Computational Review of Technology-Assisted Medical Evidence Synthesis through Human-LLM Collaboration: A Case Study of Cochrane

Yiping Ding[1], Xiaorui Jiang[1*], Opeoluwa Akinseloyin[2]

[1] School of Information, Journalism and Communication, University of Sheffield, 2 Whitham Road, Sheffield, S2 10AH, United Kingdom

[2] Centre for Computational Sciences and Mathematical Modelling, Coventry University, Puma Way, Coventry, CV1 2TT, United Kingdom

[*] Corresponding to Xiaorui Jiang (xiaorui.jiang@sheffield.ac.uk)

Abstract Medical evidence synthesis, typically done by systematic reviews, requires extensive manual effort across stages such as searching, screening, extraction, and synthesis, making them slow and costly. These limitations hinder timely updates and rapid responses during health crises. Interests in technology-assisted evidence synthesis have been increasing, driven by artificial intelligence (AI) and large language models (LLMs). In 2024, four major networks including the Cochrane, Campbell Collaboration, Joanna Briggs Institute and Collaboration for Environmental Evidence jointly launched an AI Methods Group to advance automation in evidence synthesis. This chapter presents a large-scale computational analysis of technology-assisted MES across 7,271 Cochrane reviews (2010–2024), identifying computer tools—software, packages, or algorithmic implementations—used at different review stages via an LLM-human collaborative annotation pipeline. A multi-LLM mechanism combining suggestion, verification, and self-critical questioning achieved high-recall tool extraction. Evaluation against five "gold-standard" tool lists showed major gains: approximately 100 additional tools were identified compared to each existing review-based, database-based, and Cochrane-curated gold standards. Eventually, a list of in total 514 tools was compiled. Two annotators verified all candidates within two days, demonstrating notable efficiency. A follow-up bibliometric analysis provides the first computational map of technology use in Cochrane evidence synthesis, revealing trends across time, domains, and regions.

1 Introduction

Medical research results are being published at a rapid speed. For instance, after just one year of the COVID-19 outbreak, PubMed already registered more than 100,000 related publications, nearly the total size of influenza study in the past 200 years. Medical evidence synthesis, or evidence synthesis in short, is the process of systematically searching, filtering, extracting, assessing, and combining findings from all existing studies about a specific medical or healthcare research question to create a comprehensive overview of the current knowledge on it. Systematic review (SR), also called systematic literature review (SR), is a de facto standard approach for performing a reliable and trustful medical evidence synthesis. In most cases, a systematic review follows a principled pipeline of strictly defined steps including literature search, literature screening—identifying the primary studies that are relevant to the re-search topic of a review (the focus of this project), study quality assessment, data extraction from eligible studies, data synthesis and meta-analysis, insight generation and report writing [1]. Maintaining an up-to-date evidence base effectively and efficiently is critical for addressing future public health emergencies on a scale similar to COVID-19. For example, the COVID-NMA project maintains a living synthesis of evidence on the effectiveness of interventions for COVID-19 prevention and treatment [2].

Doing rigorous SRs is very labour-intensive, slow and expensive. One estimate was that it took approximately 67.3 weeks to finish an SR. "Each SLR costs approximately" $141k, and for the ten largest pharmaceu-tical companies and ten major academic institutions, "on average, the total cost of all SLRs per year to each academic institution amounts to" about $25m and for each pharmaceutical company is" approximately $5.9m [3]. Technology has since played an increasingly important role in making evidence synthesis more effective and efficient. On the one hand,

computer tools such as EPPI-Reviewer[1], Rayyan[2], Colandr[3] and Covidence[4] and CADIMA[5] have been widely used for efficient collaboration and management of an evidence synthesis project. On the other hand, AI techniques such as machine learning (ML) [4], natural language processing (NLP) [5] and more recently Large Language Models (LLM) [6] have high potentials for assisting human reviewers to automate or semi-automate the SR pipeline or certain core SR steps to make SRs more affordable, not only faster and cheaper but also more robust and accurate. AI is often an integrative component of a computer tool such as those aforementioned, but standalone tools also exist, such as Abstractr[6], SWIFT-Review[7], RobotAnalyst[8] and ASReview[9]. These developments are called technology-assisted review (TAR) [7] or automated systematic review (ASR) [8]. In this chapter, technology-assisted medical evidence synthesis covers computer packages, software or tools for both review management and review automation.

Since Cohen's ground-breaking work on machine learning for automated abstract screening in 2006 [9], two decades of research and development have resulted in a lot of computer tools for technology-assistance evidence synthesis. It was widely agreed that ASR tools "will likely have a increasingly important role in high-quality and timely reviews", but "one of the key barriers to the use and adoption of automation tools remains the lack of knowledge about their existence." [10] Two noteworthy efforts were seen by the community to promote the awareness and dissemination of tools for evidence synthesis. Established in 2015, the International Collaboration for the Automation of Systematic Reviews[10] has formed a global network of researchers working to make systematic reviews faster and more efficient through technology, holding annual meetings with the aim to establish shared principles, protocols, and resources to facilitate this work. More recently, in Feb 2025 a new, joint AI Methods Group[11] was formed between four world-leading evidence synthesis organisations, including the Cochrane Collaboration[12], the Campell Collaboration[13], the Joanna Briggs Institute[14] and the Collaboration for Environmental Evidence (CEE)[15], with the aim to embrace disruptive role of Large Language Models (LLMs), AI and automation tools and spreading the adoption of technology and automation in evidence synthesis. This AI Methods Group was established based on the success of Cochrane's "Artificial Intelligence (AI) Methods in Evidence Synthesis" training series[16] that has run since early 2024.

To facilitate this trend and improve the evidence synthesis workflow, it is of critical value to produce a comprehensive mapping of the landscape of Technology-Assisted Medical Evidence Synthesis (TAMES), which is however lacking and challenging. There are a few recent surveys about software tools and/or AI/ML tools in conducting systematic reviews or evidence synthesis. Ironically, while these surveys focus on AI-based automation and tools in

[1] https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914
[2] https://www.rayyan.ai/
[3] https://www.colandrapp.com/
[4] https://www.covidence.org/
[5] https://www.cadima.info/
[6] https://effectivehealthcare.ahrq.gov/products/abstractr/
[7] https://www.sciome.com/swift-review/
[8] https://www.nactem.ac.uk/robotanalyst/
[9] https://asreview.nl/
[10] https://icasr.github.io/
[11] https://methods.cochrane.org/ai/
[12] https://www.cochrane.org/
[13] https://www.campbellcollaboration.org/
[14] https://jbi.global/
[15] https://environmentalevidence.org/
[16] https://www.cochrane.org/learn/webinars/collection/2929

evidence synthesis, none of them was done with the assistance of automation tools. Instead, they were all done in the same way as performing a traditional SR [11-20], i.e. by manually searching databases of published SRs using a set of AI/ML keywords, screening out the SRs that do not use any tool, and extracting and reporting the information about the AI/ML tools. At the time of writing the current chapter, the most up-to-date review was (online) published in the "AI for Evidence Synthesis" themed collection of Cochrane Evidence Synthesis and Methods on 28 Aug 2025 [21]. While the authors made a cross-organisational investigation of Cochrane, Campbell and CEE, they also followed the search-screening-extract pipeline similar to performing an SR. Due to the approach's high time cost, this review was constrained to evidence syntheses published between 2017 and 2024, resulting in only 2271 studies for data extraction and analysis and an incomplete picture of the historical development of TAMES. Regarding the sources of searching, the mapping review, there was an exception where the authors also searched several software repositories to identify ML tools. This however did not change the manual nature of their method.

Different from these prior efforts, the current chapter is the first study of using a human-LLM collaborative approach to conduct a computational bibliometric review of the evolving use of technology (computer tools and/or AI tools) in conducting, automating or semi-automating medical evidence synthesis, demonstrating how screening and data extraction of more than 7100 studies could be done within two days through the assistance of LLMs. Instead of constraining the candidate studies by pre-defined keywords or significantly limiting the time range for scale reduction, this chapter processes all Cochrane reviews within a wider time range between 2010 and 2024 and extracts computer tools and tool usage with the assistance of LLMs' capabilities of information extraction. LLMs' successes have been recorded for scientific, medical and clinical information extraction were recorded [22-23], demonstrating their comparable or even stronger performances than traditional deep learning models in various subject areas [24-27] and their particular usefulness in resource-constrained settings at a little sacrifice of accuracy [28]. Some prior research showed that LLMs may have a tendency of over-extracting information such as potentially relevant medical evidence that supports clinical coding [29], resulting in a trade-off of precision for recall. To handle this situation a human-LLM collaborative approach is adopted by signposting LLM-based extractions to human annotators for verification and correction [22]. To further reduce the workload of human annotators (i.e. increase the precision of LLMs) and increase the coverage of reported tool usage (i.e. increase the recall of LLMs), a multi-LLM collaborative pipeline was designed [30]. The pipeline consists of three stages: too and tool usage suggestion (to generate initial annotations), debating [31] and voting by peer LLMs [32] (for achieving better precision), and questioning and self-criticising the voting results [33] (aiming at improved recall).

2 Materials and Methods

2.1 Dataset and Preprocessing

The Cochrane Library was used because of its open access nature. Evidence syntheses published between 2010 and 2024 were all downloaded for processing and analysis. The time range was selected based on two facts. Frist, one of the most influential ground-breaking research in applying machine learning to automate the abstract screening step in SR was conducted by Cohen et al in 2006. It embarked community interests in this research direction in the following years, but some famous tools like EPPI-Reviewer 4 [34-35] and Abstrakr [36] were developed or went online around 2010. We conjecture that this time range would have a comprehensive coverage of computer or automation tools used in TAMES. In total 7,550 Cochrane reviews were collected. No bespoke eligibility criteria were defined as the current study will use LLMs to analyse the Methods section of each Cochrane review and "automate" the screening process.

Cochrane reviews are written in a quite consistent template, most with a Methods section summarising the protocol of performing an evidence synthesis. A sample of 750 reviews were randomly picked to check where tools are

mentioned. This was done by checking the existence of a list of known tools manually curated from 10 existing systematic or scoping reviews of (AI) tool usage in evidence synthesis [11-20] (See Sect. 3.1 for details). Results showed that the computer tools used for assisting an evidence synthesis are almost certainly described in its Methods section and they rarely appear in other sections of a Cochrane review. Therefore, in the study of this chapter, only reviews with a Methods section are included and only Methods sections are preprocessed for analysis. It was found that there were 279 reviews for which the Methods section could not be extracted so they were excluded. This happened mainly because these reviews were withdrawn after their first publication. So, in total 7,271 Cochrane reviews remained in the current study. The subsections of each review's Methods section were extracted, and the sentences of each subsection (or section) were segmented and tokenised using Standford University's Stanza NLP toolkit[17]. Preprocessed Methods sections were saved in a JSON format. Computer tools and their usage would be extracted on a sentence basis. After preprocessing, there are 827,360 sentences.

2.2 Multi-LLM Collaboration

Our approach falls in the multi-LLM collaboration paradigm, which is based on the hypothesis that the collaboration between several LLMs could possibly resolve the biases and limitations of a single LLM. Recall that the No. 1 purpose of the current study is to make the most comprehensive coverage of computer tools supporting evidence synthesis that existing methods based on manual searches cannot achieve. Multi-LLM collaboration has a better potential to fulfil this purpose by instructing LLMs to critique others' initial results and make new proposals that might have been overlooked. The proposed approach contained three stages that constitute a pipeline. In summary, tools were first suggested by a strong reasoning LLM (*suggestion* stage). Then, the initial suggestions were verified by several lightweight LLMs, whose results were combined (*verification* stage). After that, the verifier LLMs further *questioned* o3's decisions based on human feedback on the real tools. New tool proposals were again combined (*questioning* stage) for the purpose of discovering the tools that were missed in the prior two stages. The proposed multi-LLM collaboration pipeline employed three techniques to gradually improve the performance of tool extraction. First, *debating* was applied in the verification stage by asking lightweight LLMs to debate on the decisions made by a strong reasoning LLM [31]. Second, *self-reflection* (or *self-criticism*) was applied in the questioning stage to instruct LLMs to reflect on, criticise and correct their own decisions [33]. Third, *ensembling* was applied in both the verification and questioning stages to combine the decisions of multiple LLMs [32]. The following subsections present the details of each stage of our pipeline.

