Eight AI tools for evidence synthesis

Case studies and comparisons

October 2025

Queen Mary University of London
Health Equity Evidence Centre

Contents

About this document

This document provides an independent assessment of eight Artificial Intelligence (AI) tools used to support evidence synthesis. Each case study outlines the tool's purpose, features, research evidence, and the Health Equity Evidence Centre (HEEC) team's practical experience of using it. The case studies are followed by a comparative section, including a "best for..." guidance, to help readers identify which tool may be most suitable for their needs, and a summary chart for side-by-side comparison.

How this document was produced: The document was informed by a published scoping review (1) of AI tools for automating evidence synthesis, the HEEC team's experience of using these tools, and an exploration of eight specific AI software platforms.

Reference: 1. Harasgama S. JMIR Preprints. [cited 2025 Aug 11]. Artificial intelligence tools for automating evidence synthesis: A scoping review. Available from: https://preprints.jmir.org/preprint/81597

Acknowledgements: This project is supported by the Health Foundation, an independent charitable organisation working to build a healthier UK (ref no: FR-0006738).

Disclaimer: This case study was developed independently by the Health Equity Evidence Centre (HEEC). HEEC has not received funding from any software developer. All evaluations reflect the authors' own analysis and interpretation. HEEC currently use EPPI-Reviewer software under a paid licence to produce living evidence maps which are available online.

CASE STUDY

ASReview

Key facts

Developed by: Utrecht University, Netherlands

Released: 2018 (latest update version 1.6.6 March 2025)

Type(s) of AI employed: Machine learning, active learning

Stage of evidence synthesis: Title and abstract screening

Open source? Yes. The source code is available on Github under an Apache 2.0 licence

Current accessibility: Free to use web-app requiring Python download for installation

What is ASReview?

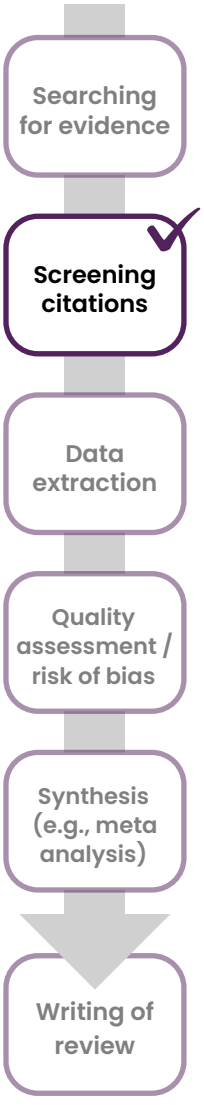
ASReview (Automated Systematic Review) is an open-source software tool designed to accelerate the title and abstract screening phase of systematic literature reviews using machine learning and active learning techniques. ASReview allows users to select the feature extractor and classifier, as well as refine the active learning pipeline. It is possible to export the project file containing all the information to fully reproduce the entire screening phase, which aligns with high-quality systematic reviewing methodology and compliance with reporting standards like PRISMA. The data is stored locally on your own computer, which ensures privacy and data security. It is designed to be extensible, allowing third parties to add modules that enhance the pipeline with new models, data, and other extensions. The tool can also connect with other software, such as reference managers (like Zotero or EndNote) and databases (via RIS, CSV, etc.).

How does it work?

The main ways machine learning is used within EPPI-Reviewer:

1. **Feature extraction:** ASReview transforms the title and abstract (T-A) text of each record into numerical vectors using NLP techniques. The default method uses a combination of TF-IDF vectorisation or word embeddings (e.g., Doc2Vec, fastText) depending on the selected model.
2. **Model selection:** Users can choose from several machine learning classifiers to guide the screening process. Common options include Naïve Bayes, Logistic Regression, Random Forest, and neural networks. These models learn to distinguish between relevant and irrelevant studies based on user labels.
3. **Custom training via active learning:** ASReview is built around an active learning loop. As users screen records and mark them as "relevant" or "irrelevant", the system retrains the classifier in real time. This allows it to reprioritise the remaining unscreened studies, pushing likely inclusions to the top of the queue.
4. **Simulation model:** ASReview includes a simulation feature that allows users to test different model and feature extraction combinations on pre-labelled datasets. This helps researchers benchmark performance and understand how many relevant records can be identified with minimal screening effort.
5. **Model explainability and reproducibility:** While ASReview prioritises transparency (e.g., logging every model decision and ranking), most of the models themselves are relatively simple compared to deep learning systems. This design choice supports reproducibility and interpretability in systematic review workflows.

How can it assist me with evidence synthesis?



What does the research say?

ASReview can significantly reduce screening time and workload while maintaining high accuracy in identifying relevant papers (1,2). In health economics, ASReview identified all data extraction papers within the top 10% of ranked articles (1). For three orthopaedic systematic reviews, all relevant papers were identified after screening 30–40% of the total papers meaning potentially saving 60-70% of screening work (3). Comparisons with other tools suggest ASReview has great potential for improving systematic review efficiency (4). Another study also found that using ASReview resulted in much time saved: only 23% of the articles were assessed by the reviewer (5), resulting in a highly accelerated literature selection process. A study by Nedelcu and colleagues (6) showed that manual screening workload could be reduced by approximately 28% without significantly compromising sensitivity.

How confident can I be in the software?

ASReview employs active learning to assist in prioritising records for screening, but it does not make inclusion or exclusion decisions. The researcher remains in control, labelling each record as relevant or irrelevant. The software learns from these labels to reorder the remaining records, presenting those most likely to be relevant at the top. As the screening progresses and fewer relevant records are found, researchers can decide when to stop, confident that they have likely identified the majority of relevant studies.

What was our experience of using it?

While the initial setup required installing Python and operating through the command prompt – a step that might be unfamiliar to those without a programming background – the comprehensive installation guides provided by ASReview made this process manageable. Once installed, the user-friendly web interface of ASReview LAB made the subsequent steps straightforward. The default 'Oracle' mode was used on a set of 1800+ search results, supplying the software with only three relevant and three irrelevant articles as initial training data. Within minutes, ASReview processed the dataset and prioritised the remaining articles based on their predicted relevance.

Why should I choose this tool?

- Reduces screening time, with flexibility over feature extractor and classifier.
- Excellent user interface and user tutorials.
- Numerous extensions available, such as ASReview Insights, which offers valuable tools for plotting the recall and extracting the statistical results of several performance metrics, such as the Work Saved over Sampling (WSS), the proportion of Relevant Record Found (RRF), the Extra Relevant records Found (ERF), and the Average Time to Discover (ATD).
- Transparent, reproducible and free.

What are the tool's limitations?

- AI-assisted screening is on title-abstract only.

1. Oude Wolcherink MJ, Pouwels XGLV, van Dijk SHB, Doggen CJM, Koffijberg H. Can artificial intelligence separate the wheat from the chaff in systematic reviews of health economic articles? Expert Rev Pharmacoecon Outcomes Res. 2023;23(9):1049–56.
2. van de Schoot R, de Bruin J, Schram R, Zahedi P, de Boer J, Weijdema F, et al. An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell. 2021;3(2):125–33.
3. Pijls BG. Machine Learning assisted systematic reviewing in orthopaedics. J Orthop. 2024 Feb;48:103–6.
4. Pellegrini M, Marsili F. Evaluating software tools to conduct systematic reviews: a feature analysis and user survey. Formre - Open J Formazione Rete. 2021 Jul 31;21(2):124–40.
5. van Dijk SHB, Brusse-Keizer MGJ, Bucsán CC, van der Palen J, Doggen CJM, Lenferink A. Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open. 2023 Jul 7;13(7):e072254.
6. Nedelcu A, Oerther B, Engel H, Sigle A, Schmucker C, Schoots IG, et al. A Machine Learning Framework Reduces the Manual Workload for Systematic Reviews of the Diagnostic Performance of Prostate Magnetic Resonance Imaging. Eur Urol Open Sci. 2023 Oct;56:11–4.

