Introducing Responsible AI use in evidence SynthEsis (RAISE)

James Thomas, Professor of Social Research & Policy, EPPI-Centre, UCL
Kaitlyn Hair, Research Fellow, EPPI-Centre, UCL

Overview

1. Introduce the three papers in the RAISE collection and what they cover
2. We will discuss:
    1. The work of the RAISE initiative, how the group came to their conclusions and recommendations, and how the guidance will continue to shape the field of AI in SRs
    2. How RAISE builds upon the work of other groups (e.g., ISCAR)
    3. What problem are we trying to solve with AI in systematic reviews? Where are the biggest bottlenecks today that AI can (or can be expected to) address (search, screening, extraction, synthesis)
    4. What constitutes “acceptable validation” of an AI tool

Introducing **RAISE**

Responsible AI use in evidence SynthEsis

Developing RAISE

- The team initially led by individuals from the International Collaboration for Automation in Systematic Reviews (ICASR), Cochrane, and Campbell (JT, EF, ANS, WM). Subsequently joined by > 30 diverse authors from different evidence synthesis organisations globally
- Step 1: online survey
  - Raise awareness
  - 164 responses
- Step 2: first version posted on OSF in September 2024 (1 paper)
  - Online responses
  - Webinars
- Step 3: second version posted on OSF summer 2025 (3 papers)
  - Updated twice since
  - Now submitted to Research Synthesis Methods

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

Introducing **RAISE 1**

Responsible AI use in evidence SynthEsis: recommendations in practice

We need to support the wider adoption of AI

We need cross-field standards *and an evidence base*

We are part of an ecosystem made up of individuals, collaborations, and organisations

Each has a role to play in developing and using AI in a responsible way

One person / organisation may play multiple roles

Recommendations for evidence synthesists

1. Remain ultimately responsible for the evidence synthesis
2. Report AI use in your evidence synthesis manuscript transparently
3. Ensure ethical, legal and regulatory standards are adhered to when using AI
4. Contribute to the ecosystem to help all roles continue to develop and grow

Recommendations for methodologists

1. Adhere to open science practices when researching and evaluating AI systems
2. Commit to objective and impartial evaluations and validation of AI systems
3. Develop best practice standards – and link with developers
4. Contribute to the ecosystem to help all roles continue to develop and grow

Recommendations for organisations producing evidence synthesis

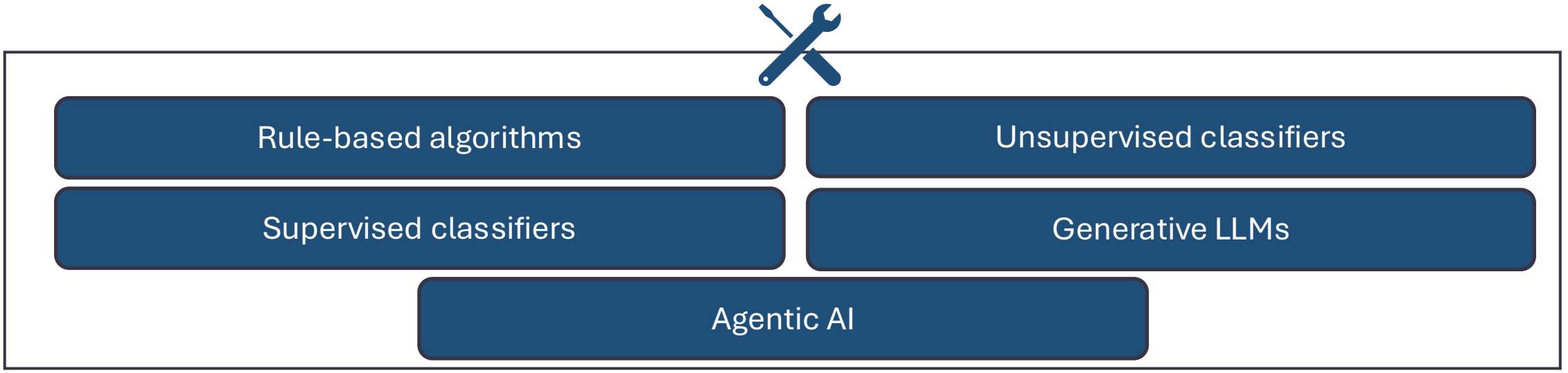
1. Ensure best practice standards for responsible AI use are clear and integrated in your policies and guidelines
2. Promote, guide and support responsible AI use in your evidence synthesis activities
3. Monitor the development and use of AI within your organization
4. Contribute to the ecosystem to help all roles continue to develop and grow

