governance evidence, and a curated reading list. Its search and coding limits, AI assistance, and evidence trail are disclosed in a minimal public record.

1. Introduction

A systematic review is itself a pipeline — search, screening, data extraction, quality appraisal, synthesis, and reporting — and every stage of that pipeline is now a target for automation by large language models (LLMs) and agent systems. The result is a fast-growing methodology literature that evaluates, systematizes, and regulates this automation, scattered across medicine (where evidence synthesis is core infrastructure), software engineering (SE, which imported the method), and general venues.

This paper surveys that literature as an exploratory systematic map (Petersen et al., 2008). Its standing searches center language models; adjacent AI/ML automation entered through broader initial screening and citation chasing. The result aims at consistent vocabulary and classified coverage of the retained catalog, not exhaustive retrieval or pooled effect estimates. It asks four questions:

- **RQ1:** How are retained works distributed across workflow stage, contribution type, evidence type, and setting?
- **RQ2:** What do the selected evidence notes show about performance measurement and reported results?
- **RQ3:** What norms and disclosure instruments appear among the selected evidence notes?
- **RQ4:** Among the selected evidence notes, which designs address reviewer independence and multi-model ensembling?

The paper contributes:

[1]The byline names the accountable human author, who directed and gated the work and takes responsibility for its content. OpenAI Codex systems (through GPT-5.6 Sol) provided substantial assistance with evidence organization, adversarial review, synthesis, manuscript drafting and editing, and repository tooling. Anthropic Claude systems (through Fable 5) provided substantial assistance with search, screening, classification, deep reading, synthesis, and manuscript drafting and revision. AI output is not treated as evidence; literature claims rest on the cited sources. The survey record linked in the title metadata documents the working evidence and synthesis trail.

1. a terminology and four-dimensional taxonomy fixed for this survey, with a source-grounded account of which of the field's own terms are contested and which are not (Section 3);
2. a faceted map of 776 included works (2020–2026) under single-pass, unvalidated, abstract-level coding, after resolving known version aliases and retractions (Section 5);
3. a scoped synthesis of what the selected evidence shows about performance measurement (Section 6), norms (Section 7), and multi-model design (Section 8); and
4. a curated, annotated reading list of the works that anchor this map, organized by the taxonomy and maintained on the survey's landing page.

Section 2 positions the survey against prior reviews. Section 3 fixes terminology and the taxonomy. Section 4 describes how the survey was made. Four sections then answer the RQs, Section 9 collects the findings and open problems, and Section 10 states limitations.

2. Background and Related Surveys

The pre-LLM baseline is well documented. van Dinter et al.'s landscape review (2021; 41 automation studies, 2006–2020) found every study in Kitchenham's conducting-the-review phase (2007) — screening dominant, appraisal nearly empty, planning and reporting untouched, and exactly one deep-learning study in the corpus. Napoleão et al.'s cross-domain mapping (2021) quantified the adoption gap between medicine and software engineering: eight practice-adopted screening tools in medicine versus two in SE. The classic method canon — review guidelines (Kitchenham & Charters, 2007), mapping-study procedure (Petersen et al., 2008), snowballing (Wohlin, 2014), and PRISMA 2020 reporting (Page et al., 2021) — predates LLMs and supplies both the vocabulary of this survey and the method now under automation pressure. Earlier primary classifiers are covered mainly through those reviews, but the broad automation vocabulary also admitted several 2020–2022 primary ML works and a 2022 RobotReviewer RCT. The map is LLM-centered but includes adjacent automation lineage without comprehensively searching that broader field.

Four deep-read secondary sources illustrate narrower views of the field. Luo et al. (2024) survey potential LLM roles stage by stage as a viewpoint; the International Collaboration for the Automation of Systematic Reviews reports community progress and evaluation threads (O'Connor et al., 2024); Song et al. (2026) inventory tools for living evidence synthesis; and Madeyski et al. (2026) review 29 LLM-screening evaluations with a methodological focus. Among these four sources, each centers one stage, community, or lifecycle; this survey instead uses one vocabulary to organize stages, settings, and norms. This is a framing difference among the deep reads, not a literature-wide novelty claim.

3. Terminology and Taxonomy

The field's vocabulary is unstable in a specific place. What the stages of a secondary study are **called** is close to shared; what the study itself is called, and how many stages it has, is not. We fix the following usage for this survey and mark where the literature diverges.

The umbrella and the object. We use *evidence synthesis* as the umbrella genre — covering systematic reviews, systematic maps, scoping reviews, rapid reviews, and living reviews alike — and *secondary study* for an individual work of that kind, over primary studies. The literature supplies at least four competing umbrellas: evidence synthesis in the Cochrane-adjacent line (Gartlehner et al., 2025), knowledge synthesis (Hamel et al., 2021), secondary studies in software engineering (Napoleao et al., 2021), and systematic literature studies (Wohlin, 2014). The object nouns vary as much: (Kitchenham & Charters, 2007) governs the systematic literature review and admits mapping studies only by contrast, while (Petersen et al., 2008) argues the systematic map is a genre in its own right; (Page et al., 2021) governs systematic reviews and treats **living** as a mode rather than a genre. This survey is an updatable systematic map in (Petersen et al., 2008)'s sense: *evidence synthesis* names its subject, while *systematic map* names its study type.

