Working towards the responsible use of AI in evidence synthesis: the RAISE project

UHMLG Spring Forum April 2026

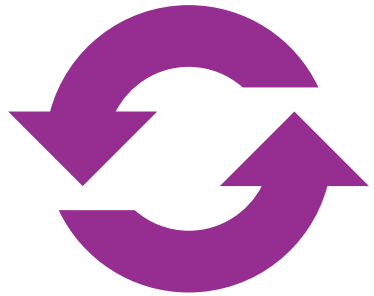
Anna Noel-Storr, Head of Evidence Pipeline & Data Curation, Cochrane

Trusted evidence.
Informed decisions.
Better health.

Core principles of evidence synthesis

Rigor

Transparency



Replicability

High stakes

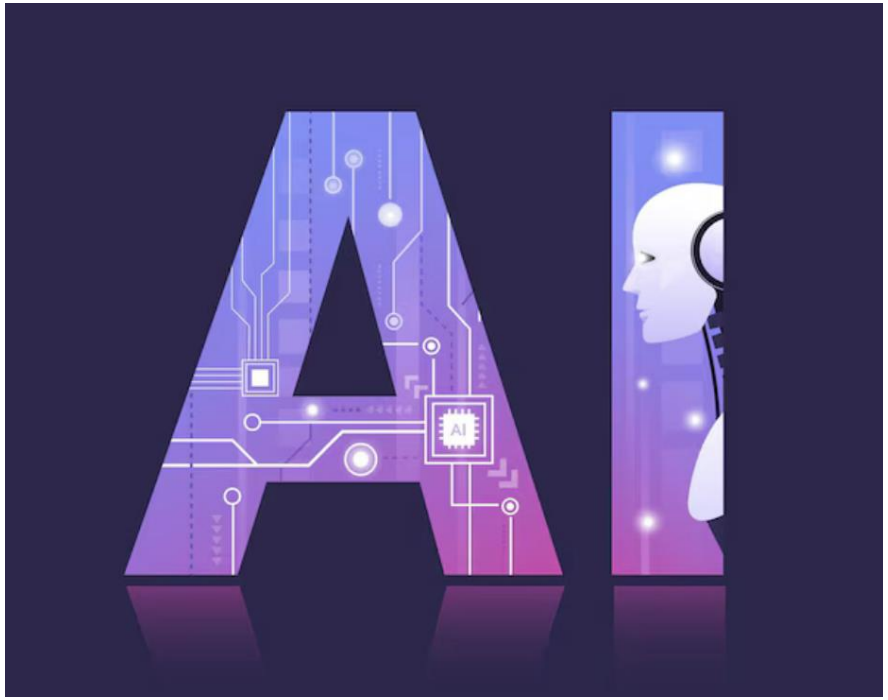
The stakes are high. Decisions that affect people's lives are made based on evidence synthesis

Plenty of excitement (and uncertainty) about AI

What's so special about LLMs?

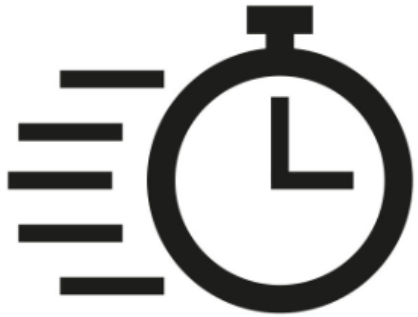
Generative AI uses Large Language Models (LLMs) which have been trained on vast datasets.

LLMs are able to recognize, summarize, translate, predict, and generate text without any specific task training or only a few instructions as a form of prompts

