Artificial Intelligence Review (2024) 57:259
https://doi.org/10.1007/s10462-024-10902-3

Artificial intelligence for literature reviews: opportunities and challenges

Francisco Bolaños[1] · Angelo Salatino[1] · Francesco Osborne[1,2] · Enrico Motta[1]

Accepted: 6 August 2024 / Published online: 17 August 2024

Abstract
This paper presents a comprehensive review of the use of Artificial Intelligence (AI) in Systematic Literature Reviews (SLRs). A SLR is a rigorous and organised methodology that assesses and integrates prior research on a given topic. Numerous tools have been developed to assist and partially automate the SLR process. The increasing role of AI in this field shows great potential in providing more effective support for researchers, moving towards the semi-automatic creation of literature reviews. Our study focuses on how AI techniques are applied in the semi-automation of SLRs, specifically in the screening and extraction phases. We examine 21 leading SLR tools using a framework that combines 23 traditional features with 11 AI features. We also analyse 11 recent tools that leverage large language models for searching the literature and assisting academic writing. Finally, the paper discusses current trends in the field, outlines key research challenges, and suggests directions for future research. We highlight three primary research challenges: integrating advanced AI solutions, such as large language models and knowledge graphs, improving usability, and developing a standardised evaluation framework. We also propose best practices to ensure more robust evaluations in terms of performance, usability, and transparency. Overall, this review offers a detailed overview of AI-enhanced SLR tools for researchers and practitioners, providing a foundation for the development of next-generation AI solutions in this field.

Keywords Systematic literature reviews · Literature review · Artificial intelligence · Large language models · Natural anguage processing · Usability · Evaluation framework

✉ Francisco Bolaños
francisco.bolanos-burgos@open.ac.uk

Angelo Salatino
angelo.salatino@open.ac.uk

Francesco Osborne
francesco.osborne@open.ac.uk

Enrico Motta
enrico.motta@open.ac.uk

[1] Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes MK7 6AA, UK

[2] Department of Business and Law, University of Milano Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan 20126, Italy

1 Introduction

A Systematic Literature Review (SLR) is a rigorous and organised methodology that assesses and integrates previous research on a specific topic. Its main goal is to meticulously identify and appraise all the relevant literature related to a specific research question, adhering to strict protocols to minimise biases (Higgins 2011; Moher et al. 2009). This methodology originally emerged within the realm of Evidence-Based Medicine Sackett et al. (1996), and it was subsequently adapted and employed in diverse research disciplines including social sciences Petticrew and Roberts (2008), engineering and technology Keele et al. (2007), education Gough et al. (2017), environmental sciences Pullin and Stewart (2006), and business and management Tranfield et al. (2003).

SLRs are recognised for being time-consuming and resource-intensive. This is due to several factors, including the lengthy process that can extend beyond a year Borah et al. (2017), the necessity of assembling a team of domain experts Shojania et al. (2007), significant financial implications from database subscriptions, specialised software, and personnel remuneration Shemilt et al. (2016), the growing number of publications Bornmann and Mutz (2015), and the periodic need for updates to maintain relevance Moher et al. (2007).

Over the past decades, numerous tools have been developed to support and even partially automate SLRs, aiming to address these challenges. Many of these tools have adopted Artificial Intelligence (AI) solutions (van den Bulk et al. 2022; Kebede et al. 2023), particularly for the screening and data extraction phases. The incorporation of AI into SLR tools has been further propelled by the emergence of more sophisticated AI techniques in Natural Language Processing (NLP), such as Large Language Models (LLMs), which have the potential to revolutionise these systems Robinson et al. (2023). While a significant body of research has examined SLR tools (Carver et al. 2013; Feng et al. 2017; Napoleão et al. 2021; Cierco Jimenez et al. 2022; Jesso et al. 2022; Khalil et al. 2022; Wagner et al. 2022; Ng et al. 2023; Robledo et al. 2023), relatively few studies have explored the role of AI in this domain (Cowie et al. 2022; Burgard and Bittermann 2023; de la Torre-López et al. 2023; Robledo et al. 2023). Furthermore, these studies focused on a limited selection of AI features, as we will discuss in Sect. 4.

This survey aims to address the existing gap by rigorously examining the application of AI techniques in the semi-automation of SLRs, within the two main stages of application, namely *screening* and *extraction*. For this purpose, we first conducted an analysis of eight prior surveys and identified the most prominent features examined in the literature. Next, we defined a framework of analysis that integrates 23 general features and 11 features pertinent to AI-based functionalities. We then selected 21 prominent SLR tools and subjected them to rigorous analysis using the resulting framework. We extensively discuss current trends, key research challenges, and directions for future research. We specifically focus on three major research challenges: (1) integrating advanced AI solutions, such as large language models and knowledge graphs, (2) enhancing usability, and (3) developing a standardised evaluation framework. We also propose a set of best practices to ensure more robust evaluations regarding performance, usability, and transparency. Finally, we performed an additional analysis on 11 recent tools that utilise the capabilities of LLMs (predominantly ChatGPT via the OpenAI API) for searching the literature and aiding academic writing. Although these tools do not cater directly to SLRs, there is potential for their features to be integrated into future SLR tools. In conclusion, this survey seeks to offer scholars a thorough insight into the application of Artificial Intelligence in this field, while also highlighting potential avenues for future research.

The remainder of this paper is structured as follows. Sect. 2 includes a description of the SLR stages and their relationship with AI. Section 3 outlines the methodology we employed to identify the SLR tools discussed in the survey. Section 4 provides a meta-review of previous surveys about SLR tools that analysed AI features. Section 5 provides an in-depth examination of the 21 tools. Section 6 discusses the key research challenges and proposes some best practices for the evaluation of AI-enhanced SLR tools. Section 7 analyses the latest generation of LLM-based systems designed to assist researchers. Finally, Sect. 8 concludes the paper by summarising the contributions and the main findings.

2 Background

In this section, we examine the various stages of a SLR and the extent of support they receive from AI in the current generation of tools. Here, the term 'AI' specifically denotes weak or narrow AI, which includes systems designed and trained for specific tasks like classification, clustering, or named-entity recognition Iansiti and Lakhani (2020). In the context of SLR, these methodologies are predominantly utilised to semi-automate tasks like screening and data extraction (Cowie et al. 2022; Burgard and Bittermann 2023).

The SLR methodology consists of six distinct stages (Keele et al. 2007; Higgins 2011): (i) Planning, (ii) Search, (iii) Screening, (iv) Data Extraction and Synthesis, (v) Quality Assessment, and (vi) Reporting. Each stage plays a pivotal role in ensuring the comprehensiveness and rigour of the review process.

The *planning* phase is foundational to the entire review process, as it involves formulating a set of precise and specific research questions that the SLR seeks to address O'Connor et al. (2008). A detailed protocol is also developed during this stage, outlining the appropriate methodologies that will be adopted to carry out the review Fontaine et al. (2022). This protocol ensures consistency, reduces bias, and enhances the transparency and reproducibility of the review.

The *search* phase aims to identify relevant papers using search strategies, snowballing, or a hybrid approach. Search strategies are typically implemented by creating a query based on a combination of terms using boolean operators (Team 2007; Glanville et al. 2019). This query is then executed on designated search engines. In snowballing, the researcher examines the references and citations of an initial group of papers (also known as seed papers) to identify additional articles. This process is iteratively repeated until no new relevant scholarly documents are found (Webster and Watson 2002; Wohlin 2014). The hybrid approach is the combination of search strategy and snowballing (Mourão et al. 2017; Wohlin et al. 2022). Traditionally, the search phase had not been significantly supported by artificial intelligence techniques Adam et al. (2022). Nevertheless, there are some emerging tools, which we will examine in Sect. 7, that have begun to incorporate LLMs in academic search engines, often within a Retrieval-Augmented Generation (RAG) framework Lewis et al. (2020). This innovative approach allows for the formulation of precise questions and complex queries in natural language, surpassing the capabilities of traditional keyword-based searches.

The *screening* phase uses a set of inclusion and exclusion criteria to further filter the paper obtained from the search stage. It typically consists of two stages: (i) title and abstract screening and (ii) full-text screening. In the first step, the reviewers screen the relevant papers according only to the title and abstract Moher et al. (2009). The second step entails a detailed evaluation of the content of each paper, a task that demands significantly

more effort but leads to a more thorough assessment. It is also customary to document the rationale for excluding any given paper during this process. The predominant application of AI in SLR regards this phase. It usually involves employing machine learning classifiers, which are trained on an initial set of user-selected papers and then used to identify additional relevant articles Miwa et al. (2014). This process frequently involves iteration, where the user refines the automatic classifications or selects new papers, followed by retraining the classifier to better identify further pertinent literature.

In the *data extraction and synthesis* phase, all the pertinent information is systematically extracted from the selected studies. The techniques for data extraction vary greatly depending on the research field and the objective of the researcher. For example, in the biomedical field, protocols like PECODR Dawes et al. (2007) (Patient-Population-Problem, Exposure-Intervention, Comparison, Outcome, Duration, and Results) and PIBOSO Kim et al. (2011) (Population, Intervention, Background, Outcome, Study Design, and Other) are used to identify key elements from clinical studies, while the STARD checklist Bossuyt et al. (2003) supports readers in assessing the risk of bias and evaluating the relevance of the results. Following the extraction, the data is aggregated and summarised (Munn et al. 2014; Garousi and Felderer 2017). Depending on the nature and heterogeneity of the data, the resulting synthesis might be qualitative or quantitative. This phase is also occasionally supported by AI solutions. Commonly, the relevant tools employ classifiers to identify articles possessing specific characteristics Marshall et al. (2018) or implement named-entity recognition for extracting specific entities or concepts Kiritchenko et al. (2010) (e.g., RCT entities Moher et al. (2001), entities pertaining environmental health studies Walker et al. (2022)).

The *quality assessment* phase evaluates the rigour and validity of the selected studies (Project 1998; Wells et al. 2000; Von Elm et al. 2007; Higgins and Altman 2008). This analysis provides evidence of the overall strength and the level of trustworthiness presented in the review (Zhou et al. 2015; Chen et al. 2022).

Finally, the *reporting* phase involves presenting the findings in a structured and coherent manner within a research paper. This presentation typically follows an established format comprising sections like introduction, methods, results, and discussion, but this may differ depending on the journal in which the manuscript will be published (Stroup et al. 2000; Page et al. 2021). Historically, this stage did not benefit from the use of artificial intelligence techniques (Justitia and Wang 2022; Li and Ouyang 2022). However, as we will discuss in Sect. 7, recent advancements have led to the development of tools based on LLMs designed to support academic writing, which can be particularly useful in this phase. These tools typically enable users to draft an initial outline of the desired document and iteratively refine it.

3 Methodology

We adopt the standard PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology Page et al. (2021) for conducting and reporting the systematic review and the meta-analysis. The PRISMA checklist is linked in the supplementary material at the end of the paper.

The primary objective of our analysis was to examine the application of Artificial Intelligence in the current generation of SLR tools to identify trends and emerging research directions. In order to identify the set of AI-enhanced SLR tools, we first formulated three

inclusion criteria and two exclusion criteria. Specifically, the inclusion criteria are the following:

IC 1: The SLR tool must incorporate AI techniques to semi-automate the screening or extraction process, while still maintaining the user's capacity to make the final decision Tsafnat et al. (2014);
IC 2: The tool must possess a user interface that facilitates paper screening or information extraction by the user;
IC 3: The tool should not require advanced technical expertise for installation and execution.

The exclusion criteria are:

EC 1: The tool is under maintenance;
EC 2: The tool has not been updated in the last 10 years.

The PRISMA diagram, depicted in Fig. 1 illustrates the main phases of the process. We utilised three main strategies for identifying the tools.

First, we conducted a search of previous survey papers on SLR tools and extracted the tools they analysed. This was accomplished using Scopus,[1] a leading bibliographic database Pranckutė (2021). We selected Scopus over other alternatives because it is widely recognised as the preferred source for conducting systematic literature reviews due to its high-quality metadata, reliable citation tracking, and extensive coverage of scientific documents, including journals, conference proceedings, and books (Baas et al. 2020; Visser et al. 2021). Specifically, we used the search string: *("Literature Reviews" OR "Systematic Review") AND ("Tools" OR "Automation" OR "Semi-Automation" OR "Semiautomation" OR "Software")*.[2] Since this field lacks standardised vocabulary (Dieste et al. 2009; Grant and Booth 2009), we aimed to maximise recall by using broad terms, planning to refine the results at a later stage. Additionally, we filtered the results by selecting only 'review' as the 'Document type' in the Scopus interface.