Stages. We use six: *search* (query design and study identification), *screening* (title/abstract and full-text selection), *extraction* (structured data capture), *appraisal* (quality and risk-of-bias assessment), *synthesis* (qualitative or quantitative aggregation), and *reporting* (writing and disclosure). This six-stage set is our collapse of the canon rather than any source's own scheme: (Kitchenham & Charters, 2007) specifies twelve steps in three phases, (Petersen et al., 2008) five steps — one of them, keywording of abstracts, with no counterpart in review vocabulary — and the

LLM-era stage models run from four phases (Song et al., 2026) to eight (Degen et al., 2024) and nine (Fernandes et al., 2026). The substantive disagreements are boundary questions — whether protocol development, registration, discrepancy resolution, or publication update are stages of their own — not disputes about what the shared stages are named.

The vocabulary comparison draws on the sources cited below and the method canon; it is not an audit of every selected note. Those sources do not support a community-specific split in stage vocabulary: both communities use *screening* and *selection* (the distinction is stage-outcome versus operation, not medicine versus software engineering), and no extraction-versus- collection dispute appears. The one community difference we can put a count behind is metric vocabulary, not stage vocabulary — WSS@95% appears in eleven medicine studies against one in software engineering (Napoleao et al., 2021).

One genuine term split does survive that test, and it sits at appraisal. The software-engineering canon says quality assessment (Kitchenham & Charters, 2007; van Dinter et al., 2021); the Cochrane line says risk of bias (Arno et al., 2022; Hirt et al., 2021; Huang et al., 2026; Rose et al., 2025); and a third sense scores appraisal instruments on reviews and trials rather than judging bias at all — PRISMA, AMSTAR, and PRECIS-2 in (Woelfle et al., 2024). Our single `appraise` facet spans all three, so a row coded `appraise` should not be read as a risk-of-bias judgment specifically.

Assistance configurations. We use *LLM assistance* for a single model performing a bounded task under prompting, and *agent* loosely for an LLM-based system acting with some autonomy inside the pipeline. Multi-model designs occurring in the literature: *ensembles* (multiple models vote, or a union/OR rule pools their includes), *agreement-gated deferral* (decisions where model and human agree are accepted; disagreements go to a second human), and *end-to-end systems* (agentic pipelines spanning most or all stages, typically with task decomposition and human gates). The guidance literature distinguishes the AI-as-*secondary-reviewer* role (quality-assurance checks on human decisions — a repeatedly sanctioned configuration in the selected guidance) from AI-as-*primary-reviewer* (the model makes first-pass decisions a human oversees).

Taxonomy. The map assigns each row one abstract-level primary-focus value on four dimensions (Table 1). The scheme was built by keywording — extracting candidate facet values from abstracts (Petersen et al., 2008) — and allowed to evolve during classification. Stage uses the primary objective or evaluation endpoint; *meta* takes works about the review process or field, and *end-to-end* requires at least four operational stages. The evidence facet records the comparison described in an abstract, not study quality or certainty.

Dimension	Values
Primary-focus stage	search, screen, extract, appraise, synthesize, report; *end-to-end* for whole-pipeline systems; *meta* for works about the process or field as a whole (surveys, guidance, community reports)
Contribution	method, system/tool, evaluation/benchmark, guideline/norms, position
Evidence label	human/reference-label agreement, benchmark-only, none; not a quality rating
Setting	medicine and evidence-based medicine, software engineering, general

Table 1: The four-dimensional, single-valued abstract-coding scheme.

4. How This Survey Was Made

Agent passes performed search, screening, classification, and deep reading under an author-set protocol and human gates. The title-page note owns the authorship, assistance, evidence-use, and public-record disclosure; this section records only the procedure and its limits.

4.1. Coverage and selection

The map covers English-language work from 2020 through 2026. Searches ran against OpenAlex, Crossref, Semantic Scholar, and arXiv, relevance-sorted and capped at 50 results per query. The initial eleven-query set had ten successful queries; a later seventeen-query set added mapping-study, scoping-review, and living-evidence vocabulary and had sixteen successful queries. The standing searches center language models. Broader AI, ML, deep-learning, and automation work entered through initial screening vocabulary and citation chasing, so the retained map is LLM-centered rather than a comprehensive automation denominator.

The initial 412 candidates received two screening passes with adjudication. One backward and forward citation round (Wohlin, 2014) then used a title-vocabulary pre-filter; the 881 retained candidates received one screening pass plus a verification pass over includes. The later batch screened 585 candidates with an eligibility-first and an exclusion-first pass, adjudication, and a human gate. Twelve designated critical works were chased in both directions. Three defective backward-index results were replaced by publisher-deposited reference lists. The searches and chases were not iterated to saturation.

4.2. Coding and evidence

Every include received one primary-focus value on each taxonomy dimension from a truncated abstract. The coding was single-pass and unvalidated. Thirty-one works were selected purposively for evidence notes to cover the research questions, taxonomy contrasts, closest prior maps, quantitative anchors, disclosure instruments, and multi-model designs. Selection was iterative, with no fixed score, random sample, or saturation rule. Twenty-five notes are full-text, five abstract-only, and one secondary-only; the secondary-only note supports no finding. The catalog's evidence facet records the comparison a work claims or plans, not study quality and not confirmation that the comparison was completed.