Recommendations for AI tool development teams

1. Adhere to open science practices when designing, building, testing and validating tools
2. Be transparent about when the AI works best, its limitations and any interests
3. Commit to continued learning, development and monitoring of the AI system
4. Contribute to the ecosystem to help all roles continue to develop and grow

Introducing **RAISE 2**

Building and evaluating AI evidence synthesis tools



Build phase:

- Context-specific training data
- Size and composition of training datasets
- Awareness of biases within training data

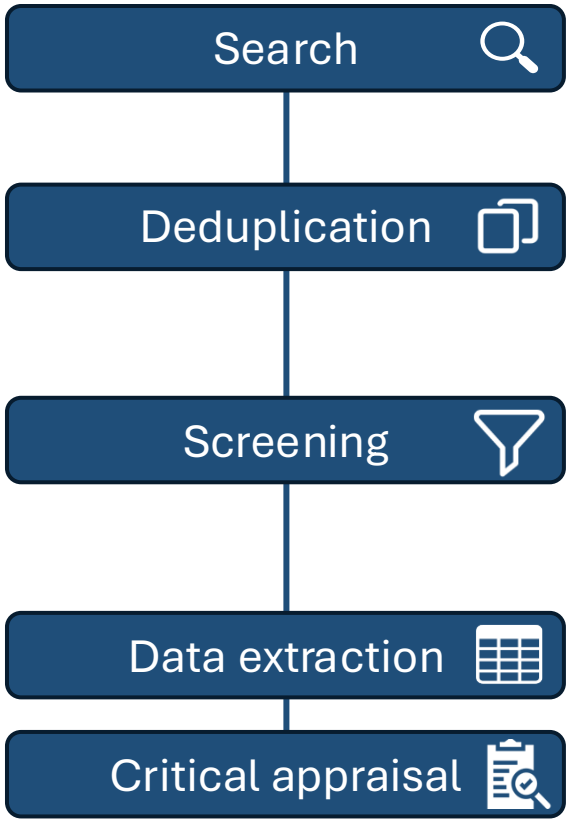
Validation phase:

- Context-specific validation data
- Appropriate metrics

LLM-specific considerations

- Smaller "training" datasets
- Separation of datasets for prompt development versus evaluation
- Response stability
- Hallucinations
- Explainability and interpretability
- Limitations and generalisability

Conducting evaluations to build a cumulative evidence base



Does an AI tool capture all records relevant to a review?
- Comparisons with "gold standard" records to assess sensitivity/specificity

Does an AI tool successfully identify and remove duplicate records?
- Datasets used for evaluation should be separate from development sets
- Duplicate groups (vs pairs) complicate evaluation tasks

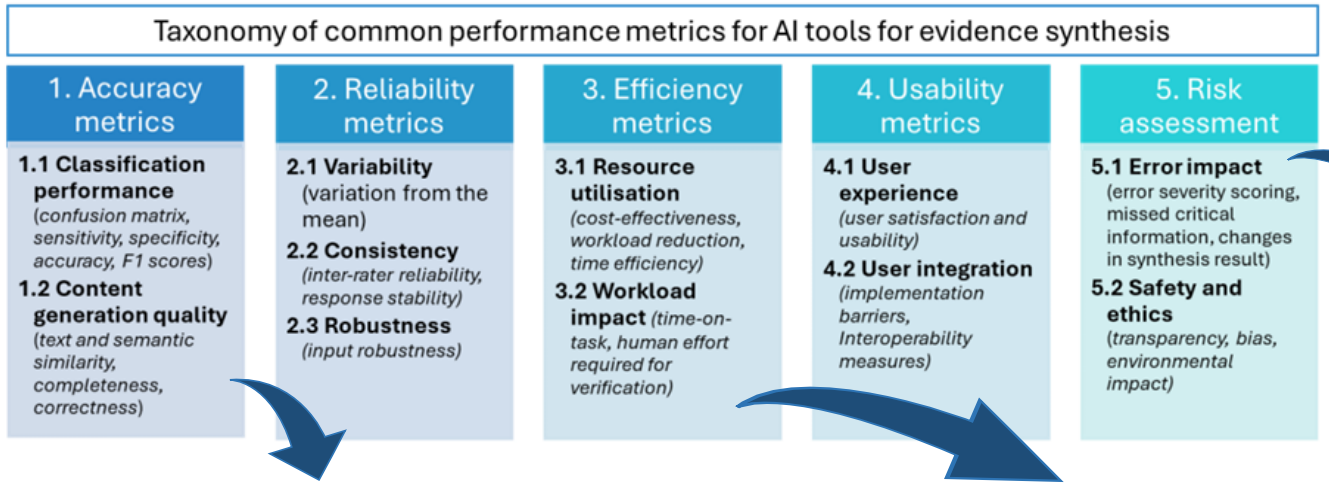
Does an AI tool successfully classify relevant/irrelevant records?
- Evaluation requires "gold standard" dataset covering full scope of use case
- Considerations of expected sensitivity and the potential impact of missed relevant studies