The search yielded 356 review papers. From this set, we identified the surveys focusing on SLR tools. This selection process was conducted in two stages. Initially, the first author, who has eight years of experience in teaching evidence-based medicine and bibliometric analysis, identified a preliminary set of 14 papers based on their titles and abstracts. Subsequently, all four authors collaboratively examined the shortlisted papers by reviewing the full texts. Potentially ambiguous cases were discussed among the authors to achieve consensus. This process yields five survey papers. We then applied a snowballing search Webster and Watson (2002) to identify additional surveys. This involved examining the references and citations of the five survey papers. As before, we implemented a two-stage selection process, screening titles and abstracts first, and followed by a full-text analysis of potentially relevant papers. This resulted in the inclusion of three additional papers. Overall, this procedure yielded a total of 8 survey papers. An analysis of these surveys led to the identification of 23 tools. Among these, 17 were associated with academic papers, whereas 6 were not.

[1] Scopus - https://www.scopus.com/

[2] The reader may notice that this query retrieves also documents about "Systematic Literature Reviews".

As a second source, we adopted the Systematic Literature Review Toolbox Marshall and Brereton (2015), a repository in which SLR tools are published and updated. This platform is highly regarded in the field and was adopted as a source in five of the eight previous surveys (Kohl et al. 2018; Van der Mierden et al. 2019; Harrison et al. 2020; Cowie et al. 2022; Robledo et al. 2023). Specifically, we utilised the advanced search functionality to retrieve all tools under the "software" category. The query returned 236 tools that were manually analysed. Similar to the analysis of the papers, this process was conducted in two phases. Initially, the first author selected a preliminary set of 45 tools based solely on their descriptions in the repository. Next, all authors evaluated these tools by examining their websites, tutorials, and the tools themselves. When necessary, the first author collected additional information through interviews with the developers. As before, any ambiguous cases were discussed among the authors until a consensus was reached. The analysis identified a total of 21 tools. Of these, 16 were associated with academic papers, while 5 were not.

As a third source, we adopted the Comprehensive R Archive Network (CRAN),[3] a well-regarded repository for R packages widely used by the statistical and data science communities. Specifically, we employed the *packagefinder* library Zuckarelli (2023) and used the query: *("systematic literature review" OR "systematic review" OR "literature review")*. We chose this library due to its proven effectiveness in retrieving relevant applications in various domains, such as ecology and evolution Lortie et al. (2020). This search yielded 329 tools, which we then evaluated using the same two-stage selection process previously applied for the tools produced by searching the SLR Toolbox. However, none of these tools incorporated AI solutions while also providing a visual interface. Consequently, we discarded all of them.

After deduplicating the results from the SLR toolbox and previous surveys, we identified 17 tools linked to research papers and 8 tools without associated papers. To validate and expand our findings, we conducted a snowballing search using the 17 papers linked to the tools as the initial seeds. Our aim was to identify additional papers associated with a relevant tool. This search was conducted on Semantic Scholar,[4] chosen for its extensive coverage in Computer Science, especially for snowballing searches Hannousse (2021). To facilitate this, we developed a custom script to interface with the Semantic Scholar API, enabling the efficient retrieval of references from seed papers and the papers citing them.[5] This process led to the identification of 584 references and 8,009 papers citing the seeds for a total of 8,593 papers. After removing duplicates, 7,304 papers remained. The authors analysed these papers using the same two-stage selection process that was employed for survey papers. This analysis yielded 15 of the papers included in the initial seeds as well as the 8 previously identified surveys, but it did not reveal any new tools or papers. The lack of new findings, despite the comprehensive snowballing process, suggests that the tool section identified earlier is exhaustive.

In conclusion, the process led to a collection of 17 tools with associated papers and 8 without (Covidence, DistillerSR, Nested Knowledge, Pitts.ai, Iris.ai, LaserAI, DRAGON/litstream, Giotto Compliance). We then excluded Giotto Compliance, DRAGON/litstream, and LaserAI from our study due to the lack of available information.[6] We also consolidated

[3] CRAN repository - https://cran.r-project.org/

[4] Semantic Scholar - https://www.semanticscholar.org/

[5] Code used for the Snowballing search - https://zenodo.org/records/11154875

[6] Specifically, there were no associated research papers or comprehensive documentation for these tools, and attempts to contact the developers for further details were unsuccessful.

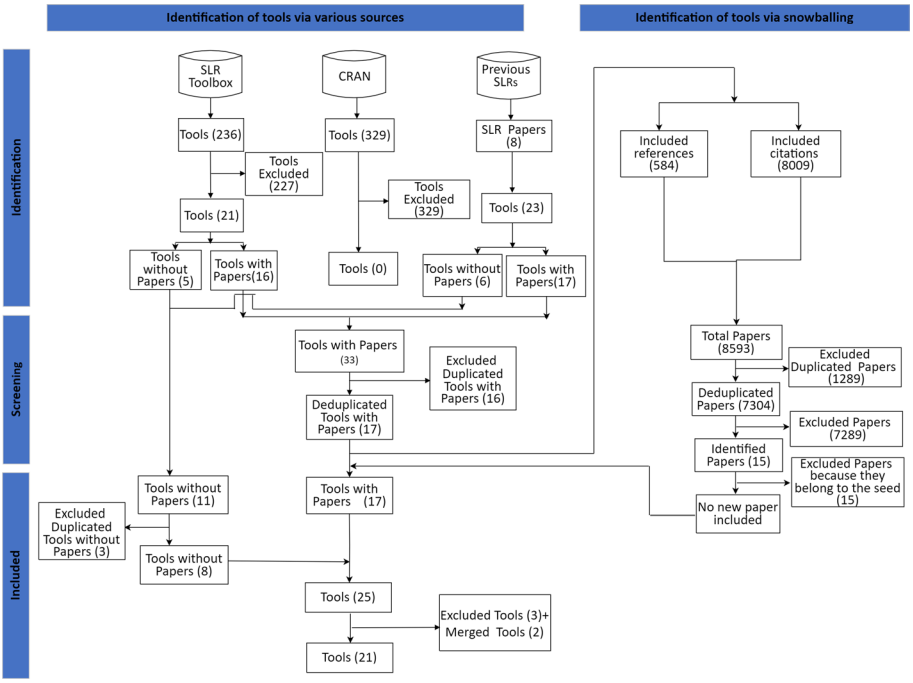


Fig. 1 PRISMA diagram of our SLR about AI-enhanced SLR tools

RobotReviewer and RobotSearch into a single entry, recognising their shared algorithmic basis. Consequently, the final set of tools considered in our study included 21 distinct tools.

Table 1 lists the 21 selected tools. The majority of them (17) were primarily designed for the purpose of screening. Two tools, namely Dextr and ExaCT, focused exclusively on extraction. The remaining two tools, Iris.ai and RobotReviewer/RobotSearch, had a dual focus on both screening and extraction. In summary, 19 tools can be used for the screening phase and 4 tools for the extraction phase. The majority of the tools (19) are web applications, while only 2, namely SWIFT-Review and ASReview, need to be installed locally. Furthermore, only 4 tools (Colandr, ASReview, FAST2, and RobotReviewer) release their code under an open license.

4 Meta-review of previous surveys

This section provides a brief meta-analysis of how the previous systematic literature reviews have described the tools in relation to Artificial Intelligence. We focus on four surveys that analysed AI features (Cowie et al. 2022; Burgard and Bittermann 2023; de la Torre-López et al. 2023; Robledo et al. 2023).

The previous survey papers characterised AI according to five main features:

1. **Approach:** dentifies the method used for performing a specific task. This is the most examined feature, receiving attention from four studies (Cowie et al. 2022; Burgard and Bittermann 2023; de la Torre-López et al. 2023; Robledo et al. 2023).

Table 1 The 21 SLR tools analysed in this survey

ID	Tool	Stage SLR	Mode	OS	References
1	Abstrackr	Screening	Web	No	Wallace et al. (2012)
2	ASReview	Screening	Desktop	Yes	Van De Schoot et al. (2021)
3	Colandr	Screening	Web	Yes	Cheng et al. (2018), Cheng and Augustin (2021)
4	Covidence	Screening	Web	No	–
5	DistillerSR	Screening	Web	No	–
6	EPPI-Reviewer	Screening	Web	No	Thomas et al. (2010), Machine Learning Functionality in EPPI-Reviewer (2019)
7	FAST2	Screening	Web	Yes	Yu and Menzies (2019)
8	LitSuggest	Screening	Web	No	Allot et al. (2021)
9	Nested Knowledge	Screening	Web	No	–
10	PICOPortal	Screening	Web	No	Agai (2020), Minion et al. (2021)
11	Pitts.ai	Screening	Web	No	–
12	Rayyan	Screening	Web	No	Ouzzani et al. (2016)
13	Research Screener	Screening	Web	No	Chai et al. (2021)
14	RobotAnalyst	Screening	Web	No	Przybyła et al. (2018)
15	SWIFT-Active Screener	Screening	Web	No	Howard et al. (2020)
16	SWIFT-Review	Screening	Desktop	No	Howard et al. (2016)
17	SysRev	Screening	Web	No	Bozada et al. (2021)
18	Dextr	Extraction	Web	No	Walker et al. (2022)
19	ExaCT	Extraction	Web	No	Kiritchenko et al. (2010)
20	Iris.ai	Both	Web	No	–
21	RobotReviewer/RobotSearch	Both	Web	Yes	Marshall et al. (2017, 2018)

OS open scource

The features covered by all tools (100%) are highlighted in bold

2. **Text representation**: escribes the processes employed to convert text into suitable input for the algorithm (e.g., BoW Zhang et al. (2010), LDA topics Blei et al. (2003), word embeddings Wang et al. (2020)). This feature was analysed by two previous surveys (de la Torre-López et al. 2023; Burgard and Bittermann 2023).
3. **Human interaction**: pecifies how users engage with a tool, detailing the operations and options available to them, as well as the characteristics of the user interface. This is among the least explored features with just one previous study de la Torre-López et al. (2023).
4. **Input**: pecifies the type of content full-text or just title and abstract) the tool will need to train its model. Alongside *Human interaction*, this is the least explored feature, with just one previous study de la Torre-López et al. (2023).
5. **Output:** epresents the outcome generated by the trained algorithm, and it has been analysed in three studies (Burgard and Bittermann 2023; de la Torre-López et al. 2023; Robledo et al. 2023).

Table 2 Analysis of SLR tools based on AI features, as conducted by previous surveys (Cowie et al. 2022; Burgard and Bittermann 2023; de la Torre-López et al. 2023; Robledo et al. 2023)

ID	Tool	Approach	Text representation	Human interaction	Input	Output	Papers
1	Abstrackr	Y	Y	Y	Y	Y	Cowie et al. (2022), Burgard and Bittermann (2023), de la Torre-López et al. (2023)
2	ASReview	Y	Y	N	N	Y	Burgard and Bittermann (2023), Robledo et al. (2023)
3	Colandr	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023)
4	Covidence	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023)
5	DistillerSR	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023)
6	EPPI-Reviewer	Y	Y	Y	Y	Y	Cowie et al. (2022), Burgard and Bittermann (2023), de la Torre-López et al. (2023)
7	FASTREAD	Y	Y	Y	Y	Y	Burgard and Bittermann (2023), de la Torre-López et al. (2023)
8	Giotto Compliance	Y	N	N	N	N	Cowie et al. (2022)
9	Nested Knowledge	Y	N	N	N	N	Cowie et al. (2022)
10	PICOPortal	Y	N	N	N	N	Cowie et al. (2022)
11	Rayyan	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023), Robledo et al. (2023)
12	Research Screener	Y	Y	N	N	Y	Burgard and Bittermann (2023)
13	RobotAnalyst	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023)
14	RobotReviewer/RobotSearch	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023)
15	SWIFT-Active Screener	Y	Y	N	N	Y	Cowie et al. (2022), Burgard and Bittermann (2023)
16	SWIFT-Review	Y	Y	N	N	Y	Burgard and Bittermann (2023), Robledo et al. (2023)
17	SysRev	Y	N	N	N	N	Cowie et al. (2022)

The tools are listed in alphabetical order, with the reviews conducting the analysis cited in the final column. Y = Yes, N=No

Table 2 summarises the analysis of the four systematic literature reviews and shows how 17 tools have been reviewed according to the five AI features. Only three tools (FASTRED, EPPI-Reviewer, and Abstractr) were actually assessed according to all five AI features. For ten tools (ASReview, Colandr, Covidence, DistillerSR, Rayyan, Research Screener, Robot-Analyst, RobotReviewer/RobotSearch, SWIFT-Active Screener, and SWIFT-Review) only three features named *approach*, *text representation*, and *output* have been assessed. The remaining tools were assessed by using only one feature (*approach*).

Upon examining the four survey papers, it is apparent that there is a limited exploration of AI features. De la Torre-López et al. de la Torre-López et al. (2023) provide the most comprehensive analysis, utilising all five specified features to examine seven tools. In contrast, Burgard and Bittermann (2023) employed only three features: *text representation*, *approach*, and *output*. Robledo et al. (2023) focused on just two features: *approach* and *output*. Cowie's et al. (2022) conducted the most restricted analysis, considering only one feature (*approach*). Furthermore, the five reported features only offer a narrow perspective on how AI can support SLRs.

In summary, the previous systematic reviews offer a relatively limited analysis of the expanding ecosystem of AI-enhanced SLR tools and their characteristics. In the next section, we will address this gap by introducing a comprehensive set of 11 AI features and applying them to evaluate the 21 SLR tools identified in Sect. 3.