2.2.1. Initial Tool Extraction – The Suggestion Stage

Many prior studies have reported that LLMs performed very well in information extraction tasks of specialised domains such as medical research [X] and clinical documentation [26-28]. Because the current study aims to extract mentions of computer tools used in medical evidence synthesis. LLMs are expected to be helpful for assisting this task. Performances of LLMs seem to correlate with their sizes and reasoning capabilities: bigger LLMs typically performed much better and closer to human performance (e.g., GPT-4o vs. GPT-4o mini) [27], and reasoning LLMs may perform better than general-purpose LLMs [37] (possibly because of their stronger context understanding and instruction following capabilities). Based on the above insights, the current study prompted a strong reasoning model OpenAI's o3 to perform the first round of extracting (i.e. suggesting) computer tool mentions from each sentence.

Due to the large data size, o3 was instructed in a zero-shot fashion with a properly engineered prompt. In-context learning (ICL) with several demonstrations (a.k.a. few-shot learning) [38-39], especially demonstrations similar to a test sample that are dynamically and properly selected [40-41], would very likely improve extraction precision. However, ICL significantly increases the input size and costs. For example, a 5-shot setting roughly means 5 times higher expense. As a cost-effective alternative, the current study proposed multi-LLM collaboration in verification

[17] https://stanfordnlp.github.io/stanza/

and questioning stage to improve both precision and recall. Figure 1 shows the prompt for tool suggestion. Particularly, "tools" to be extracted by an LLM are defined as any "method, software, workflow, algorithm, package, …" for the purpose of maximising the recall by stimulating the LLM to extract as many potential tools as possible. In total, 37,733 sentences were judged as containing tools by GPT-o3 and 794 potential tools were reported. Note that there are duplicates in this initial list of candidate tools because the same tool might be mentioned by different authors in slightly different forms or different versions of a tool were used by different authors, which should be seen as the same tool. So it was necessary to perform tool name normalisation, which was part of the job of human annotation (see Sect. 2.3 for details)

Prompt 1: Tool Suggestion

You will receive a sentence to analyze. Your task is to:
Check if the sentence mentions **any** tool (e.g. method, software, workflow, algorithm, package ….)
If it does: identify the tool mentioned.

Input
paper_id: {entry["paper_id"]}
position: {entry["position"]}
sentence: \"{entry["sentence"]}\"

Output format
tool: \"{entry["tool"]}\"

If no tool is mentioned in the sentence, do not output anything. Move to the next.
Do not output anything other than the result in the specified format.

Fig. 1 Prompt for Extracting Computer Tool Mentions.

2.2.2. Improving Precision of Tool Extraction – The Verification Stage

Some prior studies have reported that in certain cases LLMs may have a tendancy of over-extraction, especially when prompted with short input sources such as sentences, paragraphs or short documents and prompted in a way not to miss any potentially relevant information [29]. In such cases, LLMs may exhibit high recall compared to traditional methods at a significant sacrifice of precision [42]. On the contrary, LLMs have shown exceptional performance in excluding irrelevant information in various medical or clinical application areas, such as sentences in discharge summaries unable to provide supporting evidence for clinical coding [42] or abstracts of candidate studies that do not match the selection criteria of a systematic review [43]. To improve the precision, which would have a huge impact on the efficiency of human-LLM collaboration (see Sect. 3.3), five lightweight mainstream LLMs of mode heterogeneity from different LLM families developed by different companies [44] were employed to debate the results generated in the suggestion stage and make consensual decisions to rule out the obvious over-extracted non-tool phrases. The five debating LLMs are: OpenAI's GPT-4.1-mini ("gpt-4.1-mini"), Claude's Haiku 3 ("claude-3-haiku-20240307"), Google's Gemini 2 Flash ("gemini-2.0-flash"), DeepSeek's Chat model ("DeepSeep-V3-0324"), and Alibaba's Qwen Plus, which is based on the Qwen 2.5 foundation model ("qwen-plus-2025-04-28").

Figure 2 shows the same prompt used for all five lightweight LLMs to debate on the sentences that were judged as containing tools and the corresponding tool names that are identified in the suggestion stage. After that the opinions of the five debaters were combined. Classical methods include majority voting on all debaters or averaging the confidences of all debaters (called soft voting) [45], however this may accidentally miss some real tools causing additional efforts in later stages to recover them. As the priority of this study is to retrieve as many tools as possible from published evidence syntheses, a simple rule was adopted to combine the new decisions of the five debaters: if any one agreed with GPT-o3 that there was a tool mention in a sentence, then this tool and its context (review id, subsection and sentence) were saved in a candidate tool list and sent to future stages for further analysis and annotation. In total 545 candidate tools were recorded after the verification and voting process. Again, this list might contain duplicates, which would be normalised later by human annotators.

Prompt 2: Tool Debating

```
You are a helpful assistant.  You will receive a single JSON record with
four keys:
  • paper_id – unique paper identifier
  • position – sentence location inside the paper
  • sentence – full sentence to analyse
  • term – candidate tool name to evaluate

### Record
paper_id: {entry['paper_id']}
position: {entry['position']}
sentence: {entry['sentence']}
term: {entry['term']}

### Task
1.Determine whether **term** is a *tool*
2.If it does:
   - Identify the tool mentioned
   - Extract the **surrounding context**: 5 words before and 5 words after
the tool mention, outputting the verb phrases or predicate phrases
centred around the tool mention

### Output rules
1. **If the term *is* a tool**, output **all five** lines exactly in this order:
{{Tool}}: TOOL_NAME
{{Usage}}: Yes / No
{{Explanation}}: VERB_PHRASE
{{paper_id}}: {entry['paper_id']}
{{position}}: {entry['position']}
```

Fig. 2 Prompt for Debating on and Verifying Tool Mentions.

2.2.3. Increasing Recall of Tool Extraction – The Questioning Stage

To further increase the recall of tool extraction, each lightweight LLM was prompted to further question o3's based on the results of the first round of human intervention that happened at the end of the verification stage. Figure 3 illustrates the task description and detailed instructions to complete the task (the first part of the complete prompt) and Figure 4 shows the contents and formats of the inputs and outputs (the second of the complete prompt). Note that in the collaborative annotation pipeline proposed by the current study, there is human intervention directly before the ends of the verification and questioning stages to judge which LLM-extracted tools are true tools and which not (to be detailed in Sect. 3.3). So, the result of the verification stage was the Tool List Version 1 (a list of true tool names that are mentioned in Cochrane evidence syntheses after human intervention, see Sect 2.3 for details of LLM-human collaboration). Because the purpose of the question stage is to find as many overlooked tools as possible, LLMs were explicitly prompted to ignore any name in Tool List Version 1, as shown in Figure 3. As in the verification stage, the potential new tools suggested by each LLM in the questioning stage were ensembled by simple rule: If an LLM suggested a tool then this tool would be included for human judgement. In total, 499 tool mentions were voted NO by all five LLMs, i.e. no usage of this potential tool being reported in the corresponding review. On the contrary, 1198 additional potential tools were suggested by at least one LLM. Both were judged by humans to form the Tool List Version 2—the final list of computer tools that have reported being used in Cochrane reviews. 8273 sentences suspicious of having new tools that were overlooked by o3.

2.3 Human-LLM Collaboration

On top of multi-LLM collaboration was human-LLM collaboration, which put human annotators in the loop to gatekeep the accuracy of tool extraction. Although the candidate tools could be judged by human annotators only once at the end of the pipeline, in the current study human intervention happened twice before the ends of the tool verification and tool questioning stages respectively. This allowed for quantitatively evaluating the performances and justifying the necessities of each stage of the proposed pipeline. Also, we need to note that the current study does not aim to replace human by fully automating the data extraction process. Instead, the current study aims to promote a more responsible approach that enhances human reviewers with the capabilities of LLMs so that both the comprehensiveness and efficiency of a computational bibliometric review are significantly improved. Thus, the adopted approach emphasises always involve humans in the loop and leaving final verdicts to humans. Despite inevitable effort of human intervention, the efficiency could be tremendously improved. For example, in the presented case study on Cochrane reviews, the number of sentences that human annotators needed to verify was significantly reduced from over 827 thousand to a few hundred or thousand, because it was typically enough for human annotator to read one piece of context to unjustify a false tool name.

In the verification stage, in total 545 candidate tools were presented to human annotators (the corresponding author of the chapter) for judgement. Most cases were quite straightforward without the need of referring to the contexts of mentioning, such as "CRISPR method" or "Cochrane Anaesthesia Review Group method". To facilitate the judgement of some more complicated cases their locations of mentioning and the surrounding contexts were provided to human annotators, such as "GRADE method" for which the corresponding context indicated that it had the same role as "GRADE Working Group method", i.e. a guideline rather than a computer tool, and "PEPI" for which without the context it is hard to understand what it is but it was able to know that it meant the "Portable, Extensible Photogrammetry Instrument" using the corresponding context complimented by Web searching. The result of annotation was the Tool List Version 1, which would be used to quantify the performance of the human-LLM collaborative annotation approach proposed by the current study (see Sect. 3.3 for details).

In the questioning stage, 499 sentences were voted having no additional tool mention by all five questioning LLMs. They were all double-checked by the corresponding author and it was confirmed that indeed none of the 499 sentences contained mentions of tool names that did not appear in Tool List Version 1. On the other hand, the questioning process reported an additional 1,198 suspicious tools. They were also manually checked by the

corresponding author following a similar procedure as in the verification stage. The new tools that were confirmed by manual verification but were overlooked by Tool List Version 1 were added to the latter, resulting in Tool List Version 2, which would also be used for quantitative evaluation (see Sect. 4.3).

Prompt 3-1: Tool Questioning Based on Self-criticism – Task

You will receive one sentence at a time, plus a predefined Tool List 1.

Your goal is **only** to discover **new tool names** that are **not** in Tool List 1.

> Common forms of “tool names” include (but are not limited to):

> - software

> - platforms

> - algorithms

> - frameworks

> - workflow

Steps:

1. Ignore any names that appear in **Tool List 1** (case-insensitive match)to ensure you only identify new tools.

2. Scan the sentence and extract every candidate you believe to be a new tool name.

- Matching is case-insensitive; treat hyphens and spaces equivalently (e.g. “My-Tool” vs “My Tool”).

- Refer to the common forms above, but do not limit yourself to them.

3. For each candidate, extract up to 5 words before and up to 5 words after its mention (if fewer words exist, take as many as available), showing only the words (no extra punctuation).

-- The remainder is the second part of the prompt containing the input and output formats

Fig. 3 Prompt for Questioning and Suggesting Possibly Overlooked Tools – Part One about Task Description and Detailed Instructions.

Prompt 3-2: Tool Questioning Based on Self-criticism – Input and Output

-- Prior to the following is the first part of the prompt containing task description and detailed instructions for completing the task

Input Template:

paper_id: {entry["paper_id"]}

position: {entry["position"]}

sentence: "{entry["sentence"]}"

Tool List 1:

[{tool_list_str}]

Expected Output Format (exactly, no extra text):

- Only output new tools **not** in Tool List 1.

- If no new tool is found in the sentence, output nothing and move on to the next sentence.

- Strictly follow this format. The tool name must be the bare name, without any "software", "platform", "framework", or "workflow" suffixes or prefixes:

{{tool}}: TOOL_NAME

{{explanation}}: [$WORD_{-5}$ … $WORD_{-1}$] TOOL_NAME [$WORD_{+1}$ … $WORD_{+5}$]

{{paper_id}}: {entry["paper_id"]}

{{position}}: {entry["position"]}

- If multiple new tools appear in one sentence, repeat the block for each in the order they appear.

Write your answer **below the line**:

Fig. 4 Prompt for Questioning and Suggesting Possibly Overlooked Tools – Part Two about Input and Output Rules.