The record retains the exact standing queries, current dispositions, source notes, syntheses, claim/evidence bindings, and an append-only event log. It does not retain unfiltered result sets, departing from Kitchenham and Charters' save-for-reanalysis guidance (Kitchenham & Charters, 2007). Historical campaign phases survive only as reconciled aggregates because candidate-level provenance was pruned; later rows carry decided keys. Seventy-six catalog rows remain parked for re-screening on a future update.

Table 2 and Table 3 separate retained historical aggregates from the candidate-level later batch. Their quantities describe this record, not retrieval completeness.

Campaign phase	In	Out
Search (11-query set; 10 succeeded)	—	419 unique
Dedup (arXiv–DOI)	419	412
Dual-pass screen + adjudication	412	139 includes
Snowball round (pre-filtered)	139 seeds	1,204 new
Vocabulary pre-screen	1,204	881
Wave-2 screen + verification	881	533 includes
Campaign close	1,291 rows	672 included rows
Integrity correction	672 included rows	646 works

Table 2: Historical campaign funnel through 2026-08-08. Quantities are reconciled aggregates because candidate-level provenance was pruned.

Update and current ledger	In	Out
Search (17-query set; 16 succeeded)	—	517 rows
Critical-set chases (24 + 3 primary)	12 seeds	1,191 rows
Vocabulary pre-screen (chase rows)	1,191	595 retained
Dedup, enrichment, park	1,112	585 screened
Dual-pass screen + adjudication	585	132 includes

Update and current ledger	In	Out
Date-rule coding (primary chases)	41	39 before-window
Integrity (aliases, artifacts)	26 rows	13 E6, 13 dropped
Current catalog	1,975 rows	**776 include-level; 1,123 excluded; 76 parked**
Post-ledger outputs	776	facet map; 31 evidence notes

Table 3: Candidate-level update through 2026-08-09, recoverable from the event log. Its 132 include decisions produced 130 new include-level rows after one rediscovery and one deep-read reclassification.

5. RQ1 — Landscape of the retained map

Stage	Med	SE	Gen	Total	Hum	Bench	None
Search	32	2	30	64	15	19	30
Screen	153	10	93	256	141	64	51
Extract	90	2	29	121	61	39	21
Appraise	30	0	3	33	21	5	7
Synthesize	25	2	22	49	14	9	26
Report	12	3	8	23	9	4	10
End-to-end	54	2	40	96	32	10	54
Meta	67	2	65	134	10	8	116
Total	**463**	**23**	**290**	**776**	**303**	**158**	**315**

Table 4: Single-pass abstract coding: primary-focus stage × setting (medicine, SE, general) and stage × evidence label (human/reference-label agreement, benchmark-only, none). The evidence labels are not quality assessments.

Table 4 summarizes the retained map. Screening is the primary-focus label for 256 of 776 works, about a third; reporting (23) and appraisal (33) remain the smallest categories. The map also contains 47 guideline-contribution works and 96 end-to-end rows. Van Dinter et al.'s differently scoped 2006–2020 review has no directly comparable categories for those values (van Dinter et al., 2021). By primary contribution, evaluation/benchmark works are most common (363), followed by methods (171), systems/tools (132), positions (63), and guidelines (47). These are unvalidated abstract-level labels, so their exact differences describe the retained coding rather than population prevalence.

Among the deep reads, MedSR-Copilot (Huang et al., 2026) is a preprint evaluated on its authors' own benchmark. Its four subagents, fine-tuned risk-of-bias model, and deterministic synthesis engine reached 63.6% end-to-end conclusion accuracy against a 45.3% best baseline on 100 reviews. The system uses task decomposition, structured intermediate artifacts, and human review, with no debate, voting, or agent redundancy. Its tested ablations attribute −14.9 percentage points to removing two-stage extraction, −7.6 to removing tournament reranking, and −3.3 to removing retrieval augmentation (its baselines were author-run); human review and redundancy were not ablated, so the study does not establish either as the causal source of performance. Its screening F1 of about 0.44–0.51 also qualifies the meaning of "end-to-end." In Song et al.'s 34-tool living-evidence inventory, one tool served the publication-update phase; the authors warn that their "living evidence" terminology may undercount that phase (Song et al., 2026).

The setting facet assigns 463 works to medicine and 23 to software engineering. These incommensurable publication counts do not show that the practice-adoption gap measured by Napoleão et al. has widened (Napoleao et al., 2021); they show only this map's medicine-heavy distribution. Exactly one included row is jointly coded `setting=se` and `contribution=guideline`, and it is a practitioner experience report rather than an instrument; no software-engineering row is coded `appraise` at all. Mapping relayed by the field's coordination body found AI use rarely disclosed in the education and climate/health domains it covered (O'Connor et al., 2024). Mughal et al. provide one detailed disclosed-adoption exemplar among the deep reads (Mughal & Bilal, 2026).

6. RQ2 — Performance and measurement

The abstract-coded map labels 303 of 776 works `human-agree` and 158 `benchmark` (Table 4). These labels record the comparison a work **claims or plans**, not one verified to have been performed: one deep read coded `human-agree` turns out to be a protocol whose comparison has not yet been run (Rose et al., 2025). They record comparison type, not risk-of-bias or certainty, and the survey performed no formal quality appraisal. The deep reads support narrower claims about measurement practice and generalization.