5 Survey of SLR tools

We analysed the 21 SLR tools according to 34 features (11 AI-specific and 23 general) by examining the relevant literature (see Table 1), their official websites, and the online tutorials. When necessary, we sought additional information by reaching out to the developers through email or online interviews.

Section 5.1 describes the full set of features, paying particular attention to the new AI features that we first introduced for this survey. Section 5.2 reports the results of the review. Section 5.3 discusses the most suitable systems for specific use cases. Finally, Sect. 5.4 outlines the threats to validity of our analysis.

5.1 Features overview

5.1.1 AI features

To analyse the extent of AI usage within SLR tools, we considered a total of eleven features. This evaluation included the five features previously described in Sect. 4 (approach, text representation, human interaction, input, and output) along with six new features unique to this study. These additional features were identified through a review of the relevant literature (Cowie et al. 2022; Burgard and Bittermann 2023; de la Torre-López et al. 2023; Robledo et al. 2023) and a preliminary analysis of the tools. The six novel features are as follows:

- **SLR Task:** hich categorises the tasks for which the AI approach is used (e.g., paper classification, paper clustering, named-entity recognition);

- **Minimum requirement**: hich refers to the minimum number of relevant and irrelevant papers required to effectively train a classifier tasked with selecting pertinent papers;
- **Model execution**: which evaluates whether the models operate in real time (synchronously) or later, typically overnight (asynchronously);
- **Research field**: hich identifies the research domains in which the tools can be effectively employed;
- **Pre-screening support**: which specifies the application of AI techniques to assist users in manually selecting relevant papers, typically by highlighting key terms or grouping similar papers (e.g., topic maps Howard et al. (2020) based on LDA Blei et al. (2003), clustering approaches Przybyła et al. (2018));
- **Post-screening support**: which refers to the application of AI techniques to conduct a final review of the screened papers (e.g., summarisation Shah and Phadnis (2022)).

Five of the eleven features (minimum requirement, model execution, human interaction, pre-screening support, and post-screening support) are exclusive to the screening phase and will not be considered when analysing the extraction phase.

5.1.2 General features

We analysed the non-AI characteristics of SLR tool based on 23 features. We derived these features from previous studies (Marshall et al. 2014; Kohl et al. 2018; Van der Mierden et al. 2019; Harrison et al. 2020; Cowie et al. 2022) after a process of synthesis and integration. Table 3 shows the description of each feature with its category. To facilitate the systematic analysis, we grouped them into six categories: Functionality (F), Retrieval (R), Discovery (Di), Documentation (Do), Living Review (L), and Economic (E).

The *functionality* category includes features for auditing and evaluating the technical aspects of the tools. The *retrieval* category covers features related to the acquisition and inclusion of scholarly documents. The *discovery* category consists of features that facilitate the inclusion, exclusion, and management of references during the screening phase. The *documentation* category includes features that support the reporting of the findings. The *living review* category captures the ability of tools to incorporate new relevant documents based on AI techniques. Lastly, the *economic* category reflects the financial considerations associated with the tools.

5.2 Results

In this section, we present the results of our analysis. Section 5.2.1 describes the tools for the screening phase through the AI features. Section 5.2.2 presents the tools for the extraction phase also through the AI features. Finally, Sect. 5.2.3 describes the full set of 21 SLR tools according to the general features.

5.2.1 The role of AI in the screening phase

As reported in Table 1, 19 tools use AI for the screening phase. In the following, we analyse them according to the 11 AI features. To eliminate repetitions, this discussion combines the features *input* and *text representation* into the category *Input Data and Text Representation*. Moreover, the *output* feature is discussed within the context of the *SLR Task*,

Table 3 Description of the SLR features

#	SLR feature	Description
1	Authentication (F)	Ability of the tool to authenticate the users involved in the project
2	Multiplatform (F)	Ability of the tool to be run on different platforms (e.g., web, desktop)
3	Multiple user roles (F)	Ability of the tool to allow the user to have different roles (e.g., reviewer, admin) within and between projects
4	Multiple user support (F)	Ability of the tool to allow multiple users to work on the same project
5	Project auditing (F)	Ability of the tool to track all the changes done in the project
6	Project progress (F)	Ability of the tool to determine the overall progress of the annotation with respect to the total number of papers to annotate
7	Status of the software (F)	The extent to which the tool is actively maintained and has a stable release
8	Automated full-text retrieval (R)	Ability of the tool to support full-text retrieval from bibliographic databases
9	Automated search (R)	Ability of the tool to support literature search through the integration of APIs
10	Manual reference importing (R)	Ability of the tool to allow the user to enter papers manually, typically via a form
11	Manually inserting full-text (R)	Ability of the tool to allow the user to manually add full-text papers
12	Reference importing (R)	Ability of the tool to import papers using a variety of formats (BibTeX, RIS, CSV)
13	Snowballing (R)	Ability of the tool to support the automated retrieval of the citations from bibliographic databases (snowballing)
14	Deduplication (Di)	Ability of the tool to support the automatic deduplication of the references
15	Discrepancy resolving (Di)	Ability of the tool to handle differences of opinion between screeners, e.g., by allowing comments or assigning the problematic papers to a senior screener
16	In-/excluding references (Di)	Ability of the tool to allow the user to comment on reference inclusion and exclusion
17	Reference labelling & comments (Di)	Ability of the tool to allow the user to write additional comments on the references (e.g., 'to double check')
18	Screening phases (Di)	Ability of the tool to allow the user to perform the different stages screening phase
19	Exporting results (Do)	Ability of the tool to allow exporting the screened references
20	Flow diagram creation (Do)	Ability of the tool to provide the PRISMA diagram of the SLR process
21	Protocol (Do)	Ability of the tool to provide the user with pre-defined protocol templates (e.g., the Cochrane guidelines)
22	Living/updatable (L)	Ability of the tool to update the screened references by automatically including recent and relevant articles
23	Free to use (E)	It determines whether the tool is available for free or requires payment

as it is contingent upon the specific task requirements. The Table in Appendix (Table 6) reports detailed information on how each of the 19 tools addresses the 11 AI features.

Research field: Twelve tools utilise general AI solutions that are applicable across various research fields. The other seven tools employ specific AI solutions designed to support biomedical studies, typically by identifying Randomised Controlled Trials (RCTs) through the use of dedicated classifiers Noel-Storr et al. (2021). Notably, EPPI-Reviewer, PICOPortal, and Covidence offer both a general mode and a specialised setting for biomedical studies. Conversely, Pitts.ai, RobotReviewer/RobotSearch, SWIFT-Review, and LitSuggest are exclusively dedicated to the biomedical domain.

SLR task: Fifteen tools utilise artificial intelligence for only one task, most often to classify papers as relevant/irrelevant. The other four tools (Covidence, PICOPortal, and EPPI-Reviewer, Colandr) undertake two AI-related tasks. They all classify papers as relevant/irrelevant, but also execute an additional task, such as identifying a specific type of paper (e.g., economic evaluation, randomised controlled trials, etc.) or categorising papers according to a set of entities defined by the user. For the sake of clarity, in our discussion of the subsequent features, we will systematically address the first group (one task) followed by the second group (two tasks). In the first group, twelve tools focus on selecting relevant papers given a set of seed papers. Typically, each paper is assigned an inclusion probability score, usually ranging from 0 to 1. Of the remaining three, two of them (Pitts.ai and RobotReviewer/RobotSearch) identify RCTs based on a pre-built classification model, while the third (Iris.ai) clusters similar papers to build topic maps that assist users in selecting the relevant papers. In the second group, all four systems classify pertinent papers using a set of seed papers as a reference. However, they vary in their secondary AI-driven tasks. Specifically, Covidence and PICOPortal identify RCTs using a predefined classification model. EPPI-Reviewer can identify various types of studies, including RCTs, systematic reviews, economic evaluations, and COVID-19 related studies. Finally, Colandr, in addition to the standard identification of relevant papers, enables users to define their own set of categories (e.g., "water management") and subsequently performs a multi-label classification of articles based on them Cheng et al. (2018). It also maps individual sentences to the user-defined categories and provides a confidence score for each classification.

AI approach: In the group of tools performing one task, the twelve tools focused exclusively on categorising relevant papers employed various types of machine learning classifiers. The most adopted approach is Support Vector Machine (SVM) Hearst et al. (1998), which aligns with the findings of prior studies Schmidt et al. (2021). Four of the tools (Abstractr, FAST2, Rayyan, RobotAnalyst) exclusively rely on SVM. Distiller supports both SVM and Naive Bayes. ASReview allows the user to select a vast range of methods, including Logistic Regression, Random Forest, Naive Bayes, SVM, and a Neural Networks classifier. Litsuggest use logistic regression, while SWIFT-Review and SWIFT-Active Screener use a method based on log-lineal regression. Pitts.ai and RobotReviewer/RobotSearch also use an SVM for identifying RCTs Marshall et al. (2018). Finally, Iris.ai identifies and groups similar papers based on the similarity of their 'fingerprint', a vector representation of the most meaningful words and their synonyms extracted from the abstract Wu et al. (2018).

With regards to the four tools that perform two AI tasks, Covidence, EPPI-Reviewer, and PICOPortal also identify relevant papers by using a SVM classifier. In contrast, Colandr employs a method where it identifies papers by searching for keywords that are related to a set of user-defined search terms Cheng et al. (2018). For instance, it can recognise terms commonly associated with 'Artificial Intelligence' and select papers containing these

terms. Covidence also implements a machine learning classifier based on SVM with a fixed threshold for the identification of RCTs following the Cochrane guidelines Thomas et al. (2021). EPPI-Reviewer utilises a range of proprietary classifiers trained on various databases to identify papers with distinct characteristics.[7] It uses the Cochrane Randomized Controlled Trial classifier Thomas et al. (2021) to identify RCTs. It employs a classifier trained with the NHS Economic Evaluation Database (NHS EED) Craig and Rice (2007) for identifying economic evaluations and another trained on the Database of Abstracts of Reviews of Effects La Toile (2004) to identify systematic reviews in the biomedical field. Finally, it uses a classifier trained on the 'Surveillance and disease data on COVID-19'[8] for identifying research related to COVID. PICOPortal employs instead an ensemble of machine learning classifiers, which combines both decision trees and neural networks Onan et al. (2016). Finally, for the identification of the category attributed to the paper by the user, Colandr used a combination of Named Entity Recognition for extracting entities relevant to the categories and a classifier based on logistic regression Cheng and Augustin (2021).

Input data and text representation: The AI techniques employed by these tools take as input the title, abstract, or full text of papers. All the tools analysed need only titles and abstracts as input, with the exception of Colandr, which requires the full text of papers. The tools generate different representations of the papers to input into the AI models. Specifically, of the 15 tools dedicated to classifying relevant papers, the majority (8 out of 15) use a Bag of Words (BoW) approach Zhang et al. (2010), while the remainder employ various word embedding techniques Wang et al. (2020). Pitts.ai and RobotReviewer/RobotSearch use SciBERT embeddings Beltagy et al. (2019). Research Screener employs Doc2Vec embeddings Le and Mikolov (2014). ASReview offers multiple options, including Sentence-BERT Reimers and Gurevych (2019) and Doc2Vec Le and Mikolov (2014). Iris utilises a unique representation called fingerprint Wu et al. (2018), which is a vector characterising the most meaningful words and their synonyms extracted from the abstract. In the second group, Covidence and EPPI-Reviewer adopt a BoW representation, while PICOPortal uses both BoW and the BioBERT embeddings Lee et al. (2020). Finally, Colandr uses both word2vec Mikolov et al. (2013) and GloVe Pennington et al. (2014) embeddings. Overall, much like other NLP applications, these tools are evolving from traditional text representations like BoW to a range of more modern word and sentence embeddings.

Human interaction: We identified three main types of interfaces. The *first* and most typical one, implemented by 16 tools, regards the classification of paper as relevant. These graphical interfaces typically feature similar templates that allow users to upload and examine papers. Some tools (Rayyan, SWIFT-Active Screener, SysRev, Covidence, and PICOPortal) also offer a menu with additional functionalities like ranking or filtering papers based on specific criteria. A few tools (Rayyan, SWIFT-Active Screener, Research Screener, Pitts.ai, SysRev, Covidence, PICOPortal) also enable multiple users to collaboratively perform this task, allowing them to add comments for discussion about problematic papers or to delegate challenging papers to a senior reviewer. For illustration, Fig. 2a and b depict the interfaces used by ASReview and RobotAnalyst, respectively, for selecting relevant papers. The ASReview interface enables users to classify papers as either relevant or

[7] EPPI-Reviewer Documentation - https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3772

[8] COVID surveillance - https://tinyurl.com/229vcpyd

irrelevant. In contrast, the RobotAnalyst interface provides options for users to categorise papers as included, excluded, or undecided.