Does an AI tool successfully extract data or assess risk of bias?
- Evaluation requires "gold standard" datasets which are time-consuming to obtain
- Challenging to evaluate using automated methods
- Often requires manual review to determine error type and severity
- Variation in performance likely across domains/extraction targets
- Great interest in LLMs but concerns around data contamination and hallucinations

Multi-stage

Need to evaluate each tool independently

Use of metrics for evaluation



5.1 Error impact

Error severity scoring	
What does it measure	The potential downstream impact of an error made by an AI system (if left uncorrected).
Considerations for use	• Highly dependent on task context • Less well-studied in review stages beyond data extraction • According to Gartlehner et al [84], mistakes during data extraction can be characterised into major errors (could lead to erroneous conclusions), minor errors (influence on data quality, but less severe), and inconsequential errors (no meaningful impact on interpretation).
Example(s) of its use	• Gartlehner et al. evaluates LLMs for data extraction and classifies errors into "major" and "minor" errors, hallucinations, and missing data [84]

1.2 Content generation quality

BLEU (BiLingual Evaluation Understudy) score [66]	
Related terms / metrics	SacreBLEU, GLEU (Google BLEU)
What does it measure	Evaluates text generation quality by comparing overlap between generated and reference texts. Originally designed for machine translation evaluation.
Considerations for use	• Focuses on exact word matches and may miss semantically equivalent but differently worded content • Schmidt et al (11) evaluated *BLEU* and *ROUGE* scores for PICO extraction tasks in systematic reviews and found they were poor indicators of accuracy. This limitation applies broadly to word-overlap based metrics when semantic accuracy is more important than exact wording. • Better suited for tasks where exact wording matters more than semantic meaning
Example(s) of its use	• Comparison of human-extracted descriptions of PICOs used in randomised trials with those extracted by an AI tool by Schmidt et al (11)

3.2 Workload impact

Time-on-task	
Related terms / metrics	Task completion time
What does it measure	The length of time it takes a human to perform a given task.
Considerations for use	It is difficult to collect time-on-task data automatically (e.g., from web use logs) because it is not possible to tell whether an evidence synthesist is actively working on a given record, or has actually switched away to do something else.
Example(s) of its use	• A SWAR protocol for evaluating time-on-task to screen a sample of 100 records using two different approaches by Devane et al. [79]

Reporting the building and/or evaluation of an AI tool

Introduction:
- Existing tools/knowledge
- AI tool details
- Objectives

Methods:
- Study design
- Setting
- Data sources
- Data selection
- Data preprocessing
- Labelling of input/validation data
- Type of AI tool/model
- Tool development
- Performance accuracy
- Validation methods
- Performance errors
- Interpretation

Discussion:
- Strengths/limitations
- Bias
- Practical value of the tool
- Implications for use

Other:
- Ethical statement
- Availability of evaluation protocol
- Sources of support
- Declarations of interest
- Availability of data, code, materials
- Replicability
- Environmental impacts

