Elicit

Elicit for systematic literature reviews

ESG WORKING GROUP · 17 SEPTEMBER 2026

Why systematic review matters to Elicit

01 Elicit grew out of Ought, a nonprofit AI research lab with a mission to scale up good reasoning.

02 When the stakes are high, outcomes may take years to become clear, or there are few precedents to learn from, we need systematic reasoning processes we can inspect, challenge, and improve. We don't expect AI to develop these reliably by default.

03 SLR (and evidence synthesis) uses that approach: comprehensively gathering evidence, critically assessing it, and reaching conclusions through explicit steps. These principles can be used for many decisions.

04 Getting each step right helps researchers advance science today and gives us building blocks for better AI reasoning.

Elicit products



SLR workflow

Search, screening, extraction, and reporting



Research agent

Research, analysis, writing, and custom outputs



Reports

A topic overview before further investigation



API

Elicit within users’ own workflows and agents

Elicit’s review workflow

Question & protocol → **Search & records** → **Screening** → **Extraction**

CORE REVIEW WORK

Gather papers, apply criteria, and collect structured data.

Inspect source support, correct outputs, and export the results.

ADDITIONAL STAGES

The generated overview is optional orientation.

Dedicated RoB and GRADE support are not available; pairwise MA is preliminary.

Assessing Elicit through RAISE

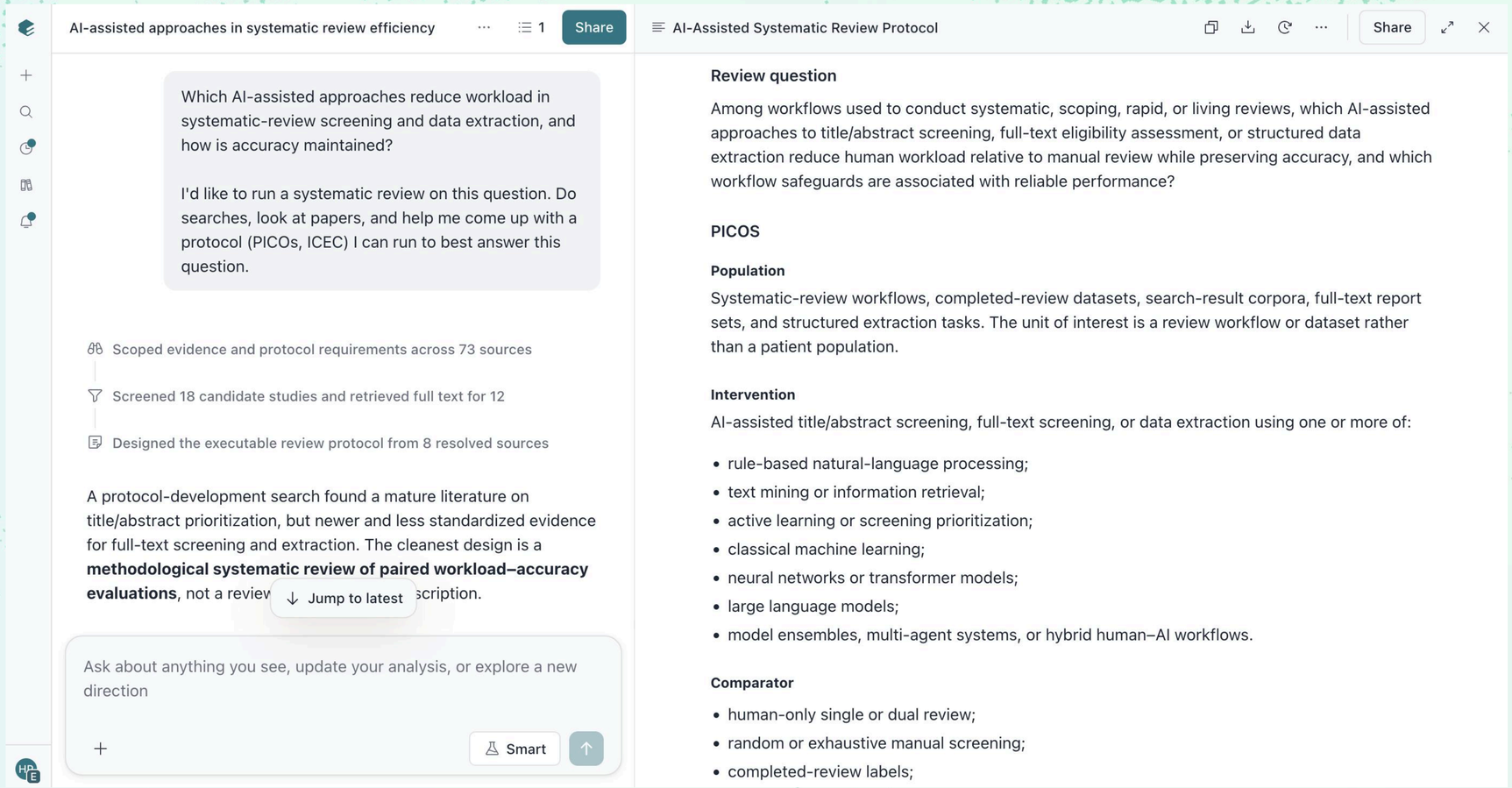
WHAT REVIEWERS NEED TO ASSESS	WHAT WE'LL COVER
Purpose and methodological fit	What Elicit does, how it works, which review methods it supports, and where its use is limited
Data and applicability	Which sources Elicit can access, what its evaluations cover, and where findings may not generalize
Performance and reliability	Accuracy, omissions, and other errors; comparisons with humans; effort after checking; performance as models change
Human oversight and practical use	How reviewers validate, inspect, correct, and document outputs
Transparency and governance	Reporting and traceability, current limitations, copyright, privacy, and training use

How to use Elicit today

STAGE	RECOMMENDED USE TODAY
Protocol	**Exploratory use:** develop ideas and refine questions; reviewers own the protocol.
Search	**Semantic: exploratory / supplementary. Use keyword search** for conventional searches, or import results from other databases.
Record management	**Verify** duplicate and PDF matches. DOI/title matching is straightforward, but Elicit's implementation has not been validated.
Title/abstract screening	**Verify** every decision; **validate** performance within the review before relying on AI screening decisions.
Full-text screening	**Verify** every decision; **validate** performance within the review before relying on AI screening decisions.
Extraction	**Verify** extracted answers against the source, including outcome data.
Synthesis	**Exploratory** orientation only. Do not use generated cross-study conclusions as the review's synthesis.
Risk of bias / GRADE	Dedicated support is not available yet.
Reporting	**Supplement** to reviewer audits; check completeness and accuracy, and add any required reporting they do not cover.
Updates	Apply the same safeguards at each stage; preserve prior outputs and document changes.

Exploratory

Question & protocol



Question and protocol

Question / protocol → **LLM suggestions** → **Reviewer decisions** → **Criteria & fields**

WHAT ELICIT OFFERS

Explore the literature via searches.

Suggest eligibility criteria from literature exploration.

Use an existing protocol and refine the criteria and questions for the review.