The *second* interface type, offered by Colandr, enables users to define specific categories to assign to the papers. This approach offers greater flexibility compared to the traditional binary classification of relevant or not relevant papers. For instance, in Fig. 3 Colandr suggests that for the given paper, the shown sentences are classified with a confidence level of high, medium or low in the category "land/water management", previously defined by the user. The user can accept, skip or reject the suggested classification.

The *third* interface type, offered by Iris.ai, is based on a *topic map*, a visualisation technique that clusters papers based on thematic similarities. The user initiates the search process with a brief description of the user's search intent (typically 300 to 500 words), a title, or the abstract of a paper. The system then clusters the papers according to their topics and generates a topic map, such as the ones depicted in Fig. 4c. These interactive visualisations enable users to effectively navigate and select papers relevant to their research. The user can iteratively repeat the clustering process until they are satisfied that all pertinent papers have been incorporated into the analysis. Iris.ai also enables users to further filter the papers according to a variety of facets.

Minimum requirements: Generally, the accuracy of a classifier improves with an increasing number of annotated papers, but this also increases the time and effort required from researchers. Most methods need between 1–15 relevant papers and typically the same number of irrelevant ones. This is a relatively low number that should allow researchers to quickly annotate the initial set of seed papers. However, the necessary quantity varies a lot across tools. For instance, ASReview, SWIFT-Active Screener, and SWIFT-Review require just one relevant and one irrelevant paper to begin classification. Covidence and Rayyan require two and five papers, respectively. Other tools require a larger number of papers. For example, Colandr needs 10 seed papers, while SysRev requires 30.

Model execution: Thirteen tools employ a real-time model execution strategy, wherein the training and classification of the model occur immediately after the user selects the relevant and irrelevant paper. Conversely, SysRev and SWIFT-Active Screener adopt a delayed-model-execution approach in which the training and classification steps are conducted at predetermined intervals. Specifically, SysRev executes these operations overnight, whereas SWIFT-Active Screener updates its model after every thirty papers, maintaining a minimum two-minute interval between the most recent and the currently used model.

Pre-screening support: Among the 19 tools evaluated, eight implement standard techniques for pre-screening support, such as keyword search, boolean search, and tag search. ASReview, Covidence, DistillerSR, and SWIFT-Active Screener only enable the user to filter the paper by keyword. Rayyan and EPPI-Reviewer enhance this functionality by highlighting keywords in their visual interface. Additionally, Colandr and Abstrackr offer the feature of colour-coding keywords based on their relevance level. Rayyan incorporates a boolean search feature, allowing users to combine keywords with operators like *AND*, *OR*, and *NOT*. For example, a boolean search such as *"literature review" AND "tools"* will retrieve scholarly documents containing both keywords in their titles or abstracts. Rayyan also provides options to search by author or publication year. EPPI-Reviewer, on the other hand, offers a tag search function, where users can tag papers with specific keywords and then search based on these tags.

RobotAnalyst, SWIFT-Review, and Iris.ai also support topic modelling. The first two use LDA Blei et al. (2003), which probabilistically assigns a topic to a paper based on the most recurrent terms shared by other papers. RobotAnalyst presents the topics

in a network, as shown in Fig. 4a in which each node (circle) represents a topic, and its size is proportional to the frequency of the terms that belong to it. SWIFT-Review uses a simpler approach displaying the topics and their terms in a bar chart, as shown in Fig. 4b. Iris.ai clusters the papers according to a two-level taxonomy of global topics and specific topics. For instance, in Fig. 4c we can observe a set of global topics in the background, which include 'companion', 'labour', 'provider', and 'woman'. Whereas in the cyan section there are the second-level specific topics, in this case concerning the 'labor' global topic, such as 'woman', 'companion', 'market', 'management', and 'care'.

RobotAnalyst offers a cluster-based search functionality. This feature employs a spectral clustering algorithm Ng et al. (2001) to group papers. It also incorporates a statistical selection process for identifying the key terms characterising each cluster Brockmeier et al. (2018). The resulting clusters are presented to the user, emphasising the most representative terms.

Finally, Nested Knowledge, PICOPortal, Rayyan, and RobotReviewer/RobotSearch provide PICO identification, which uses distinct colours to highlight the *patient/population*, *intervention*, *comparison*, and *outcome*. Rayyan also enhances search capabilities by extracting topics and enriching them with the Medical Subject Headings (MeSH) Lipscomb (2000). Furthermore, it enables users to select biomedical keywords and phrases for inclusion or exclusion.

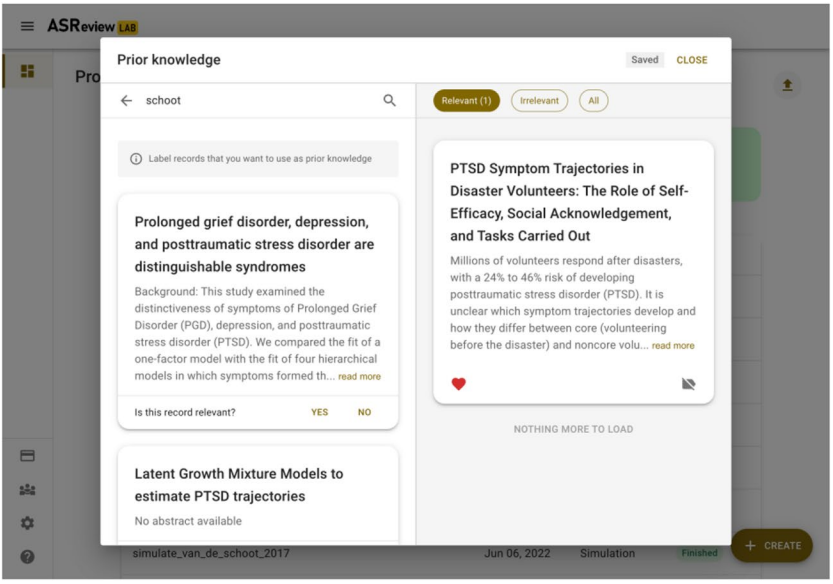
Post-screening support: Only two tools offer support for post-screening: Iris.ai and Nested Knowledge. Specifically, Iris.ai generates summaries from either a single document, multiple abstracts, or multiple documents. It employs an abstractive summarisation technique Shah and Phadnis (2022), where the summary is formed by generating new sentences that encapsulate the core information of the original text. The system also provides users with the flexibility to adjust the length of the summary, ranging from a brief two-sentence overview to a more comprehensive one-page summary. Nested Knowledge allows users to create a hierarchy of user-defined tags that can be associated with the documents. For instance, in Fig. 5a, *Mean Diastolic blood pressure* was defined as a sub-tag of *Patient Characteristics*. The user can also visualise the resulting taxonomy as a radial tree chart, as shown in Fig. 5b.

5.2.2 The role of AI in the extraction phase

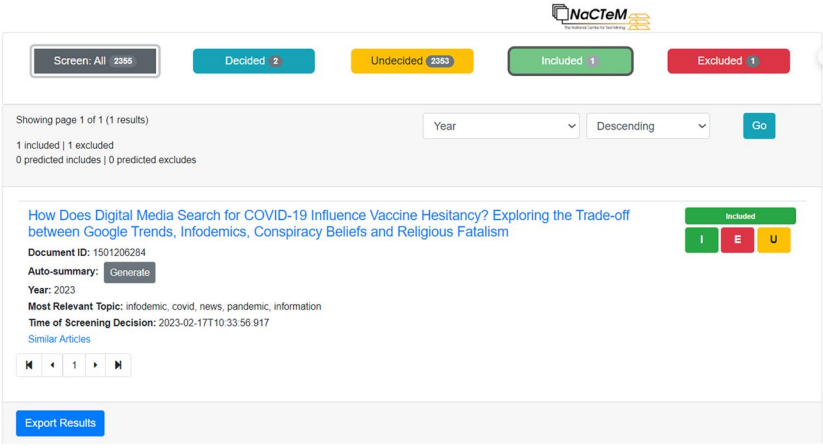
In this section, we describe the four tools that support the extraction phase (Dextr, ExaCT, Iris.ai, and RobotReviewer/RobotSearch) with a focus on the six AI features relevant to the extraction phase. We apply the same feature grouping of Sect. 5.2.1. The table in Appendix (Table 7) reports how each of the 4 tools addresses the relevant features.

Research field: RobotReviewer/RobotSearch and ExaCT focus on the medical field, whereas Dextr covers environmental health science. In contrast, Iris.ai can be employed across various research domains.

SLR task: ExaCT, Dextr, and Iris.ai perform Named Entity Recognition (NER) Nasar et al. (2021) to extract various types of information from the relevant articles. Specifically, ExaCT identifies RCT entities based on the CONSORT statement Moher et al. (2001). It returns the top five supporting sentences for each extracted RCT entity, ranked according to relevance. Dextr detects data entities used in environmental health experimental animal studies (e.g., species, strain) Walker et al. (2022). Finally, Iris.ai allows users to customise entity extraction by defining their own set of categories and associating them with a set



(a) ASReview



(b) RobotAnalyst

Fig. 2 Examples of interfaces for paper classification

of exemplary papers. This is done by filling in a form called Output Data Layout (ODL), which is essentially a spreadsheet detailing all the entities that need to be extracted. Finally, RobotReviewer/RobotSearch categorises biomedical articles according to their assessed risk of bias and provides sentences that support these evaluations.

AI approach: The tools perform the NER tasks with a variety of algorithms. ExaCT applies a two-step approach Kiritchenko et al. (2010). First, it identifies sentences that are

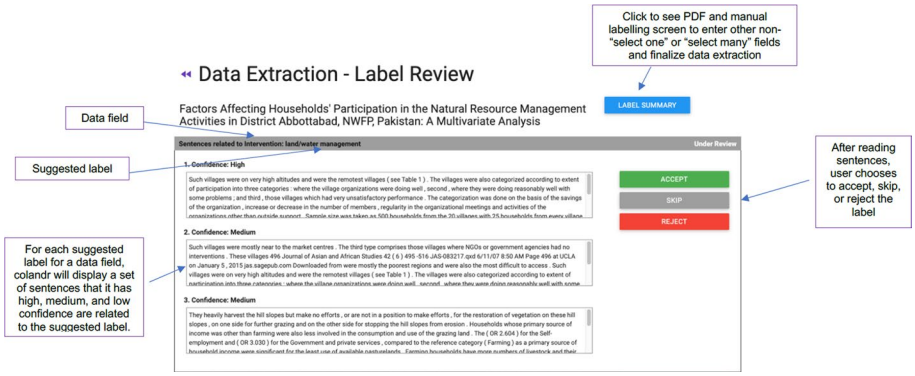


Fig. 3 Examples of the tagging process of Colandr. This figure is courtesy of Colandr (n.d.)

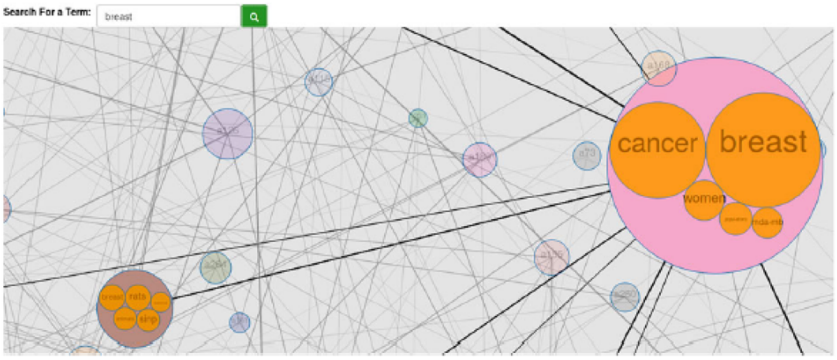
predicted to be similar to those in the pre-trained model, using a SVM classifier. Next, it extracts from these sentences a set of entities via a rule-based approach, relying on the 21 CONSORT categories Moher et al. (2001). Dextr employs a Bidirectional Long Short-Term Memory - Conditional Random Field (BI-LSTM-CRF) neural network architecture Nowak and Kunstman (2019; Walker et al. (2022). Iris.ai does not share specific information about the method used for NER. Finally, RobotReviewer/RobotSearch employs an ensemble classifier, combining multiple CNN models Krichen (2023) and soft-margin Support Vector Machines Boser et al. (1992) in order to categorise articles based on their risk of bias assessment (either low or high/unclear) and concurrently extract sentences that substantiate these judgements. The final score for each predicted sentence is the average of the scores obtained from each model.

Input data and text representation: The majority of the models accept the full-text document as input, except for Dextr, which utilises only titles and abstracts. The format requirements vary across these tools. Dextr, Iris.ai, and RobotReviewer/RobotSearch, Iris.ai process papers in PDF format. Dextr also supports input in RIS or EndNote formats. ExaCT encodes papers as HTML. The methods for text representation also differ across tools. Dextr encodes text using two pre-trained embeddings: GloVe Pennington et al. (2014) (Global Vectors for Words Representations) and ELMo Peters et al. (2018) (Embeddings from Language Models). Iris.ai utilises the same fingerprint representation Wu et al. (2018) discussed in Sect. 5.2.1. ExaCT uses a simple BoW representation. Finally, RobotReviewer/RobotSearch uses BoW for the linear model and an embedding layer for the CNN model.