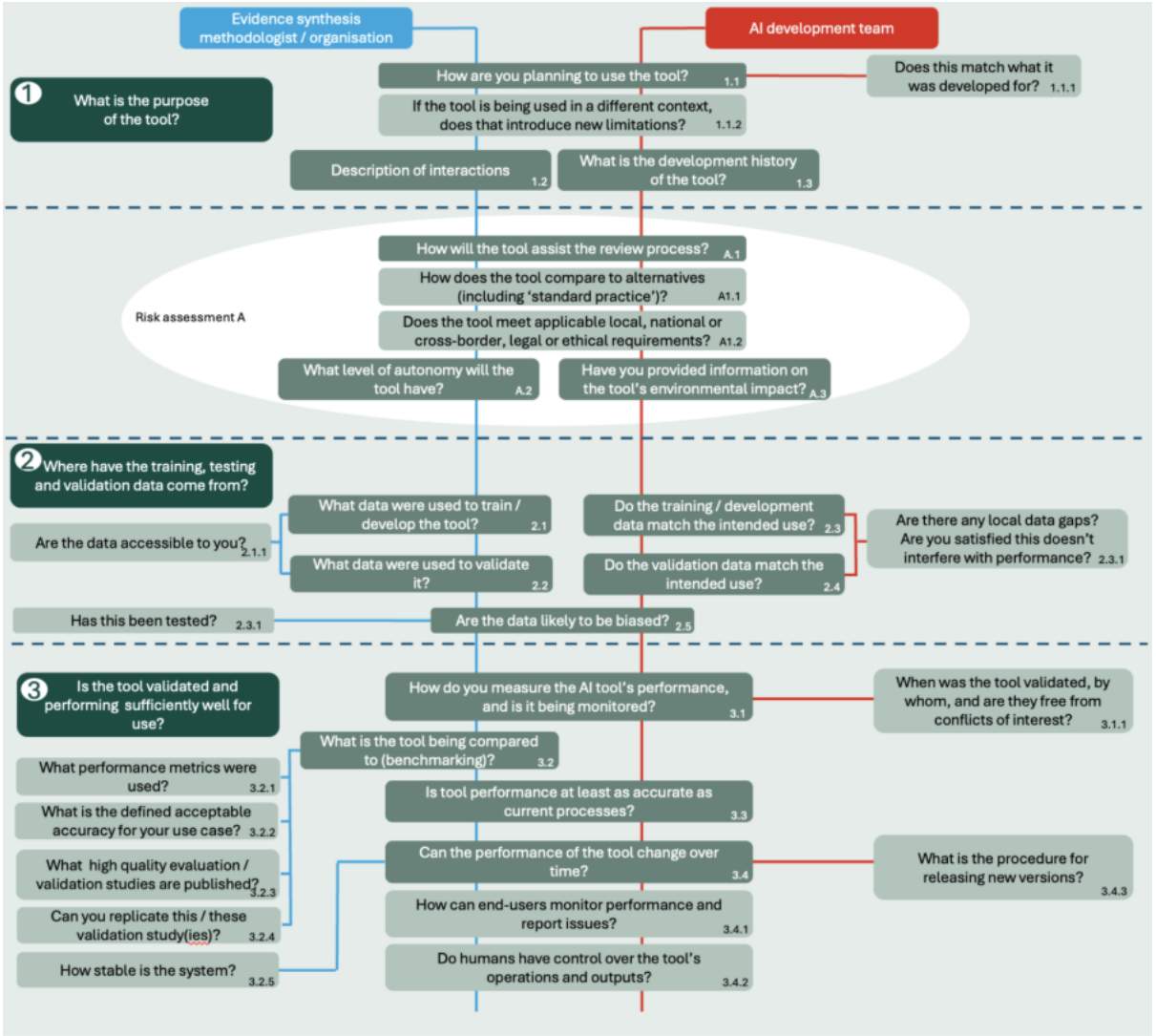
Introducing **RAISE 3**

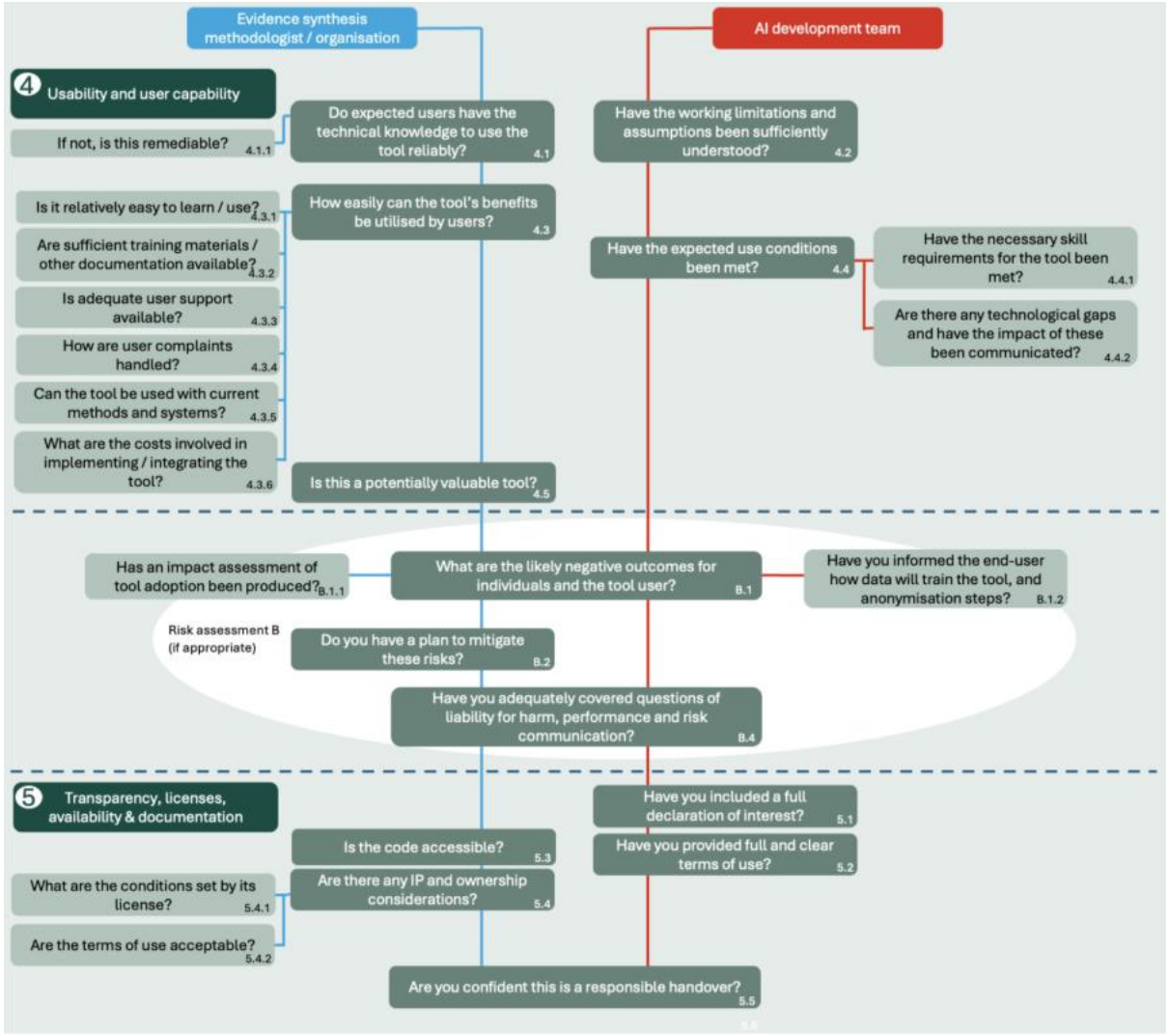
Selecting and using evidence synthesis tools

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 perform sufficiently for use?**
- **Usability and user capability**
- **Transparency, licenses, availability and documentation**