WHAT REVIEWERS NEED TO DO

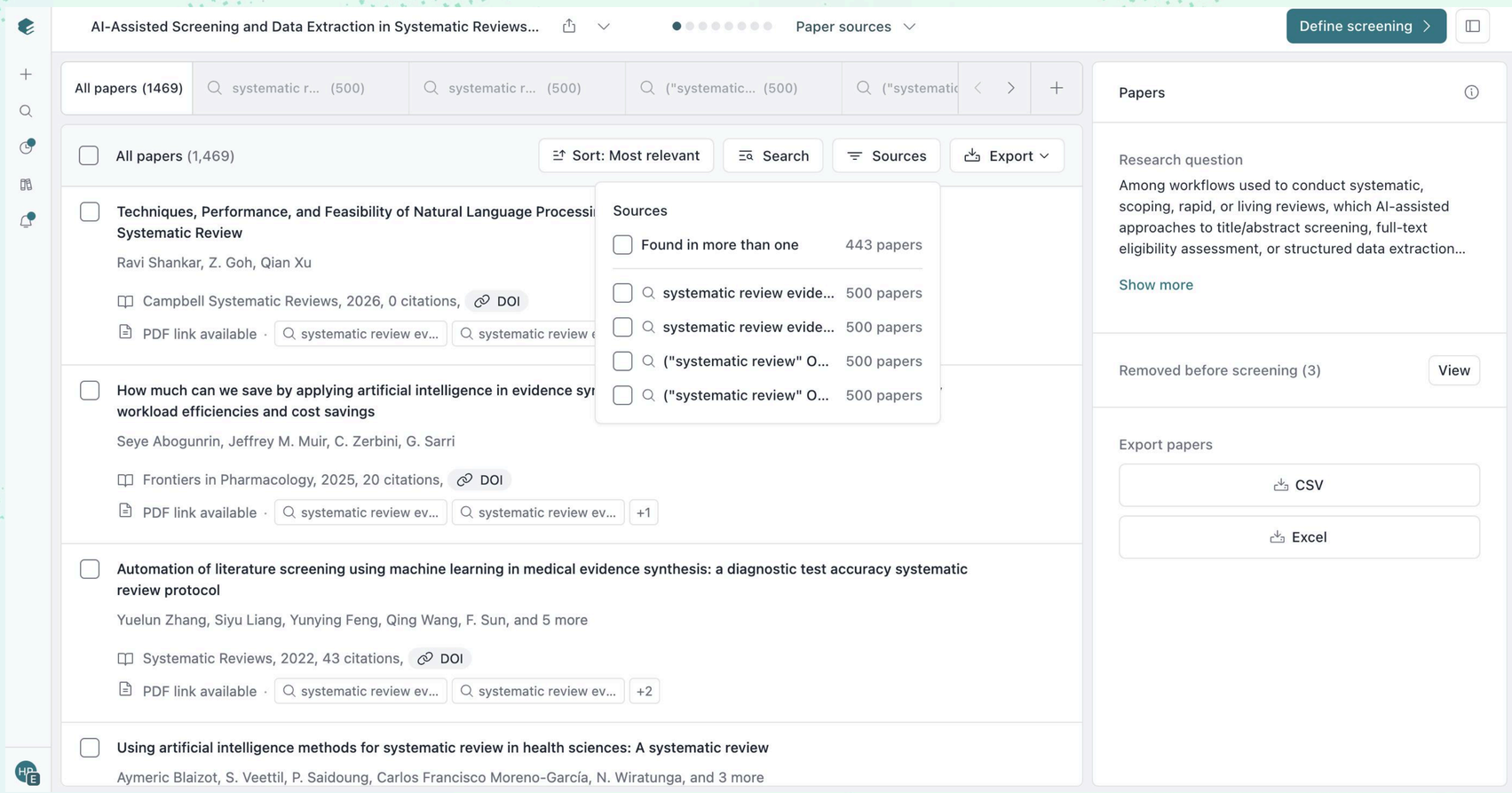
Develop and choose the review question, eligibility criteria, and methods.

Design and finalize the comprehensive search strategy.

Maintain the authoritative protocol, registration, and amendment record.

Exploratory

Search



Search

Semantic query	→	**Elicit corpus (embeddings)**	→	**Metadata-based ordering**	→	**Papers**
Keyword query	→	**Elicit / PubMed / ClinicalTrials.gov**	→	**Keyword matching**	→	**Papers**

WHAT ELICIT OFFERS

Search Elicit's corpus using semantic search.

Use keyword search over Elicit's corpus, PubMed, and ClinicalTrials.gov.

Combine search results and inspect papers.

Retain queries, dates, and filters for reporting.

Return up to 10,000 results per search and 40,000 results in total per SLR.

Import ris/bib/pdf from other databases/sources.

WHAT REVIEWERS NEED TO DO

Design and finalize the comprehensive search strategy.

Conduct searches in other databases.

Check for important missing evidence, including by inspecting results within Elicit.

Search

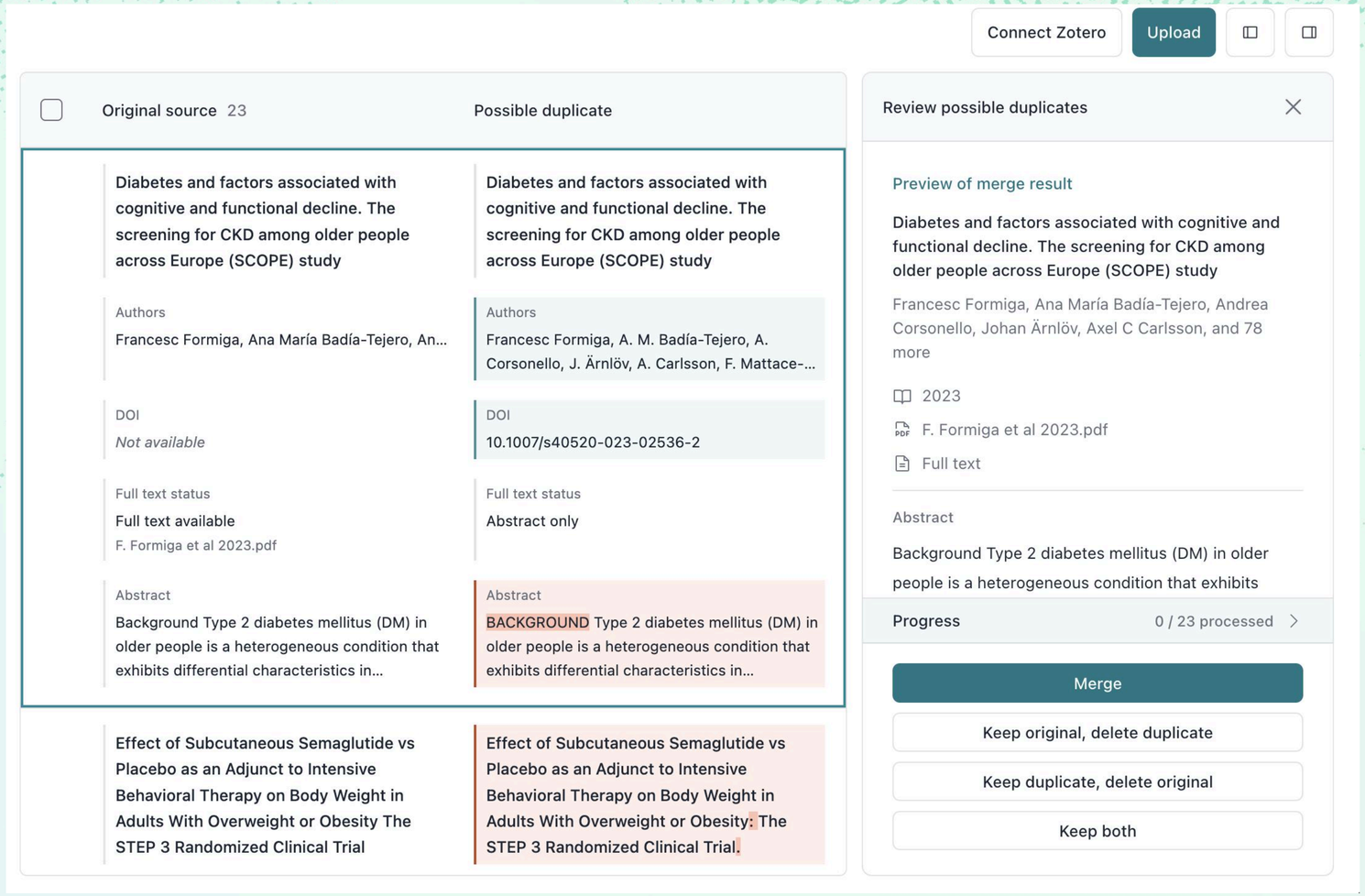
FULL EVIDENCE → INTERNAL BENCHMARK →

EVIDENCE	FINDING	SCOPE
Internal Cochrane benchmark	Corpus coverage: 95.0% (7,152/7,532 DOI reference records). At 10,000 results, mean recall: 95.0% within the corpus; 90.1% across all DOI targets. Retrieval depth matters: pooled all-target recall rose from 68.5% at 500 results to 88.1% at 5,000.	One title-based semantic query per review, with 10k results
Lau & Golder (2025)	Mean search recall: 53.6%. Recovered 25/56 (44.6%), 23/50 (46.0%), 36/48 (75.0%) and 20/41 (48.8%). Missed 31/56 (55.4%), 27/50 (54.0%), 12/48 (25.0%) and 21/41 (51.2%).	4 reviews; historical 500-result cap. Retrieval of original inclusions, before screening. *Internal 10k exact-query rerun: 87.0% mean retrieval recall (3 reconstructed review sets).
Wong et al. (2026)	Mean search recall: 33.3%. Recovered 6/15 (40.0%), 58/142 (40.8%) and 4/21 (19.0%). Missed 9/15 (60.0%), 84/142 (59.2%) and 17/21 (81.0%).	3 reviews; historical 500-result cap. All known-paper losses occurred in retrieval. *Internal 10k exact-query rerun: 81.0% mean retrieval recall (3 reviews).

> Reserve semantic search for exploratory and supplementary use; retain keyword search as the core search workflow. Deeper semantic retrieval improves recall.