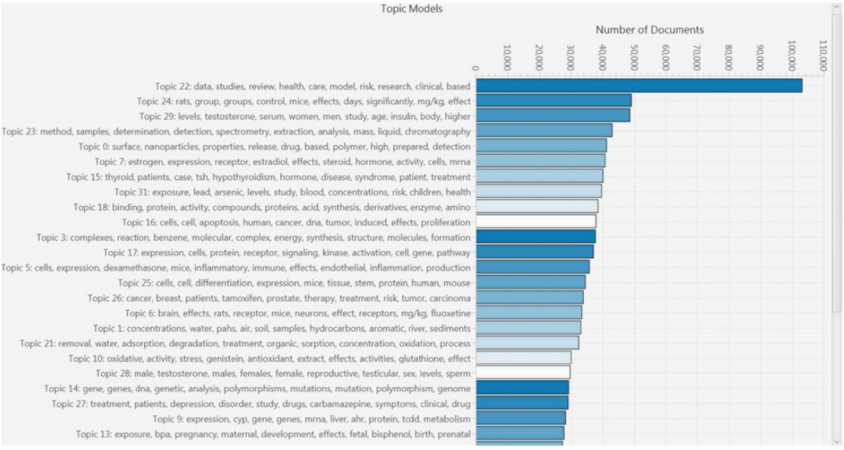
5.2.3 General features

Table 4 provides an overview of the proportion of tools covering each of the 23 features. These features are categorised across the six categories outlined in Sect. 5.1.2. The Table in Appendix (Table 8) provides a more general analysis, detailing how the 21 tools address the 23 features.

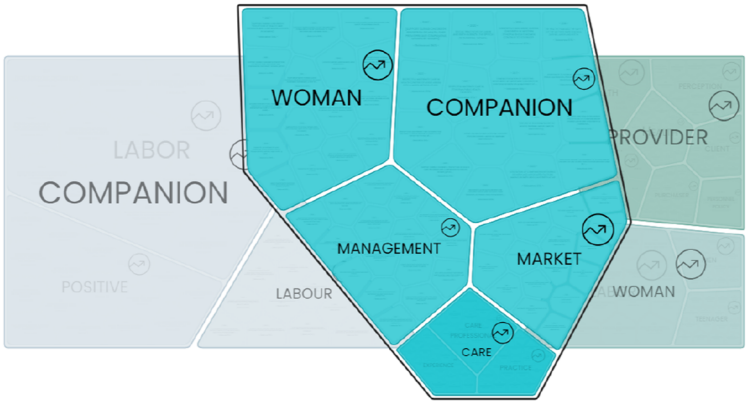
The *functionality* category exhibits the highest degree of implementation, with 5 out of 7 features being effectively executed by all the tools. The remaining two features, namely *authentication* and *project auditing*, are implemented by 18 and 9 tools, respectively. The other categories present a more heterogeneous scenario. Within the *retrieval* category, only



(a) RobotAnalyst: Topic Modelling.

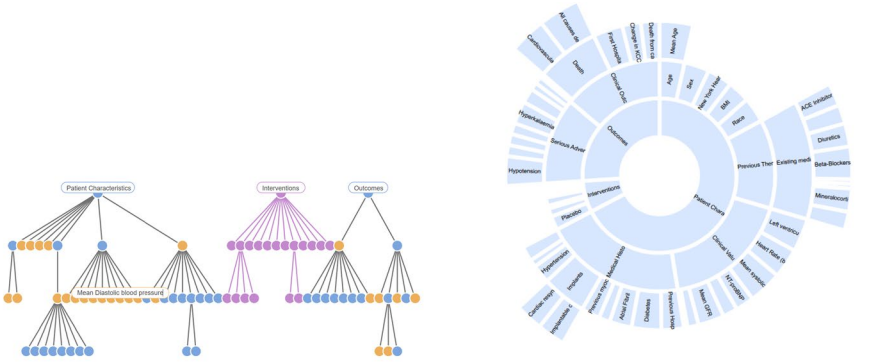


(b) SWIFT-Review: Topic Modelling.



(c) Iris.ai: Specific Topics.

Fig. 4 Examples of interactive interfaces for pre-screening



(a) Nested Knowledge: Hierarchical Ontology.

(b) Nested Knowledge: Qualitative Synthesis.

Fig. 5 Examples of interactive interface for the post-screening

reference importing is implemented by all tools. Interestingly, no tools provide the ability to automatically retrieve the reference of a paper from bibliographic databases. The tools also offer limited support for the feature within the *discovery* category. Notably, approximately 50% of the tools lack basic functionalities such as reference deduplication, options for manual annotation and exclusion of references, and features for labelling and commenting on the references. Regarding the *documentation* category, only 4 tools (DistillerSR, Nested Knowledge, Rayyan, and Covidence) provide the PRISMA diagram of the entire SLR process or the protocol templates. Significantly, LitSuggest stands out as the sole tool providing a *living review*, enabling users to easily update their earlier analyses by automatically adding recent papers that exhibit a high degree of similarity to the previously selected ones. In terms of *economic* aspects, the majority of the tools (13 out of 21) are accessible for free.

In summary, only eight of the evaluated tools implement at least 70% of the designated features. Specifically, DistillerSR, Nested Knowledge, Dextr, and ExaCT lead with the highest feature coverage at 82%. They are followed by PICOPortal and Rayyan, each with 78%, EPPI-Reviewer with 74%, and SWIFT-Active Screener with 70%. Among the remaining 13 tools, eight cover between 50 and 70% of the features, while the last five cover between 35 and 50%.

5.3 Outstanding SLR tools

Comparing our results with the previous studies in the literature (Marshall et al. 2014; Van der Mierden et al. 2019; Harrison et al. 2020; Cowie et al. 2022), we observed that many SLR tools have undergone significant development and advancements in the last few years. Particularly, the features in the *functionality* category have received more attention and are now considered standard functions. These include capabilities for tracking and auditing projects, multiple user support, and multiple user roles. Additionally, the management of references has seen considerable enhancement. As discussed in the previous section, the more complete tools in terms of feature coverage include DistillerSR, Nested Knowledge, Dextr, ExaCT, PICOPortal, and Rayyan. However, in practical scenarios, the selection of these tools should be guided by the user's specific needs and use cases. In the following,

we provide a brief analysis of some tools that our evaluation has identified as particularly suited to certain scenarios. However, it is important to recognise that there is no single best solution in this complex landscape. Therefore, we encourage researchers to experiment with these tools and determine which ones best meet their requirements.

In non-biomedical fields, ASReviewer stands out for its comprehensive range of methods for selecting relevant articles, including Logistic Regression, Random Forest, Naive Bayes, and Neural Networks classifiers. This makes it a potentially optimal choice for this phase of research. Iris.ai and Colandr are also strong contenders that may enable the greatest flexibility since they allow users to respectively cluster documents based on their semantic similarity, and create specific categories for paper classification. Moreover, they offer user-friendly interfaces for analysing the resulting data. Both platforms feature user-friendly interfaces that facilitate the analysis of the resulting data. These features are especially beneficial for exploratory studies aiming to progressively deepen understanding of a domain.

In the biomedical field, Covidence, PICOPortal, EPPI-Reviewer, RobotReviewer/RobotSearch, and Rayyan are all reliable tools. Covidence, PICOPortal, and EPPI-Reviewer have also the capability to identify Randomised Controlled Trials (RCTs) using a predefined classification model. Among these, EPPI-Reviewer offers the most flexibility, since it can be customised to identify a broader range of studies, including systematic reviews, economic evaluations, and research related to COVID-19. RobotReviewer/RobotSearch stands out as the only tool that offers automated bias analysis. This feature makes it an ideal choice for researchers who require this specific functionality. Finally, Rayyan offers a suite of biomedical features, such as PICO highlighting and filtering, the capability to extract study locations, and topic extraction enriched with MeSH terms. It also allows users to define a set of biomedical keywords and phrases for inclusion and exclusion, which is beneficial for identifying specific RCTs.

5.4 Threats to validity

This section outlines various threats to the validity of this study. We examined four primary categories of validity threats: internal validity, external validity, construct validity, and conclusion validity Wohlin et al. (2012). We considered and mitigated them as follows.

Internal validity. Internal validity in systematic literature reviews concerns to the rigour and correctness of the review's methodology. To ensure the replicability of our review, we meticulously developed a methodologically sound protocol, which incorporated systematic and transparent procedures for the selection of studies and software tools. We also adopted the PRISMA guidelines, known for their robustness and reproducibility (the PRISMA checklist is available in the supplementary material). The protocol for this SLR was developed by the first author and reviewed by co-authors to establish a consensus before initiating the review process. We identified relevant tools using two prominent software repositories (the Systematic Literature Review Toolbox and the Comprehensive R Archive Network) supplemented by manually searching relevant surveys for additional tools. Additionally, we employed a snowballing search strategy to further extend and validate our results. The selection process involved multiple stages to ensure rigorous evaluation and minimise selection bias. Initially, the first author filtered the tools based on the description in the repositories. Next, all authors participated in a more thorough review of the shortlisted tools. In cases where information was unclear or missing, the first author contacted the tool developers directly through email or online interviews. All related

Table 4 Proportion of the 21 tools implementing the 23 generic features

Category	Feature	Yes	No
Functionality	Multiplatform	**21** (100%)	0 (0%)
	Multiple user roles	**21** (100%)	0 (0%)
	Multiple user support	**21** (100%)	0 (0%)
	Project auditing	9 (43%)	12 (57%)
	Project progress	**21** (100%)	0 (0%)
	Authentication	18 (86%)	3 (14%)
	Status of software	**21** (100%)	0 (0%)
Retrieval	Automated full-text retrieval	3 (16%)	16 (84%)
	Automated search	8 (42%)	11 (58%)
	Snowballing	0 (0%)	19 (100%)
	Manual reference importing	5 (26%)	14 (74%)
	Manually inserting full-text	8 (42%)	11 (58%)
	Reference importing	**21** (100%)	0 (0%)
Discovery	Deduplication	8 (42%)	11 (58%)
	Discrepancy resolving	12 (57%)	9 (43%)
	In-/excluding references	13 (68%)	6 (32%)
	Reference labelling & comments	10 (53%)	9 (47%)
	Screening phases	19 (100%)	0 (0%)
Documentation	Exporting results	**21** (100%)	0 (0%)
	Flow diagram creation	4 (21%)	15 (79%)
	Protocol	4 (21%)	15 (79%)
Living systematic review	Living/updatable	1 (5%)	18 (95%)
Economic	Free to use	13 (62%)	8 (38%)

The features covered by all tools (100%) are highlighted in bold

publications were thoroughly reviewed to inform the development of the features. Despite the systematic process, biases could still emerge due to the subjective decisions made by researchers when applying inclusion and exclusion criteria. To mitigate this, we collaboratively reviewed the inclusion or exclusion of the shortlisted tools, thereby reducing the influence of individual biases. Another potential threat to the internal validity arises from the fact that the SLR Toolbox has been offline since March 2024. Although the developers have indicated that it will be operational again soon, there is a possibility that the tool may not be available for future surveys. Nevertheless, we believe that including its results remains valuable, given that this system was utilised in five (Kohl et al. 2018; Van der Mierden et al. 2019; Harrison et al. 2020; Cowie et al. 2022; Robledo et al. 2023) of the eight previous surveys identified in Sect. 3.

In conclusion, while the replication of this study by another research team might yield slight variations in the tools and studies included, the robust, systematic methodology employed and the collaborative nature of the review process lend a high degree of internal validity to our findings.

External validity. External validity refers to the degree to which the findings of this systematic literature review are generalisable across various environments and domains. To mitigate threats to external validity, we used multiple sources for selecting the SLR

tools. Despite these efforts, the selection of search engines and the formulation of search strings might have impacted the completeness of the tool identification. It is possible that some tools were missed because they were not described using the selected keywords or were absent from the targeted repositories and previous surveys. To counteract this limitation, several strategies were implemented. First, search strings were iteratively refined to enhance coverage and ensure a more exhaustive identification of potential tools. Second, a thorough snowballing method was employed. Finally, interviews were conducted with developers of several tools to further ensure the inclusiveness of the tool selection.

Concerning the inclusion and exclusion criteria, we identified two main potential threats to external validity. The first threat stems from the exclusion of tools that do not feature user interfaces. This criterion was set to focus on tools that are readily adoptable by the average researcher. However, earlier studies involving prototypes without interfaces still align with many of our findings. For instance, these studies also conclude that most SLR tools employ relatively outdated AI techniques Schmidt et al. (2023, 2023), as we will discuss more in detail in Sect. 6.1. The second threat concerns the exclusion of tools that were either under maintenance and unavailable for evaluation or had not been updated in the past ten years. This exclusion criterion might have omitted tools that, despite being inaccessible at the time of the review, could otherwise fulfil the inclusion criteria. These exclusions could potentially restrict the generalisability of our findings.