3 Quantitative Evaluation

3.1 "Gold Standards"

To evaluate the performance of the proposed human-LLM collaboration pipeline, five "gold standards" were curated and published in the GitHub following project page: https://github.com/xiaoruijiang/tool_in_cochrane.

Gold Standard Tool List Version 1 (**GS-TL-V1**) comprised the tools identified by ten major prior SRs of a similar purpose [11-20]. Since the first quasi-review of AI and tool usage in systematic review was published in 2014 [8], the time range for selecting the source reviews to build GS-TL-V1 is 2019 and 2024. The details of these SRs are presented in Table 1, including the number of tools that were found by each review paper about tool usage. Only

reviews (or SRs) that applied a systematic searching method were included. Because of this criterion, [17] and [19] were both included although they are not titled systematic reviews or scoping reviews. Reviews or SRs about AI techniques were excluded if there was no significant discussion or summary of tools, such as [46-47]. Tools that support automation or semi-automation of SRs do not necessarily fall in the subcategory of AI/ML tools or contain an explicit AI/ML element at the core. However, AI/ML has been obviously the central topic of ASR, especially in the scientific literature, so the source for constructing GS-TL-V1 also include SRs about AI if they had compiled a list of AI or AI-enhanced tools, such as [12-13, 17, 19]. In total, 247 computer tools were identified in GS-TL-V1. All possible forms of mention of the same tool that appeared in the ten source reviews were manually normalised by the corresponding author of this chapter.

Table 1 Ten source review articles for creating Gold Standard Tool List V1.

Ref.	Title	Journal	Year	#Tool	AI only?
[11]	Usage of automation tools in systematic reviews (doi: 10.1002/jrsm.1335)	Research Synthesis Methods	2019	34	No
[12]	Using artificial intelligence methods for systematic review in health sciences: A systematic review (doi: 10.1002/jrsm.1553)	Research Synthesis Methods	2022	9	Yes
[13]	Machine Learning Tools and Platforms in Clinical Trial Outputs to Support Evidence-Based Health Informatics: A Rapid Review of the Literature (doi: 10.3390/biomedinformatics2030032)	BioMedInformatics	2022	48 (32+16)	Yes
[14]	Tools to support the automation of systematic reviews: a scoping review (doi: 10.1016/j.jclinepi.2021.12.005)	Journal of Clinical Epidemiology	2022	57	No
[15]	The use of artificial intelligence for automating or semi-automating biomedical literature analyses - A scoping review (doi: 10.1016/j.jbi.2023.104389)	Journal of Biomedical Informatics	2023	62	No
[16]	An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review (doi: 10.1186/s12874-024-02320-4)	BMC Medical Research Methodology	2024	41	No
[17]	Artificial intelligence for literature reviews: opportunities and challenges (doi: 10.1007/s10462-024-10902-3)	Artificial Intelligence Review	2024	25	Yes
[18]	Automation tools to support undertaking scoping reviews (doi: 10.1002/jrsm.1731)	Research Synthesis Methods	2024	55	No
[19]	Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review (doi: 10.1007/s10462-024-10844-w)	Artificial Intelligence Review	2024	44	Yes
[20]	Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed (doi: 10.1186/s13643-024-02592-3)	Systematic Reviews	2024	50	No

The Gold Standard Tool List Version 2 (**GS-TL-V2**) contained computer tools from the well-known Systematic Review Toolbox[18,19] [48]. Our search was done on 18 July 2025 by sending an empty search string on its website,

[18] https://systematicreviewtools.com/

[19] Reference [11] also gathered candidate tools from the "Systematic Review Toolbox" website, but at the time of their search, the website only "listed 111 tools". "Based on publicly available information and personal experience, tools that automated any part of the systematic review process were selected … after disagreements were resolved, 31 tools remained." Three more tools were manually added by the authors, resulting in 34 tools. Because the article was comparatively old and the tool list was obviously incomplete compared to the current landscape in 2025, it was not chosen as a separate gold standard.

which returned 347 tools at the time of searching. Non-computer tools were annotated by the corresponding author, including those that were used as *database* (e.g., source for searching studies such as COVID-NMA, PDQ-Evidence), *guideline* (such as AMSTER, a measurement tool for assessment of multiple systematic reviews), or *knowledge hub* (such as the website named "Which screening method should I use?"). There were three exceptions made to ensure the consistency between GS-TL-V2 and GS-TL-V1. Epistemonikos, Trialstreamer and MedTerm Search Assistant were treated as both *database* and *computer software* (thus tool in our definition) because they also provide search facilities that systematic reviewers repetitively referred to as a tool according to the reviews included in GS-TL-V1. Additionally, there was a debate on whether models like BERT, its variants like RoBERTa, domain-specific and task-specific versions (such as BioBERT and srBERT respectively) could be called a tool. Inconsistencies existed in prior studies. For example, srBERT—a pretrained BERT model on article abstracts aimed at systematic review automation—was treated in [20] as tool but other pretrained models like BioBERT and RoBERTa were not. The decision of the current study was to treat them as *method* rather than *tool*. Thus, they were excluded from evaluation and analysis. GS-TL-V2 contained 216 tools.Gold Standard Tool List Version 3 (**GS-TL-V3**) contained names of computer programs that appeared in references of Cochrane reviews that were marked as "Computer program". Thus, it matches the definition of tool in the current study. But there were obviously other entity types that were marked as Computer program, including *organisation* (such as IARC – International Agency For Research on Cancer), *data source* (such as PubMed, ResearchGate), *guideline* (such as CDC Surgical Site Infection Guideline), *intervention* (such as CogniSpeed – CogniSpeed Brain Training), *assessment* (such as Computerized Neurocognitive Function Test), and others that were hard to classify (such as Veritas Health Innovation and Ayush 2007). These were all manually annotated by the corresponding author and excluded from evaluation and analysis. GS-TL-V3 contained 105 tools.

The Gold Standard Tool List Version 4 (**GS-TL-V4**) and Version 5 (**GS-TL-V5**) each came from a recent systematic review about AI/ML in evidence synthesis [3, 21]. Reference [3] was chosen as a separate source of gold standard GS-TL-V4 because the authors not only chose tools from research articles but also integrated tools that list in several software repositories, including the Systematic Review Toolbox. It is worth noting that GS-TL-V4 also contained *programming packages*, a type of tool (computer software) that was not included in the Systematic Review Toolbox and any other gold standards. This made GS-TL-V4 the most similar to the aim of the current study in nature. GS-TL-V4 contained 177 tools (collated from Figure 1 and Table 1 in [3], among which 59 that were included in GS-TL-V4 were marked as ML tools according to [3][20]). On the contrary, Reference [21] was added as a separate gold standard because of its most recency (published on 25 August 2025, just after the current study completed all analyses). This was an impressive article and the largest study of this kind so far, containing 2271 evidence synthesis studies published on Cochrane, Campbell Collaboration, and Environmental Evidence Reviews between 2017 and 2024. However, their data extraction and analysis were all made manually, while the current study presented a case study of semi-automating a large-scale bibliometric review in assistance with LLMs. Despite the large size, only 63 tools were reported in [21] according to its supplementary material file. After excluding Baidu as a general-purpose search engine, the size of GS-TL-V5 was 62.

The overall quantitative evaluation process is illustrated in Figure 5, and will be described in detail in the Results section (Sect. 3.3)

[20] According to [3], there were 63 ML tools (including source codes of included studies), but four were excluded by the current study because they did not include detailed instructions for most systematic reviewers to apply in real-world practice. Additionally, the excluded tools (indeed source codes of published papers) did not overlap with any other gold standards.

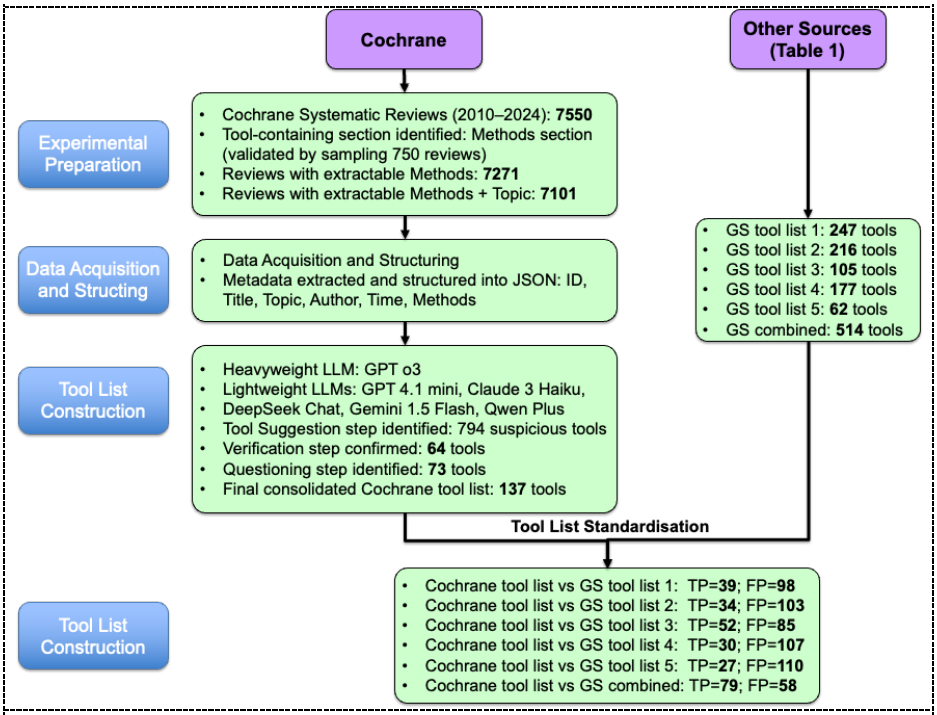


Fig. 5 The flowchart for quantitative evaluation.

3.2 Evaluation Metrics

Three basic metrics could be borrowed from machine learning for evaluation, i.e. *true positive* (TP) – tools extracted that were indeed in a gold standard (common tools), *false positive* (FP) – tools extracted that were not in a gold standard (unique tools), *false negative* (FN) – tools in a gold standard that were not extracted (missed tools). There was no way to identify true negatives and it was meaningless to do so. Based on them, two more metrics could be defined. *Precision* is the percentage of our identified tools that were in a gold standard: $Precision = TP/(TP + FP)$. *Recall* is the percentage of the tools in a gold standard that were successfully identified by our approach: $Recall = TP/(TP + FN)$. Note that the priority of LLM-assisted tool extraction is to identify as many as possible the tools that have been used in prior evidence syntheses. To facilitate this analytical purpose, FP has the most value. Because tools were all extracted from Cochrane reviews and all manually checked by the chapter authors, they were real tools. False positives mean the tools that were identified by the current study but overlooked by a gold standard, so a high FP signifies the value of the proposed approach.

Human annotation happened at the ends of the verification and questioning stages, resulting in two lists of extracted tools for evaluation. Each will be evaluated against the five gold standards to demonstrate (1) the performances of the proposed LLM-Human collaborative annotation pipeline and (2) the unique contributions by each component of the pipeline. It is worth noting that there must be tools that were missed by the current approach, which will be discussed in the Discussions and Limitations section.

3.3 Results

3.3.1. Overall Results of Tool Identification

In the Verification stage, in total 794 candidate tools were returned by GPT-o3, of which 545 were verified by the five lightweight LLMs. Human annotation confirmed that **64** tools were indeed tools identified in the verification stage. During annotation, the number of tool names were more than that because the same tool may appear in different articles in slightly different forms. When creating GS-TL-V1, the authors of the current study maintained a dictionary of tool name normalization. This dictionary was used to assist normalising the surface forms of the same tool during the manual annotation after the verification and questioning stages, and it was gradually adapted when the other four gold standards were created. In the Questioning stage, in total 1697 text segments were marked as potential tools, of which 499 were voted by all five lightweight LLMs as non-tool, and 1198 received at least one vote for being a tool. After manually checking these 1198 candidates, another **73** tools were identified in the questioning stage, forming a total of 137 tools that were extracted by the proposed LLM-human collaborative pipeline from Cochrane reviews published between 2010 and 2024. The complete list of the identified tools, the lists of five "gold standards" and the comparisons between them were published on the project page of this study at https://github.com/xiaoruijiang/tool_in_cochrane. The fact that about 53.3% of the tools were identified in the questioning stage on the one hand justifies critical role of a multi-stage approach comprising proposal and self-reflection but on the other hand highlights the limitations of the proposed approach. More discussions are presented in the Limitations section.