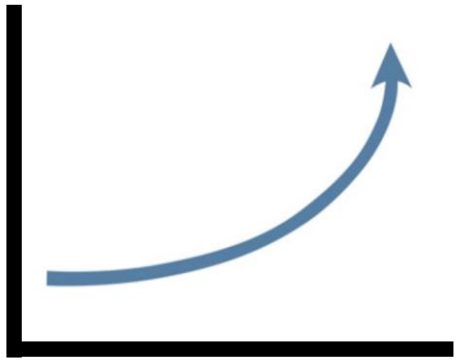


Plenty of excitement (and uncertainty) about AI

Increases capacity and brings new capability



Speed

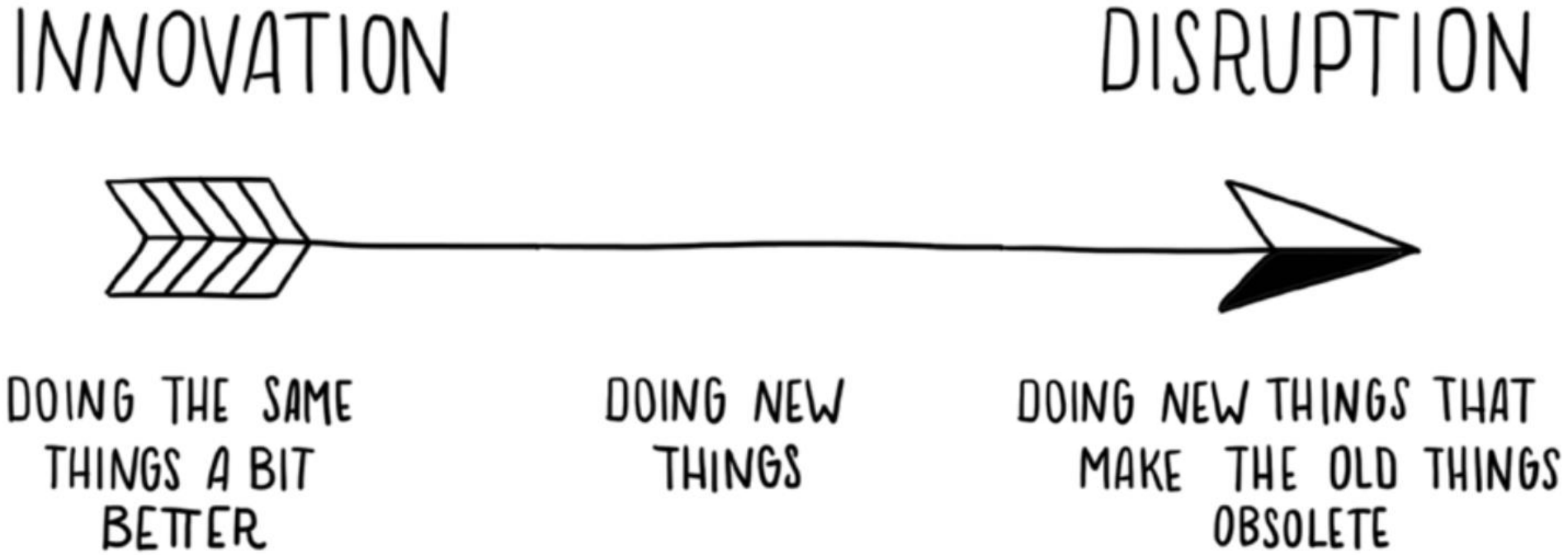


Scale

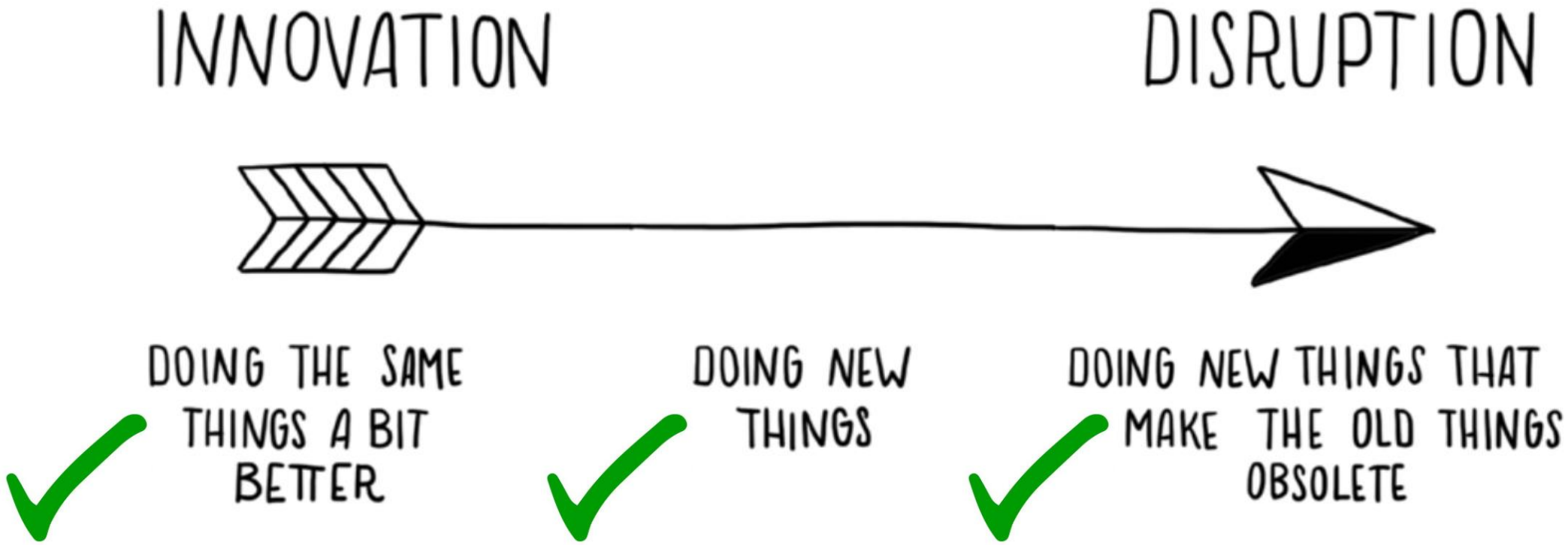
Breadth

LLMs have dramatically opened-up the potential task space both in terms of capacity and capability