Measurement practice in the selected evidence. Madeyski et al.'s convenience sample of 29 LLM-screening evaluations found 24% reporting complete confusion matrices, 10% reporting the Matthews correlation coefficient (MCC), and 59% reporting accuracy. In one 9,695-record reanalysis, the accuracy-best model lost 63.3% of relevant evidence where the authors' cost-weighted choice lost 5.8% (Madeyski et al., 2026). Within SESR-Eval, pooled versus per-review aggregation also changed model comparisons on the same data (Huotala et al., 2025). The starkest demonstration needs no model comparison at all: assessing 190 nursing trials with one tool against Cochrane judgments, Hirt et al. report Cohen's κ of 0.60 for allocation concealment, 0.52 for randomization, 0.43 for blinding of personnel, and 0.04 — near chance — for blinding of outcome assessors, alongside sensitivity spanning 0.44–0.88 and positive predictive value 0.25–0.79 (Hirt et al., 2021, abstract-only). Which domain is reported can therefore move the agreement verdict from near-chance to moderate within a single tool and corpus; sensitivity and predictive value expose different failure modes in the same data.

What individual studies show. In SESR-Eval, no nondegenerate operating point among nine tested models across 24 SE reviews met the authors' proposed bar of recall ≥ 0.95 at precision about 0.50 (Huotala et al., 2025). In one Gargari et al. review, prompt variants moved GPT-3.5 sensitivity from 62% to a level the authors compared with a junior reviewer (Kohandel Gargari et al., 2023), though most of that work's per-prompt figures are directional prose in an unverified supplement with no significance testing. Across five highly imbalanced sepsis questions under one prompt, Oami et al. found GPT-4 Turbo specificity of 0.98 versus 0.51 for GPT-3.5, while sensitivity changed from 0.83 to 0.85 without a significant difference (Oami et al., 2025). Within SESR-Eval, study effects exceeded differences among the larger tested models (Huotala et al., 2025). In Syriani et al.'s high-conflict MobileMDE corpus (52.7% recorded human conflict), recall was 0.327 versus 0.738–0.947 in the other four corpora; the association does not establish conflict as the cause (Syriani et al., 2024). Across Woelfle et al.'s three appraisal instruments, each individual LLM scored below each individual human and human inter-rater κ ranged from 0.84 to 0.29 (Woelfle et al., 2024). Among the selected extraction deep reads are an SE proof-of-concept reporting 87.8% accuracy (Felizardo et al., 2024, abstract-only), a 23-study social-science inventory with no pooled benchmark (Legate et al., 2024), and secondhand error ranges of 4–31% (Gartlehner et al., 2025).

Read against human baselines. Fagerberg et al. (2025) summarize prior estimates of single-human-reviewer screening sensitivity at ~87–92% (range 42–100%). Gartlehner et al. (2025) cite prior reports that human extraction errors reach 50% of data elements. A rare RCT-grade automation study found noninferiority, not superiority, with inconclusive time savings (Arno et al., 2022, abstract-only); it remains the only randomized evaluation among the selected evidence notes. The map's facets do not record study design, so we did not check the wider catalog for others. Agent evidence should be calibrated against these imperfect baselines, not an idealized perfect reviewer.

The metric prescription remains unsettled. Oami et al. treat specificity as the deciding workload metric when sensitivity is comparable (Oami et al., 2025). Madeyski et al. reject specificity as a primary metric under class imbalance — an exclude-everything classifier can score it perfectly — and instead require the confusion matrix, lost evidence, and MCC or cost-weighted MCC (Madeyski et al., 2026). Both positions reject recall alone, but they prescribe different ways to trade missed evidence against downstream screening work. This survey does not resolve that methodological disagreement.

Ensemble gains, in incommensurable units. Two selected deep reads report ensembles beating their own best member, and their effect sizes cannot be placed on one scale. A cross-vendor three-agent vote reached mean average precision 0.341 against constituents at 0.271, 0.266, and 0.182, with WSS@95% of 0.680 (Akinseloyin et al., 2026); a five-model same-family BERT ensemble reached F1 89.16% against a best standalone 88.53% on document-type triage at roughly 70/30 balance, against labels from a single crowdsourced annotation team (Knafou et al., 2023) — a less imbalanced task than the screening evidence above. Relative gain, absolute F1 points, and ranking-

based precision are different measurement families, and the metric-fragmentation problem this section documents within studies recurs between them.

The second of those studies also repeats a pattern first visible in appraisal: requiring unanimity plus a probability threshold lifts triage F1 to about 98.5% at roughly 99% recall while deciding only about half the corpus (Knafou et al., 2023), just as consistency-gated appraisal reached accuracy whose confidence interval merely overlapped human performance, and only on the items it did not defer (Woelfle et al., 2024). Higher conditional performance bought with deferred coverage appears at two stages and in two technology generations, which makes it a shape worth naming — though the two designs share no corpus, metric, or model family, so this is a recurring pattern rather than a replication.

7. RQ3 — Norms and disclosure instruments

The guidance sources repeatedly address tool identity and version, stage or task, human role, configuration, and verification, but no one item set is common to every instrument. FRAISR (Degen et al., 2024) records only stage, tool name/version, and input parameters; PRISMA-trAIce (Holst et al., 2025) and HAICO-SLR (Fernandes et al., 2026) add human-role and oversight fields. Among the conduct guidance, the Cochrane-family statements (Gartlehner et al., 2025) and HAICO-SLR (Fernandes et al., 2026) keep a human decision in every stage and add human accountability (no AI authorship). The pre-LLM screening guidance (Hamel et al., 2021) is narrower and makes a distinction later guidance can blur: it labels fully autonomous score-threshold screening an inappropriate use, while separately ranking options for the unscreened remainder after a human-chosen truncation point, with AI-only exclusion the highest-risk option. In the Cochrane-family guidance the sanctioned role for AI is a *secondary* quality-assurance reviewer, re-checking single-reviewer exclusions and extractions; HAICO-SLR goes further, sanctioning AI first-pass filtering and drafting under human validation (Fernandes et al., 2026).