CASE STUDY

ChatGPT

Key facts

Developed by: Open AI

Released: November 2022, with a major update in May 2024 with the release of GPT-4o

Type(s) of AI employed: Large language model

Stage of evidence synthesis: All stages

Open source? No

Current accessibility: Free/$20 per month/$200 per month depending on subscription level

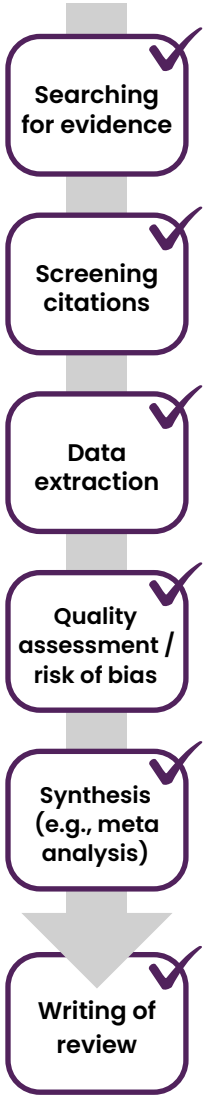
What is ChatGPT?

ChatGPT is increasingly being used to speed up the evidence synthesis process and can be used at all major stages of the literature review. However, it is not designed specifically for evidence reviewing, and as such the models have been trained on a broad corpus, not limited to scientific research. Furthermore, limitations such as the generation of inaccurate or fabricated information, known as hallucinations, mean that caution is required when integrating these tools into evidence synthesis. While ChatGPT and similar tools may offer a more rapid evidence synthesis, their outputs should be critically appraised to ensure accuracy and reliability.

How does it work?

1. **Pretraining corpus:** ChatGPT is trained on a broad dataset that includes websites, books, Wikipedia, forums, and some academic content. However, it is not specifically fine-tuned on peer-reviewed health literature or systematic review datasets.
2. **Transformer model with embeddings:** ChatGPT uses a transformer-based architecture (GPT) that represents text as high-dimensional embeddings. These embeddings capture semantic relationships between concepts (e.g., linking "myocardial infarction" with "heart attack"), enabling the model to interpret prompts and retrieve contextually relevant information. However, the model does not retain links to original sources in its training data. It cannot verify claims or reliably cite specific studies unless provided with source material during the interaction.
3. **Natural language generation:** Based on the input and embeddings, ChatGPT generates fluent, context-aware text. It can draft summaries, rephrase content, structure frameworks, or respond to open-ended questions in natural language. However, the model may occasionally generate inaccurate or fabricated information ("hallucinations"), especially when asked to cite sources or summarise complex material. Outputs should be checked for factual accuracy.
4. **Extraction and synthesis from input text:** When provided with abstracts, structured summaries, or full-text content, ChatGPT can identify and extract information such as study design, sample size, interventions, and outcomes.

How can it assist me with evidence synthesis?



What does the research say?

The use of ChatGPT in evidence synthesis is well-documented in the literature, with significantly more published examples and evaluations compared to other large language models. ChatGPT shows promise in automating article screening with high sensitivity and workload savings (1,2) and has demonstrated high accuracy in data extraction for systematic reviews (2). However, when used for literature searches, ChatGPT's performance was inferior to human experts (3). Some studies have reported ChatGPT's potential to streamline clinical review processes (4) and improve research article quality (5). The effectiveness of ChatGPT depends on the user's skill and the quality of prompt engineering (6), which shape the accuracy of its outputs and mitigate biases. Concerns remain regarding research integrity and ownership when using AI-generated text (5).

How confident can I be in the software?

The underlying technologies are not open source and therefore it is difficult to be fully confident in the software. The LLM was trained on a broad corpus and therefore without careful prompt engineering the results may reflect popular science narratives rather than published research. Users must be aware of hallucinations, such as non-existent references.

What was our experience of using it?

In our experience, ChatGPT proved valuable in the initial stages of evidence synthesis. It assisted in refining our research question and suggesting a coherent structure for the literature review. The model provided potential thematic areas, which we explored further through iterative prompting to gain more comprehensive insights. However, we encountered challenges in ensuring the comprehensiveness of the literature search. ChatGPT's outputs lacked transparency regarding search strategies and inclusion criteria, making it difficult to ascertain the completeness of the evidence base. Additionally, while the model generated references to support its summaries, manual verification revealed inconsistencies and inaccuracies. It should also be noted that we were not harnessing its full potential; leveraging Python could have significantly improved the workflow, and given a more comprehensive and efficient literature search, a structured analysis and summarisation. Therefore, the efficacy of this tool seemed to be more dependent on the skill of the researcher than some of the other tools.

Why should I choose this tool?

- Can assist with all stages of the review.
- With skilled prompt engineering, it can reduce workload and maintain quality (with human oversight).
- User-friendly interface.

What are the tool's limitations?

- General purpose LLM, not trained specifically on scientific research.
- Knowledge cut-off (currently Oct 2023), but can browse the web if this has been enabled by the user.
- As with other LLMs, uses large amounts of energy and water, so less sustainable than more 'traditional' AI techniques.

1. Issaiy M, Ghanaati H, Kolahi S, Shakiba M, Jalali AH, Zarei D, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol. 2024;24(1):11.
2. Motzfeldt Jensen M, Brix Danielsen M, Riis J, Assifuah Kristjansen K, Andersen S, Okubo Y, et al. ChatGPT-4o can serve as the second rater for data extraction in systematic reviews. PLoS One. 2025;20(1):e0313401.
3. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, et al. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. JMIR Med Inform. 2024 May 14;12:e51187.
4. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. J Med Internet Res. 2024 Jan 12;26:e48996.
5. Khlaif ZN, Mousa A, Hattab MK, Itmazi J, Hassan AA, Sanmugam M, et al. The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation. JMIR Med Educ. 2023 Sep 14;9:e47049.
6. Bansal P. Prompt Engineering Importance and Applicability with Generative AI. JCC. 2024;12(10):14–23.

CASE STUDY

Claude

Key facts

Developed by: Anthropic

Released: First released 2023; latest release (Claude 4) May 2025

Type(s) of AI employed: Large language model

Stage of evidence synthesis: All stages

Open source? No

Current accessibility: Free/USD$20 per month/USD$30 per month depending on subscription level

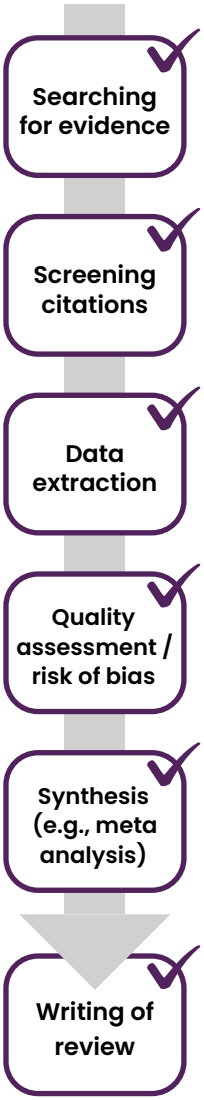
What is Claude?