Incredible potential



Incredible potential

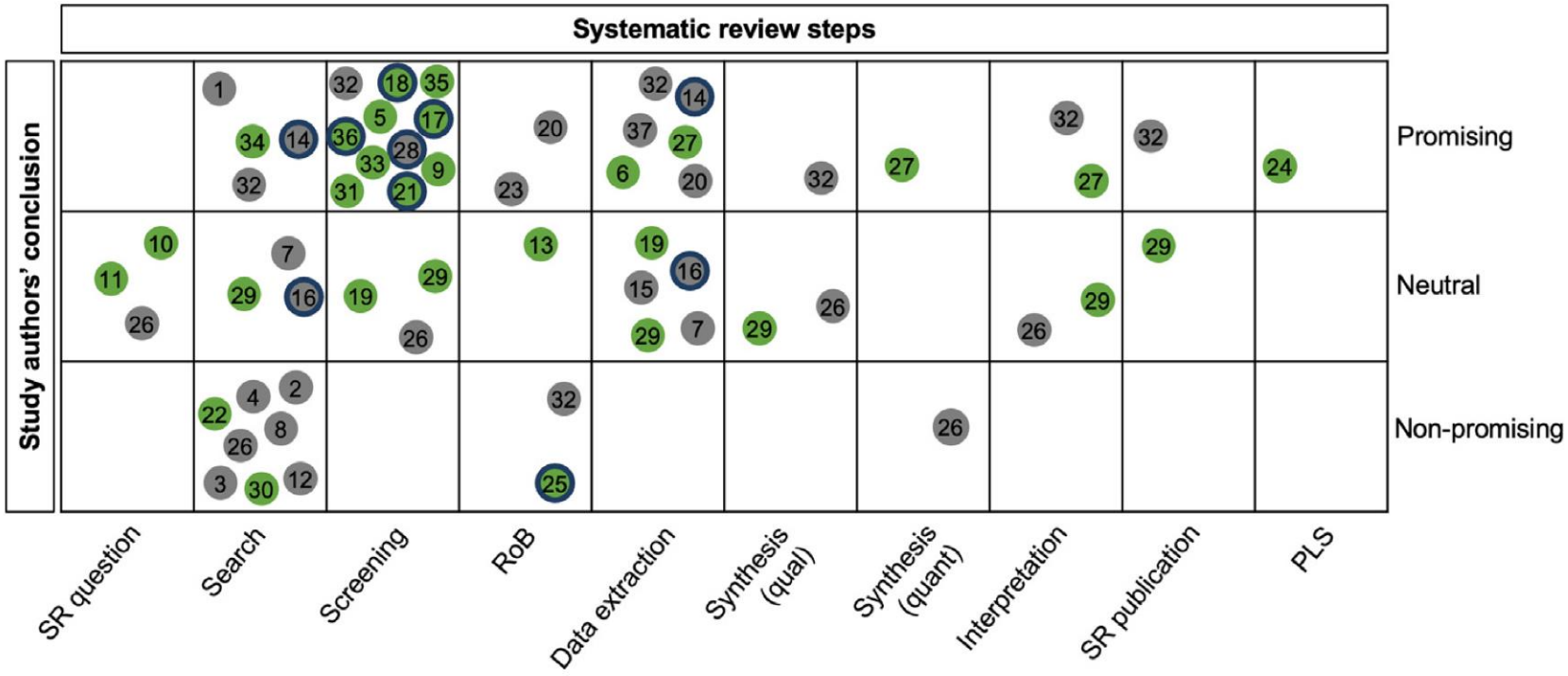


Makes if quite challenging to know where to focus effort

Lieberum JL, Töws M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, Böhringer D, Meerpohl JJ, Eisele-Metzger A. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol. 2025 Feb 26;181:111746