Verify

Record management and obtaining full texts

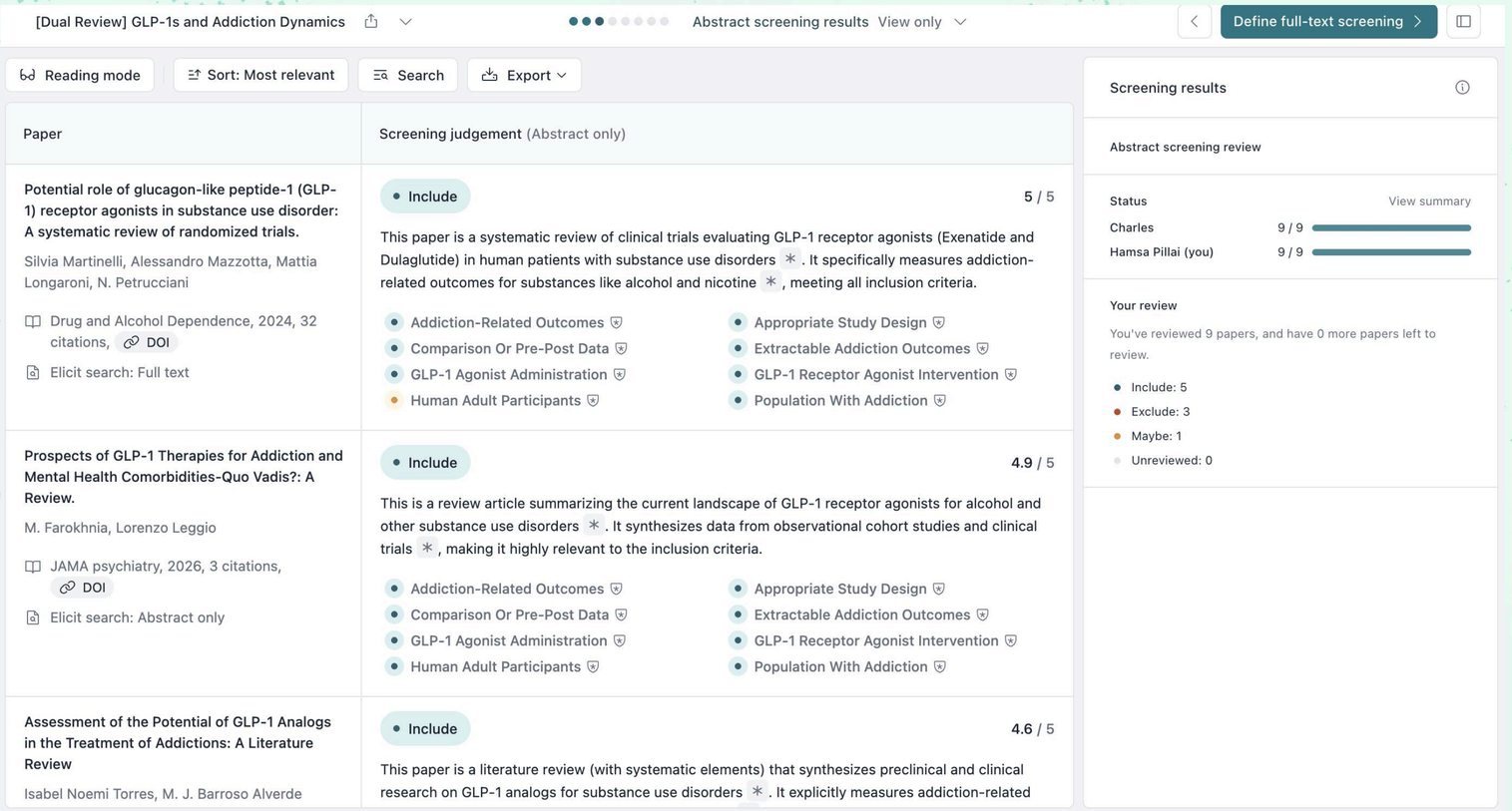


Record management and obtaining full texts

WHAT ELICIT OFFERS	WHAT REVIEWERS NEED TO DO
Combine retrieved records and import external search results.	Review duplicate matches and verify or correct PDF-to-record matches.
Identify potential duplicates by DOI and title; optionally check semantic similarity. Matches are reviewable and overridable.	Identify and track separate reports of the same study.
Retrieve open-access full texts online or use institutional credentials.	Obtain and upload missing full texts; resolve remaining access and permission questions.
List all papers with abstracts only so reviewers can obtain and upload the missing full texts.	
Automatically match uploaded PDFs to records; every match can be reviewed and corrected.	

Verify / Validate

Title/abstract screening



Title/abstract screening

Criteria + paper → **LLM selects relevant passages** → **LLM reasoning** → **Criterion recommendations, quotes, score**

WHAT ELICIT OFFERS

Suggest inclusion/exclusion criteria from the question, or use an existing protocol.

Recommend criterion-level and paper-level screening decisions, with quotes, rationale, and exclusion reasons.

Support dual review, disagreement adjudication, and a log of each reviewer's decisions.

Provide pilot screening and agreement statistics for calibration and validation.

Use scores for optional prioritization, or choose another sort order.

WHAT REVIEWERS NEED TO DO

Verify the criteria and validate screening for the review.

Verify or override Elicit's screening recommendations in every review workflow.

In dual review, each reviewer verifies or overrides every recommendation; adjudicate disagreements.

Title/abstract screening

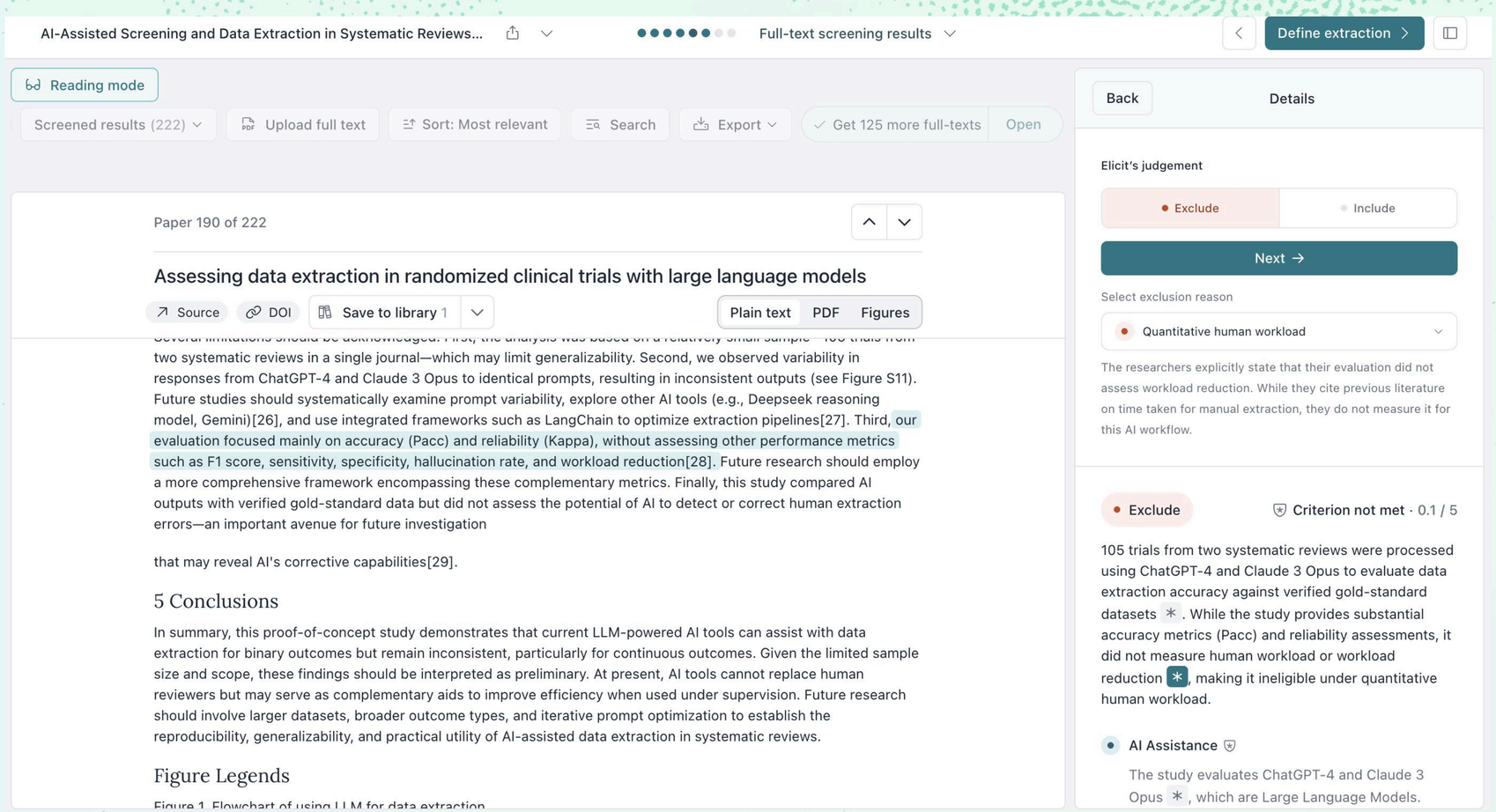
FULL EVIDENCE → INTERNAL BENCHMARK →

EVIDENCE	FINDING	SCOPE
Internal benchmark 108 reviews	Sensitivity: 96.9% (902/931 retained; 29 missed). Specificity: 92.5% (4,777/5,162 excluded).	Reference inclusions from published reviews; negative labels inferred from PubMed records.
Cao et al. (2025) 5 reviews ▪ preprint	Weighted screening sensitivity: 88.5% Elicit; 87.3% dual human review.	1,767 sampled records for Elicit; competitor-authored study.
Lau & Golder (2025) 4 reviews	Eligible papers wrongly excluded: 11/25 (44.0%), 11/23 (47.8%), 4/36 (11.1%) and 8/20 (40.0%).	Known eligible papers already retrieved. Authors suggest ambiguous/complex PICO contributed; exclusions involved uncertain eligibility details.
Gartlehner et al. (2020) Human-only study	Human screening sensitivity: 86.6% single; 97.5% paired.	Separate experiment; paired results modeled from individual decisions.