Three unvalidated disclosure proposals coexist: PRISMA-trAIce's 14 items and human/AI-split flow diagram (Holst et al., 2025), FRAISR's per-stage machine-readable table (Degen et al., 2024, preprint), and HAICO-SLR's dual conduct-and-reporting tables (Fernandes et al., 2026, preprint). A fourth instrument fragments the picture along a second axis rather than competing on the first: the RDAL checklist prescribes what an active-learning-aided review must **store** — random seeds, labeling order, per-iteration model identity and training-set size — rather than what its authors must report, motivated by the observation that storing every relevance score scales quadratically with corpus size (Lombaers et al., 2024). It is equally unvalidated; its only application is a worked example against a tool two of its three authors develop; in that example the tool stores nine of the fifteen items outright and one partly, leaves four to author pre-registration, and treats one as optional (item numbering reconstructed from the paper's prose, since the checklist table was not machine-readable). A layer of position statements sits beside them: the Cochrane-family statements endorse the RAISE guidance (Gartlehner et al., 2025). The proposals disagree on PRISMA-AI's history: Holst et al. describe it as announced but unpublished, whereas Fernandes et al. say it was never developed (Fernandes et al., 2026; Holst et al., 2025). They agree on the point that matters here — no usable PRISMA-AI instrument was available. PRISMA 2020 itself asks for automation details in study selection and data collection (Page et al., 2021); Luo et al. characterize that coverage more narrowly as screening only (Luo et al., 2024). The three proposals' evidence notes report no validation or adoption evidence; coexistence alone does not establish a standards race. O'Connor et al. relay rare disclosure in the education and climate/health domains they discuss (O'Connor et al., 2024). One detailed practice exemplar among the deep reads names the model, cites a PRISMA item, publishes a validation table beside the flow diagram, and revisits residual risk in limitations (Mughal & Bilal, 2026).

Where the map can watch norms actually being made, the difficulty is visible in the panel's own report of where its agreement failed. A three-round Delphi of 29 experts on living evidence synthesis reached consensus on 19 of 23 statements, but its authors report that agreement ran lowest precisely on the objective, actionable use of automation and digital tools, and highest on general statements (Golob et al., 2025, preprint) — the authors' own characterization of their round data rather than a published vote breakdown, since the statement tables were not machine-readable. The panel's only automation-relevant statement, that software and automation should be validated and justified, did reach consensus and is neutral on the role question these guidance sources disagree about. The one place this map watches norms being formed produced an operational commitment only at that general validate-and-justify level.

8. RQ4 — Reviewer independence and multi-model design

No selected evidence record identifies a definition of what makes two agent passes *independent* in the sense dual human review requires — a criterion for when one pass counts as independent of another (Hamel et al., 2021). The nearest observed construction is procedural rather than architectural: a registered protocol stipulates that two different people each run the model in separate sessions (Rose et al., 2025), defining operator independence while leaving sampling, context isolation, and session leakage unaddressed. The selected set contains one correlation measurement (Akinseloyin et al., 2026), but neither search wave ran a targeted independence query or coded an independence facet, so this cannot establish a literature-wide absence.

The evidence beneath that finding spans designed multi-model and human–model comparisons, indirect signals, and one system with no redundancy. This is an exploratory subset, not an exhaustively searched class:

- A cross-vendor OR ensemble (GPT-5 Thinking + Gemini 2.5 Pro, two runs each, 736 Cochrane citations) reached 99.7% sensitivity and 49.3% specificity on the authors' adjudicated labels; against the original Cochrane labels its sensitivity was 94.0–94.5% (Fagerberg et al., 2025). All 18 relabels ran Include⟶Exclude, were decided by two same-team adjudicators, and every one of them moved measured sensitivity upward; the authors flag the specificity figure as a conservative lower bound. The preprint's post-hoc vaccine subgroup had seven positives: GPT-5 sensitivity was 43% while Gemini flagged the same ambiguous records. The design has no matched same-family arm and does not isolate model family, model identity, duplicate runs, and the OR rule as causes.
- A 9-run consistency ensemble reached appraisal accuracy whose confidence interval overlapped human performance, only on items surviving near-unanimous agreement, deferring 74–88% of items; the *deferral* design — human + LLM score, send disagreements to a second human — beat humans-alone in 8 of 10 human–LLM pairings on the two easier retrospective appraisal instruments (best pairs 95–96% accuracy, range 89–96%) while sparing ~65–70% of second-reviewer item count; on PRECIS-2 it beat humans-alone in 1 of 10 pairings, reaching 80–86% while sparing about 29% (Woelfle et al., 2024). The reference was two-rater consensus, prompts differed by model, and time savings were not measured. This supports deferral for these instruments, not a general design law.
- An abstract-only human–AI framework spans screening through thematic analysis and routes decisions by an AI-confidence threshold under human oversight (Brîncoveanu et al., 2025, abstract-only). It reports no extractable threshold, error, workload, or human-checkpoint values, so it establishes a configuration rather than its effectiveness.
- A cross-vendor three-agent vote (GPT-4o Mini, Claude 3 Haiku, Gemini 1.5 Flash, with a fourth model adjudicating) beat every constituent — mean average precision 0.341 against 0.271, 0.266, and 0.182 — and beat its own debate variants; one adjudication variant matched it (0.345) at many times the cost (Akinseloyin et al., 2026). Alone among the selected evidence records it measures a proxy for independence, reporting Spearman correlations of 0.48–0.56 between its agents' scores and concluding that model heterogeneity is what makes aggregating weak screeners work. Its three agents differ in vendor, size, and training corpus at once, with no same-family arm, so the causal claim outruns the design.
- A five-model same-family ensemble reached F1 89.16% against a best standalone 88.53% and reports no correlation or error-overlap statistic between its members at all (Knafou et al., 2023) — a configuration documented without a mechanism.
- An abstract reports that open-weight models screened more conservatively than GPT-4.1 across 25k titles — a model-behavior contrast awaiting full-text numbers (Safarpour et al., 2026, abstract-only).
- High self-consistency coexists with mediocre accuracy — run-to-run Fleiss κ 0.82–0.97 on the two corpora tested (Syriani et al., 2024) — stability is not validity.
- One selected end-to-end preprint uses no redundancy (Huang et al., 2026); that architecture is descriptive and supplies no comparison of independence mechanisms.