Construct validity. Construct validity concerns the extent to which the operational measures used in a study accurately represent the concepts the researchers intend to investigate. In our systematic literature review, a primary concern is whether the 34 features identified to evaluate SLR tools cover all relevant characteristics, particularly concerning the integration of Artificial Intelligence. To address potential gaps identified from previous studies, we developed a set of 11 AI-specific features aimed at capturing aspects previously overlooked. Despite these efforts to create a thorough framework for analysis, AI remains a rapidly evolving field, and our feature set might not encapsulate all current and emerging dimensions. To mitigate this issue, the authors collaboratively developed the feature definitions, striving to create a comprehensive representation that incorporates both established dimensions identified in prior surveys and emerging trends noted in recent publications and software developments. Nevertheless, it is acknowledged that some relevant aspects may still be absent from our analysis.

Conclusion validity. Conclusion validity in systematic literature reviews refers to the extent to which the conclusions drawn from the review are supported by the data and are reproducible. In our review, we focused on mitigating threats to conclusion validity by employing a systematic process for identifying relevant software tools and extracting pertinent data for analysis. To ensure accuracy and consistency in data collection, we developed a data extraction form based on the general and AI-specific features identified during our meta-review and feature analysis. The first author applied this form to a small subset of tools to test its effectiveness. Subsequently, all authors independently used the same form to extract data for the same subset of tools. Comparative analysis of the extracted data revealed a high degree of consistency among authors, thereby validating the data extraction process. Following this validation, the first author continued with the data extraction for the remaining tools. Throughout the data analysis and synthesis phases, we engaged in multiple rounds of discussions to refine our categorisation and representation of the features. This collaborative approach aimed to reduce bias and enhance the reliability of our findings.

A persistent threat to conclusion validity in the context of software tool reviews is the dynamic nature of software development Ampatzoglou et al. (2019). Software

tools frequently evolve, acquiring new functionalities that may not be documented in the published literature. To address this, we supplemented our literature review with comprehensive examinations of websites, tutorials, and relevant academic papers. Additionally, we reached out directly to developers to obtain updated or missing information. This proactive approach frequently provided crucial clarifications and additions, which we incorporated into our final review, thereby strengthening the reliability of our conclusions. However, the field of AI is evolving rapidly, particularly in areas such as Generative AI Brynjolfsson et al. (2023) and Large Language Models Min et al. (2023). As a result, it is expected that many tools will soon incorporate new AI features. Therefore, while our findings offer a snapshot of the current landscape, they may not fully represent the ongoing advancements.

6 Research challenges

The current generation of SLR tools can demonstrate significant effectiveness when utilised properly. Nonetheless, these tools still lack crucial abilities, which hampers their widespread adoption among researchers. This section will discuss some of the key research challenges identified from our analysis that the academic community will need to address in future work. It is not intended to provide a systematic review like the one in Sect. 5, but rather to explore some of the most compelling research directions and open challenges, aiming to inspire researchers in this area. Section 6.1 analyses the current challenges associated with integrating AI within SLR tools and discusses the potential social, ethical, and legal risks associated with the resulting systems. Section 6.2 addresses usability concerns, which represent a major barrier to the adoption of these tools. Finally, Sect. 6.3 discusses the challenges in establishing a robust evaluation framework and suggests some best practices.

6.1 AI for SLR

As previously discussed, several SLR tools now incorporate AI techniques for supporting in particular the screening and extraction phases. However, current approaches still suffer from several limitations. Consistent with prior research (Kohl et al. 2018; Burgard and Bittermann 2023; Schmidt et al. 2021), our study reveals that the majority of SLR tools still depend on possibly outdated methodologies. This includes the use of basic classifiers, which are no longer considered state-of-the-art for text and document classification. Likewise, several tools continue to employ BoW methods for text representation, although some of the most recent ones (Walker et al. 2022; Van De Schoot et al. 2021; Marshall et al. 2018) have shifted towards adopting word and sentence embedding techniques, such as GloVe (Pennington et al. 2014), ELMo Peters et al. 2018), SciBERT Beltagy et al. 2019), and Sentence-BERT Reimers and Gurevych (2019). Therefore, the first interesting research direction regards incorporating advanced NLP technologies, particularly the rapidly evolving Large Language Models (LLMs) Min et al. (2023). LLMs represent the state of the art for many NLP tasks and demonstrated remarkable proficiency in classifying and extracting information from documents (Dunn et al. 2022; Xu et al. 2023). However, integrating these models presents several challenges Ji et al. (2023). Firstly, LLMs are trained on general data, resulting in less effective performance in specialised fields and languages with fewer resources. Secondly, LLMs may generate inaccurate or fabricated

information, known as "hallucinations". Finally, understanding the decision-making process of LLMs is complex, and their outputs can be inconsistent. A possible solution to these issues is the integration of LLMs with different types of knowledge bases that can provide verifiable factual information Meloni et al. (2023). This is typically achieved through the Retriever-Augmented Generation (RAG) framework Lewis et al. (2020), which allows LLMs to retrieve information from a collection of documents or a knowledge base. For example, the recent CORE-GPT Pride et al. (2023) utilises a vast database of research articles to assist GPT3 (Brown et al., 2020) and GPT4 OpenAI (2023) in generating accurate answers. In addition, the extraction phase in particular could be enhanced by also incorporating modern information extraction methods such as event extraction Li et al. (2022), open information extraction Liu et al. (2022), and relation prediction Tagawa et al. (2019).

A second interesting research direction regards interpretability. Indeed, current classification methods for the screening phase typically operate as 'black boxes', not giving much additional information on why a certain paper was deemed as relevant. One important research challenge here is to improve this step by including interpretability mechanisms such as fact-checking Vladika and Matthes (2023) or argument mining Lawrence and Reed (2020) to provide further insights. Such techniques would provide deeper insights into the screening process, enhancing the reliability and credibility of the tools. In the field of explainable AI Linardatos et al. (2020), significant research has been conducted to improve our understanding of the processes models use to generate specific outputs. Specifically, in the context of LLMs, various prompting techniques have been developed to enhance the models' ability to explain their reasoning and justify their decisions. These techniques include Chain-of-Thought (CoT) Wei et al. (2023), Tree of Thoughts (ToT) (Long 2023; Yao et al. 2023) and Graph of Thoughts (GoT) Besta et al. (2023).

A third promising research direction involves the use of semantic technologies Patel and Jain (2021), particularly knowledge graphs, to enhance the characterisation and classification of research papers Salatino et al. (2022). Knowledge graphs consist of large networks of entities and relationships that provide machine-readable and understandable information about a specific domain following formal semantics Peng et al. (2023). They typically organise information according to a domain ontology, which provides a formalised description of entity types and their relationships Hitzler (2021). In recent years, we saw the emergence of several knowledge graphs that offer machine-readable, semantically rich, interlinked descriptions of the content of research publications (Jaradeh et al. 2019; Salatino et al. 2019; Angioni et al. 2021; Wijkstra et al. 2021). For instance, the latest iteration of the Computer Science Knowledge Graph (CS-KG)[9] details an impressive array of 24 million methods, tasks, materials, and metrics automatically extracted from approximately 14.5 million scientific articles Dessí et al. (2022). Similarly, the Open Research Knowledge Graph (ORKG)[10] provides a structured framework for describing research articles, facilitating easier discovery and comparison Jaradeh et al. (2019). ORKG currently includes about 25,000 articles and 1,500 comparisons. This survey is also available in ORKG (https://orkg.org/review/R692116). In a similar vein, Nanopublications[11] allow the representation of scientific facts as knowledge graphs Groth et al. (2010). This method has been recently applied to support "living literature reviews",

[9] Computer Science Knowledge Graph - http://w3id.org/cskg/

[10] https://www.orkg.org/

[11] https://nanopub.net/

which can be dynamically updated with new findings Wijkstra et al. (2021). The integration of these knowledge bases offers significant possibilities. It allows for a more detailed and multifaceted analysis of document similarity, and aids in identifying documents related to specific concepts. For instance, it would enable the retrieval of articles that mention particular technologies or that utilise specific materials.

Other SLR phases, such as appraisal and synthesis, received relatively little attention. This gap offers a substantial research opportunity for the application of AI techniques in these areas. In the appraisal phase, incorporating AI-driven scientific fact-checking tools to evaluate the accuracy of research claims could provide significant benefits Vladika and Matthes (2023). For the synthesis phase, the use of summarisation techniques Altmami and Menai (2022) and text simplification methods Sikka and Mago (2020) has the potential to enhance both the efficiency of the analysis and the clarity of the final output.

Finally, we recommend that the research community participates to scientific events and initiatives in this field, such as ICASR[12] (Beller et al. 2018; O'Connor et al. 2018, 2019, 2020), ALTAR[13] Di Nunzio et al. (2022), and the MSLR Shared Task[14] Wang et al. (2022). These initiatives are focused on discovering the most effective ways in which AI can improve the SLR stages.

6.1.1 AI impact assessment

The importance of evaluating the impact of AI systems has grown significantly, particularly with the recent enactment of the European Commission's Artificial Intelligence Act, which establishes specific requirements and obligations for AI providers. In this context, it is crucial to assess the potential impact of AI-enhanced SLR tools, considering both the relevant literature and the new regulatory framework (Renda et al. 2021; Ayling and Chapman 2022).

Stahl et al. (2023) propose an impact assessment model consisting of two main steps: (1) determining whether the AI tool is expected to have a social impact, and (2) identifying the stakeholders who might be affected by the AI system. We can apply this model to the SLR tools discussed in this survey.

Regarding social impact, SLR tools aim to support the identification, analysis, and synthesis of findings that are pertinent to specific research questions. The information generated by these tools is typically incorporated into research papers and, in some cases, may influence policy development Birkland (2019). The primary concern here is the dissemination of inaccurate scientific information and how such information might be used by the community and policymakers.

Regarding potentially impacted stakeholders, we consider three main groups. The first group consists of authors who use these tools for literature reviews. These individuals face the risk of including incorrect studies and drawing inaccurate conclusions, potentially jeopardising the quality of their work and their careers. To mitigate these risks, it is crucial to use tools that demonstrate high performance and transparency, especially in terms of the datasets used and potential biases. Additionally, these tools should provide mechanisms that allow users to inspect, interpret, and override the tool's choices. The second group includes the readers of these literature reviews. They are primarily at risk of being exposed

[12] International Collaboration for the Automation of Systematic Reviews (https://icasr.github.io/)

[13] Augmented Intelligence for Technology-Assisted Reviews Systems (https://altars2022.dei.unipd.it/)

[14] Multidocument Summarisation for Literature Review (https://github.com/allenai/mslr-shared-task)

to and subsequently disseminating incorrect or biased information. In addition to the strategies previously mentioned, the scientific community itself plays a crucial role in mitigating this risk by reproducing and correcting earlier results Munafò et al. (2017). The third group becomes relevant when policy development is involved. In these instances, targeted populations might be affected by policies based on incorrect or biased analyses Young (2005). To mitigate this risk, policymakers shall conduct additional analyses to verify the accuracy of the information and use multiple sources.

In conclusion, while SLR tools carry some inherent risks, these can generally be managed through responsible use and adherence to validation and correction strategies Myllyaho et al. (2021). A major challenge remains in enhancing the trustworthiness of these tools through robust evaluation mechanisms O'Connor et al. (2019). As we will discuss in Sect. 6.3, the current landscape lacks high-quality evaluation frameworks.

In the context of the recent EU Artificial Intelligence Act,[15] it is important to note that if we classify SLR tools as "specifically developed and put into service for the sole purpose of scientific research and development", they would be explicitly exempt from this legislation. Nevertheless, it is still worthwhile to examine how these tools might be categorised under the four risk categories outlined by the AI Act: Unacceptable Risk, High Risk, Limited Risk, and Minimal Risk. After a detailed analysis of the current draft of the legislation, it seems that a typical AI-enhanced SLR tool would most likely be classified as 'Limited Risk'. This classification primarily concerns potential issues regarding transparency Larsson and Heintz (2020), which may become more pronounced as these tools begin to utilise generative AI Brynjolfsson et al. (2023). According to the AI Act, these systems should be "developed and used in a way that allows appropriate traceability and explainability while making humans aware that they communicate or interact with an AI system as well as duly informing users of the capabilities and limitations of that AI system and affected persons about their rights."

6.2 Usability

The current generation of SLR tools remains underutilised Marshall et al. (2018). Most researchers continue to depend on manual methods, often supported by software like Microsoft Excel, or reference management tools Marshall et al. (2015) such as Zotero[16] and Mendeley.[17] Recent studies Van Altena et al. (2019), suggest that this limited usage primarily stems from usability issues, in addition to a few other relevant factors: (i) *steep learning curve*, as researchers may be unfamiliar with the tools' functionalities Scott et al. (2021), (ii) *misalignment with user requirements*, as many of these software deviate from the guidelines set forth by SLR protocols and exhibit limited compatibility with other software systems (Thomas 2013; Arno et al. 2021), (iii) *distrust*, as there is uncertainty about the reliability and the mechanisms of these tools (O'Connor et al. 2019; Haddaway et al. 2020), and (iv) *financial obstacles*, predominantly arising from licensing expenses, along with feature restrictions in trial versions Dell et al. (2021). This suggests that usability and

[15] EU Artificial Intelligence Act - https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

[16] Zotero - https://www.zotero.org/

[17] Mendeley - https://www.mendeley.com/

accessibility should be prioritised in the design process to encourage wider adoption of these tools (Hassler et al. 2014, 2016; Al-Zubidy et al. 2017).