Claude is a large language model (LLM) tool developed by Anthropic that has shown potential in supporting evidence synthesis tasks. In a recent scoping review of LLMs used in evidence synthesis, Claude was the second most-researched model after ChatGPT, reflecting growing interest in its application within academic and systematic review contexts. While Claude shares many core capabilities with ChatGPT, such as natural language understanding and summarisation, Claude has a notably large context window – useful for handling long or complex documents – and a tendency toward more cautious responses.

How does it work?

1. **Pretraining corpus and alignment:** Claude is trained on a broad dataset that includes websites, books, Wikipedia, forums, and some academic content. However, it is not specifically fine-tuned on peer-reviewed health literature or systematic review datasets. Claude's training includes Anthropic's Constitutional AI approach, which aims to align the model with human values and reduce harmful or unhelpful outputs.
2. **Transformer architecture and embeddings:** Claude is based on a transformer neural network that encodes text into high-dimensional embeddings, capturing semantic relationships between terms and concepts. However, the model does not retain links to its training data and cannot verify claims or cite specific sources.
3. **Natural language generation:** Claude produces fluent, structured responses across a wide range of tasks, such as drafting summaries, rephrasing text, or supporting protocol development. As with other LLMs, Claude may hallucinate information, particularly when summarising detailed content or generating citations. Outputs should be checked for factual accuracy.
4. **Extraction and reasoning over long inputs:** When provided with abstracts, structured summaries, or full-text content, Claude can identify and extract information such as study design, sample size, interventions, and outcomes.

How can it assist me with evidence synthesis?



What does the research say?

Claude has been applied to evidence reviews of randomized controlled trials for both data extraction and risk-of-bias assessments. In two studies, data extraction with Claude 2 achieved an accuracy of 96.3% (with test-retest reliability of 95–97%)(1, 2). In a direct comparison, Claude 2 performed at 96.3% accuracy, while GPT-4 scored 68.8% when relying on a third-party PDF parsing tool; when provided selected text, accuracy increased to 98.7% for Claude 2 and 100% for GPT-4. A study by Lai and colleagues (3) demonstrated that Claude-3.5-sonnet can enhance the accuracy and efficiency of data extraction and risk-of-bias (RoB) assessments, particularly when combined with human oversight. Using Claude only (ie without human oversight) gave a high accuracy of 96.2% but using a hybrid approach of Claude plus human oversight improved the accuracy to >97%, surpassing the conventional manual approach, which had an accuracy of 95.3%. In terms of efficiency, Claude significantly reduced processing time, averaging 82 seconds per RCT for data extraction and 41 seconds for RoB assessment, compared with 86.9 minutes and 10.4 minutes, respectively, for a manual method. However, in a study comparing risk-of-bias assessment between Claude and Cochrane authors, the overall agreement was found to be only 41% (4), and the authors concluded that currently Claude's risk-of-bias judgements cannot replace human risk-of-bias assessment.

How confident can I be in the software?

The underlying technologies are not open source and therefore it is difficult to be fully confident in the software. The LLM was trained on a broad corpus and therefore without careful prompt engineering the results may be inaccurate.

What was our experience of using it?

As with other conversational AI assistants such as ChatGPT, Claude was user-friendly and intuitive. Similarly to ChatGPT, Claude gave a helpful suggested structure for a review article and adapted this according to further instructions. By breaking it down into sections, Claude was able to produce an evidence-based summary with references one section at a time. No obvious errors or hallucinations were noted, but compared to a tool such as Elicit, Claude lacked the ability to 'one-click' to the citation statement and context of the references.

Why should I choose this tool?

- Large context window, so it can process longer documents, maintain more context in conversations, and generate more coherent, informed outputs than some other LLM-based tools.
- Can assist with all stages of the review.
- With skilled prompt engineering, it can reduce workload and maintain quality (with human oversight).

What are the tool's limitations?

- Not trained specifically on scientific research.
- Knowledge cut-off (currently Oct 2024), and cannot browse the web.
- No direct access to academic databases or full-text articles.
- As with other LLMs, uses large amounts of energy and water, so less sustainable than more 'traditional' AI techniques.

1. Konet A, Thomas I, Gartlehner G, Kahwati L, Hilscher R, Kugley S, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods. 2024 Sep;15(5):818–24.
2. Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Res Synth Methods. 2024 Jul;15(4):576–89.
3. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. npj Digit Med. 2025 Jan 31;8(1):1–8.
4. Eisele-Metzger A, Lieberum JL, Toews M, Siemens W, Heilmeyer F, Haverkamp C, et al. Exploring the potential of Claude 2 for risk of bias assessment: Using a large language model to assess randomized controlled trials with RoB 2. Research Synthesis Methods. 2025 Mar 12;1–18.

CASE STUDY

Copilot

Key facts

Developed by: Microsoft

Released: 2023 (initially as Bing Chat)

Type(s) of AI employed: Large language model

Stage of evidence synthesis: All stages

Open source? No

Current accessibility: Free/$20 per month/$30 per month depending on subscription level

What is Copilot?

Microsoft Copilot, integrated within the Microsoft 365 suite, offers general-purpose AI support that can assist with certain aspects of evidence synthesis, such as suggesting a report structure, drafting summaries, and data extraction. Its strengths lie in its integration with tools like Word and Excel, enabling users to streamline repetitive tasks and structure content efficiently. Compared to tools specifically designed for literature reviews, Copilot lacks tailored workflows for screening and risk-of-bias assessment, but it may serve as a complementary aid for improving productivity and clarity in documentation.

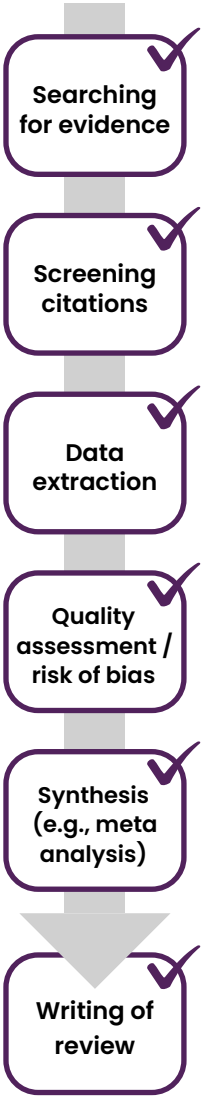
How does it work?

1. **Transformer-based language model:** Copilot uses a large language model (LLM), typically GPT-4 (depending on the application or tasks), built on transformer neural network architecture. These models are trained on a mixture of private and publicly available data, including academic and technical content.
2. **Retrieval-augmented generation** (RAG)**:** When a user submits a prompt (e.g. "Summarise this paragraph"), Copilot first retrieves relevant context from the user's Microsoft 365 environment, such as the open document, recent emails, or files stored in OneDrive or SharePoint. This context is added to the prompt to produce more relevant responses.
3. **In-product integration:** Copilot is embedded directly into Microsoft 365 apps like Word, Excel, Outlook, and Teams, enabling users to generate content, summarise discussions, or automate tasks without leaving their workflow.