The measurement that arrived supplies the set's clearest association between score diversity and ensemble performance. Allowing that study's agents to debate raised their inter-agent correlation and lowered ensemble performance relative to parallel voting. Score diversity therefore tracks performance in that comparison, but debate also changes the interaction itself, so the design isolates neither independence nor error correlation as a

cause. No selected evidence record reports the quantity a theory would need: error correlation measured within versus across model families.

9. Synthesis and open problems

RQ	Finding	Evidence
1	The retained coding is screening-heavy, medicine-heavy, and thin in appraisal and reporting; these are map counts, not adoption measures	map; (Napoleao et al., 2021)
2	Metric choice, aggregation, and study effects materially change performance verdicts; one benchmark found study effects larger than differences among its larger models	(Huotala et al., 2025; Madeyski et al., 2026)
3	Disclosure elements recur without a common item set; four unvalidated instruments divide reporting and reproducible storage	(Degen et al., 2024; Fernandes et al., 2026; Holst et al., 2025; Lombaers et al., 2024)
4	Selected OR-ensemble and human–LLM-deferral results do not supply a matched causal account or a definition of agent-reviewer independence	(Fagerberg et al., 2025; Hamel et al., 2021; Woelfle et al., 2024)
§3	Genre nouns and stage granularity vary more than the recognizable shared stage names; appraisal has the clearest terminology split	registered comparison; canon

Table 5: One bounded synthesis per research question, plus the terminology result from Section 3.

Table 5 condenses descriptive, differently scoped findings. The selected studies show heterogeneous absolute performance, several workflow structures (decomposition, human review, deferral, OR ensembling), and recurring measurement deficiencies under class imbalance. Their designs and datasets are not commensurate enough to rank structure, scale, redundancy, and methodology as causal bottlenecks.

Priorities suggested by the retained map, rather than demonstrated literature absences, include: a targeted definition and measurement study of reviewer *independence* for agents (Section 8); deeper evidence for the small appraisal, synthesis, and reporting categories; work in the map's underrepresented SE setting; publication-update tools beyond the terminology-sensitive inventory (Song et al., 2026); validation and adoption studies for the four instruments; and norms for agent-primary configurations. The Cochrane-family guidance keeps AI as a secondary checker, while HAICO-SLR sanctions first-pass roles under human validation (Fernandes et al., 2026).

10. Limitations

Search coverage is bounded by relevance-sorted top-50 caps, English-only retrieval, and one failed query in each wave's set. The snowball rounds apply a title-vocabulary pre-filter, which reintroduces the terminology dependence snowballing exists to escape (Wohlin, 2014) — and does so asymmetrically, since the filter reads titles only. This is a recall bound on every chase, and the record cannot quantify how many in-scope works it cost. The initial campaign's model-vocabulary list and candidate-level provenance were not retained, and the later search retains logged queries rather than unfiltered result sets.

Registrar defects shaped the update materially. Three of twelve backward chases returned bibliographies the citation index could not supply, requiring publisher-deposited reference lists; 87 screened candidates carried no registrar abstract and were judged on title, venue, and year; 39 rows were coded before-window mechanically from publication years without a screening pass. Seventy-six records remain parked as undecidable.

Screening and classification read truncated abstracts (600–900 characters); every facet was single-pass, with no retained model- or prompt-level audit trail. The screening passes are repeated automated checks under a human gate, not independent human review, so correlated classification errors may survive.

The disclosure exemplar we cite is validated by its own authors, on a stratified sample whose single observed miss carries a wide interval; the randomized appraisal study we lean on completed 7 of 15 recruited teams.