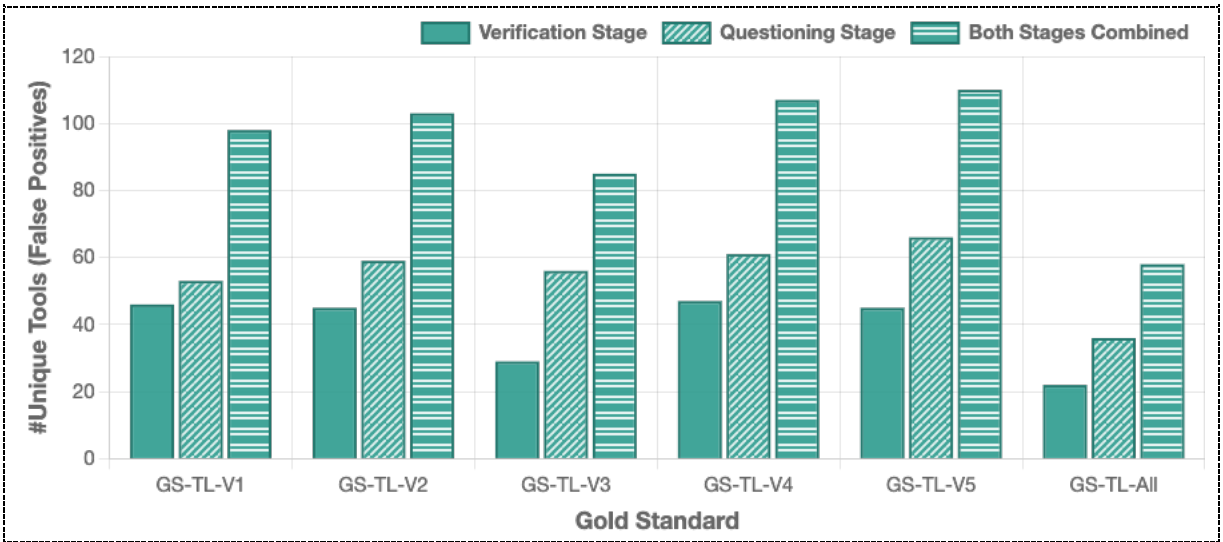


Fig. 6 The performance of identifying the (unique) tools overlooked by the existing manually curated tool lists.

3.3.2. How Many Unique Tools Were Discovered?

The most prominent value of the proposed approach is its effectiveness and efficiency of identifying tools from a quasi-exhaustive search from a collection of evidence syntheses and its capability of identifying the unique tools that were largely overlooked by prior studies that were manually done. This capability was quantified by comparing the tool lists extracted by our pipeline (TL-V1 and TL-V2) against the five "gold standards", as shown in Figure 6. The gold standards were manually curated through systematic searches by prior studies. Treating them as gold standards, the unique tools that were identified by the proposed pipeline were seen as *false positives*. Compared to GS-TL-V1, the most comprehensive list of 246 tools, 98 new tools were successfully identified by the proposed pipeline, 46 and 53 from the verification and questioning stages, respectively. This increased the toolset by 39.8%.

There were several main (overlapping) categories of overlooked tools. (1) The first overlooked category was *single-stage tools* or single-purpose tools, either as standalone software or as programming packages. Tools for data synthesis and meta-analysis were most prevalent in this category. Standalone software included the Comprehensive Meta-Analysis software, CiNeMA (Confidence in Network Meta-Analysis), MetaInsight, WinBUGS, etc. There were also a lot of programming packages such as BRugs, BUGSnet, gemtc, meta, metafor, netsplit, all of which are R packages. Other single-purpose tools existed, such as tools for data analysis and extraction like Anthro (WHO Anthro Survey Analyser), tools for reference management like RefMan and ProCite, tools for data extraction like GetData Graph Digitizer and xyExtract Graph Digitizer, and tools for graphing and visualisation production. (2) The second category was statistical analysis software such as GraphPad, SAS, SPSS, Stata, S-PLUS. (3) The third category was *plug-ins of existing software or tool* like metan for Stata, METADAS for SAS, MetaView for RevMan, and Grab It! XP for Excel. While most tools dot not fall in the AI/ML tool category that prior studies focused on, they fall in the wider category of automation tools according to prior studies. When compared to GS-TL-V1, several important tools that are for either managing the whole evidence synthesis/systematic review process or AI-enhanced were identified, such as RevMan Web (the Web version of Review Manager, RevMan in short), the Cochrane Task Exchange (platform for connecting people needing help with reviews), the TSA software (Trial Sequential Analysis). More "trivial" tools included Access, Excel, Word, Google Form, Baidu Translate (considering Google Translate was selected by prior studies as a valid tool), PDFTron (for PDF processing), OmniPage (for OCR—Object Character Recognition), etc. Similar patterns of the newly identified tools could be observed when comparing against other gold standards.

GS-TL-V2 comes from Systematic Review Toolbox (SRToolbox), which is so far the most comprehensive manually maintained list of SR automation tools. Using the proposed pipeline, 103 additional tools were identified (45 from the verification stage and 59 from the questioning stage). The unique tools were largely similar to those when compared to GS-TL-V1. This was because several sources of GS-TL-V1, such as [17] and [18], also searched SRToolbox. Comparatively, three important additional tools of automation were identified, including the Cochrane RCT Classifier, SARA (System for Automatically Requesting Articles), and TerMIne. It was surprising that the extremely popular RCT Classifier was not included in SRToolbox, considering the fact that two gold standards included it (GS-TL-V1 and GS-TL-V4). SARA was validated in real-world review, which allowed a large review to be completed within two weeks [49]. TerMIne is a widely used term extraction tool developed by the NaCTeM (the National Centre for Text Mining), appearing also in two gold standards (GS-TL-V1 and GS-TL-V5). Different from GS-TL-V2, the third gold standard consists of the manually curated computer programs in the Cochrane Library, so in a sense it should be the most relevant baseline.

Comparatively, 85 tools that were unique to GS-TL-V2 were identified, which was amazing. Among them, 29 new tools were identified in the verification stage while 56 in the questioning stage. Important tools that were overlooked by GS-TL-V3 included Ovid, OvidSP, PubReMiner, Publish or Perish, RefWorks (all by tool suggestion and verification), as well as 2dsearch, Citavi, Cochrane Crowd, Consensus, ExaCT, GATE, Cochrane RCT Classifier, RevManHAL, RobotSearch, SARA, Cochrane Task Exchange, TerMIne, Voyant and Zotero (all by tool questioning). It was surprising to see that the famous platforms or tools developed and promoted by the Cochrane were not in GS-TL-V3, like the Cochrane Crowd, Cochrane Task Exchange, and particularly the Cochrane RCT Classifier considering the fact that "the most employed ML technique identified … was the RCT classifier" (see GS-TL-V5, [21]). Some tools were overlooked by GS-TL-V3 probably because the Cochrane Library treated them as platforms, like 2dsearch (an advanced platform for literature searching), Consensus (an AI search engine of scientific publications) and RobotSearch. But the Cochrane Library might had missed several popular and famous offline tools that should be marked as "Computer program", including ExaCT (Extracts Characteristics of Studies from RCTs), GATE (General Architecture for Text Engineering), RevManHAL (an add-on which "helps auto-generate the abstract, results and discussion sections of RevMan-generated reviews"), TerMIne, Voyant (for text and corpus analyses) and Zotero (for reference management).

GS-TL-V4 was special. On the one hand, it was one of the most comprehensive systematic reviews of SR automation tools combining systematic searches of scientific publication databases as well as software repositories, although the focus was on ML-assisted SR automation tools [3]. On the other hand, it not only searched SR Toolbox but also famous software repositories for various mainstream programming languages including The Comprehensive R Archive Network (CRAN), The Python Package Index (PyPI), etc. So, it was natural to expect that a large number of tools found by the current study would appear in GS-TL-V4. However, 107 unique tools were identified when compared to GS-TL-V4. This was probably due to the authors of GS-TL-V4 [21] explicitly used "ML-assisted" as their inclusion criteria. Indeed, when the tools in GS-TL-V4 were manually investigated most of them were to some extent based on ML or having an ML element with only few exceptions or vague cases such as Cochrane Register of Studies, which was treated as Data Source by the current study, and Epistemonikos, which could be seen either as a Data Source or an intelligent searching method, and Cochrane Screen4Me, which is essentially a service for crowd-based screening. Tools that were potentially missed by GS-TL-V4 (i.e., potentially AI tools or tools have an AI element) included GetData Graph Digitizer (for data extraction), GATE (for text mining), Elicit (for automating screening and data extraction), SARA (System for Automatically Requesting Articles) and TerMIne.

In contrast to [21], GS-TL-V5 was curated by searching evidence syntheses published by Cochrane, Campell Collaboration and Environmental Evidence Reviews [3]. Except time range, the exclusion criteria were only about publication type rather than AI/ML keywords. The authors categorised tools used in evidence syntheses into four groups: manual tool, automation tool, ML-enabled tool, and ML-embedded tool. In this sense, this most update-to-date review bears high similarity to the current study. Surprisingly, 109 new tools that were unique to GS-TL-V5 were identified using the proposed pipeline, 45 in the verification stage and 64 in the verification stage. Again, these unique tools were mainly statistical software or plug-ins for existing software or tools mainly for data extraction, synthesis and meta-analysis purposes, or general-purpose tools what probably could be categorised as manual tools according to [21], such as Access, Excel, Word, Google Form, etc. The verification stage identified AI tools including Citavi, Data Abstraction Assistant, EXaCT (both for data extraction), GATE, RobotSearch (for literature searching), SARA, SR-Accelerator's Polyglot Search Translator (for literature searching), and Voyant (for text analysis). Surprisingly, although the authors of [21] claimed that "the most employed ML technique identified in the preliminary review was the RCT classifier", it was not included in the software/tool list but was manually added back to GS-TL-V5 by us.

Overall, combining all five gold standards resulted in a total of 514 tools, and the current study could further identify 58 tools (22 and 36 in the verification and questioning stages, respectively), all of which are *automation tools* according to [21]. The results demonstrated that the proposed pipeline could help identify tools that were easily overlooked by manual approaches even if a systematic approach was applied and implied that, when tool was talked about, researchers of past reviews of a similar purpose to the current study paid more attention to ML-enabled and ML-embedded tools than other types of automation tools. Enhanced by an LLM-assisted annotation pipeline, it was possible to draw a more comprehensive picture of TAMES.

3.3.3. How Effective When Compared to Manual Reviews

A striking fact is that the proposed LLM-human collaborative pipeline proved to have an extremely complementary effective when compared to prior systematic reviews that formed GS-TL-V1. Figure 7 shows the identified tools that are unique to or common with each source of GS-TL-V1. The numbers of common tools were quite small, so we listed the common tools in Table 2 for easier investigation. Indeed, most common tools were the top AI-enabled SR tools such as (in alphabetical order) Covidence, DistillerSR, EPPI-Reviewer, Rayyan, RevMan, RobertReviewer, and the Cochrane RCT Classifier. The data extraction tool ExaCT was also a very popular SR automation tool. There were several categories of automation tools that do not explicitly emphasise on AI/ML: (1) The collaborative platforms such as Cochrane Crowd and Cochrane Screen4Me, which have been promoted by the Cochrane Collaboration for several years; (2) A number of mainstream reference management systems including EndNote, Mendeley,

RefWorks, and Zotero, which have also incorporated AI in certain components; (3) Tools for later stages of SR beyond screening such as DAA (for data extraction), GRADEpro (for quality assessment). Voyant was treated as a corpus analysis tool rather than an AI tool that we expect to make or suggest automated decisions. Another striking fact is that there were NOT A SINGLE common tool about data synthesis and meta-analysis. It seems that human reviewers had a strongly bias towards AI or automation software that have a graphical user interface, but ignored the packages which either act as a software plug-in or require programming expertise. General-purpose statistical software was also excluded even if they were heavily used in meta-analysis such as SAS, SATA, Stata.