What does the research say?

There is a paucity of evidence evaluating the use of Microsoft Copilot in evidence syntheses. One study explored the feasibility of using Bing Chat, which was built on the same technology as Copilot but served as a chatbot rather than an in-app AI assistant, as a supplementary tool for data extraction in systematic reviews (1). The authors propose a method where Bing Chat acts as a "second reviewer" to verify data items initially extracted by a human reviewer. The authors suggest that this technique may serve as an additional verification process, particularly beneficial when resources are limited or for novice reviewers. However, they emphasised that it should not replace established double-independent data extraction methods without further evaluation.

How can it assist me with evidence synthesis?



How confident can I be in the software?

The underlying technologies are not open source and therefore it is difficult to be fully confident in the software. The LLMs were not trained specifically on scientific research and therefore without careful prompt engineering the results may be inaccurate.

What was our experience of using it?

Compared with other LLM-based tools such as ChatGPT and Claude, Microsoft Copilot is more task-focused – it feels less like a conversation and more like an AI feature inside apps (Word, Excel, etc). So while it's powered by conversational AI, the interaction is often less "chatty" and more action-focused. The biggest advantage was the ability to use it alongside other Microsoft 365 tools such as Word and Excel.

Why should I choose this tool?

- Integration with Microsoft software (Excel, Word).
- Searches the web in real time.

What are the tool's limitations?

- Not trained specifically on scientific research.
- Knowledge cut-off (currently April 2023), however, Copilot can access real-time information from the web, allowing it to provide up-to-date responses beyond its training data.
- No direct access to academic databases or full-text articles.
- As with other LLMs, uses large amounts of energy and water, so less sustainable than more 'traditional' AI techniques.

1. Hill JE, Harris C, Clegg A. Methods for using Bing's AI-powered search engine for data extraction for a systematic review. Res Synth Methods. 2024 Mar;15(2):347–53.

CASE STUDY

Elicit

Key facts

Developed by: Elicit Research PBC

Released: 2022

Type(s) of AI employed: Large language model

Stage of evidence synthesis: Most major stages

Open source? No

Current accessibility: Price plans vary from free up to $79/month

What is Elicit?

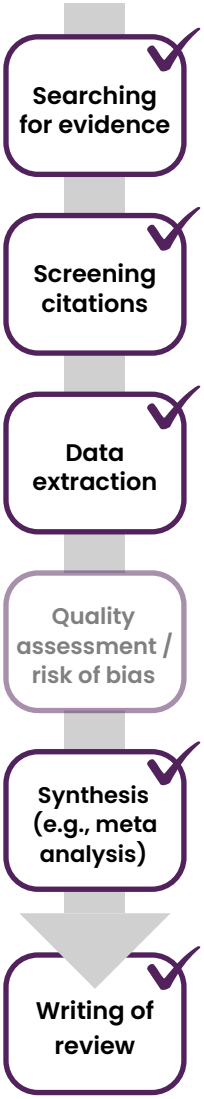
Elicit is an AI-based research tool that allows users to search over 126 million academic papers from the Semantic Scholar corpus. It uses semantic searches, which focus on understanding the meaning of a query rather than relying solely on keyword matching, as is common in tools like Google Scholar. A notable feature of Elicit is its data extraction capability, which includes support for adding custom columns to organise specific information across multiple papers. Users can upload their own PDFs to a personal library and extract data directly from these papers. The platform also includes a "Chat with Papers" function for interacting with individual documents, as well as a "List of Concepts" feature that identifies key topics related to a research query and provides associated references.

How does it work?

Elicit supports evidence synthesis and systematic reviews by combining semantic search, machine learning, and large language models (LLMs) to search, screen, extract, and summarise scientific literature.

1. **Corpus & Coverage:** Indexes 126M+ papers from Semantic Scholar, including journal articles, preprints (arXiv, bioRxiv), and conference papers.
2. **Semantic Embedding:** The title and abstract of every paper in the Semantic Scholar corpus (126M+) are converted into vector embeddings using a pretrained transformer (e.g., SciBERT, all-MiniLM), which capture semantic meaning. The model learns relationships between words and concepts, so that "myocardial infarction" and "heart attack" are recognised as similar even if the exact wording differs.
3. **Semantic Search:** Embeds user queries and retrieves papers based on how closely their meaning matches, even if different words are used.
4. **Process-Based Query Handling:** Breaks tasks into stages (search → screen → extract → summarise), using ML for screening and LLMs for extraction and synthesis. By contrast, ChatGPT answers questions end-to-end in a single step; its Deep Research feature tries to emulate multi-step reasoning but is not tailored for systematic evidence synthesis.
5. **Screening:** Uses LLMs to rank papers by relevance, reducing manual screening workload.
6. **Extraction:** Elicit uses LLMs to extract structured information: outcomes, interventions, sample sizes, populations, etc.
7. **Summarisation:** Uses LLMs to synthesise findings, highlight limitations, and generate summaries with source citations to the exact sentence and paper where each finding comes from.

How can it assist me with evidence synthesis?



What does the research say?

In a comparative study of an AI-assisted versus human-only evidence review, two AI tools (Elicit and Consensus) were used for finding papers for the AI-assisted review. The study found surprisingly little overlap in the list of references for the manual and automated review (1). Even when using the same search ("What is the impact of technology diffusions on growth and productivity in the UK?"), there was no overlap of the top five papers of the Google Scholar search vs Elicit. In another comparative study (2), the results from an umbrella review conducted independently of AI were compared with the results of Elicit searching using the same criteria. Elicit demonstrated moderate reliability with partial overlap in included and excluded studies compared to manual methods; there were three common articles, three exclusively identified by Elicit and 17 exclusively identified by the AI-independent umbrella review search, suggesting that the manual search method was more comprehensive than the Elicit search. Elicit also showed limited repeatability with notable variation across trials. A recent study (3) found that Elicit can effectively automate data extraction for structured information such as study design, but for nuanced or interpretive data, human reviewers are still necessary.

How confident can I be in the software?

The underlying technologies are not open source and therefore it is difficult to be fully confident in the software. However, in comparison to some other LLM-based tools, with Elicit it is easier to check the information because there are direct links to the citation statements.

What was our experience of using it?

The Elicit platform was easy to use and intuitive, making the research process more accessible. The iterative approach allowed us to refine our queries and screening criteria but this meant the process wouldn't meet the criteria for conducting a formal systematic review. One feature we found particularly helpful was being just one click away from the original text quotation. This made it easy to check for inaccuracies (of which there were some), particularly of papers being misquoted. There was a reliance on abstracts rather than full-text at the screening and data extraction, which sometimes limited the quality of the extracted data. Elicit automated the thematic analysis but it is unclear how well it could perform with quantitative data and statistical analysis.

Why should I choose this tool?

- Find relevant papers even if they don't match keywords, and optionally combine these semantic searches with keyword searches.
- Gives custom summaries of the abstract that are specific to your question.
- Flexibility to adjust screening criteria midway without repeating or increasing workload.
- Data extraction table gives links to direct quotes from text.

What are the tool's limitations?

- The search method does not meet the criteria for a systematic review, so a traditional search may be required in addition to Elicit's search output.
- The AI-assisted screening only screens on title-abstract.
- As with other LLMs, hallucinations are possible.
- As with other LLMs, uses large amounts of energy and water, so less sustainable than more 'traditional' AI techniques.