The integrity review resolved known version aliases across both waves but cannot prove that no semantically renamed duplicate remains. The facet labels are not quality assessments. Full-text notes disagree with abstract-level facets for a minority of evidence notes, including one work whose coded evidence label describes a comparison its protocol has not yet performed; the map remains as coded and the notes are authoritative for those works. Evidence extraction had no second pass. The 31 works were selected purposively rather than sampled; five notes are abstract-only and one is secondary-only and supports no finding. This set does not support literature-wide absence claims — RQ4 in particular had no targeted query or coding facet. Medicine accounts for 463 of 776 include-level rows against 23 for software engineering, so thresholds and examples are medicine-calibrated, and terminology claims about software engineering rest on a small deep-read stratum rather than on the map.

11. Conclusion

In this corrected exploratory map, screening is the most common primary-focus label, appraisal and reporting are the smallest, and the catalog is medicine-heavy. The selected performance studies report heterogeneous performance and show that metric choice under class imbalance can change model rankings. Four deep-read instruments across two genres — three disclosure checklists and one reproducible-storage checklist — cover recurring but nonidentical elements, and none of their evidence notes reports validation or adoption. The selected multi-model studies document ensemble and deferral configurations; one measures inter-agent correlation and argues from it that heterogeneity drives the gain, but none isolates the mechanism with a matched design. No selected evidence record identifies a definition of independent agent reviewers, and the one construct the set contains is operator-level rather than agent-level; establishing whether the wider literature does better requires a targeted search.

The map data behind this survey, the per-work evidence notes, and a curated reading list organized by the taxonomy of Section 3 are maintained on the survey's landing page and in the public survey record.

References

Akinseloyin, O., Jiang, X., & Palade, V. (2026). Large language model-based multiagent collaboration for abstract screening toward automated systematic reviews. *Biology Methods and Protocols*, *11*(1). https://doi.org/10.1093/biomethods/bpag006

Arno, A., Thomas, J., Wallace, B., Marshall, I. J., McKenzie, J. E., & Elliott, J. H. (2022). Accuracy and Efficiency of Machine Learning–Assisted Risk-of-Bias Assessments in "Real-World" Systematic Reviews: A Noninferiority Randomized Controlled Trial. *Annals of Internal Medicine*, *175*(7), 1001–1009. https://doi.org/10.7326/m22-0092

Brîncoveanu, C., Carl, K. V., Witzki, A., & Hinz, O. (2025). Augmenting Systematic Literature Reviews: A Human-AI Collaborative Framework. In *KI 2025: Advances in Artificial Intelligence* (pp. 3–17). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-02813-6_1

Degen, B., Vogel, S., & Rzejak, D. (2024). *Leveraging Artificial Intelligence for Systematic Reviews: The FRAISR Reporting Framework and guidance for researchers*. https://doi.org/10.31219/osf.io/ju8dk

Fagerberg, P., Sallander, O., Patil, K. V., Berg, A., Nyman, A., Borg, N., & Lindén, T. (2025). *Dual-Model LLM Ensemble via Web Chat Interfaces Reaches Near-Perfect Sensitivity for Systematic-Review Screening: A Multi-Domain Validation with Equivalence to API Access*. https://doi.org/10.1101/2025.11.03.25339455

Felizardo, K. R., Steinmacher, I., Lima, M. S., Deizepe, A., Conte, T. U., & Barcellos, M. P. (2024). Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '24*, 407–413. https://doi.org/10.1145/3674805.3690743

Fernandes, B., Schleder, M. O., Oliveira, O. C. de, Jaeger, R. de B., Nascimento, L., & Oliveira, F. dos S. de. (2026). *HAICO-SLR Guide: Conducting and Reporting Human-AI Collaboration in Systematic Literature Reviews*. https://doi.org/10.2139/ssrn.7048543

Gartlehner, G., Nussbaumer-Streit, B., Hamel, C., Garritty, C., Griebler, U., King, V. J., Devane, D., & Kamel, C. (2025). Responsible Integration of Artificial Intelligence in Rapid Reviews: A Position Statement From the Cochrane Rapid Reviews Methods Group. *Cochrane Evidence Synthesis and Methods*, *3*(6). https://doi.org/10.1002/cesm.70063

Golob, M. M., Livingstone-Banks, J., Vandvik, P. O., Michaels, M., Hodder, R., Munn, Z., Thomas, J., Rada, G., Guyatt, G., & Nunan, D. (2025). *Best Practice Methods for Living Evidence Synthesis in Health Care: An International Modified Delphi Survey*. https://doi.org/10.1101/2025.11.07.25339719

Hamel, C., Hersi, M., Kelly, S. E., Tricco, A. C., Straus, S., Wells, G., Pham, B., & Hutton, B. (2021). Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01451-2

Hirt, J., Meichlinger, J., Schumacher, P., & Mueller, G. (2021). Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews. *Journal of Nursing Scholarship*, *53*(2), 246–254. https://doi.org/10.1111/jnu.12628

Holst, D., Moenck, K., Koch, J., Schmedemann, O., & Schüppstuhl, T. (2025). Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist. *JMIR AI*, *4*, e80247. https://doi.org/10.2196/80247

Huang, H., Zheng, Q., Qiu, P., Zhao, W., Zhang, Y., Xie, W., Wang, Y., Zhang, X., & Wu, C. (2026). *A PRISMA-Aligned Agentic Framework for Medical Systematic Reviews and Evidence Synthesis*. https://doi.org/10.64898/2026.07.30.26359375

Huotala, A., Kuutila, M., & Mäntylä, M. (2025). SESR-Eval: Dataset for Evaluating LLMs in the Title-Abstract Screening of Systematic Reviews. *2025 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 1–12. https://doi.org/10.1109/esem64174.2025.00053

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering* (Technical Report Nos. EBSE-2007–1).