The literature has given limited attention to the usability of SLR tools. To the best of our knowledge, only a few studies focused on this aspect. For instance, Harrison et al. Harrison et al. (2020) conducted an experiment where six researchers were tasked with using six different tools in trial projects. Findings indicated that two tools also presented in this study, Rayyan and Covidence, were perceived as the most user-friendly. Van Altena et al. (2019) conducted a survey involving 81 researchers about the usage of SLR tools and found that the primary reasons cited by participants for discontinuing the use of a tool included poor usability (43%), insufficient functionality (37%), and incompatibility with their workflow (37%). In the same study, a set of SLR tools was assessed using the System Usability Scale (SUS) questionnaire Lewis (2018). The tools demonstrated comparable usability, with scores ranging from 66 to 77. These scores correspond to a 'C' to 'B' grade, indicating satisfactory but not outstanding performance.

Therefore, a critical challenge in this field lies in the need for more comprehensive research focused on usability. This involves conducting in-depth studies to understand the various aspects of usability, such as effectiveness, efficiency, engagement, error tolerance, and ease of learning Quesenbery (2014). The goal is to gather empirical data and user feedback that can provide insights into how users interact with tools, identify common usability issues, and understand the specific needs and preferences of different user groups. Based on these findings, it is essential to develop robust, evidence-based usability guidelines Schall et al. (2017). These guidelines should offer clear and actionable recommendations for designing user-friendly interfaces and functionalities in future tools.

6.3 Evaluation of SLR tools

A robust evaluation framework is essential for comparing SLR tools and supporting their continuous improvement National Academies of Sciences (2019). In the following subsections, we will first discuss the shortcomings of existing evaluation methods and then propose a set of best practices as an initial step towards developing a high-quality evaluation framework.

6.3.1 Lack of standard evaluation frameworks

The assessment of SLR tools presents a significant challenge due to the absence of standard evaluation frameworks and established benchmarks. Existing literature includes various evaluations of SLR tools that focus on individual phases of the SLR process (Liu et al. 2018; Yu et al. 2018; Burgard and Bittermann 2023). However, these evaluations are not directly comparable due to variations in datasets and evaluation methodologies. Moreover, most SLR tools are tested using small, custom datasets, which may not provide a realistic representation of their performance in typical usage scenarios Burgard and Bittermann (2023). Additionally, leading commercial providers of SLR tools typically do not make evaluation data available, which complicates comparisons with both existing competitors and new prototypes developed by the research community.

Another concern is related to the performance metrics. Indeed, canonical metrics like precision, recall, and F1-score may not suffice to assess these tools. For instance, for the screening phase, it is critical to minimise the costs of screening while preserving a high recall. For this reason, it was suggested to adopt the F2 score Sumbul et al. (2021) instead

of the F1 score. The F2 score is computed as the weighted harmonic mean of precision and recall. In contrast with the F1 score, which assigns equal importance to precision and recall, the F2 score places greater emphasis on recall compared to precision. The Work Saved over Sampling (WSS) Van De Schoot et al. (2021) is another metric that proved to be quite effective in assessing the screening phase Burgard and Bittermann (2023). However, Kusa et al. (2022) point out that this measure depends on the number of documents and the proportion of relevant documents in a dataset, making it difficult to compare the performance of different screening tasks performed over different systematic reviews. To address this, they introduced the Normalised Work Saved over Sampling (nWSS) metric Kusa et al. (2023), which facilitates the comparison of paper screening performance across various datasets.

Another limitation arises from the restricted range of dimensions assessed during the evaluation of SLR tools. *Performance* is only one of several aspects that should be considered. *Usability*, as discussed in Sect. 6.2, is another crucial factor. Trustworthiness is also a vital dimension O'Connor et al. (2019). Although trustworthiness is partially reliant on performance, it also involves reliability, transparency, and ethical integrity, all of which can influence researchers' willingness to use these tools. Indeed, while automation might boost the efficiency of the review process, it also carries the risk of introducing errors. These errors could lead to the omission of pertinent studies or the inclusion of inappropriate ones, which could substantially alter the research results. To address these issues, O'Connor et al. (2019) propose two main strategies to enhance trust in SLR tools. The first strategy is to undertake detailed studies comparing the precision of automated tools with conventional review methods. The second strategy involves encouraging reputable teams or funding agencies to support the use of these tools. Wang et al. (2023) recommend that creators of AI-driven tools should investigate different affordances to enhance user trust. Specifically, they identify three essential design elements: (i) clear communication about AI capabilities to set appropriate user expectations; (ii) availability of user settings to adjust and tailor preferences related to AI-generated recommendations; and (iii) inclusion of indicators that explain the mechanisms of the underlying models, helping users evaluate the AI's suggestions. Bernard et al. Bernard and Balog (2023) further expand on this by advocating for the assessment of fairness, accountability, transparency, and ethics (FATE) aspects. They also explore definitions, approaches, and evaluation methodologies aimed at developing trustworthy information retrieval systems. A promising avenue for future research is to further explore the application of these concepts to the emerging generation of SLR tools and, more in general, to AI tools designed to support research activities.

In the realm of AI-enhanced SLR tools, *transparency* is one of the greatest challenges in building user trust. This is primarily because most contemporary AI models operate as black boxes, making their internal processes difficult to comprehend Castelvecchi (2016). Additionally, as shown in Table 2, only 4 out of the 21 tools analysed operate under open licenses, which exacerbates the lack of transparency. To mitigate this issue, existing research suggests several approaches. These include making the AI model, its training data, and the corresponding code openly accessible for examination by users and experts of (National Academies of Sciences et al. 2018; Abbasi 2023). Furthermore, it is recommended that developers offer thorough evaluations of any potential biases and perform ablation studies to determine common error types Ntoutsi et al. (2020). It is also advised to integrate explainable AI methods Linardatos et al. (2020).

6.3.2 Towards an evaluation framework: best practices

In this section, we aim to present some best practices to establish a robust evaluation framework for SLR tools, informed by our surveys and subsequent analysis. While developing a comprehensive evaluation framework is beyond the scope of this paper, we aim to contribute to ongoing discussions by proposing an initial theoretical framework.

We propose a set of best practices centred around three critical aspects of SLR systems: *performance*, *usability*, and *transparency*. First, we suggest developing replicable methodologies to assess the performance of various algorithmic components designed to address the tasks outlined in the previous section. Second, we recommend conducting a comprehensive and unbiased assessment of usability. Finally, we emphasise the importance of improving the trustworthiness of these tools by disclosing essential information about their capabilities and limitations, sharing knowledge bases and models, and adopting explainable AI solutions.

We do not claim that this set of principles is exhaustive; rather, it represents an initial effort to introduce a few principles that could make evaluations in this domain more reproducible and transparent. The principles we outline precede specific implementation decisions and are not tied to any particular technology, standard, or method. It is also important to recognise that these principles are not novel but reflect established guidelines used by various communities facing similar challenges (Liu et al. 2018; National Academies of Sciences et al. 2018; O'Connor et al. 2019; Linardatos et al. 2020; Abbasi 2023; Bernard and Balog 2023; Wang et al. 2023). However, as noted previously, the community that develops SLR tools does not consistently adhere to these practices, leading to a lack of comparability among these systems.

The proposed best practices are outlined in the following.

Performance: All models and algorithms employed by an SLR tool for specific tasks should undergo formal evaluation. These evaluations must adhere to established benchmarks and best practices recognised by the relevant scientific community. We thus recommend the following practices.

1. *Detailed Documentation*: Provide a comprehensive description of all the algorithms employed for the different functions within the system.
2. *Standardised Evaluation*: Evaluate these algorithms against standard metrics and benchmarks that are widely accepted within the scientific community relevant to the tasks being performed.
3. *Benchmark Disclosure*: Publicly release the benchmarks used for evaluating these methods to facilitate comparison with alternative approaches.
4. *Benchmark Adoption*: Whenever possible, opt to reuse established benchmarks, especially those that are recognised and have previously been used for evaluating SLR tools.
5. *Code Availability*: Ensure that the code for both the algorithms and the evaluation process are persistently available on an online repository to promote accessibility and reproducibility.

Usability: The evaluation of usability should be comprehensive, replicable, and conducted in environments that closely resemble the diverse settings in which the system will operate, involving various types of potential users. To ensure a thorough assessment, we recommend the following practices.

1. *Representative User Participation*: Conduct detailed user studies with participants who accurately represent the system's target user base.
2. *Diverse Usability Factors*: The user studies should comprehensively evaluate various usability aspects discussed in the literature such as effectiveness, efficiency, engagement, error tolerance, and ease of learning.
3. *Standard Questionnaires*: To facilitate comparisons with other systems, the evaluation should also employ established usability questionnaires such as the System Usability Scale (SUS) Brooke (1996), the User Experience Questionnaire (UEQ) Laugwitz et al. (2008), or the Usability Metric for User Experience (UMUX) Finstad (2010).
4. *Accessibility*: Evaluate usability for individuals with diverse disabilities, including visual, auditory, physical, speech, cognitive, language, learning, and neurological. Adopting the Web Content Accessibility Guidelines (WCAG) 2.1, developed by the World Wide Web Consortium (W3C), is recommended to guide this process.
5. *Availability of Materials*: Publish all materials related to the usability evaluation in a third-party repository to ensure reproducibility.

Trasparency: In line with the AI Act and the necessity of enhancing user trust O'Connor et al. (2019), transparency is essential for AI-driven SLR tools. It is important to incorporate transparency also in the evaluation process, ensuring traceability and explainability, and clearly defining the tools' capabilities and limitations. Although proprietary systems might emphasise confidentiality to preserve a competitive advantage, it is crucial to balance commercial interests with the broader imperative for accountability and trust in AI technologies. We recommend the following practices to enhance transparency.

1. *Availability of Training Data*: Since the training dataset influences the model's behaviour and can perpetuate biases, ensuring its availability is essential.
2. *Availability of Knowledge Bases*: Many systems utilise various knowledge bases, such as taxonomies and vocabularies of research areas, to enhance performance. These resources should be made accessible for user inspection.
3. *Availability of Models*: Trained models should be made available to facilitate further analysis of their performance and potential biases.
4. *Explainability*: The tool should, wherever possible, provide clear explanations for its decisions, aligning with the principles of explainable AI.
5. *Comprehensive Documentation*: All functionalities of the software should be documented clearly and in user-friendly language.
7. *Clarify the Limitations*: Developers should clearly communicate the limitations of the software, indicating where the tool is expected to perform well and where it may not meet expectations.

We aim for these best practices to serve as an initial step in establishing a comprehensive evaluation framework. We hope that this effort will be expanded through dedicated theoretical and empirical research, promoting wider implementation of recognised best practices within this field.

7 Emerging AI tools for literature review

Since 2023, a new generation of AI tools designed to assist researchers has emerged. This development is largely influenced by the advancements in Large Language Models Sanderson (2023). Several leading bibliographic search engines are currently introducing LLM technology. For instance, Scopus and Dimensions[18] are working on their own chatbot engine and are planning to release them throughout 2024 (Van Noorden 2023; Aguilera Cora et al. 2024). Similarly, CORE,[19] a search engine providing access to 280 million papers, has recently presented the prototype CORE-GPT, an enhanced version that can answer natural language queries by extracting information from these documents Pride et al. (2023). These LLM-based tools do not directly support specific SLR phases as the applications that we reviewed in Sect. 5. Nevertheless, their functionalities can aid researchers in conducting literature reviews and are expected to be integrated into future SLR tools. Therefore, when discussing the advancement of AI-driven SLR tools and identifying research challenges in this domain, it is essential to consider these tools and their features. A comprehensive analysis of emerging LLM-based tools designed to assist with literature reviews and scientific writing would require an extensive survey. This section aims to present an initial exploratory study that provides insights into how this new generation of LLM-based tools is being used to assist research and what functionalities could potentially be integrated into SLR tools. Since this is an exploratory study rather than a systematic review, we adopted a straightforward search strategy focusing on tools available as online services. Therefore, we excluded tools that are solely described in academic papers and not available for practical use.

We used TopAI Tools,[20] a renowned search engine indexing more than 11K AI systems and searched for the following relevant terms: "literature review", "systematic review", "scientific research", "search engine", and "writing assistant". This search returned 164 tools, which were processed using the same two-stage selection process described in Sect. 3. We first screened the tools by using their short descriptions and then all authors performed a thorough examination of 18 candidate tools. This process yielded 11 tools in this domain. Table 5 reports an overview of these tools.