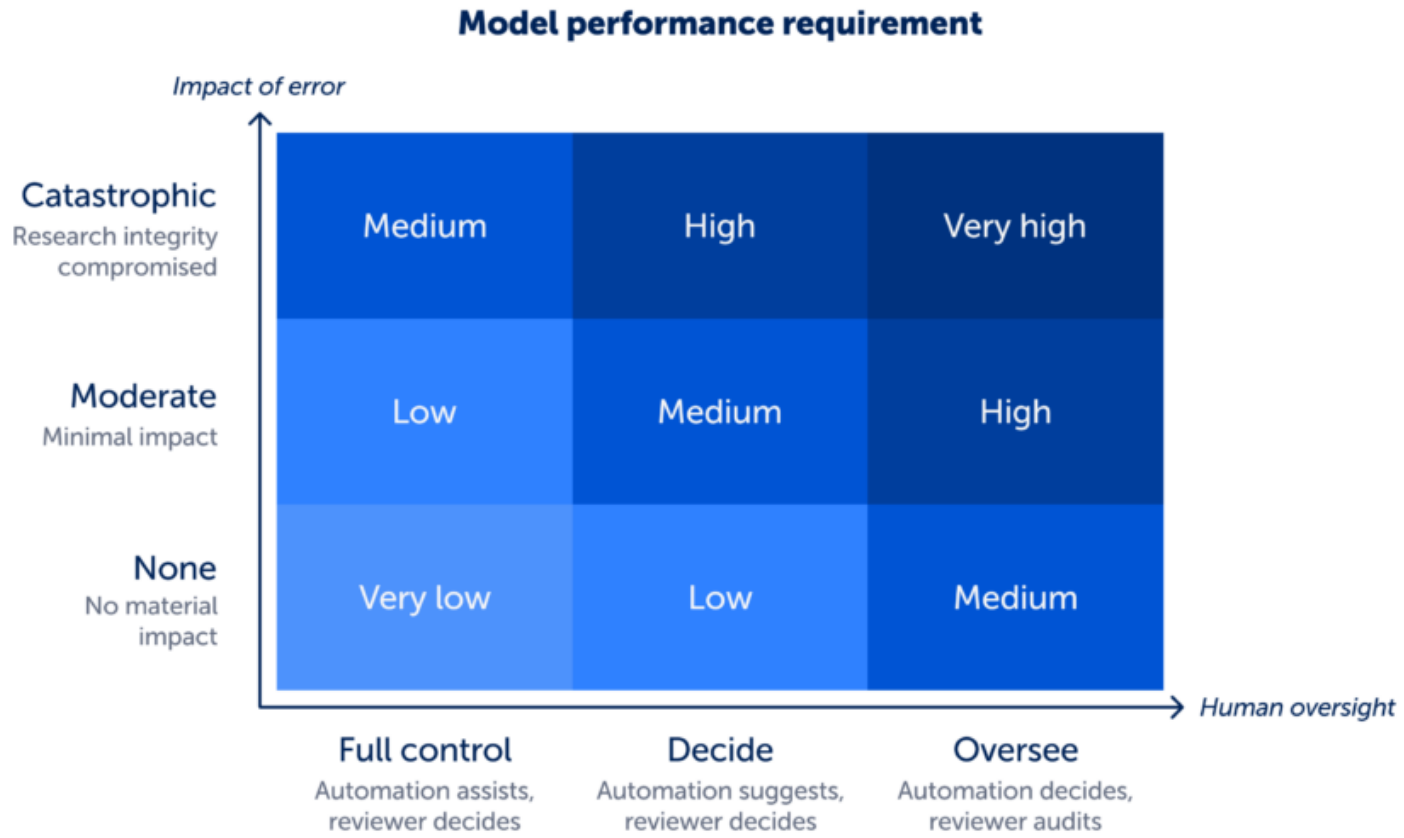
From RAISE 3


Evidence synthesis methodologist / organisation
AI development team
1 What is the purpose of the tool?
How are you planning to use the tool? 1.1
Does this match what it was developed for? 1.1.1
If the tool is being used in a different context, does that introduce new limitations? 1.1.2
Description of interactions 1.2
What is the development history of the tool? 1.3
Risk assessment A
How will the tool assist the review process? A.1
How does the tool compare to alternatives (including 'standard practice')? A1.1
Does the tool meet applicable local, national or cross-border, legal or ethical requirements? A1.2
What level of autonomy will the tool have? A.2
Have you provided information on the tool's environmental impact? A.3
2 Where have the training, testing and validation data come from?
Are the data accessible to you? 2.1.1
What data were used to train / develop the tool? 2.1
What data were used to validate it? 2.2
Do the training / development data match the intended use? 2.3
Do the validation data match the intended use? 2.4
Are there any local data gaps? Are you satisfied this doesn't interfere with performance? 2.3.1
Has this been tested? 2.3.1
Are the data likely to be biased? 2.5
3 Is the tool validated and performing sufficiently well for use?
How do you measure the AI tool's performance, and is it being monitored? 3.1
When was the tool validated, by whom, and are they free from conflicts of interest? 3.1.1
What is the tool being compared to (benchmarking)? 3.2
What performance metrics were used? 3.2.1
What is the defined acceptable accuracy for your use case? 3.2.2
What high quality evaluation / validation studies are published? 3.2.3
Can you replicate this / these validation study(ies)? 3.2.4
How stable is the system? 3.2.5
Is tool performance at least as accurate as current processes? 3.3
Can the performance of the tool change over time? 3.4
What is the procedure for releasing new versions? 3.4.3
How can end-users monitor performance and report issues? 3.4.1
Do humans have control over the tool's operations and outputs? 3.4.2