Contrary to the sparse common tools, the numbers of unique tools were significant. It is worth noting that three recent references in 2024 [16], [18] and [20] focused on automation tools, not only AI tools, but the current study identified 120, 128 and 122 unique tools respectively, amongst which AI-related tools included 2dsearch (for literature searching), Consensus (AI-powered search engine), Elicit (find, analyse and summarise relevant academic papers), RobertSearch, SRA-Polyglot Search Translator (for searching), TerMIne (terminology mining), Yale MeSH Analyzer. However, similar to the results in Sect. 3.3.2, the majority of these tools were standalone software, programming packages or tool plug-in for certain SR steps including quality assessment, data synthesis and meta-analysis, with meta-analysis being the dominant realm of automation tools.

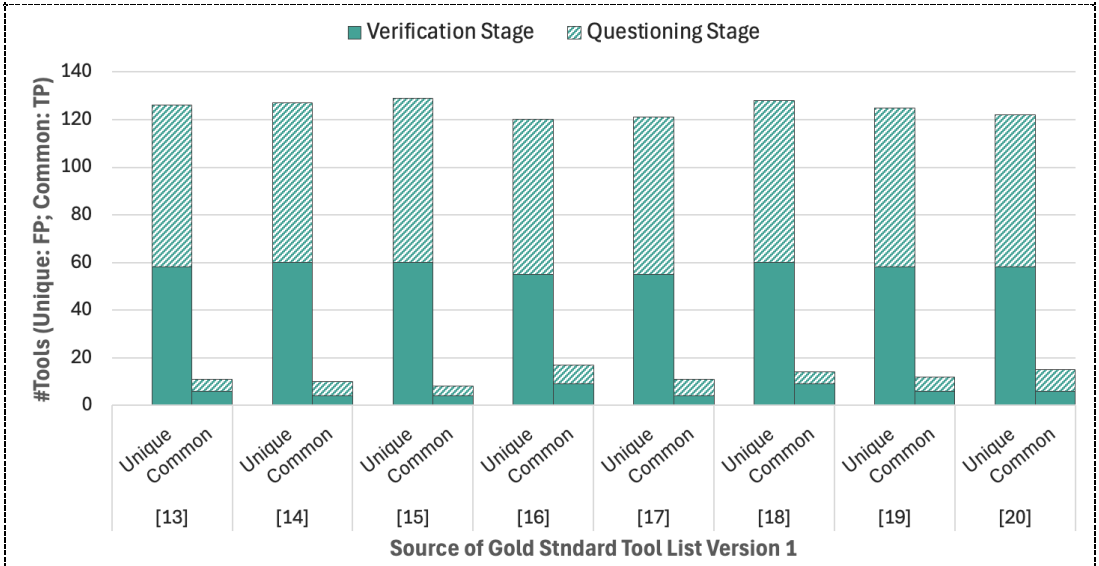


Fig. 7 The performance of identifying tools compared with prior systematic reviews of published evidence syntheses that were done manually.

Table 2 Lists of identified tools from Cochrane reviews that were common with prior systematic reviews of published evidence syntheses.

Ref.	Common Tools Seen as AI Tools	Other Common Tools	AI only?
[13]	V-Stage: Covidence; DistillerSR; EPPI-Reviewer; Rayyan; RobotReviewer	EndNote; WebPlotDigitizer SARA	Yes
	Q-Stage: ExaCT; RevManHAL; RobotSearch	--	
[14]	V-Stage: DistillerSR; EPPI-Reviewer; Rayyan; RevMan; RobotReviewer	--	No
	Q-Stage: ExaCT; RevManHAL; RobotSearch	Google Translate; SARA	
[15]	V-Stage*: DistillerSR; EPPI-Reviewer; Rayyan; RobotReviewer	--	No
	Q-Stage: ExaCT; Cochrane RCT Classifier; RobotSearch	--	
[16]	V-Stage: Covidence; DistillerSR; EPPI-Reviewer; PubReMiner; Rayyan; RobotReviewer	EndNote; PlotDigitizer; RefWorks; Cochrane Screen4Me	No
	Q-Stage: ExaCT; Cochrane RCT Classifier; Yale Mesh Analyzer	Cochrane Crowd; DAA (Data Abstract Assistant); Zotero	
[17]	V-Stage: Covidence; DistillerSR; EPPI-Reviewer; Rayyan; RobotReviewer	--	Yes
	Q-Stage: Consensus; Elicit; ExaCT; RobotSearch	--	
[18]	V-Stage: Covidence; DistillerSR; EPPI-Reviewer; PubReMiner; Rayyan; RevMan; RobotReviewer	EndNote; Excel; GRADEpro; MAGICapp; Mendeley	No
	Q-Stage: DeepL; ExaCT; Cochrane RCT Classifier; TerMIne	Google Translate; Zotero	
[19]	V-Stage: Covidence; DistillerSR; EPPI-Reviewer; PubReMiner; Rayyan; RobotReviewer	GRADEpro	Yes
	Q-Stage: 2dsearch; ExaCT; Cochrane RCT Classifier; Yale Mesh Analyzer	--	
[20]	V-Stage: Covidence; DistillerSR; EPPI-Reviewer; PubReMiner; Rayyan; RobotReviewer	EndNote	No
	Q-Stage: ExaCT; GATE; Cochrane RCT Classifier; RobotSearch; TerMIne; Yale Mesh Analyzer	SARA; Voyant	

* V-Stage: Verification Stage; Q-Stage: Questioning Stage.

3.4 Discussions

The results in Sect. 3.3.2 and 3.3.3 showed that the proposed pipeline had strong capability of identifying tools, including manual, AI and automation tools according to [21], and showed exceptional complimentary values to manually compiled tool lists. It is also necessary to know whether the pipeline is good at identifying the tools that have been widely used. Table 3 lists the top-25 tools according to the 10 source reviews of GS-TL-V1, sorted in descending order of the number of times of being mentioned. The "Found" and "Existing" columns represent whether the tool was identified by the proposed pipeline and whether the tool indeed existed in the Cochrane Library (which can be verified by searching Cochrane). The most mentioned tools that indeed existed in Cochrane reviews were all identified by the proposed pipeline, except RobotAnalyst. But it was not a mistake by the pipeline. In fact, RobotAnalyst only appeared once in Cochrane (by searching the latter at the time of writing, search on 5 Nov 2025) but it was not in the original list of reviews that was downloaded before data processing and analysis started. The tools that were not identified were not used by the authors of Cochrane reviews. To some extent, the analysis implied the potential high recall of tool extraction by the proposed pipeline: While the (recalls of) "true positives" were low based on the five gold standards, these low recalls could not be unjustify the effectiveness of the proposed pipeline; they were caused by the fact that Cochrane reviews only used a much smaller set of tools. The

proposed pipeline strong complementary values to the traditional systematic searching methods for SR automation tools.

From another angle, good recalls on one or several gold standards are not enough to justify the comprehensiveness of any tool identification approach.

As Figure 8 shows the common tools and unique tools among each pair of gold standards. Compared to common tools that only constituted a small portion, each gold standard contained a large number of tools that were unique to others. Notably, GS-TL-V3 contained 71.43% unique tools compared to all other four gold standards combined. The numbers of common tools between GS-TL-V3 and other gold standards were also quite low. Recall that GS-TL-V3 contained the tools that are labelled as computer programs in Cochrane reviews' reference sections. This partially justified our prior conjecture that many SR automation tools were not used by Cochrane reviews. In the opposite direction, it also signifies that many tools that were marked as computer programs were not recognised by other. Surprisingly, even GS-TL-V2 and GS-TL-V4 contained tools from SR Toolbox and mainstream programming package repositories, their overlaps with GS-TL-V3 were still low, having only 19 and 16 tools in common respectively. This signified the difficulty of maintaining or systematically searching a comprehensive list of SR automation tools. Comparatively, the proposed pipeline was able to identify 52 common tools with GS-TL-V3, obtaining the highest recall of roughly 50% among all gold standards (see Figure 9). In total 79 tools were identified when compared to all gold standards combined but only 27 additional tools came from gold standards other than GS-TL-V3.

Table 3 Top tools in Gold Standard Tool List V1 with 4 or more mentions.

No	Tool	Cnt	Ours	CR*	Sources
1	Abstrackr	8	No	No	[13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]
2	DistillerSR	8	Yes	Yes	[13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]
3	EPPI-Reviewer	8	Yes	Yes	[13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]
4	ExaCT	8	Yes	Yes	[13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]
5	Rayyan	8	Yes	Yes	[13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]
6	RobotAnalyst	8	No	Yes*	[13]; [14]; [15]; [16]; [17]; [18]; [19]; [20]
7	RobotReviewer	7	Yes	Yes	[13]; [14]; [15]; [16]; [17]; [19]; [20]
8	SWIFT-Review	7	No	No	[13]; [14]; [15]; [16]; [17]; [19]; [20]
9	ASReview	6	No	No	[14]; [16]; [17]; [18]; [19]; [20]
10	Covidence	6	Yes	Yes	[13]; [16]; [17]; [18]; [19]; [20]
11	LitSuggest	6	No	No	[13]; [14]; [15]; [17]; [18]; [19]
12	Research Screener	6	No	No	[15]; [16]; [17]; [18]; [19]; [20]
13	RobotSearch	6	Yes	Yes	[13]; [14]; [15]; [17]; [19]; [20]
14	SRA-Polyglot Search Translator	6	Yes	Yes	[13]; [14]; [16]; [18]; [19]; [20]
15	SWIFT-Active Screener	6	No	No	[14]; [15]; [16]; [17]; [19]; [20]
16	Colandr	5	No	No	[13]; [15]; [16]; [17]; [19]
17	Dextr	5	No	No	[13]; [15]; [16]; [17]; [20]
18	RCT Classifier	5	Yes	Yes	[15]; [16]; [18]; [19]; [20]
19	RCT Tagger	5	No	No**	[13]; [14]; [15]; [19]; [20]
20	SRA-Deduplicator	5	No	No	[13]; [14]; [16]; [18]; [20]
21	SRA-Helper	5	No	No	[13]; [14]; [16]; [18]; [20]
22	Bibot	4	No	No	[13]; [14]; [15]; [18]
23	EndNote	4	Yes	Yes	[13]; [16]; [18]; [20]
24	SRA-WFA	4	No	No	[13]; [14]; [18]; [20]
25	R package revtools	4	No	No	[14]; [15]; [19]; [20]

* RobotAnalyst appeared once in Cochrane Reviews but not in our dataset.

** Only appeared in Cochrane Central Register of Clinical Trials, not in Cochrane Reviews.

	GS Null	GS-TL-V1	GS-TL-V2	GS-TL-V3	GS-TL-V4	GS-TL-V5	Other Four	Unique%
GS-TL-V1	247	0	160	228	168	215	130	52.63%
GS-TL-V2	216	129	0	197	97	189	67	31.02%
GS-TL-V3	105	86	86	0	89	83	75	71.43%
GS-TL-V4	177	98	58	161	0	151	39	22.03%
GS-TL-V5	62	30	35	40	36	0	19	30.65%

Fig. 8 The unique tools of each gold standard compared to others. "GS Null" means the number of tools in a gold standard.