> Clearly defined, less interpretive PICOS criteria appear more suitable for Elicit screening. Pilot and validate within the review; verify recommendations against source evidence.

Verify / Validate

Full-text screening



Full-text screening

Criteria + parsed PDF → **LLM selects relevant passages** → **LLM reasoning** → **Criterion recommendations, quotes, score**

WHAT ELICIT OFFERS

Recommend criterion-level and paper-level screening decisions, with quotes, rationale, and exclusion reasons.

Support dual review, disagreement adjudication, agreement statistics, and decision logging.

Use scores for optional prioritization, or choose another sort order.

Retrieve full texts via open access or institutional credentials.

List missing full texts; match uploaded PDFs to records, with every match reviewable and correctable.

WHAT REVIEWERS NEED TO DO

Verify the criteria and validate screening for the review.

Verify or override Elicit's screening recommendations in every review workflow.

In dual review, each reviewer verifies or overrides every recommendation; adjudicate disagreements.

Full-text screening

FULL EVIDENCE → INTERNAL BENCHMARK →

EVIDENCE	FINDING	SCOPE
Internal benchmark 74 reviews	Sensitivity: 99.5% (212/213 retained; 1 missed). Specificity: 70.1% (115/164 excluded).	377 papers; eligibility reference labels reviewed and adjusted.
Mazzali et al. (2026) One review	Sensitivity: 100% (15/15 retained). Precision: 62.5% (15/24 eligible).	Sensitivity 95% CI: 78.2–100%. 9 ineligible papers also retained.
Iselin et al. (2025) Human + Elicit vs manual	Mean decision time: 66.8 vs 84.1 seconds/paper (20.6% shorter with Elicit).	1,200 papers; includes verification and extra reading. Interpretive criteria needed more source reading. Excludes setup/adjudication.
Stoll et al. (2019) Human-only study	Eligible reports missed by individual humans: 29/242 (12.0%) and 16/242 (6.6%).	Separate study; final inclusion after discussion was the reference.

> Less interpretive criteria needed less additional source reading; human-assisted decisions were faster. Validate and verify within the review, including relevant supplements.

Verify

Extraction

AI-Assisted Screening and Data Extraction in Systematic Reviews... | Extraction pilot | Run extraction

Sort: Title (A→Z) | Search | Add a column

Paper	Arm		Accuracy metrics and results	Paired workloac
High-performance automated abstract screening with large language model ensembles R. Sanghera, A. Thirunavukarasu, M. El Khoury, Jessica O'Logbon, Yuqing Chen, and 5 more J. Am. Medical Informatics Assoc., 2025, 35 citations, DOI Elicit search: Full text	GPT-3.5 Turbo, optimal prompt		- Development subset (n=800): sensitivity 1.000; 100% ; specificity 0.393 ; balanced accuracy 0.697 ; precision/PPV 0.458 ; NPV 1.000 ; F1-score 0.628 .\n- Comprehensive dataset: sensitivity 1.000 ; specificity 0.419 ; balanced accuracy 0.710 ; precision/PPV 0.004 ; NPV 1.000 ; F1-score 0.008 .\n- Reliability across reviews: Cohen's kappa 0.487–1.000; median 0.868 * .	- 100% recall: p workflows, the was 0.458 * ; reductions acro from 37.55% to - 95% and 99% reported in the workload–accur
	GPT-4 Turbo, optimal prompt		- Development subset (n=800): sensitivity 0.605 ; specificity 0.975 ; balanced accuracy 0.857 ; precision/PPV 0.927 ; NPV 0.828 ; F1-score 0.732 .\n- Reliability across reviews: Cohen's kappa 0.870–1.000; median 0.957 * .	-

10 sources selected randomly for piloting. Elicit will extract the full set of 142 sources in the next step.
Use this pilot step to refine your columns to save on usage before the full run.

Back | Result details

Supporting quotes from paper

1 of 1 | Previous [| Next]

Table Page 15

Human/Model	Sensitivity (Recall)	Specificity	Balanced Accuracy	Precision (PPV)	NPV	F1-score
Cochrane	1.000	0.993	0.996	0.235	1.000	0.381
GPT-3.5	1.000	0.419	0.710	0.004	1.000	0.008
GPT-4o	0.904	0.949	0.926	0.038	1.000	0.074
Gemini 1.5 Pro	0.756	0.976	0.866	0.068	0.999	0.125
LLaMA 3	0.841	0.776	0.809	0.008	1.000	0.017
Sonnet 3.5	0.823	0.982	0.903	0.096	1.000	0.172

Table 3: Performance of five LLMs used with their respective optimal prompts across every record returned by replicated searches, compared to an expected performance ceiling calculated from the...

Explanation

- The paper reports model- and human-specific accuracy metrics in the 800-record development experiment (Table 2), and model-specific metrics in the comprehensive replicated-search experiment (Table 3).
- Reliability is reported using review-stratified Cohen kappa ranges and medians for each LLM and for human researchers.
- Cross-arm comparisons include the sensitivity–precision trade-off, LLM-versus-human comparisons, ensemble performance, and review-level associations.

Extraction

PDF parsing → **Text, tables & figures** → **LLM + question / schema** → **Answer + quotes / rationale**

WHAT ELICIT OFFERS

Suggest extraction fields or use fields supplied by the reviewer.

Set custom schemas for the required output formatting.

Extract from text, tables and figures, with supporting quotes and rationale.

Split trial data by arm.

Pilot extraction questions, review the answers, and revise the questions.

Inspect, edit, and export the structured extraction table.

WHAT REVIEWERS NEED TO DO

Define and pilot extraction questions; assess the answers and revise as needed.

Review extracted answers and perform calibration/validation manually; extraction agreement statistics are not available.

Resolve conflicting reporting and link reports of one study.

Arrange any required dual extraction; there is no dedicated workflow today.

Extraction

FULL EVIDENCE → INTERNAL BENCHMARK →

EVIDENCE 1 OF 2

STUDY / EVALUATION	MAIN FINDING	CONTEXT
Internal benchmark	Extraction accuracy (AI graded): 95.6% (305/319 answers).	Methods, participants, interventions; outcomes excluded.
Helms Andersen et al. (2025)	Fully correct: study design 30/30 (100%); population 28/30 (93.3%); review-specific variables 22/30 (73.3%).	80/90 fully correct (88.9%); 4/90 (4.4%) confabulations. 83/90 (92.2%) accepted with minor omissions; 30 papers × 3 fields.
Hilkenmeier et al. (2025)	Correct: population / country 41/43 (95.3%) each; study design 40/43 (93.0%); participants 38/43 (88.4%); main results 33/43 (76.7%); measures 30/43 (69.8%); constructs of interest 22/43 (51.2%).	Incorrect: 18.6% Elicit vs 13.3% human. Correct fields: 6.07/7 (86.7%) primary studies; 4.85/7 (69.3%) reviews/theoretical articles. Cited-study / construct confusion; unavailable supplements. 301 values/arm; unblinded assessment.
Bianchi et al. (2025)	70/140 (50.0%) as complete/correct as human; 64/140 (45.7%) incomplete; 6/140 (4.3%) different.	Effect details incomplete in 19/20 (95.0%) trials. 20 trials × 7 fields.

Extraction

FULL EVIDENCE → INTERNAL BENCHMARK →

EVIDENCE 2 OF 2

STUDY / EVALUATION	MAIN FINDING	CONTEXT
Meliante et al. (2025)	Cell accuracy: 51.4% (655/1,274). Missing cells: 22.4% (285/1,274).	Tables from four glaucoma reviews.
Cao et al. (2025)	Extraction accuracy (AI graded): 74.8%, weighted across 7 reviews.	495 papers; competitor-authored preprint.
Lagisz et al. (2026)	86.6% correct; 48/70 (68.6%) fields maintained pilot accuracy on new papers.	7 reviews; 536 matched outputs. Pilot success did not guarantee transfer.
Lee et al. (2026) Preprint	24.8 min/study saved, including preparation; noninferior quality with human verification.	Errors: 19.3% assisted vs 20.2% manual; fabricated data: 0.4% in both. 50 RCTs; 2,550 fields/arm; includes minor errors.

> Study design and population were stronger fields in Andersen and Hilkenmeier. Outcome data need further validation, particularly numerical results, uncertainty estimates, and results by study arm and time point. Lee found time savings with human verification.