1. GOV.UK [Internet]. [cited 2025 Jun 3]. AI-Assisted vs human-only evidence review: results from a comparative study. Available from: https://www.gov.uk/government/publications/ai-assisted-vs-human-only-evidence-review/ai-assisted-vs-human-only-evidence-review-results-from-a-comparative-study

2. Bernard N, Sagawa Jr Y, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. BMC Med Res Methodol. 2025 Mar 18;25(1):75.

CASE STUDY

EPPI-Reviewer

Key facts

Developed by: EPPI Centre at the Social Science Research Unit, University College London (UCL), UK

Released: 1993 as a desktop application ('EPIC'); 2010 web-based version available to the public (EPPI-Reviewer 4); 2024 (EPPI-Reviewer 6); most recent update 3 July 2025 (version 6.16.3.0)

Type(s) of AI employed: Machine learning, natural language processing, large language models

Stage of evidence synthesis: All stages, but not all are AI assisted

Open source? Partial; the source code for the core of EPPI-Reviewer is available on GitHub, but not yet the source code for the AI components

Current accessibility: £10 per month per user (with unlimited personal reviews), plus £35 per month whilst sharing (collaborating on) a review. (Note, if facilities expire, users still have read access to their reviews)

What is EPPI-Reviewer?

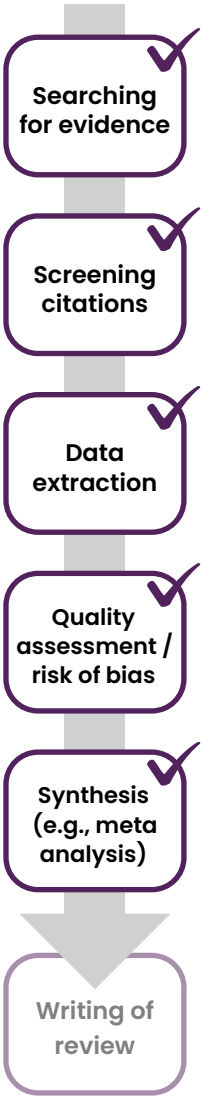
EPPI-Reviewer is a not-for-profit web-based software programme developed by the EPPI Centre (UCL) to support systematic reviews and evidence syntheses. Its wide functionality includes screening, data extraction and meta-analysis. Direct searching in PubMed and search result data transfer can be combined with automatic updates from the OpenAlex database. Articles can be classified using one of several pre-built models or custom-made models, and the results screened using ML-assisted priority screening mode. Evidence maps can be produced in EPPI-Mapper and EPPI-Visualiser. EPPI-Reviewer can be integrated with R packages (such as Metafor) for advanced statistical analyses. A recent update allows automated data extraction using a choice of LLMs (e.g. GPT-4o, DeepSeek) at additional cost.

How does it work?

The main ways machine learning is used within EPPI-Reviewer:

1. **Feature extraction:** EPPI-Reviewer converts the title and abstract (T-A) text into numerical vectors using NLP techniques, including a bag-of-words model and TF-IDF weighting. These vectors serve as the input features for machine learning models.
2. **Pre-built classifiers:** The platform includes ready-made classifiers trained on large datasets (e.g., Cochrane RCT classifier, economic evaluation classifier) that can label studies based on study type. These are especially reliable for biomedical literature and randomised controlled trials (RCTs).
3. **Custom classifier building:** Users can fine-tune their own machine learning models by manually coding a training set (e.g., "include" vs "exclude") and training a classifier to predict labels on new records. This is useful for tailoring the system to specific review topics.
4. **Priority screening (active learning):** EPPI-Reviewer employs an active learning approach where the system iteratively learns from user screening decisions. Decisions train a Support Vector Machine (SVM) classifier that reprioritises the remaining unscreened records, pushing likely relevant ones to the top of the screening list.
5. **Clustering:** Imported references can also be automatically clustered based on shared textual features, helping identify themes or group related studies (although this feature is covered mainly under EPPI-Reviewer's text mining tools).
6. **LLM data extraction:** EPPI-Reviewer has deployed OpenAI's GPT-4o model to automate data extraction from abstracts or full-texts. Outputs are clearly marked as "robot-coded" to distinguish them from human inputs. Additional OpenAI models are available though, being newly introduced, these have not yet been evaluated as extensively.

How can it assist me with evidence synthesis?



What does the research say?

EPPI-Reviewer has demonstrated potential to reduce the burden of systematic review screening (1). In a comparative study, it was found that EPPI-Reviewer could potentially reduce workload by 9% to 60%, while Abstrackr performed within a range of 4% to 49% (2). In a study by Waffenschmidt and colleagues (3), EPPI-Reviewer outperformed Rayyan by identifying 88% of relevant citations after screening 50% of the citation set, compared to 66% with Rayyan, suggesting that EPPI-Reviewer can more effectively prioritise relevant studies. Notably, even when some relevant studies were missed, the overall conclusions remained unchanged. This suggests that EPPI-Reviewer's prioritisation can enhance efficiency without compromising the reliability of review outcomes. A retrospective evaluation found that by using the Cochrane RCT classifier in EPPI-Reviewer it was possible to speed up study selection in qualitative evidence syntheses (4). Thomas and colleagues (5) found a recall of 99.5% and precision of 8% when using a Cochrane RCT Classifier in EPPI-Reviewer to automatically classify citations as likely RCTs or not, leading to a screening workload reduction of 70%.

How confident can I be in the software?

EPPI-Reviewer is a well-established software and its machine learning capabilities have been well validated and reported (1–3). The machine learning components within EPPI-Reviewer should be viewed as tools for prioritisation rather than decision making; rather than excluding studies or making definitive judgments, the algorithm simply reorders citations based on their predicted likelihood of relevance, using patterns it learns from the user's own screening decisions. Since the final inclusion decisions remain fully under the control of human reviewers, users can be confident that the integrity and comprehensiveness of the review are preserved. The LLM-based automated data extraction is awaiting validation.

What was our experience of using it?

There was a learning curve when using EPPI-Reviewer, partly due to the tool's wide range of functions and because the user interface is not as intuitive as some other tools. There is a support team who were quick and helpful. One feature of EPPI-Reviewer that we found particularly useful was the ease with which the evidence could be mapped using EPPI-Visualiser (see examples here). The automated data extraction using GPT-4o was tested on a set of 10 title-abstracts, and correctly extracted the required data (population, intervention, outcome) 85% of the time, partially correct 7.5% and incorrect 7.5% of the time. Automated data extraction can also be performed on full-texts, at a small additional cost (approx. 20p/pdf).

Why should I choose this tool?

- Reduces screening time.
- Highly customisable.
- Can be used across the evidence synthesis pathway.
- Includes classifiers such as the Cochrane RCT Classifier, which was trained on biomedical records, making it a good choice for a review of biomedical literature.
- Articles coded in EPPI-Reviewer can easily be used to build living evidence maps using EPPI-Visualiser.
- Automated data extraction on full-texts.

What are the tool's limitations?

- LLM-based automated data extraction not yet evaluated.
- The LLM-based data extraction feature is likely to have a high environmental cost, as with other LLM-based technologies.

1. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, et al. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods. 2014;5(1):31–49.
2. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev. 2020 Apr 2;9(1):73.
3. Waffenschmidt S, Sieben W, Jakubeit T, Knelangen M, Overesch I, Bühn S, et al. Increasing the efficiency of study selection for systematic reviews using prioritization tools and a single-screening approach. Syst Rev. 2023 Sep 14;12(1):161.
4. Ames HMR, Hestevik CH, Jardim PSJ, Larsen MS, Langøien LJ, Bergsund HB, et al. Can using the Cochrane RCT classifier in EPPI-Reviewer help speed up study selection in qualitative evidence syntheses? A retrospective evaluation. Cochrane Evid Synth Methods. 2025;3(1):e70012.
5. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021 May;133:140–51.

CASE STUDY

Scite

Key facts

Developed by: Josh Nicholson and Anand Desai, now owned by Research Solutions

Released: 2018, with the release of Scite Assistant in May 2023

Type(s) of AI employed: Natural language processing, large language models

Stage of evidence synthesis: Citation searching

Open source? No: classification models and Scite Index calculations are private

Current accessibility: £14.13/month when paying monthly

What is Scite?

The core feature of Scite is Smart Citations which classify citations as supporting, contrasting, or mentioning, providing researchers with contextual insights into how scientific articles are cited. Scite includes a conversational AI tool, called Scite Assistant, which was released in May 2023, which translates natural language queries into searches, summarises results, and provides relevant references. This tool enhances research by analysing over 1.2 billion citation statements, ensuring that AI-generated content aligns with existing scientific evidence. In November 2023, Research Solutions acquired Scite, aiming to integrate its capabilities into a broader suite of research tools. Subsequently, in March 2025, Scite Assistant received significant enhancements, including the deployment of an advanced reasoning AI model optimised for scientific research, further improving its ability to provide accurate and contextually relevant information.

How does it work?

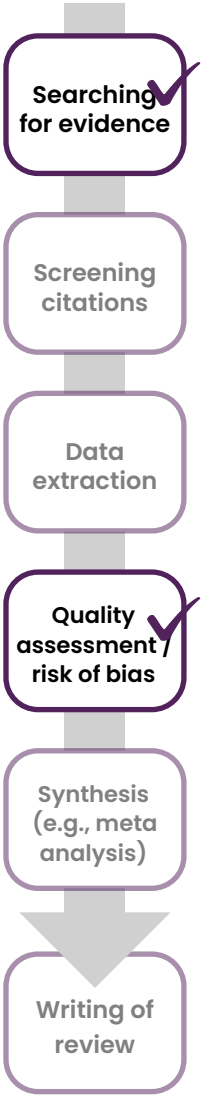
Scite's Smart Citations use machine learning and natural language processing to classify in-text citations as supporting, contrasting, or mentioning. This is done by extracting citation statements from the full text of articles and analysing them with transformer-based models like SciBERT, trained on labelled citation examples. A custom citation-matching engine accurately links each statement to its referenced paper, enabling a structured citation network that captures the context and intent behind each citation.

Scite Assistant uses large language models (LLMs) combined with Scite's structured citation data to help answer research questions and summarise scientific findings. It employs retrieval-augmented generation (RAG), retrieving relevant citation-backed content from the literature before generating responses. This allows the Assistant to provide outputs that are both context-aware and directly linked to supporting or contrasting evidence from peer-reviewed sources.

What does the research say?

Scite is a smart citation index that uses machine learning to categorise citations as mentioning, supporting, or contrasting (1). A critical evaluation by Bakker and colleagues (2) found low overall accuracy in how citations are classified by Scite, especially in distinguishing between supporting and contrasting citations. However, Rife and colleagues (3) argue that Bakker's methodology in their analysis of Scite's classifications is flawed, highlighting the necessity for rigorous, independent assessments of Scite's citation classifications.

How can it assist me with evidence synthesis?



Basumatary and colleagues explored Scite's application in conducting a contextual smart citation analysis of highly cited articles (4). Their study demonstrated that Scite can effectively trace scholarly influence and improve understanding of how knowledge is interlinked across various disciplines, indicating that the tool can enhance the citation analysis process (4). Moreover, the versatility of Scite in generating literature reviews has been assessed, with results indicating that it partially meets many established criteria, although comprehensive completion remains a challenge (5).

How confident can I be in the software?

Scite software is not open source, and this means that the underlying algorithms, data processing methods and LLM approaches are not publicly available making it challenging to fully evaluate the tool's internal workings. Without information on the training data, model architecture, or fine-tuning processes, users cannot comprehensively assess the potential biases or limitations inherent in the AI's responses.

However, Scite mitigates some of these concerns by providing direct access to the references and citation contexts it uses to generate answers. This feature allows users to verify the AI's outputs against the original sources, fostering a level of transparency and enabling critical evaluation of the information presented. Therefore, Scite can be used as a supportive tool, but a critical perspective on the AI-generated content must be maintained.

What was our experience of using it?

Both Scite and Scite Assistant felt intuitive to use. Scite assistant is broadly similar to Elicit, in that users can ask a research question in natural language, and the tool will return a short report including references. Scite, used for checking how a paper has been cited, is a little more niche in its use but the uniqueness of this feature makes it a useful addition to a researcher's toolkit.

Why should I choose this tool?

- Helps check the credibility of a paper by quickly seeing how it has been cited.
- Answers research questions with a report-style answer, with easy-to-check clickable references.
- Offers features like context display, a browser extension, and a reference check for retracted articles.

What are the tool's limitations?

- Access to a slightly smaller number of papers and databases than Google Scholar.
- As with other LLMs, uses large amounts of energy and water, so less sustainable than more 'traditional' AI techniques.

1. Nicholson JM, Mordaunt M, Lopez P, Uppala A, Rosati D, Rodrigues NP, et al. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. Quantitative Science Studies. 2021 Nov 5;2(3):882–98.
2. Bakker C, Theis-Mahon N, Brown SJ. Evaluating the Accuracy of scite, a Smart Citation Index. Hypothesis: Research Journal for Health Information Professionals [Internet]. 2023 Sep 13 [cited 2025 Jun 5];35(2). Available from: https://journals.indianapolis.iu.edu/index.php/hypothesis/article/view/26528
3. Rife S, Nicholson J, Uppala A, Rosati D. Reply to Bakker et al.: Assessing the Accuracy of the Scite Citation Classification System Requires the Same Definitions to be Used for Training as for Testing. Hypothesis: Research Journal for Health Information Professionals [Internet]. 2025 Mar 18 [cited 2025 Jun 5];37(1). Available from: https://journals.indianapolis.iu.edu/index.php/hypothesis/article/view/28018
4. Tracing the footprints of scholarly influence in academia: a contextual smart citation analysis of highly cited articles using Scite | Request PDF. ResearchGate [Internet]. [cited 2025 Jun 5]; Available from: https://www.researchgate.net/publication/381576344_Tracing_the_footprints_of_scholarly_influence_in_academia_a_contextual_smart_citation_analysis_of_highly_cited_articles_using_Scite
5. (PDF) AI literature review systems: an analysis of performance, affordances, and outputs for a complex topic in the social sciences. ResearchGate [Internet]. 2025 Mar 13 [cited 2025 Jun 5]; Available from: https://www.researchgate.net/publication/389743460_AI_literature_review_systems_an_analysis_of_performance_affordances_and_outputs_for_a_complex_topic_in_the_social_sciences

CASE STUDY

TERA (The Evidence Review Accelerator)

Key facts

Developed by: Institute for Evidence-Based Healthcare (IEBH) at Bond University in Australia

Released: 2017 (as SRA, the Systematic Review Accelerator); re-released as TERA in 2024

Type(s) of AI employed: Most tools within TERA are automated but not necessarily AI, with the exception of MechaScreener, which uses a large language model for title-abstract screening

Stage of evidence synthesis: All stages, but not all are AI assisted

Open source? Partial; code for some of the tools is on Github

Current accessibility: Free, or AU$10/month to create more than one review or screen >1000 items/month using MechaScreener

What is TERA?