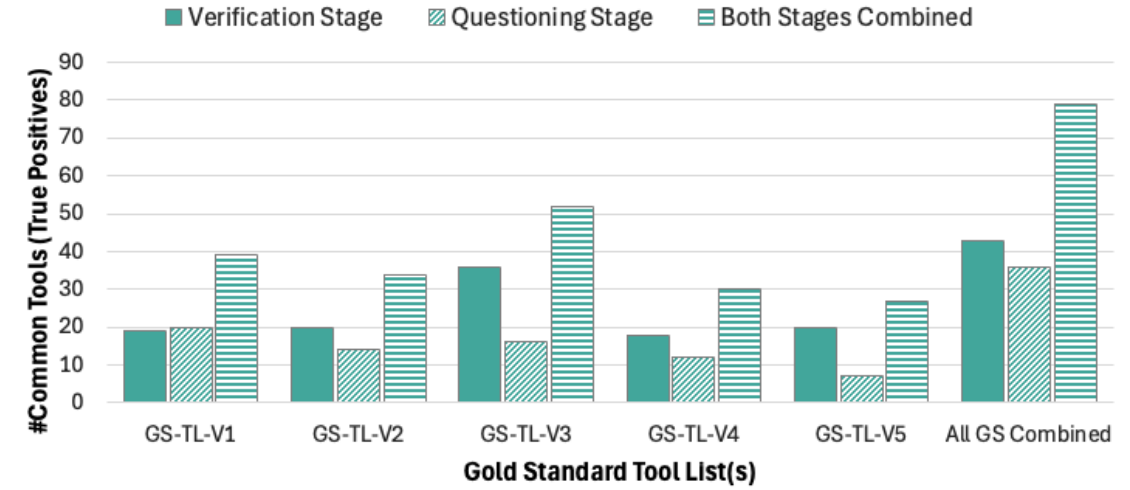


Fig. 9 The commons tools with different gold-standard tool lists.

4 Bibliometric Analysis

This chapter provides some preliminary bibliometric analyses over the Cochrane reviews that were processed and analysed in Sect. 3. Figure 10 shows the yearly distributions of evidence syntheses that were published in the Cochrane Library between 2010 and 2024 (see the blue bar charts) and the percentages of tool usage among them (see the orange curve). While the number of number of reviews fluctuated over the years, there was a clear trend of increasing use of tools in conducting systematic reviews, with the percentage of tool usage increased from around 85% to as high as 97%.

Figure 11 illustrates the trends of using tools for the top-10 topics of most published Cochrane reviews. The topics were extracted from the Cochrane Library. Note that a Cochrane review may be classified into several topics, in which case the number of reviews of each topic was increased by one. From Figure 11, it seems that most topics share a similar pattern of increasing use of tools in conducting medical evidence syntheses: Overall speaking the use of tools was increasing across all topics. In some topics like Cancer (the dark blue curve) and Pregnancy & Childbirth (the light blue curve), all reviews in recent years have used at least a tool in the review process. The Mental Health topic (the grey curve) was interesting. Its tool usage was comparatively low in the early years around 2010, but the tool usage in this topic surged in the following 2-3 years to reach the same level as other topics, and the percentage of tool usage stayed high, hitting 100% after 2022. Other topics bore a similar trend as Mental Health, like

Complimentary & Alternative Medicine (the green curve) and Heat & Circulation (the purple curve), although their starting points were higher than Mental Health. This implied a quick and prevalent transition to TAMES and, once tools had been adopted, the transition to TAMES seemed to be irreversible.

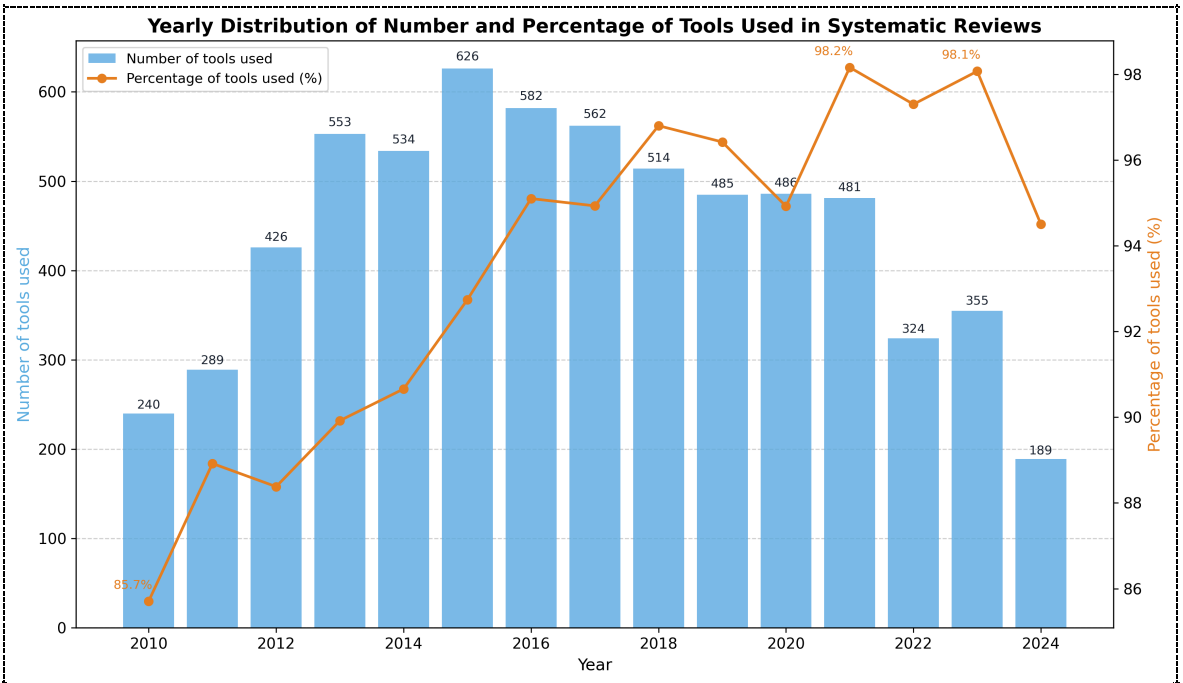


Fig. 10 Trend of tool usage in Cochrane reviews over the years.

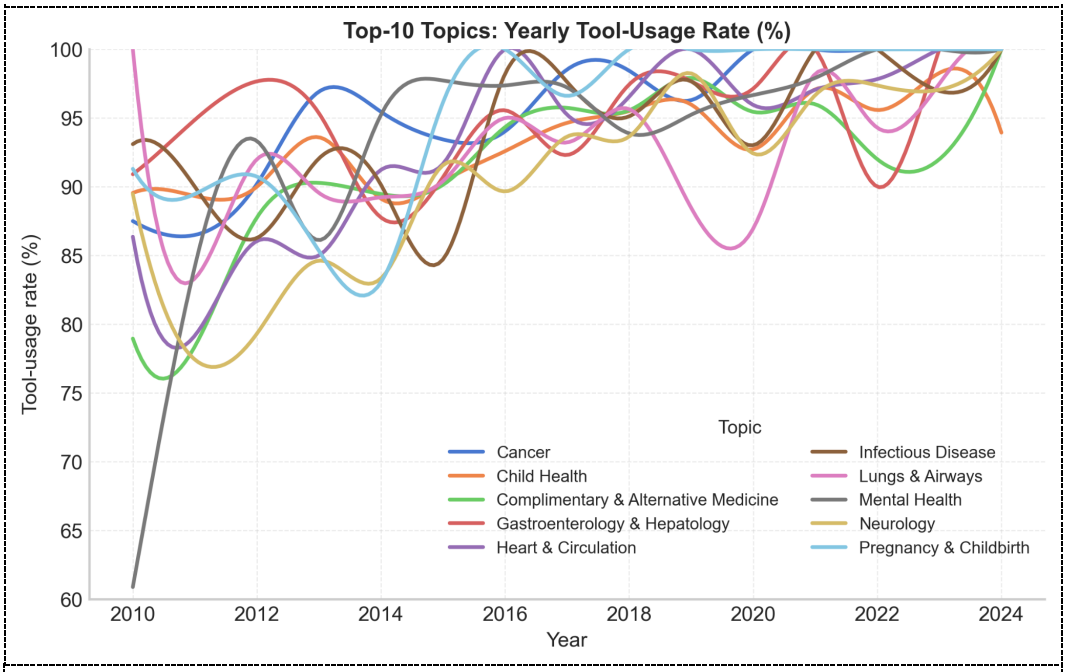


Fig. 11 Trends of tool usage among top-10 Cochrane topics over the years.

There were several pivotal time points in the development and application of (AI/ML-enabled) TAMES. 2010 was selected because around 2010 the first tools like Abstrakr, DistillSR and EPPI-Reviewer entered the market and started to be adopted. In 2020, the outbreak of COVID-19 resulted in a stronger need for adapting to remote collaboration and such tools/platforms and a potential switch of emphasis of the review topics. 2021 was an important year for SR automation, when an extremely popular and excessively cited tool ASReview was published, together with other tools like Research Screener, so it might be reasonable to expect a more universal transition to TAMES after 2021. 2023 was selected because of the birth of ChatGPT and similar LLM and Generative AI models, which stimulated heated discussion about new solutions to SR automation and the future of it. Indeed, there was a peak of tool usage after 2020 (See Figure 10), although it is hard to decide to which factor this increase should be attributed. Figure 12(a-d) show the distributions of tool-using Cochrane reviews across the top-10 topics in 2010, 2020, 2022 and 2024. It was observed that before COVID, there was an increase in the topic Heart & Circulation (from rank 9 in 2010 to rank 2 in 2020) and after COVID there had been increasing attention to the topics Infectious Disease (from rank 7 in 2020 to rank 2 in 2022 but dropped again in 2024 to rank 9) and Lungs & Airways (from rank 8 in 2020 to rank 6 in 2022 and but dropped out of top-10 in 2024). This might be connected with the temporary shift of research focus to coronavirus-related research and research in its aftermath shortly after the outbreak of COVID-19 in late 2019.

5 Limitations and Future Work

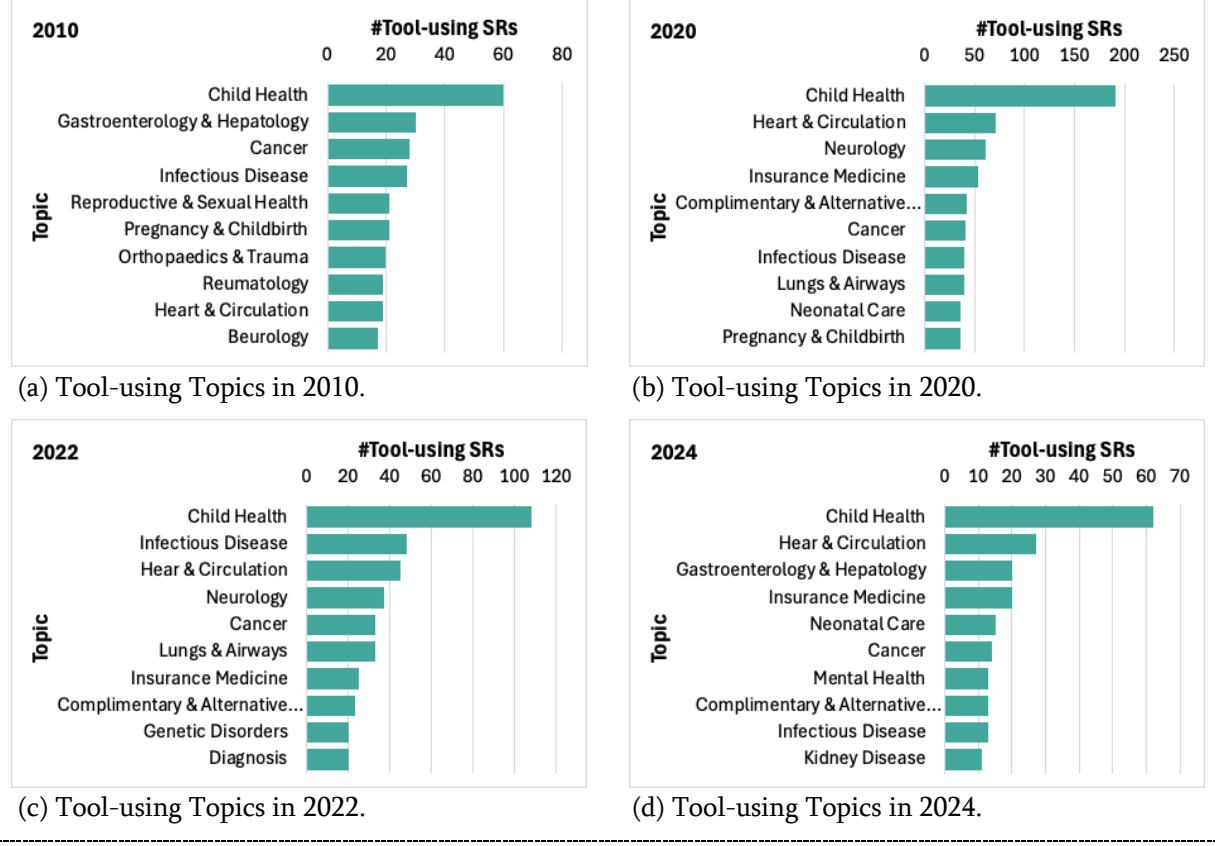
The study presented in this chapter has several limitations. The most prominent limitation is the scope. The study was limited to more than 7000 Cochrane reviews. Although the proposed pipeline showed strong performance in identifying the contexts of tool usage, there were large numbers of false negatives – tools identified by gold standards that were manually curated from various sources beyond Cochrane. Even when compared to the smallest gold standard GS-TL-V5 of only 62 tools, there were (62-27=)36 false negatives. Thus the results of the preliminary bibliometric analyses only apply to the Cochrane.