Evidence synthesis methodologist / organisation
AI development team
4 Usability and user capability
If not, is this remediable? 4.1.1
Do expected users have the technical knowledge to use the tool reliably? 4.1
Have the working limitations and assumptions been sufficiently understood? 4.2
Is it relatively easy to learn / use? 4.3.1
Are sufficient training materials / other documentation available? 4.3.2
Is adequate user support available? 4.3.3
How are user complaints handled? 4.3.4
Can the tool be used with current methods and systems? 4.3.5
What are the costs involved in implementing / integrating the tool? 4.3.6
How easily can the tool's benefits be utilised by users? 4.3
Is this a potentially valuable tool? 4.5
Have the expected use conditions been met? 4.4
Have the necessary skill requirements for the tool been met? 4.4.1
Are there any technological gaps and have the impact of these been communicated? 4.4.2
Has an impact assessment of tool adoption been produced? B.1.1
What are the likely negative outcomes for individuals and the tool user? B.1
Have you informed the end-user how data will train the tool, and anonymisation steps? B.1.2
Risk assessment B (if appropriate)
Do you have a plan to mitigate these risks? B.2
Have you adequately covered questions of liability for harm, performance and risk communication? B.4
5 Transparency, licenses, availability & documentation
Have you included a full declaration of interest? 5.1
Have you provided full and clear terms of use? 5.2
Is the code accessible? 5.3
Are there any IP and ownership considerations? 5.4
What are the conditions set by its license? 5.4.1
Are the terms of use acceptable? 5.4.2
Are you confident this is a responsible handover? 5.5

Considerations before using the framework

- Look critically; **all tools have limitations**
- Consider the acceptable model performance given the potential impact of errors and degree of human oversight
- Collaborate with those who have complementary backgrounds



Covidence, Beyond evaluation: deciding when AI is appropriate in evidence synthesis (April 2026)

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

Current state of AI tools

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	

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.

From RAISE 3, Table 1

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 protocol				
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.carrot2.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.undermind.ai/), Elicit (https://elicit.com/), Asta Find Papers (https://asta.allen.ai	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.yale.edu/), TERA WordFreq (https://tera-tools.com/word-freq), PubReMiner (https://hgserver2.a	Acceptable for use

Excerpt from RAISE 3, Table 2

Summarisation is not synthesis

Data synthesis				
Quantitative analysis	Rule-based	Statistical meta-analysis can be performed using dedicated software packages or web applications. Statistical parameters should be clearly documented and flexible for users to amend to suit their analytical choices.	R packages (e.g. Meta, Metafor), CRSRU Meta-analysis apps	Acceptable for use
	Generative LLMs	The use of LLMs to *synthesise* (as opposed to summarise) results across studies.	Google AI Mode, Elict, Perplexity, Asta	Not acceptable for use
Code drafting and troubleshooting	Generative LLMs	Generative models can provide solutions to coding problems and/or draft entire scripts based on instructions. Substantive knowledge of the coding language is recommended to ensure that the code performs as intended. All	ChatGPT, CoPilot, Claude, Gemini, DeepSeek	Human verification required

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

The question of acceptable thresholds

- What's good enough?
- Lack of community consensus on the **appropriate level of confidence** and **appropriate level of performance**