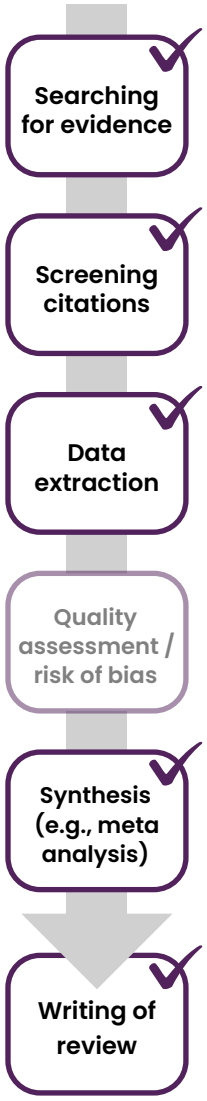
TERA (The Evidence Review Accelerator) is a suite of tools designed to streamline the literature review process, with an emphasis on maintaining rigorous, transparent and reproducible methods. It includes a 'Review Wizard' to take the user through each stage of the literature review (with different review types available). Most stages of the review predominantly use rules-based algorithms. For example, the Review Wizard generates a written methods section populated from specific data entered by the user, which while appearing to be leveraged by generative AI is actually written using a pre-written proforma that randomly alternates to provide variety. The only phase of the review process which uses AI is the screening stage; the recently-released MechaScreener uses an LLM to screen title-abstracts. A traditional non-AI screening tool, Screenatron, is also available within TERA.

How does it work?

TERA comprises several different tools to support the review process; while each is tailored to a specific task, many share similar underlying approaches:

Tool	Description	Primary Technology
Review Wizard	Helps plan and document evidence review	Rule-based (no ML)
Word Frequency Analyser	Counts term frequencies in seed articles to inform search term selection	Basic NLP techniques such as tokenisation, stopword removal and frequency counting
MeshMate	Suggests relevant Medical Subject Headings (MeSH) terms, using both keyword-based matching and semantic BERT-based models	Hybrid (Rule + ML)
SearchRefiner	Visualises search terms and results, to help refine searches	Rule-based (no ML)
Polyglot Search Translator	Converts search syntax between databases using predefined mappings	Rule-based (no ML)
Deduplicator	Identifies and merges duplicate records using string matching rules	Rule-based (no ML)
Screenatron	Supports human-only screening workflows	Manual
MechaScreener	Uses (undisclosed) LLM to screen title-abstracts	Large Language Model
Disputatron	Resolves conflicts between reviewer decisions via deterministic rules	Rule-based (no ML)

How can it assist me with evidence synthesis?



Tool	Description	Primary Technology
SpiderCite	Performs citation chaining (snowballing) using network logic	Rule-based (no ML)
TERA Farmer	Returns similar records based on PubMed's Best Match algorithm (1)	Hybrid (Rule-based+ ML)
Calculon	Calculates missing statistics using formulaic rules	Rule-based (no ML)
MetaPairwise, MetaInsight, MetaDTA	Conducts various meta-analyses; some include ML-driven model fitting	Hybrid (Rule-based+ ML)
Replicant	Generates write-ups using template-driven logic	Rule-based (no ML)

What does the research say?

In one evaluation of the Systematic Review Accelerator suite (now TERA), it was found that for the majority of SR tasks where an SRA tool was used, the time required to complete that task was reduced while methodological quality was maintained (2). For the six systematic review tasks in which times were compared, the manual team spent 2493 minutes (42 hours) on the tasks, compared to 708 minutes (12 hours) spent by the automation team. The manual team had a higher error rate in two of the six tasks, a lower error rate in one of the six tasks, and similar error rates for the two compared tasks. One task could not be compared between groups. Another study looking at accuracy measures of automated deduplication tools found SRA to be adequately accurate as a deduplication tool, and comparable with Mendeley and Rayyan (3). Forbes and colleagues (4) found the Deduplicator tool within SRA was faster and had a lower error rate than a semi-manual Endnote method. A case study (5) found that a small and experienced systematic reviewer team using Systematic Review Automation tools who have protected time to focus solely on the SR can complete a moderately sized SR in two weeks.

How confident can I be in the software?

TERA's approach combines innovation with the rigour of traditional review methods. With the exception of the MechaScreener tool, TERA tools are only using rules-based algorithms, and this limited and controlled use of computational power makes the platform trustworthy, as it enhances efficiency without compromising the reliability or accuracy of the evidence synthesis process. MechaScreener is currently being evaluated and results have not yet been published.

What was our experience of using it?

The TERA platform was easy to navigate, and we particularly appreciated how the Review Wizard guides the user step-by-step through each stage of the evidence synthesis process. Its structured workflow (and Review Plan) made it simple to stay organised and focused. One of the standout features was how seamlessly the platform incorporated all the key stages of a review – from search strategy development to screening and synthesis.

Why should I choose this tool?

- A whole suite of tools within one platform
- Guides the user step-by-step through each step of the evidence review process, helping ensure methodological consistency and completeness
- Review Wizard that is customisable to type of evidence review (from systematic to scoping)

What are the tool's limitations?

- The MechaScreener screens on title-abstract only, and not full text.

1. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. Best Match: New relevance search for PubMed. PLoS Biol. 2018 Aug;16(8):e2005343.
2. Clark J, McFarlane C, Cleo G, Ishikawa Ramos C, Marshall S. The Impact of Systematic Review Automation Tools on Methodological Quality and Time Taken to Complete Systematic Review Tasks: Case Study. JMIR Med Educ. 2021 May 31;7(2):e24418.
3. Guimaraes NS, Ferreira AJF, Silva RDR, de Paula AA, Lisboa CS, Magno L, et al. Deduplicating records in systematic reviews: there are free, accurate automated ways to do so. J Clin Epidemiol. 2022;152:110–5.
4. Forbes C, Greenwood H, Carter M, Clark J. Automation of duplicate record detection for systematic reviews: Deduplicator. Syst Rev. 2024;13(1):206.
5. Clark J, Glasziou P, Del Mar C, Bannach-Brown A, Stehlik P, Scott AM. A full systematic review was completed in 2 weeks using automation tools: a case study. J Clin Epidemiol. 2020 May;121:81–90.

Best for... choosing the right tool for the job

Refining research strategy

TERA

This suite offers specialised tools like MeshMate for suggesting MeSH terms, SearchRefiner for visualising search terms, Polyglot Search Translator for converting search syntax between databases, and TERA Farmer for finding similar records based on PubMed's Best Match algorithm.

Reproducibility & transparency

ASReview

Offers high confidence due to its open-source nature, allowing full reproducibility of the screening phase by exporting the project file and allowing researchers to quality assess outputs.

EPPI-Reviewer

A well-established software with extensively validated machine learning capabilities. Its ML components can prioritise records rather than making final decisions, ensuring human reviewers retain control and preserve review integrity.