N/A

Risk of bias

Risk of bias

FULL EVIDENCE →

PLANNED ELICIT SUPPORT

Use RoB frameworks to recommend granular judgments, with rationale and supporting quotes.

Provide κ and agreement statistics for calibration and validation.

WHAT REVIEWERS WOULD DO

Choose the appropriate framework and validate its use within the review.

Verify the recommended judgments and their source support; override them where needed.

WHAT HAS BEEN EVALUATED

INTERNAL PROTOTYPE • PRELIMINARY

Reported agreement: 74%; κ = 0.726.

Experimental workflow; not a validation of dedicated, optimized RoB support.

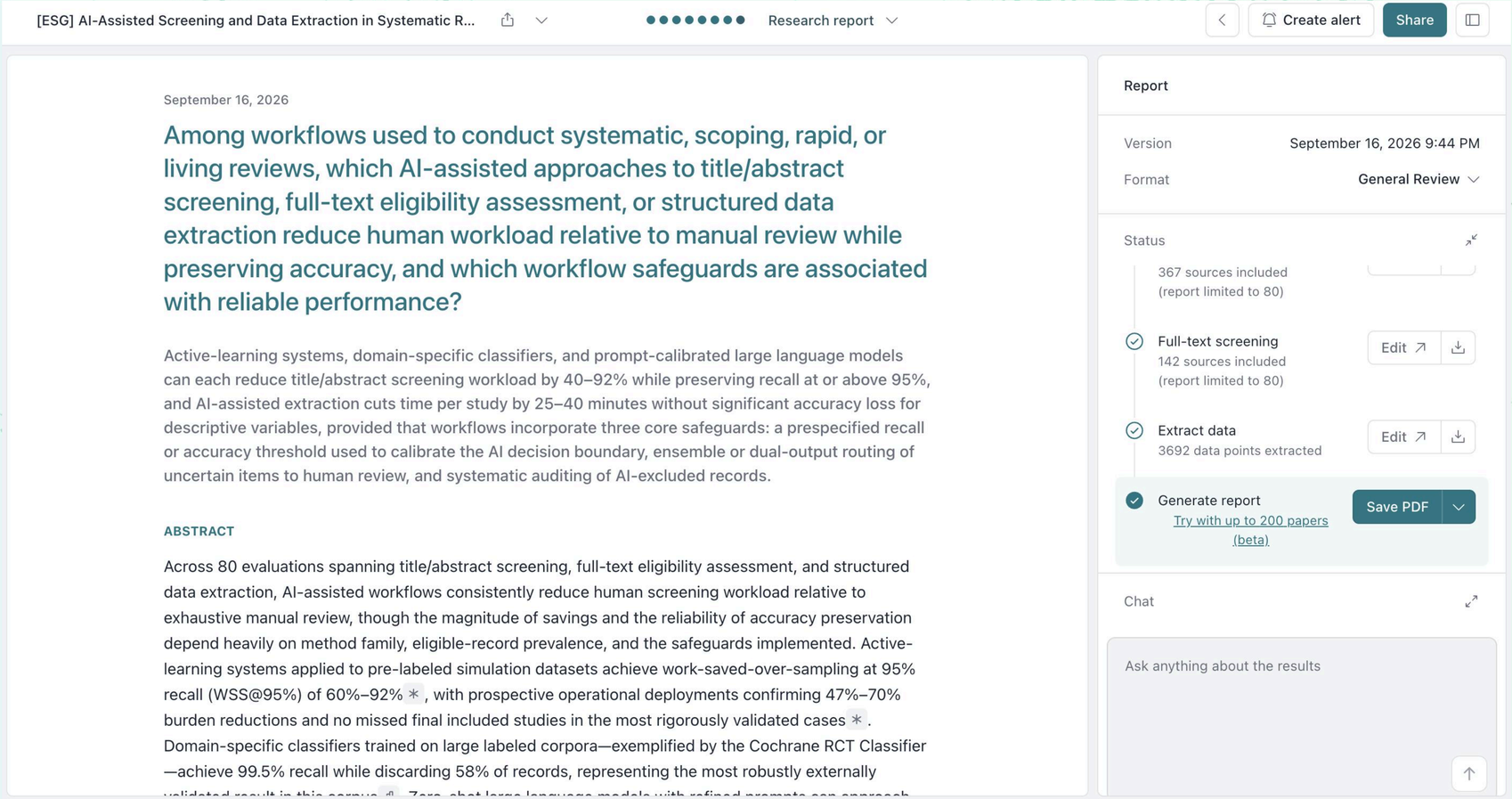
WONG ET AL. (2026) • PUBLIC MANUSCRIPT

Agreement with reference RoB judgments: mean κ = 0.04 across 3 reviews.

Configured appraisal columns showed low agreement with the reference assessments.

Exploratory

Synthesis



Synthesis: generated overview

Extracted data table → **80–200 papers (plan + screening score)** → **LLM + structured review template** → **Optional overview**

WHAT ELICIT OFFERS	WHAT REVIEWERS NEED TO DO
Generate an optional orientation summary using the extracted data and a structured review template. Use the research agent to define or modify output (for example, to emphasize particular evidence or subgroup analyses)	Treat screening decisions and extracted data as the primary outputs. Check whether report claims use the right sources. Use the overview for orientation, rather than as the review's conclusions.

Synthesis: generated overview

FULL EVIDENCE →

EVIDENCE	FINDING	SCOPE
Internal evaluation December 2025	About 84% of footnoted claims mostly or fully supported by supplied source quotes.	Automated claim-to-quote assessment: scores 4–5 on a 5-point scale. Measures supportedness, not completeness or overall review validity.
Heath / NICE (2026) Prioritisation briefings	6–8/12 (50.0–66.7%) topics judged sufficient for prioritisation briefings, depending on question wording.	NICE analyst judgments; free/basic Elicit. Failures involved outdated or non-UK evidence. Related briefing task; SLR workflow use not confirmed.
Barres et al. (2025) Scientific answers	86.4% of claims supported by citations; 60.8% both supported and relevant.	25 life-science questions; expert claim checks. Cora-developer study of scientific answers; SLR workflow use not confirmed.
Wong et al. (2026) 3 reviews ▪ manuscript	Examples of overgeneralising animal findings, emphasising unrequested subgroups, and overstating certainty.	Generated review reports; qualitative findings. Evidence sets differed, so errors cannot be isolated to report writing.

> Evaluations found useful briefing outputs and source-supported claims, alongside irrelevant claims, outdated evidence and overconfident conclusions.

Synthesis: pairwise meta-analysis

FULL EVIDENCE →

WHAT ELICIT OFFERS (PRELIMINARY)

Start from the review question/protocol and completed-review data or a supplied study-data table.

Use dedicated meta-analysis and study-data-cleanup skills to prepare and assess the analysis.

Write and run R code for pooled effects, forest plots, heterogeneity and planned sensitivity analyses.

Return a study disposition table (analysable, recoverable or not poolable), plus inputs, code and results.

WHAT REVIEWERS NEED TO DO

Specify the question, outcome, comparison and analysis plan; resolve consequential gaps.

Check study data, pooling decisions, code, statistical methods and interpretation.

Treat this as a preliminary workflow that is not ready for rigorous SLR use.

WHAT HAS BEEN EVALUATED

Internal tests: a small set of existing meta-analyses and NICE DSU material.

Beta-blocker example: random-effects OR 0.781 versus 0.78 in the reference.

Issues found: fixed-effect results remained in sensitivity-analysis prose; a prediction interval used the wrong method.

N/A

Certainty of evidence / GRADE

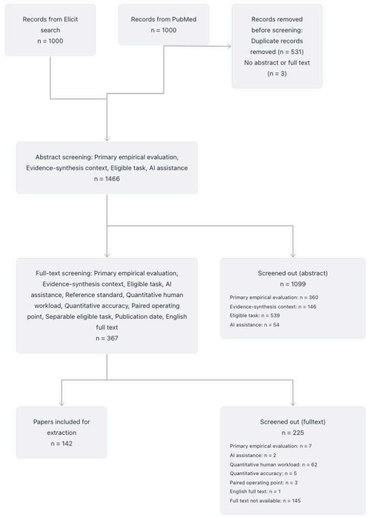
Certainty of evidence / GRADE

WHAT ELICIT OFFERS	WHAT REVIEWERS NEED TO DO
Nothing yet for formal certainty-of-evidence / GRADE assessment. We want to support this in the future.	Assess certainty using the appropriate methods externally.

Supplement

Reporting and the decision record

METHODS



Paper search

Elicit Corpus

We performed a semantic search across over 138 million academic papers from the Elicit search engine, which includes all of Semantic Scholar and OpenAlex.

We ran these queries:

- "systematic review evidence synthesis study selection title abstract screening full text screening artificial intelligence machine learning active learning text mining natural language processing large language model workload time recall sensitivity human reviewer"
- "systematic review evidence synthesis data extraction information extraction automation artificial intelligence machine learning transformer large language model human verification workload time accuracy error omission hallucination"

Dual review history

Who reviewed each paper, how conflicts were resolved, and when each step was completed. Expand any row for the papers involved, or export a CSV you can attach to your methods write-up.