Table 2. Stopping Boundaries and Decision Rules for Interim Analyses

Performance Metrics	Futility Boundaries (Point Estimate)*	Non-inferiority Margins (Upper Limit of 95% CI)*	Decision Rules
Screening			
Sensitivity	<80%**	<95%**	Stop if either boundary is crossed
Specificity (for full-text screening only)	<50%***	<60%***	Stop if either boundary is crossed
Data Extraction			
Sensitivity	<92%**	<97%**	Stop if either boundary is crossed
Major Error Proportion	>3%†	>2%†	Stop if either boundary is crossed
Usability			
System Usability Scale (score)	<57‡	<75‡	Stop if threshold is not met

Cochrane Evaluation of (Semi-) Automated Review (CESAR) Methods: Protocol for an adaptive platform study within reviews (April 2026) Gartlehner et al. MedRxiv.

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*.]

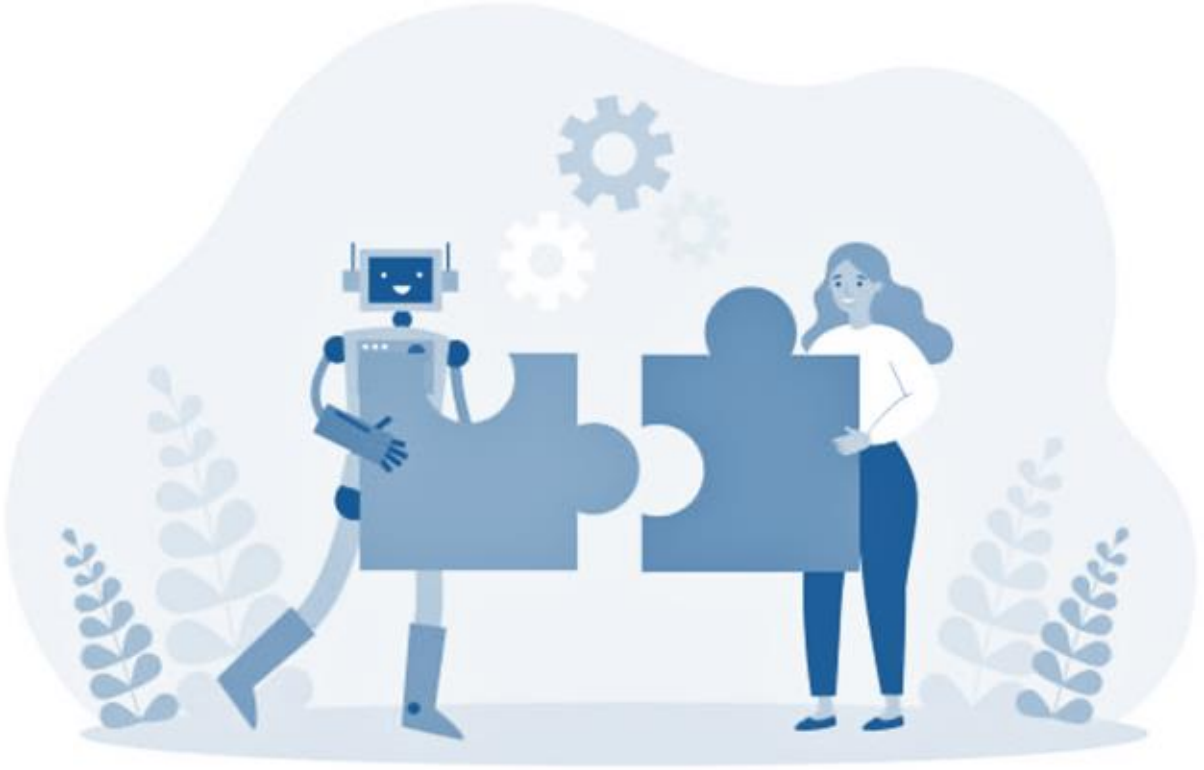
- [*describe how you have determined it is methodologically sound and will not undermine the trustworthiness or reliability of the synthesis or its conclusions (e.g., model validation, feature validation)*]
- [*describe how it has been validated or calibrated to ensure that it is appropriate for use 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

Human oversight and accountability



AI should be a companion, not a replacement

How we think the committee should address AI use in the updated standards

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

Thank you!

Questions?

RAISE https://osf.io/fwaud/