TERA

Inspires high confidence as most of its tools rely on rules-based algorithms, which, unlike complex LLMs, enhance efficiency without compromising reliability or accuracy.

Citation analysis

SCITE

Its core feature, Smart Citations, assesses the citation's impact, offering contextual insights into how articles are cited.

Managing the review process from start to finish

TERA

This is a comprehensive suite of tools designed to streamline the entire literature review process. Its structured workflow helps users stay organised and focused.

EPPI-Reviewer

Offers wide functionality and can be used across the entire evidence synthesis pathway.

Title-abstract screening

ASReview

Designed specifically to accelerate title and abstract screening using machine learning with active learning. Significantly reduce screening time and workload (60-70% savings) while maintaining high accuracy (1).

EPPI-Reviewer

Offers an ML-assisted priority screening mode. Studies show it can reduce workload by 9% to 60% and outperforms Rayyan (2). Its Cochrane RCT classifier saved 70% screening workload with 99.5% recall (3).

Data extraction

Claude

Claude 2 has high accuracy in data extraction (96.3%) (4,5). A hybrid approach combining Claude with human oversight achieved >97% accuracy and significantly reduced processing time, averaging 82 seconds for extraction versus 86.9 minutes manually (6).

Elicit

Elicit uses large language models to extract structured information like outcomes, interventions, sample sizes, and populations. Users can upload their own PDFs for direct data extraction.

User friendliness

ChatGPT

Natural language chatbot requires no prior experience or knowledge, but the quality of the output will be directly affected by the quality of the prompts.

ASReview

Despite the installation process requiring command prompts, this was easy to do with the website's how-to guide. After installation, it was easy to use with an intuitive user interface.

Adherence to systematic review guidelines

ASReview

Allows users to export project files for full reproducibility of the screening phase.

EPPI-Reviewer

Adheres to systematic review guidelines by supporting transparent, structured workflows – including screening, data extraction, synthesis, and reporting.

TERA

Built with an emphasis on maintaining rigorous, transparent, and reproducible methods using predominantly rules-based algorithms. Its 'Review Wizard' helps ensure methodological consistency and completeness.

Comparison chart

Tool	AI-Assisted Stages	Current Cost	Key Research Finding	Transparency	User Experience Highlights	Key Limitations
ASReview (ML, Active Learning)	Title-abstract screening	Free	Reduces screening time/workload while maintaining high accuracy.	High confidence: open source, project file exportable for reproducibility, data stored locally. Researcher remains in control; software prioritises, not decides.	Initial setup (Python) required, but comprehensive guides helped. User-friendly web interface. Default 'Oracle' mode effective with minimal training data. Excellent user interface & tutorials.	AI-assisted screening on T-A only.
ChateGPT (LLM)	All stages	Free – US$200/ month	Shows promise in automating screening (high sensitivity/workload savings) and data extraction (high accuracy).	Difficult to be fully confident: not open source, trained on broad corpus, outputs may reflect popular narratives. Hallucinations are a concern.	Valuable for refining research questions & structuring reviews. Search not comprehensive & inaccuracies in references. Efficacy dependent on prompt engineering. User-friendly interface. We paid US$20 month but the free version would cover most requirements.	Not trained specifically on scientific research. Knowledge cut-off (Oct 2023, though can browse web).
Claude (LLM)	All stages	Free – US$30/ month	High accuracy with data extraction, improved further with human oversight. Significantly reduced processing time.	Difficult to be fully confident: not open source, trained on broad corpus, results may be inaccurate without careful prompt engineering. Hallucinations possible.	User-friendly and intuitive. Helpful for review structure, evidence-based summaries with references (one section at a time). Large context window (processes longer documents, maintains context). We paid US$13/month but the free version would cover most requirements.	Not trained specifically on scientific research. Knowledge cut-off (Oct 2024, cannot browse web).
Copilot (LLM)	All stages	Free – US$30/ month	Limited evidence. Bing Chat (same tech) explored for data extraction verification.	Difficult to be fully confident: not open source, LLMs not trained specifically on scientific research. Hallucinations possible.	More task-focused than conversational. Biggest advantage: integration with Microsoft 365 tools (Word, Excel). Searches the web in real time. We used the free version which was adequate for our needs.	Not trained specifically on scientific research. Knowledge cut-off (Apr 2023, though can access real-time web).
EPPI-Reviewer (ML, NLP, LLM)	All stages (but not all AI-assisted	£10/ month	Can reduce workload significantly whilst maintaining high recall. Outperformed Rayyan in prioritising relevant studies.	High confidence: ML capabilities validated, prioritises not decides. Final inclusion decisions remain with human reviewers. Partially open source. LLM data extraction awaiting validation.	Highly customisable across evidence synthesis pathway. Learning curve due to wide functions/interface. Support team helpful. Ease of evidence mapping using EPPIVisualiser.	LLM-based automated data extraction not yet evaluated.
Elicit (LLM)	Most major stages	Free – US$79/ month	Moderate reliability, partial overlap with manual methods; manual search more comprehensive. Limited repeatability across trials.	Difficult to be fully confident: not open source. Easier to check info due to direct links to citation statements. Hallucinations possible.	User-friendly and intuitive. Iterative approach for refining queries. One-click access to original text quotation for checking inaccuracies. Flexibility to adjust screening criteria mid-way. We paid US$49/month to allow data extraction on up to 200 PDFs.	Search method does not meet criteria for systematic review. Screening on T-A only. Hallucinations possible.
Scite (NLP, LLM)	Citation searching and tracking	US$12/ month	Low accuracy in classifying supporting/contrasting citations in one study, but others argue methodology flawed.	Difficult to be fully confident: not open source. Mitigates concern by providing direct access to references/citation contexts for verification.	Intuitive to use. Scite Assistant similar to Elicit (report answers with references). Scite (checking citations) is niche but useful.	Slightly smaller number of papers/databases than Google Scholar.
TERA (LLM (MechaScreener only), rules based for others)	All stages (but not all AI-assisted)	Free – AU$10/ month	Reduced task time while maintaining quality. Deduplicator tool faster and lower error than semi-manual method.	High confidence: mostly rules-based; enhances efficiency without compromising reliability. Partially open source. MechaScreener awaiting published evaluation.	Easy to navigate. Review Wizard guides step-by-step, ensuring consistency. Seamlessly incorporated key stages. We used the free version which allows a single review and the use of MechaScreener for up to 10,000 items.	MechaScreener screens on title-abstract only, not full text.

References

1. Pijls BG. Machine Learning assisted systematic reviewing in orthopaedics. J Orthop. 2024 Feb;48:103–6.
2. Waffenschmidt S, Sieben W, Jakubeit T, Knelangen M, Overesch I, Bühn S, et al. Increasing the efficiency of study selection for systematic reviews using prioritization tools and a single-screening approach. Syst Rev. 2023;12(1):11.
3. Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, et al. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. J Clin Epidemiol. 2021 May;133:140–51.
4. Konet A, Thomas I, Gartlehner G, Kahwati L, Hilscher R, Kugley S, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods. 2024;15(5):818–24.
5. Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. Res Synth Methods. 2024 July;15(4):576–89.
6. Lai H, Liu J, Bai C, Liu H, Pan B, Luo X, et al. Language models for data extraction and risk of bias assessment in complementary medicine. npj Digit Med. 2025 Jan 31;8(1):1–8.