Large Language Models offer new potential

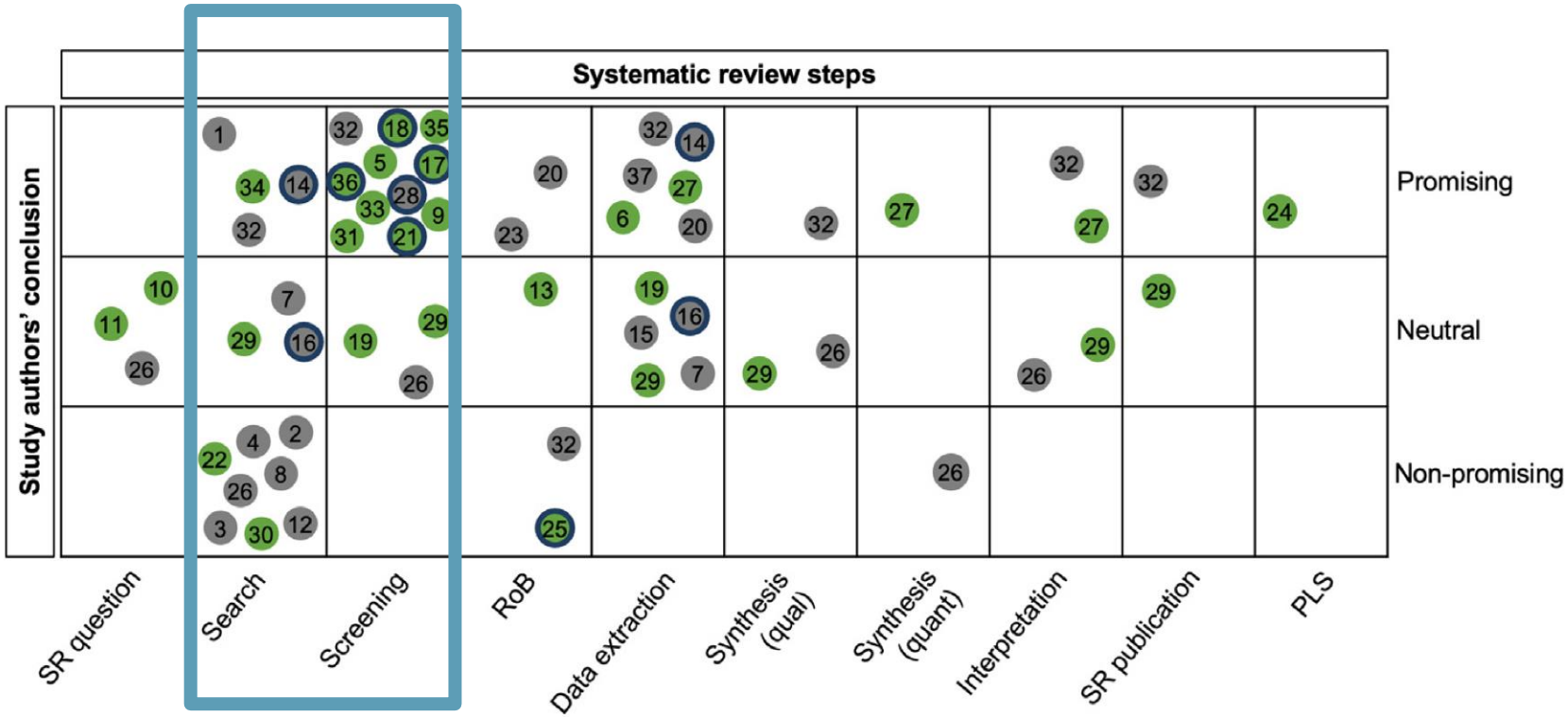
Methodological studies are emerging and have covered



Lieberum JL, Töws M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, Böhringer D, Meerpohl JJ, Eisele-Metzger A. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. J Clin Epidemiol. 2025 Feb 26;181:111746

Large Language Models offer new potential

Methodological studies are emerging and have covered



Multiple challenges

Multiple challenges to effective integration:
methodological, **ethical**, **operational**, **technological**

Using AI effectively in evidence synthesis

Integration of AI is challenging – it can feel like we're trying to fly a plane whilst still building it

A global challenge

We have a responsibility to use AI, but only if we can use it safely and responsibly.

To do this we need guidance, frameworks and infrastructure that enables us to learn and evolve as the technology itself learns and evolves.

A challenge for the whole ecosystem

Introducing RAISE (Responsible use of AI in evidence SynthEsis)

We were asked to write some guidance…

… about which tool to use, and when

But found we couldn't!

The evidence base on which to base our advice was very limited

AI tools were being developed that were not engineered to be fit-for-purpose

Recommendations and guidance

Three-paper RAISE collection

1. **Responsible AI in Evidence synthesis 1:**
   Recommendations for practice
2. **Responsible AI in Evidence synthesis 2:**
   Building and evaluating evidence synthesis tools
3. **Responsible AI in Evidence synthesis 3:**
   Selecting and using evidence synthesis tools

Paper 1: Recommendations for practice

1

- How the RAISE papers have been produced
- What **type of AI** is covered by RAISE
- Defining the **main roles in the ecosystem** and the recommendation for practice for each role

Paper 1: Recommendations for practice

- What type of AI is covered by RAISE?