The second limitation was that the data processing was limited to the Methods section because this section is the most likely section where tool usage is disclosed when describing the methodology. Because of the comparative sparsity of tool mention in other sections, limiting to the Methods section was a well-justified cost-effective choice. However, some tools were inevitably undiscoverable. From Sect. 3.4 and Figure 9, the proposed pipeline had high false negative despite that a large number of tools overlooked by the gold standards. Particularly, GS-TL-V3 and the current study were all constrained to Cochrane, so the proposed pipeline recorded the highest true positive rate (i.e. recall) with respect to GS-TL-V3. However, the false negative rate was as high as about 50%, many of which were indeed tools missed by the pipeline. The reasons were multi-fold. Often the names of computer programs were not labelled in the Cochrane Library. Instead, many computer programs were associated with a cited reference that is about the corresponding tool, and in such cases the surname of the first author of the publication followed by the publication year were used as the computer program name. False negatives also occurred when the tool was not mentioned in the Methods section. A notable example of such case was the R package robvis, which was 29 times in our time range of search but not discovered by the proposed pipeline.

After investigation, many false negatives were found to be real tools overlooked by the proposed pipeline, such as Adobe Photoshop (CD005342), R package compute.es (CD006239), SpiderCite (CD013649), etc. Although AI-enabled tools were all discovered by the proposed pipeline, most overlooked tools in Cochrane seemed to be automation or manual tools according to [21]. The bottleneck of the proposed pipeline was the tool suggestion stage, which expected a strong reasoning model, GPT o3 in the current study, to find as many suspicious cases of tool usage as possible. If GPT o3 missed a tool then it became unrecoverable. The questioning stage proved to be extremely important because it found more tool than the verification stage (Figure 6) by criticising and reflecting on GPT o3's

initial results. It is hypothesised that if self-reflection and questioning were applied to all the initial decisions of GPT o3, more tools would have been detected by doubling or tripling API costs.

Alternative solutions exist. Another important value of the results of the current study is that a high-quality dataset about tool (not limited to SR automation as many are general-purpose tools) has been created. A future direction will be training deep learning-based tool identification methods and developing a large-small language collaborative approach for tool and tool usage identification. Also, the experiments and analyses will be extended to other sources of medical evidence syntheses, so that both the dataset and the analysis can reflect the whole landscape of TAMES.



(a) Tool-using Topics in 2010.

(b) Tool-using Topics in 2020.

(c) Tool-using Topics in 2022.

(d) Tool-using Topics in 2024.

Fig. 12 Distributions of tool-using Cochrane reviews in different years.

6 Related Work

The work presented in this chapter sits at the intersection of three evolving domains: technology-assisted medical evidence synthesis, applications of LLMs for research tasks, and computational analysis of scientific practice.

6.1 Prior Reviews of Technology-Assisted Medical Evidence Synthesis

A primary objective of many reviews has been to catalogue the available automation tools. Early efforts revealed a rapidly expanding ecosystem. The mapping review in [3] uniquely incorporated software repositories, identified 63 distinct ML-based tools out of 177 automation tools (which was collated from [3] by the current study). A scoping

review in [15] identified 273 studies that were categorized into three groups: assembly of scientific evidence (47%), mining the biomedical literature (41%), and quality analysis (12%). The focus of these tools was found heavily skewed, with a majority targeting screening and a relative neglect of later phases like data extraction and quality assessment [12, 14, 17]. New tools like LLMs or those built on LLMs were also investigated in [17]. Complementing these efforts, some studies shifted from mere cataloguing to practical guidance, offering a step-by-step framework for integrating validated tools [18]. Beyond cataloguing tools, several reviews have also assessed their performance and claimed efficiency gains [16]. The mode of evaluation was often limited, as noted in both [16] and [21], that very few studies provided prospective evaluations conducted within real-time review workflows.

These studies established a crucial point: our understanding of tool usage has, until now, been based on self-reported survey data [11] or manual, small-scale synthesis of the literature [12, 17]. There has been a methodological gap in the ability to systematically and objectively measure tool adoption at scale by directly analysing the full texts of a comprehensive corpus of evidence syntheses. Prior reviews in our field have been limited by the inherent constraints of manual data extraction. The most recent study published in Aug 2025 made a significant stride by systematically analysing thousands of evidence syntheses published in three major evidence synthesis networks—Cochrane, Campell Collaboration and Environmental Evidence Reviews, but its methodology still relied on a manually curated list of tools and complete manual extractions, a process noted as a limitation [21]. Other reviews, such as [16], explicitly state that their conclusions are based on a sample of studies (i.e., published evidence syntheses) that could be feasibly processed by a human team. This manual paradigm is not scalable to the entire corpus of evidence, such as all Cochrane reviews, and might be quickly outdated by the rapid pace of tool development.

The advent of LLMs has opened new frontiers for automating and scaling complex text-analysis tasks. Within evidence synthesis, LLMs were identified by [17] and [19] as a transformative trend. Their focus on using LLMs to conduct reviews was relevant to the current study, though not on analysing reviews a scientific corpus. The essence of the work presented in this chapter is using LLMs to conduct a computational bibliometric review of tool usage in medical evidence syntheses based on an exhaustive search of studies. LLMs remarkable proficiency in information extraction and its reasoning capabilities were employed to assist the identification and annotation of SR automation tools, which presents a novel opportunity to overcome the limitations of manual scientometric analyses and exceed the current scale limits as in [21] caused by keyword-based searches, which were never perfect no matter how finetuned.

6.2 Collaboration among Multiples Large Language Models and Humans

Research has shown that LLMs are able to refine their own output without external supervision such as in Self-Refine [50], where an LLM critiques its own response using a self-generated critique and refines the original output based on this iterative internal feedback loop. However, single models inevitably have inherent limitations, which could be overcome by frameworks that allow multiple LLMs to collaborate. Communication-based approaches, such as Exchange-of-Thought (EoT) [51], enable LLMs to exchange internal reasoning steps, which is shown to enhance their collective capabilities. The concept of multi-agent debate (MAD) is also widely studied: ChatEval utilizes MAD to improve the reliability of LLM-based evaluators [52]. Research has also demonstrated that multi-agent communication may rely on peer opinions to improve the overall performance [53], this could be due to MAD encouraging more divergent and creative thinking [54]. Beyond debate, ensembling is also an effective method to enhance accuracy in tasks like citation screening for literature reviews [55]. Comprehensive surveys exist for multi-agent collaboration, focusing on collaborative strategies such as merging, ensembling, and cooperation [56-57] and combining multiple agents' decisions [58].

LLMs are increasingly used as automated annotators due to their efficiency and semantic comprehension capabilities, as also shown in the current study. However, LLMs are also prone to inherent biases and errors, which requires human oversight to gatekeep data quality [59]. A key example is the use of a human-in-the-loop (HIL) validation process for tasks such as data extraction in systematic reviews, where human validation is required to

ensure the reliability of LLM-generated outputs [60]. Studies also showed that human annotators often strongly took the LLM suggestions and improved their self-reported confidence in subjective tasks [61], rather than making them fast annotators. While the current study showed that LLMs can assist humans in achieving a task with significantly reduced time, which could not be imagined due to the scale without LLMs, we agreed with [61] it is important for human annotators to constantly reminder themselves to maintain a critical attitude towards LLM-made annotations. In scientific workflows, HIL reviewing can alleviate the heavy burden on human reviewers while mitigating risks like bias, requiring human over-sight for final editorial decisions [62]. Alternatively, LLM-as-a-Judge has emerged been a new paradigm of using LLMs to evaluate the outputs of other models [63-64], obtaining significant attention [65-66]. Research has identified self-favouritism as a notable challenge of LLM-as-a-Judge, where LLM evaluators may bias towards their own model family [67]. The current study adopted an LLM ensemble approach to alleviate the impact of self-favouritism.

7 Conclusion

This chapter presents the efforts and results of an experiment of conducting a large-scale computational bibliometric review of the practice of applying AI and technology in medical evidence synthesis (TAMES) at a scale that could not have been done in such an efficient and effective way, using over 7100 Cochrane reviews as a case study. Through collaborating with an LLM-enabled annotation pipeline, it only took less than two days for an expert human annotator to justify real tools that had been used in evidence synthesis by the time of writing from the initial annotation results of LLMs. The annotation pipeline contained a tool suggestion stage using a strong reasoning model (GPT o3 in the current study), a tool verification stage that used five lightweight LLMs to evaluate and vote on the candidate tools aiming to improve precision of tool identification, and a tool questioning stage that used the five lightweight models to find out potentially overlooked tools in previous stages aiming to improve the recall of tool extraction. Human intervention happened twice before the ends of the tool verification and questioning stages, respectively. Particularly, the tool questioning stages required LLMs to only suggest potential tools that did not appear in tool verification stage, which was a way of providing human feedback to LLMs.

Evaluation was made by comparing the identified tools against five “gold-standard” tool lists that were compiled by assembling tools from prior systematic reviews of tool usage (247 tools), the popular manually maintained systematic review toolbox (216 tools), the computer programs manually labelled in the Cochrane Library (105), one recent mapping review whose sources of search included code repositories (177 tools), and the most recent systematic review considering three evidence synthesis organisations that included Cochrane (62 tools). A significant contribution of the current study is that 98, 103, 85, 107 and 110 unique tools were identified when compared to each gold standard, and there were still 58 unique tools when compared to all five gold standards combined. Comparatively, the overlaps between tools used in Cochrane (identified by the current study) with the gold standards were rather limited, indicating that tool usage was quite diverse across difference sources of medical evidence synthesis. Particularly, when compared to the third gold standard that was compiled from Computer programs in the Cochrane library, the proposed pipeline achieved a recall of 50%. Apart from the reasons causing many Computer programs that were impossible to be identified due to lack of linguistic cues and could only be detected by analysing the references, there were still some tools that the proposed pipeline failed to identify although the number of such tools was hypothesised to be small, such as SpiderCite which was cited once in Cochrane. In overall speaking, the proposed LLM-human collaborative annotation pipeline has proved its excellent balance between performance and efficiency. When compared to each gold standard, the questioning stage helped extract more unique tools than in the verification stage, which justified the design of the proposed annotation pipeline and also highlighted that it is important to pay attention to not only the over-extraction problem that had been reported in prior studies but also the potential under-extraction issue of LLMs.