← Back to protocol

Export CSV

May 13, 2026 11:37 AM	Abstract screening was completed	Abstract screening
May 13, 2026 11:37 AM	Hamsa Pillai (you) resolved 2 conflicts (1 include, 1 exclude)	Abstract screening
May 13, 2026 11:36 AM	Abstract screening moved to conflict resolution	Abstract screening
May 13, 2026 11:35 AM	Hamsa Pillai (you) reviewed 9 papers (5 include, 3 exclude, 1 maybe)	Abstract screening

DECISIONS · Include · 5 · Exclude · 3 · Maybe · 1 · Disagreements with model · 1

PAPER	MODEL	REVIEWER	NOTE
Prospects of GLP-1 Therapies for Addiction and Mental Health Comorbidities-Quo Vadis?: A Review.	Include	Include	–
Assessment of the Potential of GLP-1 Analogs in the Treatment of Addictions: A Literature Review	Include	Include	–
GLP-1 Analogues in the Neurobiology of Addiction: Translational Insights and Therapeutic Perspectives	Include	Include	–
GLP-1 signaling and alcohol-mediated behaviors; preclinical and clinical evidence.	Include	Exclude	–

Addiction-Related Outcomes

Reporting and the decision record

WHAT ELICIT OFFERS

Export papers, screening decisions, and extracted data, all with quotes and rationale.

Export AI answers and recommendations, each reviewer's decisions, and the final adjudicated decisions.

Track every input paper, its source, and where it leaves the workflow, with PRISMA-style flow information.

Export the searches, eligibility/screening criteria, and extraction questions used in the review.

WHAT REVIEWERS NEED TO DO

Report the methods, AI use, and human checking involved.

Document protocol amendments and their rationale.

Complete disclosures beyond the available exports and decision record.

Stage-specific guidance

Updating a review

Updating a review

WHAT ELICIT OFFERS	WHAT REVIEWERS NEED TO DO
Add papers: process the new papers through screening and extraction.	Add papers: document the updated search and review the new results.
Change screening criteria: rescreen all papers in that stage.	Change screening criteria: preserve the prior log before discarding dual review; review the new recommendations.
Change an extraction question: regenerate that column for all papers.	Change an extraction question: preserve prior data and the reason for the change; check the regenerated answers.

Stability

FULL EVIDENCE →

HOW WE MANAGE MODEL CHANGES

Check stage-specific performance against our internal benchmark baseline, using paired comparisons and uncertainty estimates.

Check latency, failures and capacity; use staged rollouts.

Model updates affect future work and do not alter previously saved screening or extraction results.

REPEAT RUNS AND CURRENT LIMITS

Users cannot pin a model version today.

Small internal test: one topic, ten runs. Five named core papers appeared in all ten; one borderline paper appeared in three.

Extraction differences included wording and factual changes. The test did not establish an overall accuracy change.

Privacy, copyright and permissions

COPYRIGHT AND PERMISSIONS

Full texts can come from open-access sources, institutional access or user uploads; applicable licenses and agreements govern processing.

Open access does not mean unrestricted reuse. Uploaded subscription content depends on publisher and institutional permissions.

We are pursuing rights-checking integrations to help users determine whether their permissions cover processing in Elicit.

PRIVACY AND TRAINING

Uploaded PDFs are privately stored and are not added to Elicit's shared paper corpus.

Model providers process content to deliver results and are prohibited from using the data we send them to train their models.

Elicit may use de-identified customer data to train its own models. Training on identifiable customer data requires opt-in. Customer agreements may further restrict this use.

Development priorities

ACCESS AND WORKFLOW

More database integrations.

Rights-checking integrations.

Better tracking of protocol changes.

Improved search methods.

Extraction dual review & editing

APPRAISAL AND ANALYSIS

Dedicated RoB support if the required permissions are obtained, followed by rigorous evaluation.

Continue development and evaluation of pairwise meta-analysis.

Later: network meta-analysis and indirect treatment comparisons (ITC).

Future support for GRADE and certainty of evidence.

How to use Elicit today

STAGE	RECOMMENDED USE TODAY
Protocol	**Exploratory use:** develop ideas and refine questions; reviewers own the protocol.
Search	**Semantic: exploratory / supplementary. Use keyword search** for conventional searches, or import results from other databases.
Record management	**Verify** duplicate and PDF matches. DOI/title matching is straightforward, but Elicit's implementation has not been validated.
Title/abstract screening	**Verify** every decision; **validate** performance within the review before relying on AI screening decisions.
Full-text screening	**Verify** every decision; **validate** performance within the review before relying on AI screening decisions.
Extraction	**Verify** extracted answers against the source, including outcome data.
Synthesis	**Exploratory** orientation only. Do not use generated cross-study conclusions as the review's synthesis.
Risk of bias / GRADE	Dedicated support is not available yet.
Reporting	**Supplement** to reviewer audits; check completeness and accuracy, and add any required reporting they do not cover.
Updates	Apply the same safeguards at each stage; preserve prior outputs and document changes.

Questions?

Search: external evidence

STUDY / TEST DATE	WHAT WAS TESTED	WHAT WAS FOUND
Wong et al. Jul 2025–Apr 2026	Replicated searches for 3 reviews; maximum 500 results.	Recovered 6/15, 58/142, 4/21 (mean recall 33.3%); missed 9, 84, 17. Additional eligible papers: 2, 50, 3. Losses occurred in retrieval.
Di Febo et al. Aug 2025–Jan 2026	Top 10 results across 4 cardiology topics; expert relevance ratings.	~62–63% of retrieved papers judged relevant; ≈40% judged key. These are shares of retrieved results, not recall.
Lau & Golder Feb–Mar 2025	Search and screening for 4 reviews; maximum 500 results.	Combined search + screening: mean sensitivity 39.5%; retained 14/56, 12/50, 32/48, 12/41; missed 42, 38, 16, 29; 11 additional eligible papers.
Moens et al. Mar 2024	Recovery of known included studies in 4 reviews.	Recovered 2%, 19.05%, 0% and 0% of the known included studies in the four reviews.
Bernard et al. Apr 2023	Repeated searches for one umbrella review.	Recovered 3/17 known reviews and found 3 additional eligible reviews. Results varied between runs.
Ajith et al. (LitSearch) Date not reported	80 queries seeking specific machine-learning / NLP papers.	The target was in the top five results for 23.1% of inline-derived questions and 17.5% of author-written questions.
Fedorov et al. Date not reported	Top 10 results on one library-technology topic.	7/10 results were judged relevant. Searching took 10 minutes; this excluded checking the results.
O'Rourke et al. Apr 2025	4 nursing reviews; 3 prompts; top 10 / 100 results; Elicit Plus.	Authors report mean recovery of 3.7% of human-review inclusions. Search filters differed from review eligibility.
Kiepura et al. Date not reported	30 metadata + 30 keyword-content queries; free tier; first ≤10 results.	Precision/recall: 8%/41% (metadata), 3%/10% (content). Strict queries, benchmark-relative targets; comparator-developer study.
Meliante et al. (2025)	Four glaucoma reviews; historical search configuration.	Known-study retrieval: 50%, 6%, 43%, 10% across reviews. Exact run/version not established.
Vyas et al. (2026)	24 dermatology RCTs; concept and keyword searches.	Recovered 1/24 by concept search; 0/24 by keyword search. Historical configuration; not a test of current keyword integrations.
Mazzali et al. (2026)	Native search; one air-pollution umbrella review.	40/500 returned records overlapped the conventional search. This 8% is overlap among retrieved records, not recall.
Jones (2025)	22 dermatology reviews; first eight recommendations.	Mean overlap 2.5/8; six searches found later papers. Conference abstract; overlap is not eligible-study recall.
Enomoto et al. (2023)	One liver-cancer review; 31 known studies.	11/31 recovered (35.5%); 20 missed; one additional eligible study. Recovered 8/9 comparative studies; missed conference/single-arm reports.
Seth et al. (2025)	Thumb-arthritis RCTs; human screening of search output.	10/23 eligible trials recovered (43.5%); no additional AI-only inclusions. One topic; not an AI-screening test.
Tomczyk et al. (2024/26)	Nine e-commerce questions; top-eight results.	Mean relevance 56% (Scopus 78%, WoS 65%, SciSpace 33%); 80% unique. Top-ranked relevance, not SR recall.
Park (2025/26)	One artificial-pancreas search; free interface.	48/48 expanded results judged relevant by the author; no exhaustive target set. Correction scope remains unresolved.
Wagner et al. (2026)	Illustrative programmer-productivity search.	2/10 results judged relevant; mostly abstract-based summaries judged adequate. One illustrative query.
Aljamaan et al. (2024)	100 references across ten medical prompts.	Title/journal/author errors: 0; DOI: 4; date: 40; link: 1; keyword-relevance: 58. Component errors overlap; not 103 bad references.
Aydin et al. (2026)	Five prompts in Nov 2024 and Jan 2026.	20/20 and 50/50 citation instances authentic. Existence of a paper does not establish claim support or recall.
Krupin et al. (2026)	One bone-fracture search; Jun–Aug 2025.	500 retrieved, 38 shortlisted, six manually selected. Two reported "hallucinations" among 38; error definition unresolved.
Chandler et al. (2026)	Top-ranked retrieval diversity and source characteristics.	Elicit tended toward older, highly cited papers; ranked second on median citations per topic. No recall/accuracy estimate.
Tranfield & Caldwell (2024)	Ten STEM queries; six tools; May 2024.	Descriptive source-type and unique-yield comparison. No reliable Elicit-specific accuracy or recall estimate recovered.
Danler et al. (2024)	Eight tools; one medical-data-warehouse question.	Elicit returned four scholarly references; one shared with Consensus. Small descriptive comparison, not a completeness test.
Patterson et al. (2025)	Librarian/student ratings; free and paid tools.	Paid Elicit ranked among leading tools; summaries/relevance praised. Subjective rubric; no validated recall or accuracy score.
Gorton (2025)	Monash librarians' 22-tool comparison.	Elicit rated 8/10 on a practical feature/usability rubric. This is not 80% accuracy or recall.