In RAISE, AI is defined as follows:

"…a set of advanced technologies that enable machines to do highly complex tasks effectively – which would require intelligence if a person were to perform them"

Harwich et al, 2018

Machine learning – uses algorithms to make predictions

Natural language processing – understands and generates human language

Expert systems that use rule-based algorithms to mimic human decisions

Deep learning, including LLMs that focus on text and language

Cross-cutting generative AI to generate text, figures or code

Paper 1: Recommendations for practice

- Defining the **main roles in the ecosystem** and the recommendation for practice for each role

RAISE includes **eight distinct roles**, each with different expertise and needs, along with recommendations to define responsible AI use. Since these recommendations are interlinked, **all roles need to participate collaboratively** across the ecosystem to support adherence to the recommendations.

With each role in the ecosystem playing its part, a **virtuous cycle** is created where tools are developed that are fit-for-purpose and ready for use, making evidence synthesis faster, and cost-effective, and **ultimately leading to better decisions and improved outcomes.**

Main roles in the ecosystem

- Evidence synthesists
- Methodologists
- AI tool developers
- Organisations that produce evidence synthesis
- Funders and commissioners
- Publishers of evidence synthesis
- Trainers of evidence synthesis methods
- User of evidence synthesis

One person / organisation may play multiple roles
Each has a role to play in developing and using AI in a responsible way

Paper 2: Building and evaluating evidence synthesis tools

- Guidance on **building and evaluating** AI tools
- **Performance metrics** considerations for AI tool evaluations
- Guidance on **how to report** building and evaluating AI tools

Paper 2: Building and evaluating evidence synthesis tools

- Guidance on building and evaluating AI tools

The two main phases of building an AI tool are described: the **build** phase and the **validation** phase.

Includes issues for building and validating an LLM, including:

- **Prompt development:** Creating prompts is essentially building a model
- **Response stability:** LLMs can give different output to the same prompt on repeated tests. Evaluations should assess the consequences of this behaviour.

Paper 2: Building and evaluating evidence synthesis tools

- Performance metric considerations AI tool evaluations

Key considerations related to LLMs:

- **Stability:** Few established metrics for evaluating stability, however, estimating the frequency and seriousness of deviation, may serve as useful starting point

- **Robustness:** The ability of a system to handle unexpected inputs in a predictable manner should be one of the aspects tested and validated during evaluation

- **Hallucinations:** Range of hallucination scenarios. Various ways to test type of scenario eg posing nonsensical questions to see if the system accurately states that the information is not available

Paper 2: Building and evaluating evidence synthesis tools

Taxonomy of common performance metrics for AI tools for evidence synthesis

1. Accuracy metrics	2. Reliability metrics	3. Efficiency metrics	4. Usability metrics	5. Risk assessment
1.1 Classification performance *(confusion matrix, sensitivity, specificity, accuracy, F1 scores)*	**2.2 Consistency** *(inter-rater reliability, response stability)*	**3.1 Resource utilisation** *(cost-effectiveness, workload reduction, time efficiency)*	**4.1 User experience** *(user satisfaction and usability)*	**5.1 Error impact** (error severity scoring, missed critical information, changes in synthesis result)
1.2 Content generation quality *(text and semantic similarity, completeness, correctness)*	**2.3 Robustness** *(input robustness)*	**3.2 Workload impact** *(time-on-task, human effort required for verification)*	**4.2 User integration** *(implementation barriers, Interoperability measures)*	**5.2 Safety and ethics** *(transparency, bias, environmental impact, topic-specific safety)*

Paper 2: Building and evaluating evidence synthesis tools

- Guidance on **how to report** building and evaluating AI tools

Adapted framework by Kolbinger et al for use in the evidence synthesis context. Authors adapt the tool as needed to suit their specific use case.