Knafou, J., Haas, Q., Borissov, N., Counotte, M., Low, N., Imeri, H., Ipekci, A. M., Buitrago-Garcia, D., Heron, L., Amini, P., & Teodoro, D. (2023). Ensemble of deep learning language models to support the creation of living systematic reviews for the COVID-19 literature. *Systematic Reviews*, *12*(1). https://doi.org/10.1186/s13643-023-02247-9

Kohandel Gargari, O., Mahmoudi, M. H., Hajisafarali, M., & Samiee, R. (2023). Enhancing title and abstract screening for systematic reviews with GPT-3.5 turbo. *BMJ Evidence-Based Medicine*, *29*(1), 69–70. https://doi.org/10.1136/bmjebm-2023-112678

Legate, A., Nimon, K., & Noblin, A. (2024). (Semi)automated approaches to data extraction for systematic reviews and meta-analyses in social sciences: A living review. *F1000research*, *13*, 664. https://doi.org/10.12688/f1000research.151493.2

Lombaers, P., Bruin, J. de, & Schoot, R. van de. (2024). Reproducibility and Data Storage for Active Learning-Aided Systematic Reviews. *Applied Sciences*, *14*(9), 3842. https://doi.org/10.3390/app14093842

Luo, X., Chen, F., Zhu, D., Wang, L., Wang, Z., Liu, H., Lyu, M., Wang, Y., Wang, Q., & Chen, Y. (2024). Potential Roles of Large Language Models in the Production of Systematic Reviews and Meta-Analyses. *Journal of Medical Internet Research*, *26*, e56780. https://doi.org/10.2196/56780

Madeyski, L., Kitchenham, B., & Shepperd, M. (2026). LLM4SCREENLIT: Recommendations on assessing the performance of large language models for screening literature in systematic reviews. *Information and Software Technology*, *198*, 108204. https://doi.org/10.1016/j.infsof.2026.108204

Mughal, A. H., & Bilal, M. (2026). LLM-Based Test Oracles: Source-of-Authority Taxonomy – A Systematic Literature Review. *Arxiv Preprint Arxiv:2607.05031*. https://arxiv.org/abs/2607.05031

Napoleao, B. M., Petrillo, F., & Halle, S. (2021). Automated Support for Searching and Selecting Evidence in Software Engineering: A Cross-domain Systematic Mapping. *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, 45–53. https://doi.org/10.1109/seaa53835.2021.00015

O'Connor, A. M., Clark, J., Thomas, J., Spijker, R., Kusa, W., Walker, V. R., & Bond, M. (2024). Large language models, updates, and evaluation of automation tools for systematic reviews: a summary of significant discussions at

the eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, *13*(1). https://doi.org/10.1186/s13643-024-02666-2

Oami, T., Okada, Y., & Nakada, T.-aki. (2025). GPT-3.5 Turbo and GPT-4 Turbo in Title and Abstract Screening for Systematic Reviews. *JMIR Medical Informatics*, *13*, e64682. https://doi.org/10.2196/64682

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008, June). Systematic Mapping Studies in Software Engineering. *Electronic Workshops in Computing*. https://doi.org/10.14236/ewic/ease2008.8

Rose, C. J., Bidonde, J., Ringsten, M., Glanville, J., Berg, R. C., Cooper, C., Muller, A. E., Bergsund, H. B., Meneses-Echavez, J. F., & Potrebny, T. (2025). Using a large language model (ChatGPT) to assess risk of bias in randomized controlled trials of medical interventions: protocol for a pilot study of interrater agreement with human reviewers. *BMC Medical Research Methodology*, *25*(1). https://doi.org/10.1186/s12874-025-02631-0

Safarpour, H., Balogh, G., & Orban, A. J. (2026). Empirical Evaluation of Open Source Large Language Models for Paper Selection: Are LLMs Trustworthy Tools for Scoping Reviews?. *2026 IEEE International Conference on Software Analysis, Evolution and Reengineering - Companion (SANER-C)*, 301–308. https://doi.org/10.1109/saner-c67878.2026.00047

Song, X., Lian, Z., Wang, R., Li, R., Yang, Z., Luo, X., Feng, L., Ma, Z., Pu, Z., Wang, Q., Ge, L., Li, C., Chen, Y., Yang, K., & Lavis, J. (2026). The Phases of Living Evidence Synthesis Using AI: Living Evidence Synthesis (Version 1). *Journal of Medical Internet Research*, *28*, e76130. https://doi.org/10.2196/76130

Syriani, E., David, I., & Kumar, G. (2024). Screening articles for systematic reviews with ChatGPT. *Journal of Computer Languages*, *80*, 101287. https://doi.org/10.1016/j.cola.2024.101287

van Dinter, R., Tekinerdogan, B., & Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, *136*, 106589. https://doi.org/10.1016/j.infsof.2021.106589

Woelfle, T., Hirt, J., Janiaud, P., Kappos, L., Ioannidis, J. P., & Hemkens, L. G. (2024). Benchmarking Human–AI collaboration for common evidence appraisal tools. *Journal of Clinical Epidemiology*, *175*, 111533. https://doi.org/10.1016/j.jclinepi.2024.111533

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, EASE '14*, 1–10. https://doi.org/10.1145/2601248.2601268