The current study also analysed the tools that were used in Cochrane reviews and the tool-using review topics. The most important AI-enabled tools, especially standalone software, were successfully identified, such as DistillerSR, EPPI-Reviewer, Rayyan, RobotReviewer, Covidence. Some popular tools were not identified because they were indeed not used by Cochrane reviews, such as Abstrackr, SWIFT-Review, SWIFT-Active Learner, and especially ASReview (Lab)—so far the most cited open-source standalone software. Most unique tools were software packages or tool plug-ins for a particular systematic review step, including data extraction, quality assessment (e.g. risk of bias assessment), and especially data synthesis and meta-analysis—the stage of a significant variety of tools. Bibliometric analyses revealed that TAMES had become universal. As early as in 2010, tool utilisation rate was already as high as over 85%, and the percentage of TAMEs had increased to almost 98% in recent years. The penetration of tool seemed to have co-evolved with the continuous development and gradually maturation of SR automation tools, particularly AI tools. Once tools had penetrated a topic area, the trend seemed to be irreversible—TAMES destined to be the future of medical evidence synthesis.

References

[1] Van Dinter R, Tekinerdogan B, Catal C (2021) Automation of systematic literature reviews: A systematic literature review. Inform Software Tech 136: 106589.

[2] Borah R, Brown A.W, Capers PC et al (2017) Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the prospero registry. BMJ Open 7(2):e012545.

[3] Jimenez R.C, Lee T, Rosillo N et al (2022) Machine learning computational tools to assist the performance of systematic reviews: A mapping review. BMC Med Res Methodol 22:322.

[4] Sundaram G, Berleant D (2023) Automating Systematic Literature Reviews with Natural Language Processing and Text Mining: A Systematic Literature Review. In: Yang XS, Sherratt RS, Dey N et al (eds) Proceedings of Eighth International Congress on Information and Communication Technology, ICICT 2023. Lecture Notes in Networks and Systems, vol 693. Springer, Singapore, p 73.

[5] Lieberuma JL, Toewsb M, Metzendorf MI et al (2025) Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. J Clin Epidemiol 181:11746.

[6] Cao C, Arora R, Cento P (2025) Automation of systematic reviews with large language models. medRxiv.

[7] Kanoulas E, Li D, Azzopardi L et al (2017) CLEF 2017 Technologically Assisted Reviews in Empirical Medicine Overview. In: Cappellato L, Ferro N, Goeuriot L et al (eds) Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14 September 2017. CEUR Workshop Proceedings, vol. 1866.

[8] Tsafnat G, Glasziou P, Choong MK et al (2014) Systematic review automation technologies. Syst Rev 3:74.

[9] Cohen AM, Hersh WR, Peterson K et al (2006) Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assn 13(2):206–219.

[10] Scott AM, Forbes C, Clark J et al (2021) Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol 138:80-94.

[11] van Altena AJ, Spijker R, Olabarriaga SD (2019) Usage of automation tools in systematic reviews. Res Synth Methods;10(1):72-82.

[12] Blaizot A, Veettil SK, Saidoung P et al (2022) Using artificial intelligence methods for systematic review in health sciences: A systematic review. Res Synth Methods 13(3):353-362.

[13] Christopoulou SC (2022) Machine Learning Tools and Platforms in Clinical Trial Outputs to Support Evidence-Based Health Informatics: A Rapid Review of the Literature. BioMedInformatics 2(3):511-527.

[14] Khalil H, Ameen D, Zarnegar A (2022) Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 144:22-42.

[15] dos Santos AO, da Silva ES, Couto LM et al (2023) The use of artificial intelligence for automating or semi-automating biomedical literature analyses: A scoping review. J Biomed Inform 142:104389.

[16] Affengruber L, van der Maten MM, Spiero I et al (2024). An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review. BMC Med Res Methodol 24:210.

[17] Bolaños F, Salatino A, Osborne F et al (2024) Artificial intelligence for literature reviews: opportunities and challenges. Artif Intell Rev 57:259.

[18] Khalil H, Pollock D, McInerney P et al (2024) Automation tools to support undertaking scoping reviews. Res Synth Methods 15(6): 839-850.

[19] Ofori-Boateng R, Aceves-Martins M, Wiratunga N et al (2024) Towards the automation of systematic reviews using natural language processing, machine learning, and deep learning: a comprehensive review. Artif Intell Rev 57:200

[20] Tóth B, Berek L, Gulácsi L et al (2024) Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in PubMed. Systematic Reviews 13:174.

[21] Scotti KL, Young S, Gainey MA (2025) Artificial Intelligence and Automation in Evidence Synthesis: An Investigation of Methods Employed in Cochrane, Campbell Collaboration, and Environmental Evidence Reviews. Cochrane Evidence Synthesis and Methods 3(5): e70046.

[22] Dagdelen J, Dunn A, Lee S et al (2024) Structured information extraction from scientific text with large language models. Nat Commun 15:1418.

[23] Ntinopoulos V, Biefer HGC, Tudorache C et al (2025). Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. BMJ Health Care Inform 32:e101139.

[24] Gupta S, Mahmood A, Shetty P et al (2024) Data extraction from polymer literature using large language models. Commun Mater 5:269.

[25] Polak P, Morgan D (2025) Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nat Commun 15:1569.

[26] Wiest IC, Wolf F, Leßmann ME et al (2025) A software pipeline for medical information extraction with large language models, open source and suitable for oncology. npj Precis Onc 9:313.

[27] Balasubramanian JB, Adams D, Roxanis I et al (2025) Leveraging large language models for structured information extraction from pathology reports. Journal of Pathology Informatics. doi: 10.1016/j.jpi.2025.100521.

[28] Builtjes L, Bosma J, Prokop M et al (2025) Leveraging open-source large language models for clinical information extraction in resource-constrained settings. JAMIA Open 8(5). doi: 10.1093/jamiaopen/ooaf109.

[29] Khadka S, Jiang X, Palade V (2025) Data Quality in Clinical Coding: A Critical Analysis and Preliminary Study. medRxiv:2025.08.24.2533432.1.

[30] Khan MA, Ayub U, Naqvi SAA et al (2025) Collaborative large language models for automated data extraction in living systematic reviews. J Am Med Inform Assn 32(4):638-647.

[31] Estornell A, Liu Y (2024) Multi-LLM Debate: Framework, Principals, and Interventions. In Globersons A, Mackey L, Belgrave D et al (eds) Advances in Neural Information Processing Systems, vol 37, Curran Associates, Inc., p. 28938-28964.

[32] Chen Z, Li J, Chen P et al (2025) Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. arXiv:2502.18036.

[33] Pan L, Saxon M, Xu W (2024) Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies. T Assoc Comput Ling 12: 484-506.

[34] Thomas J, Brunton J, Graziosi S (2010) EPPI-Reviewer 4: software for research synthesis. EPPI Centre Software. London: Social Science Research Unit, UCL Institute of Education.

[35] Thomas J, McNaught J, Ananiadou S (2011) Applications of text mining within systematic reviews. Res Synth Methods 2:1-14.

[36] Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. Miami, FL, USA, 28–30 Jan 2012.

[37] Kataoka Y, Takayama T, Yoshimura K et al (2025) Automating the data extraction process for systematic reviews using GPT-4o and o3. Res Synth Methods. doi: 10.1017/rsm.2025.10030.

[38] Agrawal M, Hegselmann S, Lang H et al (2023) Large language models are few-shot clinical information extractors. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 6-10 December 2023.

[39] Dong Q, Li L, Cai D et al (2024) A Survey on In-context Learning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), 12-16 November 2024.

[40] Min S, Lyu X, Holtzman A et al (2022) Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 7-11 December 2022.

[41] Peng K, Ding L, Yuan Y et al (2024) Revisiting Demonstration Selection Strategies in In-Context Learning. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), Bangkok, Thailand, 11-16 Aug 2024.

[42] Jiang X, Khan K, Vasantha ST et al (2024). Evidence Extraction for Automated Medical Coding: Preliminary Evaluation. In: Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval (NLPIR), Okayama, Japan, 13-15 December, 2024.

[43] Akinseloyin O, Jiang X, Palade V (2025). Weakly Supervised Active Learning for Abstract Screening Leveraging LLM-Based Pseudo-Labeling. medRxiv: 10.1101/2025.08.24.25334314.

[44] Zhang H, Cui Z, Chen J et al (2025) Stop Overvaluing Multi-Agent Debate -- We Must Rethink Evaluation and Embrace Model Heterogeneity. arXiv:2502.08788.

[45] Wanawana R, Palade V et al (2006) Multi-Classifier Systems: Review and a roadmap for developers. Int J Hybrid Intell Syst 3(1):35-61.

[46] Feng Y, Liang S, Zhang Y (2022) Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. J Am Med Inform Assn 29(8):1425-1432.

[47] de la Torre-López J, Ramírez A, Romero J.R (2023) Artificial intelligence to automate the systematic review of scientific literature. Computing 105:2171-2194

[48] Johnson EE, O'Keefe H, Sutton A et al (2022) The Systematic Review Toolbox: keeping up to date with tools to support evidence synthesis. Syst Rev:11:258.

[49] Clark J, Glasziou P, Del Mar C et al (2020) A full systematic review was completed in 2 weeks using automation tools: A case study. J Clin Epidemiol 121:81-90.

[50] Madaan A, Tandon N, Gupta P et al (2023). Self-refine: Iterative refinement with self-feedback. In: Oh A, Naumann T, Globerson A et al (eds) Advances in Neural Information Processing Systems, vol 36, Curran Associates, Inc., p. 46534-46594.

[51] Yin Z, Sun Q, Chang C et al (2023) Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), Singapore, 6-10 December 2023.

[52] Chan CM, Chen W, Su Y (2024) Chateval: Towards better LLM-based Evaluators through Multi-Agent Debate. In: Proceedings of The Twelfth International Conference on Learning Representations (ICLR), Vienna, Austria, 7-11 May 2024.

[53] Chen X, Yi H, You M et al (2025) Enhancing diagnostic capability with multiagents conversational large language models. NPJ Digital Medicine, 8(1):159.

[54] Liang T, He Z, Jiao W (2024) Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA, 12-16 November 2024.

[55] Zhang Z, Nezhad MJM, Gupta P et al (2025). Enhancing AI for citation screening in literature reviews: Improving accuracy with ensemble models. Int J Med Inform 203:106035.

[56] Guo T, Chen X, Wang Y et al (2024) Large Language Model Based Multi-agents: A Survey of Progress and Challenges. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI), Jeju, South Korea, 3-9 August 2024.

[57] Lu J, Pang Z, Xiao M et al (2024) Merge, Ensemble, and Cooperate! A Survey on Collaborative Strategies in the Era of Large Language Models. arXiv:2407.06089.

[58] Chen Z, Li J, Chen P et al (2025) Harnessing Multiple Large Language Models: A Survey on LLM Ensemble. arXiv:2502.18036.

[59] Pangakis N, Wolken S (2025) Keeping Humans in the Loop: Human-Centered Automated Annotation with Generative AI. In: Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (WSDM), Hannover, Germany, 10-14 March 2025.

[60] Schroeder NL, Jaldi CD, Zhang S (2025). Large language models with human-in-the-loop validation for systematic review data extraction. arXiv: 2501.11840.

[61] Schroeder H, Roy D, Kabbara J (2025) Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks. In: Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July-1 Aug 2025.

[62] Drori I, Te'eni D (2024) Human-in-the-loop AI Reviewing: Feasibility, Opportunities, and Risks. J Assoc Inf Syst 25(1):98-109.

[63] Chiang CH, Lee HY (2023) Can large language models be an alternative to human evaluations? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers), Toronto, Canada, July 2023.

[64] Li D, Jiang B, Huang L et al (2025) From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), Suzhou, China, 4-9 November 2025.

[65] Li H, Dong Q, Chen J et al (2024) LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. arXiv:2412.05579.

[66] Gu J, Jiang X, Shi Z et al (2025) A Survey on LLM-as-a-Judge. arXiv:2411.15594.

[67] Panickssery A, Bowman SR, Feng S (2024) LLM Evaluators Recognize and Favor Their Own Generations. In: A Globerson, L Mackey, D Belgrave et al (eds) Advances in Neural Information Processing Systems, vol 37, Curran Associates, Inc., p. 68772-68802.