Search: internal evidence

Tang et al. · Pfizer + Elicit · conference draft

Measure	What / how we measured	Result
Corpus coverage	Check whether Elicit's corpus contains 7,532 DOI-identifiable papers included in Cochrane reviews.	7,152 / 7,532 95.0%
Mean recall: in-corpus DOI targets	Search by review title; average recall across 741 reviews for DOI targets already in the corpus. Cap: 10,000 results.	95.0%
Mean recall: all DOI targets	Search by review title; average recall across 747 reviews for all DOI targets, including those outside the corpus. Cap: 10,000.	90.1%

Result limit	500	1,000	5,000	10,000
Pooled recall: all 7,532 DOI targets	68.5%	76.7%	88.1%	88.3%

Cochrane review-title queries. 859 completed searches; 29 failed. Recall analysis restricted to DOI-identifiable targets; no repeatability assessment.

Title and abstract screening: external evidence

STUDY / TEST DATE	WHAT WAS TESTED	WHAT WAS FOUND
Mazzali et al. 5 Sep 2025	Title/abstract screening of records imported from a conventional search.	Retained 30/33 records advanced by humans: sensitivity 90.9% (95% CI 75.7–98.1), precision 30/70 (42.9%). The three missed records were later excluded at full text. Threshold tuned on this dataset.
Cao et al. 2025; exact run date unverified	Title/abstract screening; 1,767 sampled records from 5 reviews.	Weighted sensitivity 88.5% Elicit vs 87.3% dual humans; Elicit specificity 84.2%. Elicit sensitivity range 76.9–100%; best prompt condition selected. Competitor-authored preprint; final-inclusion reference.
Lau & Golder (2025)	Four reviews; screening of known eligible papers already retrieved.	Wrongly excluded 11/25, 11/23, 4/36 and 8/20. Rationales involved uncertain population, context, outcomes and design; clearer-PICO explanation is a hypothesis.
Gartlehner et al. (2020) Human-only benchmark	280 reviewers; 2,000 abstracts; 24,942 decisions. Paired performance modeled.	Sensitivity 86.6% single (95% CI 80.6–91.2), 97.5% paired (95.1–98.8). Separate experiment; not a head-to-head Elicit comparison.

Title and abstract screening: internal evidence

108 reviews ▪ 6,093 papers ▪ 931 reference inclusions

MEASURE	WHAT / HOW WE MEASURED	RESULT
Sensitivity	Apply eligibility criteria to titles/abstracts; measure retention of 931 reference inclusions. Pooled across papers; 29 missed.	902/931 = 96.9%
Specificity	Measure exclusion of 5,162 inferred negatives: PubMed results absent from published inclusion lists, without independent adjudication. 385 retained.	4,777/5,162 = 92.5%
Decision rule and applicability	Retain “Yes” and “Maybe”; exclude only “No”. The benchmark does not test the current dual-human-review interface.	“Maybe” retained

Full-text screening: external evidence

STUDY / TEST DATE	WHAT WAS TESTED	WHAT WAS FOUND
Iselin et al. Date unresolved	1,200 papers randomized to human screening with or without Elicit.	66.8 vs 84.1 s/paper (20.6% less), including extra reading; excludes setup/ adjudication. Information sufficient for 82.3% of interpretive self-regulation decisions vs 93.3–98.8% for other criteria; sufficiency is not accuracy.
Mazzali et al. 3 Oct 2025	Full-text eligibility decisions for one review.	Retained all 15 human inclusions: sensitivity 100% (95% CI 78.2–100). Also retained 9 ineligible papers, giving precision 15/24 (62.5%). Threshold tuned on this dataset.
Rokhshad et al. Jan–Mar 2025	Full-text screening of 59 dental reports, including 2 pilots.	Accuracy 86%, precision 82%, sensitivity 93%, specificity 80%, F1 87% (Table 1). Denominator and round aggregation remain unclear.
Arroqui et al. Date not reported	46 uploaded papers; 3 repeated screening/extraction runs; Team plan.	Initial exclusions varied 3/2/2. Human interpretation and correction ultimately selected the same 12 studies; no measured time saving.
Stoll et al. (2019) Human-only benchmark	Independent full-text decisions, compared with final inclusion after discussion.	Individual reviewers initially missed 29/242 and 16/242 ultimately included reports. Separate human study; not a head-to-head Elicit comparison.

Full-text screening: internal evidence

74 reviews ▪ 377 papers ▪ adjusted reference labels

MEASURE	WHAT / HOW WE MEASURED	RESULT
Sensitivity	Apply full-text criteria and measure retention of 213 adjusted reference inclusions. One missed; “Maybe” retained.	212/213 = 99.5%
Specificity	Measure exclusion of 164 adjusted reference exclusions. 49 retained. Overall paper accuracy: 327/377 = 86.7%.	115/164 = 70.1%
Reference-set adjustments	Remove criterion judgments after AI disagreement: 138/1,271 removed; 30 paper labels changed to include. All 377 papers retained; criterion accuracy 94.8% on 1,133 judgments.	138/1,271 criterion judgments removed

Data extraction: external evidence

STUDY / TEST DATE	WHAT WAS TESTED	WHAT WAS FOUND
Świt May 2026	7 chemistry papers; extraction for environmental assessments.	Derived greenness scores differed from expert averages by 0.01–0.12 on a 0–1 scale (calculated from rounded scores); not field accuracy.
Wong et al. Jul 2025–Apr 2026	3 reviews; extraction assessed on papers found by both approaches.	Accuracy 70.6%, 71.6%, 65.1% by review; mean 69.1% (category averages). Population/intervention confusion and omissions; only overlapping papers assessed.
Lagisz et al. Repeat Jun 2025	7 reviews; tune extraction prompts, then test them on new papers.	70/90 fields passed prompt development; 48/70 maintained ≥7/8 correct on new papers. Matched 536 outputs: 86.6% TEST, 85.6% RETEST; 88.8% identical.
Cao et al. 22 Mar 2025	495 papers across 7 reviews; high-accuracy extraction.	An AI grader judged 74.8% of values correct; results ranged from 58.8% to 83.1% across reviews. Competitor-authored preprint.
Rokhshad et al. Jan–Mar 2025	31 dental reports, including 2 pilots; data extraction.	Most category medians were 4–5 on a 0–5 response-quality rubric; “Prep-proc” fell to 0 in rounds 2–3. Not cell-accuracy percentages.
Lee et al. Dec 2024–Jan 2025	50 RCTs × 51 fields; two reviewers; checked Elicit vs manual; preprint.	Scores 84.1 vs 83.6/100; difference 0.57 (95% CI -1.29–2.43), 10-point noninferiority margin. 48.5 vs 73.3 min/study incl. setup; 24.8 min saved.
Lee (2026): errors	Final checked answers; 2,550 fields per arm; assisted / manual.	Any errors: 493/514 (19.3%/20.2%); missed 186/175; incomplete 175/174; misallocated 8/13; false 10/10 (0.4% each); calculation/inference 108/132; other 6/10.
Shaw et al. Dec 2024	53 papers, 19 fields; human assessment of AI extractions.	Human assessors strongly agreed in their judgments (AC2 = .960). This is assessor agreement, not Elicit accuracy; time savings were estimated.
Hilkenmeier et al. Autumn 2024	43 papers, 7 fields; Elicit and human extractions compared.	Correct: Elicit 245/301 (81.4%), human 261/301 (86.7%); unblinded. Empirical 6.07/7 vs other papers 4.85/7. Cited-study/ construct confusion and unavailable supplements.
Bianchi et al. Aug–Sep 2024	20 trials, 7 fields; comparison with human extractions.	70/140 outputs were as complete/correct as human data; 64 were incomplete and 6 differed. Effect details were missing in 19/20 trials.
Helms Andersen et al. 5 Jul 2024	30 papers, 3 fields; 90 extracted answers.	80/90 answers were fully correct (88.9%). Accepting minor omissions raised this to 83/90 (92.2%); four confabulations were reported.
Moens et al. Mar 2024	Study characteristics extracted across 4 reviews.	Correct by field: author 96.9%, year 62.5%, design 41.7%, sample 81.3%, age 31.3%, sex 46.9%, BMI 20%, work status 0%; different denominators.
Spillias et al. 29 Jan 2024	33 fisheries papers; 11 questions about study context.	>99% answered, but presence/absence $\kappa<0$; absent-information answers in 27% of all pairs. Subjective quality above “fair”; coverage is not accuracy.
Meliante et al. Date not reported	Reproducing tables from 4 glaucoma reviews; 1,274 cells.	655/1,274 cells were accurate (51.4%); 22.4% were missing, 13.7% imprecise and 12.5% incorrect.
Vyas et al. Date not reported	24 dermatology trials; study data and Jadad components.	13 extraction errors were reported, but the total number of checked fields is unclear. Country and derived Jadad score each wrong in 5/24.
Arroqui et al. Date not reported	Six-field extraction; three researchers; externally retrieved papers.	Wrong-field outputs and instruction failures described; no formal extraction accuracy or measured review-time saving.