Structured around 4 key sections, each with reporting requirement sub-sections:

- **Introduction**: Existing tools; AI Tool; Objective
- **Methods:** Study design; Setting; Data sources; Data selection; Data processing; Labelling in input/validation data; Type of tool/model; Tool development; Performance accuracy; Validation methods; Performance errors; Interpretation
- **Discussion:** Strengths and limitations; Bias; Practical value of the tool; Implications for use
- **Other information:** Ethical statement; Availability of protocol; Sources of support; Declarations of interest; Availability of data/code; Replicability; Environmental impacts

Paper 3: Selecting and using evidence synthesis tools

- The **current state** of the AI tools
- **Selecting and using** an AI tool
- **Assessing the suitability** of an AI tool for use in an evidence synthesis

AI use expectations for evidence synthesists

1. Evidence synthesists are ultimately responsible for their research, including the decision to use AI and how it is used
2. Evidence synthesists can use AI provided they can demonstrate it will not compromise the methodological rigor or integrity of their synthesis
3. AI use should be fully and transparently reported
4. AI should be used with human oversight

Cochrane

Expectation 1: Evidence synthesists are ultimately responsible for their research, including the decision to use AI and how it is used

Introducing the responsible handover framework for an AI tool for evidence synthesists

Responsible handover framework

- What is the purpose of the AI tool?
- Where have the training and testing data come from?
- Is the AI tool validated and performs sufficiently for use?
- Usability and user capability
- Transparency, licenses, availability and documentation

From RAISE 3

Domain 1: What is the purpose of the tool?

Main considerations for evidence synthesists:

- ❑ Is the intended use case and expected benefit clearly defined?
- ❑ Does it match your specific task and context?
- ❑ Is the level of human oversight appropriate?
- ❑ For LLMs, are the model version and prompts recorded?

Domain 2: Where have the training and testing data come from?

Main considerations for evidence synthesists:

- ❑ Are training and testing data sources disclosed and accessible?
- ❑ Do training and testing data match your evidence synthesis domain/context?
- ❑ Are there known limitations or biases (language, geography, date range, methods, etc.)?
- ❑ Is there a separation between the training, testing and validation datasets?

Domain 3: Is the AI tool validated and performs sufficiently for use?

Main considerations for evidence synthesists:

- ❑ Has performance been validated?
- ❑ Is the validation published and / or peer-reviewed?
- ❑ Is it reported in sufficient enough detail to be transparent and replicable?
- ❑ Are metrics appropriate and does performance meet your acceptable threshold?
- ❑ Can performance change over time?

Domain 4: Usability and user capability

Main considerations for evidence synthesists:

- ❑ Does your team have the skills to use it reliably?
- ❑ Is training/documentation available?
- ❑ What are the costs (financial, time, infrastructure)?
- ❑ Is user support available?

Domain 5: Transparency, licenses, availability and documentation

Main considerations for evidence synthesists:

- ❑ Are the terms of use clear and acceptable?
  - ❑ Plagiarism and provenance; copyright and intellectual property; jurisdiction and licensing; data security and usage rights; and data protection and confidentiality
- ❑ Will uploaded content be used for training? Can you opt out?
- ❑ Does the tool meet your legal and ethical requirements?
- ❑ Is there a conflict-of-interest disclosure?

Consider not proceeding if...

- No published validation in a relevant context
- Concerns about the replicability of the validation
- Concerns about performance claims based only on the developer's own validation and /or methodological limitations
- Lack of legal or policy compliance at the evidence synthesist’s or tool user’s organisation, or at the national or international level
- Terms allow use of your content for model training without opt-out
- Inappropriate level of human oversight available, e.g. no way to monitor or audit outputs
- The AI tool developer is unresponsive to questions

Decisions after using the framework

Proceed

- AI was validated for its proposed use; supporting evidence is strong
- Risk of well-understood with moderate errors or inconsequential differences
- Limitations known and reasonably manageable

Proceed with mitigations

- Tool shows promise but has gaps in evidence,
- Requires additional monitoring, or presents moderate risks that can be actively managed.

Do not proceed

- AI not validated for its proposed used; missing or weak supporting evidence
- Transparency is insufficient for meaningful assessment
- Risks cannot be adequately mitigated; AI unreliable

Expectation 2: Demonstrate AI will not compromise the methodological rigor or integrity of the synthesis

Current state of AI tools

Table 2: current (February 2026) state of AI tools

Task	Tool Class	Detail and considerations	Example tools	Recommendation
Writing a proto				
Question formulation	Generative LLMs	Asking LLMs to provide novel questions for synthesis may support early question development. However, suggestions may be incomplete, irrelevant, subject to bias (based on its sources), or overlap with past reviews.	ChatGPT, CoPilot, Claude, Gemini, DeepSeek	Human verification required
Drafting	Generative LLMs	Pre-trained LLMs can provide an outline using well-established protocol formats. Users may also provide a format / direct the LLM to resources to support this.	ChatGPT, CoPilot, Claude, Gemini, DeepSeek	Human verification required
The Search				
Exploring the literature	Unsupervised	Topic modelling tools aid in identifying clusters of evidence quickly to get a sense of key themes/areas of interest.	Carrot2 (https://search.carro 2 .org)	Acceptable for use
	Agentic AI	AI agents develop, refine, and perform searches based on natural language queries. Highly dependent on data sources the tool has access to and requires human input at each stage to guide agent. May be helpful to gain a sense of the literature at an early stage but should not be used as part of any formal evidence retrieval.	Undermind (https://www.under i nd.ai/), Elicit (https://elicit.com/), Asta Find Papers (https://asta.allen.a	Acceptable for use
Search strategy development	Rule-based	Tools analyse frequency of keywords and/or controlled vocabularies in search results. Specialised tools are required to cover indexing from different bibliographic databases. May provide additional keywords to inform search strategy but should be used in combination with other search development methods.	Yale MeSH Analyzer (https://mesh.med.y l e.edu/), TERA WordFreq (https://tera-tools.com/word-freq PubReMiner (https://hgserver2.a	Acceptable for use

Excerpt from RAISE 3, Table 2

Cochrane

Current state of AI tools

Table 1: Tool use recommendations

Recommendation	
Acceptable for use	AI outputs may be used directly within the review workflow, if any limitations or potential biases are acknowledged and accounted for.
Human verification required	AI outputs may be used to support review tasks but must be carefully checked by humans before use. The degree of checking required may vary, but typically this will require a human to read and possibly make amendments to the entirety of the output.
Requires validation within the review	AI outputs may be used if their performance is explicitly evaluated within the context of the review itself and deemed adequate (e.g. comparable to human performance).
Exploratory and supplementary use	AI outputs may be used for developing ideas or as a starting point to support understanding. All outputs should be extensively refined by human reviewers prior to use for a review task. Alternatively, outputs may be appropriate for use as an additional, supplementary approach, but without replacing established processes.
Not acceptable for use	The current state of technology means that these AI outputs have such serious limitations, that they should not be relied upon.

Table 2: current (February 2026) state of AI tools

Task	Tool Class	Detail and considerations		
Writing a protocol				
Question formulation	Generative LLMs	Asking LLMs to provide nove question development. How subject to bias (based on its		
Drafting	Generative LLMs	Pre-trained LLMs can provid formats. Users may also pro support this.		
The Search				
Exploring the literature	Unsupervised	Topic modelling tools aid in sense of key themes/areas o		
	Agentic AI	AI agents develop, refine, ar queries. Highly dependent o human input at each stage t the literature at an early sta evidence retrieval.		
Search strategy development	Rule-based	Tools analyse frequency of k results. Specialised tools are bibliographic databases. Ma strategy but should be used in combination with other search development methods.	WordFreq (https://tera-tools.com/word-freq); PubReMiner (https://hgserver2.am	

From RAISE 3, Table 1

Cochrane

A word about validation vs. verification in your review

Requires validation within the review	How did *authors* validate whether AI tool would perform well for their specific review, e.g., SWAR	For example, any AI use in screening, data extraction, quality assessment
Human verification required	How did *authors* check and verify the AI outputs were correct	For example, genAI use in question formulation, text drafting, search query translation

Expectation 3:
Full and transparent reporting of AI use

Disclosure of AI use

Name and purpose of AI tool

We will use [*AI system/tool/approach name, version, date of use*] developed by [*organization/developer*] for [*specific purpose(s)] in [the evidence synthesis process*]. The [*AI system/tool/approach*] will [*state it will be used according to the user guide, and include reference, and/or briefly describe any customization, training, or parameters to be applied*].

Disclosure of AI use

Name and purpose of AI tool

We will use [*AI system/tool/approach name, version, date of use*] developed by [*organization/developer*] for [*specific purpose(s)] in [the evidence synthesis process*]. The [*AI system/tool/approach*] will [*state it will be used according to the user guide, and include reference, and/or briefly describe any customization, training, or parameters to be applied*].

Degree of human oversight

Outputs from the [*AI system/tool/approach*] are justified for use in our synthesis because:

- [*state the degree of human oversight such as any steps taken to review, verify, or override AI-generated outputs*.]

Disclosure of AI use

Name and purpose of AI tool

Degree of human oversight

Justify use of AI system or tool (conduct of review)

We will use [*AI system/tool/approach name, version, date of use*] developed by [*organization/developer*] for [*specific purpose(s)] in [the evidence synthesis process*]. The [*AI system/tool/approach*] will [*state it will be used according to the user guide, and include reference, and/or briefly describe any customization, training, or parameters to be applied*].

Outputs from the [*AI system/tool/approach*] are justified for use in our synthesis because:

- [*state the degree of human oversight such as any steps taken to review, verify, or override AI-generated outputs*.]

- [*describe how you have determined it is methodologically sound and will not undermine the trustworthiness or reliability of the synthesis or its conclusions (e.g., model validation, feature validation)*]
- [*describe how it has been validated or calibrated to ensure that it is appropriate for use in the context of the specific evidence synthesis, to include degree of author involvement, if not covered in the user guide, evaluations or elsewhere (e.g., real-world effectiveness)*].