Data extraction: internal evidence

28 reviews · 111 study records · 319 responses

MEASURE	WHAT / HOW WE MEASURED	RESULT
AI-graded correctness	Extract methods, participants and interventions; grade against reconstructed reference questions. Correct: 101/105, 99/104, 105/110. Outcomes/numerical data excluded.	305/319 = 95.6%
Reference construction and grading	AI reconstructed reference questions; Opus 4.6 graded responses. Highly selected sample; reconstructed questions could contain source clues.	Opus 4.6 as judge
Separate human check	Human reviewers checked 42 responses in a partly error-enriched sample, separate from the final benchmark run. This does not independently verify the 95.6% estimate.	38/42 = 90.5%

Risk of bias: evidence

STUDY / TEST DATE	WHAT WAS TESTED	WHAT WAS FOUND
Wong et al. Jul 2025–Apr 2026	Appraisal judgments across 3 review replications.	Agreement 51.7%, 66.2%, 33.3%; κ .209, .062, −.149 (mean ≈.04). Configured appraisal columns; not a dedicated optimized RoB workflow.
Mazzali et al. 3 Oct 2025	AMSTAR-2 EH assessment of review quality.	17/56 additional-item judgments were wrong (30.4%). General-item errors conflict: 44/180 in the prose versus 34/180 in Table 1.
Rokhshad et al. Jan–Mar 2025	Risk-of-bias assessment of dental reports.	Reported agreement with experts: κ = 0.80, 0.52, 0.24, 0.52 for Q1–Q4. Mapping to the six appraisal criteria remains unclear.
Vyas et al. Date not reported	Jadad quality scores calculated from Elicit-extracted components for 24 trials.	Derived Jadad scores were wrong for 5/24 trials. Scores were calculated manually from extracted information.
Internal prototype Preliminary	Experimental RoB workflow; separate from the Cochrane benchmark.	Reported agreement 74%; κ .726. Original evaluation source, sample and denominators not recovered; these remain unverified preliminary figures.

Synthesis: external and internal evidence

STUDY / TEST DATE	WHAT WAS TESTED	WHAT WAS FOUND
Wong et al. Jul 2025–Apr 2026	Generated narrative reports for 3 review replications.	Reports extended animal findings beyond their scope, emphasised unrequested diagnostic subgroups, and overstated certainty about care-navigation benefits.
Krump et al. Mar 2026	Answers to 12 clinical questions; coverage of reference concepts.	Included 42/47 critical concepts (89%) and 32/44 noncritical concepts (73%). The reference concept set partly used tool outputs; this measures coverage. Preprint.
Heath et al., NICE ESG Presented Jul 2026	12 topics for rapid-prioritisation briefings.	8/12 topics were judged sufficient with one question formulation. Four insufficient topics had recency or geography problems; this is a usefulness assessment from presentation slides.
Isern-de-Val et al. Date not reported	AI summaries compared with conclusions of an angiotensin-II meta-analysis.	The abstract reports broadly similar conclusions despite limited citation overlap. Elicit-specific overlap and evaluation details require the full text.
Reich Date not reported	Three-paper classroom synthesis and design-planning exercise.	Qualitative comparison of Elicit, NotebookLM and SciSpace; ChatGPT produced preference scores, rather than independent accuracy measurements.
Barres et al. Date not reported	25 causal questions; expert claim checks, limited to 5 min/claim.	98.4% of claims cited real sources; 86.4% justified; 60.8% justified and relevant. Cora-developer benchmark.
Lozano et al. Jul 2023	Clinfo.ai benchmark; automated medical-answer scoring; 146 in tables, 145 in text.	UniEval consistency .352 and relevance .902; BERTScore .807. Automated text scores, not clinical correctness; developer-led study.
Bond et al. (2024)	AI-in-environmental-impact-assessment review.	Elicit citations sometimes supported broader AI claims rather than the specific EIA claim. Qualitative case; mixed-tool workflow; no error-rate denominator.
Wagner et al. (2026)	Programmer-productivity illustrative test.	Mostly abstract-based Elicit summaries judged adequate; only 2/10 retrieved papers relevant. Small example, not summary-accuracy validation.
Zhao (2024)	Library audit; sleep/weight-gain question.	Source checks found weak/indirect support and older sources (latest 2019). One query; no overall accuracy denominator.
Park (2025/26)	Free-interface search/summary inspection.	Side-by-side summaries/source text judged useful. No scored extraction or claim-support dataset; correction scope unresolved.
Hancz (2026)	Applied livestock review; 500-paper retrieval.	Some output entries had mismatched source identities. Qualitative observations; no interpretable error-rate denominator.
Raposio & Baldelli (2025)	Migraine-surgery answer; nine-tool comparison.	Elicit gave four references and medium detail; authors judged broad consistency. Small clinical-answer example; no validated accuracy rubric.
Strasser et al. (2026)	Thrombophilia panel vs expert consensus; abstract-level access.	AI agreement reported, but no verified Elicit-only estimate. Clinical recommendations are not extraction or synthesis accuracy.
Patterson et al. (2025)	Librarian/student rubric and comments.	Paid Elicit rated highly; summaries/relevance praised, transparency and waiting time criticized. Subjective utility, not claim accuracy.
Danler et al. (2024)	One medical-data-warehouse question; eight tools.	98-word Elicit answer, four scholarly references; qualitative source/critical-perspective observations. No accuracy denominator.
Internal Reports (Dec 2025)	40-, 80- and 200-paper report configurations.	≈84% of footnoted claims mostly/fully supported (automated score 4–5/5 against quotes). Supportedness, not completeness or review-conclusion validity.
Internal Reports (Nov 2025)	20 reports/version; writing changed; upstream work fixed.	Reported support improved 74%→85%; minimum report 54%→74%. Raw denominator/aggregation still needs verification; historical version comparison.
Internal human review (2025)	Interim contractor review: 21 evaluations, 12 people.	Human accuracy/usefulness/preferences assessed; final corrected cross-tool results unverified. No quantitative ranking carried forward.
Internal pairwise MA (2026)	Fixed input bundles; archived 56868ef; beta-blocker example.	22-trial random-effects OR 0.781 (0.686–0.890) vs reference 0.78 (0.68–0.89). Both fixed/random results reproduced; model preference differed.
Internal pairwise MA (2026)	Vaping headline pooling and heterogeneity checks.	RR 2.706 (1.334–5.490), $I^2$47.7%, and RR 1.323 (1.192–1.470), $I^2$0% matched references. Small benchmark; preliminary, not rigorous-SLR validation.
Internal MA: methods checks	Archived grading: heterogeneity, sensitivity and interpretation.	Found sensitivity prose retaining a fixed-effect estimate after a random-effects switch, and a prediction claim inconsistent with its prespecified interval. No separate NICE DSU pass rate.

Repeat-run evidence

Output consistency and correctness are separate measures.

EVALUATION	FINDING	LIMITATION
Internal · May 2026	One GLP-1/AUD question; 10 SLR runs; minor whitespace / punctuation changes.	No established frozen corpus or pipeline; not byte-identical input repeats.
Included-paper examples	Five core papers in 10/10 runs; one further paper in 9/10; one borderline paper in 3/10.	Selected examples, not a population sensitivity or error rate.
Extraction variation	Paraphrases; different analysis populations; some factual differences in dates and counts.	No systematic per-run accuracy or change-in-conclusion assessment.
External · Lagisz	476/536 (88.8%) identical outputs; correctness separately 86.6% and 85.6%.	Historical configuration; successfully developed prompts and selected fields.

← Stage summary