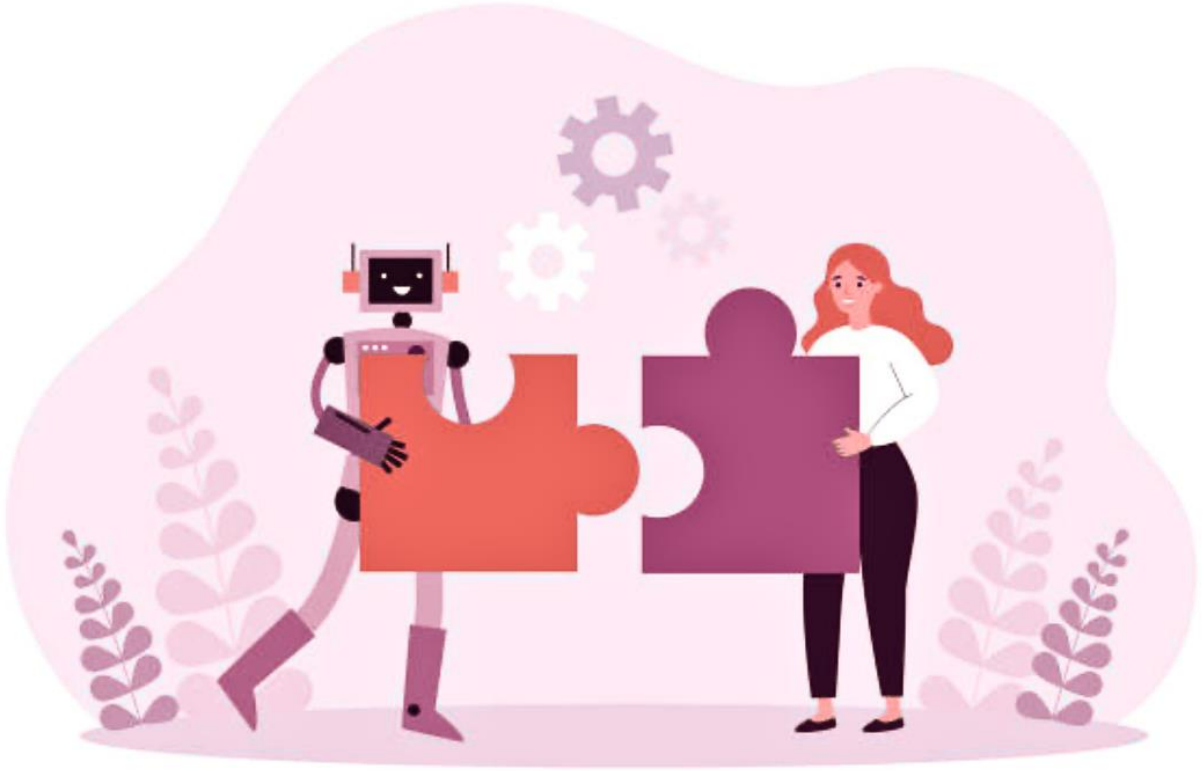
Limitations [*of the AI system/tool/approach*] include [*describe known limitations, potential biases, and ethical concerns*]/ [*are included as a supplementary material*]. [*If applicable*] A detailed description of the methodology, including parameters and validation procedures, is available in [*supplementary materials*].

Other considerations for reporting AI use

- Methods:
  - PRISMA items include a note about reporting any automation tools
- Discussion > Limitations of the review process:
  - detail any limitations or potential biases, consider the potential impact of errors, limitations or generalizability
- Declaration of interest
  - declare any financial and non-financial interests related to the tool, including any relevant interests in the organization(s) that own or fund them

Expectation 4:
AI should be used with human oversight

Human oversight and accountability



AI should be a companion, not a replacement

AI use expectations for evidence synthesists

1	• Evidence synthesists are ultimately responsible for their research, including the decision to use AI and how it is used	Apply the responsible handover framework
2	• Evidence synthesists can use AI provided they can demonstrate it will not compromise the methodological rigor or integrity of their synthesis	Via evidence, verification methods, or real-world validation methods
3	• AI use should be fully and transparently reported	Adhere to reporting expectations
4	• AI should be used with human oversight	Via all of the above

Recommendations and guidance

Three-paper RAISE collection

1. **Responsible AI in Evidence synthesis 1:**
   Recommendations for practice
2. **Responsible AI in Evidence synthesis 2:**
   Building and evaluating evidence synthesis tools
3. **Responsible AI in Evidence synthesis 3:**
   Selecting and using evidence synthesis tools

Thank you!

Questions?