The eleven systems that we identified typically employ LLMs (mostly via the OpenAI API[21]) often enhanced with a RAG framework Lewis et al. (2020) to integrate knowledge from scientific and technical documents. As discussed in Sect. 6.1, the RAG framework enhances LLMs by enabling them to retrieve relevant information from a knowledge base or a collection of documents. This information is then incorporated into the context of the LLMs, allowing them to rely on verifiable sources and thereby reducing inaccuracies and hallucinations Ji et al. (2023).

We classified the 11 tools into two categories: search engines and writing assistants. Search engines enable users to enter a query using natural language and provide a list of related research papers and their summaries. Their main contribution is the ability to use natural language rather than keywords for searching research papers. On the other hand, writing assistants accept a description of a document, such as "Survey paper about knowledge graphs", and generate pertinent text that can be then iteratively refined by a researcher.

[18] Dimensions- https://www.dimensions.ai/

[19] CORE - https://core.ac.uk/

[20] TopAI Tools- https://topai.tools/

[21] OpenAI API - https://openai.com/blog/openai-api

Seven tools were categorised as search engines and three as writing assistants. Textero.ai was the only identified tool fitting into both categories.

7.1 Search engine tools

The tools in this category allow users to formulate a natural language query and generate a list of relevant research papers sourced from online repositories. Generally, these tools also provide concise summaries of the most prominent papers. Beyond the natural language query functionality, some tools incorporate additional search features. For instance, EvidenceHunt allows users to locate papers using keywords, medical specialisations, or filters specific to PubMed searches. Similarly, Scite offers the capability to conduct keyword searches in titles and abstracts, and uniquely, to search for specific terms within 'citation statements' Nicholson et al. (2021), i.e., segments of text that include a citation Ding et al. (2014). Additionally, Scispace and Elicit allow users to automatically extract information from papers based on predefined categories. For instance, a user can request the extraction of all references to 'technologies' within a text. However, the quality of the extracted results can vary significantly.

The bibliographic databases employed by these tools differ. Elicit, Consensus, and Perplexity utilise Semantic Scholar.[22] EvidenceHunt relies on PubMed.[23] Scite sources its content from Semantic Scholar and a broader array of publishers, such as Wiley, Sage, Europe PMC, Thieme, and Cambridge University Press. The bibliographic databases used by Scispace, Textero.ai, and MirrorThink are not documented. Scite and Consensus process full-text papers, while Elicit and EvidenceHunt only use titles and abstracts.

The majority of the tools (6 out of 8) are versatile and applicable across different research fields. EvidenceHunt is specifically tailored for use in biomedicine, while Elicit is designed to cater to both biomedicine and social sciences.

The specific details of the implementation for many of these tools remain undisclosed, as they are proprietary commercial products. However, it appears that a majority of them employ the OpenAI API, utilising various prompting strategies and often integrating a RAG framework Lewis et al. (2020) to incorporate text from pertinent articles. Notably, two of the tools explicitly state their models: Elicit and Perplexity; both of which leverage OpenAI's GPT technology.

7.2 Writing assistant tools

These tools enable the user to describe the document they want to generate and then iteratively refine it. Jenni.ai is a highly interactive tool that enables collaborative editing between the user and the AI. Initially, the user provides a step-by-step description of the desired text. Subsequently, the system generates a template for the document and progressively incorporates new sections. These sections can be edited by the user in real-time, facilitating a dynamic and iterative writing process. Textero.ai operates similarly. Users are required to input the title and description of the text they wish to create. They can then request the tool to gather pertinent references for integration and select a citation style, such as MLA or APA. The generated text can be further refined by the user either manually or through various AI functions designed to enhance or summarise sections of

[22] Semantic Scholar - https://www.semanticscholar.org/

[23] PubMed - https://pubmed.ncbi.nlm.nih.gov/

the text. Additionally, a panel on the right side provides convenient access to the list of cited references, with each paper accompanied by a brief summary. For this reason, we categorised this tool also as a search engine. Silatus can operate in four distinct modes: *question answering*, which generates a specific answer; *research report*, producing a comprehensive explanation of a research topic; *blog post*, creating content suitable for blogs; and *social media post*, tailored for social media platforms. In each mode, the user is prompted to provide a concise initial prompt to initiate text generation. Optionally, the user can instruct Silatus to retrieve and integrate pertinent references into the generated text.

As before, most systems do not disclose their technologies, yet they appear to incorporate different versions of the OpenAI API, augmented with specific prompting techniques. Silatus employs GPT-4, while Jenni.ai uses a combination of GPT-3.5 and its proprietary AI technologies. It remains unclear whether any of them have fine-tuned their models for writing-related tasks.

The quality of the text produced by these systems varies significantly, even when using very informative prompts. Presently, these tools may be more beneficial for master's students who are required to write brief essays rather than for researchers. Nonetheless, as the technology continues to evolve, it is anticipated that a new generation of tools will emerge, offering substantial assistance in academic writing. These AI systems could potentially automate several complex tasks, such as generating comprehensive literature reviews Hope et al. (2023), recommending citations (Ali et al. 2020; Buscaldi et al. 2024), and identifying new scientific hypotheses (Sybrandt et al. 2020; Borrego et al. 2022).

8 Conclusion

In this survey, we performed an extensive analysis of SLR tools, with a particular focus on the integration of AI technologies in the screening and extraction phases. Our study includes a detailed evaluation of 21 tools, examining them across 11 AI-specific features and 23 general features. The analysis extended to 11 additional applications that leverage LLMs to aid researchers in retrieving research papers and supporting the writing process. Throughout the survey, we critically discussed the strengths and weaknesses of existing solutions, identifying which tools are most suitable for specific use cases. We also explored the main research challenges and the emerging opportunities that AI technologies present in this field.

Our findings paint an exciting picture of the current state of SLR tools. We observed that the existing generation of tools, when used effectively, can be highly powerful. However, they often fall short in terms of usability and user-friendliness, limiting their adoption within the broader research community. Concurrently, a new generation of tools based on LLMs is rapidly developing. While promising, these tools are still in their infancy and face challenges, such as the well-documented issue of hallucinations in LLMs. This highlights the need for the research community to focus on knowledge injection and RAG strategies to ensure the generation of robust and verifiable information.

The challenges identified in our survey represent a vibrant and evolving area of interest for researchers. It is anticipated that in the next five years, we may see the emergence of a novel generation of AI-enabled research assistants based on LLMs. These AI-enabled research assistants could support researchers by performing a variety of crucial

Table 5 Literature review tools based on LLMs

ID	Tool	Mode	Type	Website
1	Scite. [Brody (2021), Rife et al. (2021), Nicholson et al. (2021)]	Web	Search engine	https://scite.ai/
2	Elicit. [Kung (2023)]	Web	Search engine	https://elicit.com/
3	Consensus	Web	Search engine	https://consensus.app/
4	EvidenceHunt	Web	Search engine	https://evidencehunt.com/
5	MirrorThink	Web	Search engine	https://mirrorthink.ai/
6	Perplexity	Web/App	Search engine	https://www.perplexity.ai/
7	Scispace	Web	Search engine	https://typeset.io/
8	Jenni.ai	Web/App	Writing assistant	https://jenni.ai/
9	ResearchBuddies	Web	Writing assistant	https://researchbuddy.app/
10	Silatus	Web	Writing assistant	https://silatus.com/
11	Textero.ai	Web	Both	https://textero.ai/

tasks such as generating comprehensive literature reviews, identifying new scientific hypotheses, and fostering crucial innovation in research practices. The research community bears the crucial task of steering the growth of AI, minimising bias, and upholding strict ethical standards. With the AI revolution impacting many fields, it is essential to remember that human critical thinking and creativity are still vital and remain a core responsibility of the researchers.t

Appendix A: Systematic literature review tools analysed through AI and generic features

In this appendix, we report three tables that describe the 21 systematic literature review tools examined according to both generic and AI-based features. In Appendix Tables 6 and 7, we present the analysis of the AI features for the screening and the extraction phases, respectively. In Appendix Table 8, we report the analysis of the tools according to the generic features. Due to space constraints, only a summarised version of these tables is included here. The full version is available online on both GitHub (https://angelosalatino.github.io/ai-slr/) and the Open Research Knowledge Graph (https://doi.org/10.48366/R692116).

Table 6 Screening phase of systematic literature review tools analysed through AI features

Tool	Research field	SLR task	Text representation	Input	Minimum requirement
Abstrackr	Any	Classification of relevant papers.	Bag of words.	Title & Abstract	–
ASReview	Any	Classification of relevant papers.	Bag of words. Embeddings: SentenceBERT, doc2vec.	Title & Abstract	Relevant papers: 1. Irrelevant papers: 1.
Colandr	Any	Task 1: Classification of relevant papers. Task 2: Identification of the category attributed to the paper by the user.	Task 1: Embeddings: Word2vec. Task 2: Embeddings: Glove	Task 1: Title & Abstract Task 2: Full content	Task 1: 10 relevant papers and 10 irrelevant papers. Task 2: Minimum 50 papers.
Covidence	Any	Task 1: Classification of relevant papers. Task 2: Identification of biomedical studies (RCTs).	Bag of words for both tasks: ngrams.	Task 1: Title & Abstract Task 2: Title & Abstract	Task 1: 2 relevant papers and 2 irrelevant papers. Task 2: Not Applicable.
DistillerSR	Any	Classification of relevant papers.	Bag of words.	Title & Abstract	Relevant papers: 10. Irrelevant papers: 40.
EPPI-Reviewer	Any	Task 1: Classification of relevant papers. Task 2: Identification of biomedical studies (RCTs, Systematic Reviews, Economic Evaluations, COVID-19 categories, long COVID).	Task 1: Bag of words (ngrams). Task 2: The Cochrane RCT classifer uses bag of words. For the other approaches the information is not available.	Task 1: Title & Abstract Task 2: Title & Abstract	Task 1: 5 relevant papers. Number of irrelevant papers not available. Task 2: Not Applicable
FAST2	Any	Classification of relevant papers.	Bag of words.	Title & Abstract	–
Iris.ai	Any	Clustering of Abstracts	Embeddings.	Title & Abstract	Not Applicable
LitSuggest	Biomedicine	Classification of relevant papers.	Bag of words.	Title & Abstract	–
Nested knowledge	Any	Classification of relevant papers.	–	Title & Abstract	–

Table 6 (continued)

Tool	Research field	SLR task	Text representation	Input	Minimum requirement
PICOPortal	Any	Task 1: Classification of relevant papers. Task 2: Identification of biomedical studies (RCTs).	Embeddings for Task 2: BioBERT. No information regardin Task 1.	Task 1: Title & Abstract Task 2: Title & Abstract	–
pitts.ai	Biomedicine	Identification of biomedical studies (RCTs).	Embeddings: SciBERT	Title & Abstract	Not Applicable
Rayyan	Any	Classification of relevant papers.	Bag of words: ngrams	Title & Abstract	Relevant papers: 5. Irrelevant papers: 5.
Research screener	Any	Classification of relevant papers.	Embeddings: paragraph embedding	Title & Abstract	Relevant papers: 1. Irrelevant papers: Information not available.
RobotAnalyst	Any	Classification of relevant papers.	Bag of words.	Title & Abstract	–
RobotReviewer/RobotSearch	Biomedicine	Identification of biomedical studies (RCTs).	Embeddings: SciBERT	Title & Abstract	Relevant papers: NA. Irrelevant papers: NA.
SWIFT-active screener	Any	Classification of relevant papers.	Bag of words.	Title & Abstract	Relevant papers: 1. Irrelevant papers:1.
SWIFT-review	Biomedicine	Classification of relevant papers.	Bag of words.	Title & Abstract	Relevant papers: 1. Irrelevant papers:1.
SysRev.com	Any	Classification of relevant papers.	–	Title & Abstract	Relevant papers: 30. Irrelevant papers: 30.

Table 7 Extraction phase of systematic literature review tools analysed through AI features

Tool	Research field	SLR task	Approach	Text representation	Input	Output
RobotReviewer/ RobotSearch	Biomedical	Identifies risks of bias: how reliable are the results?	ML classifier, combining a lineal model and a Convolutional Neural Network (CNN) model. These models are trained on a dataset containing manually annotated sentences stating the level of bias.	Bag of word: ngrams. Embeddings: embedding layer from CNN Model.	Full-text paper.	Risk of bias classification (as Low, High, Unclear)
ExaCT	Biomedical	NER of Randomised Controlled Trials	Task 1: ML classifier based on SVM to identify sentences regarding a control trial. Task 2: Rule base detection to identify the 21 CONSORT categories.	Bag of words: ngrams.	Full-text paper.	Possible RCT entities
Dextr	Environmental Health Science	Task 1: NER of animal studies. Task 2: Entity linking of animal studies.	Task 1: ML Classifier implementing a neural network model based on bidirectional LSTM with a Conditional Random Field (BI-LSTM-CRF) architecture. Task 2: Linking according to a customised ontology	Task 1: Embeddings: GloVe, ELMo. Task 2: Not Applicable.	Title and Abstracts	Task 1: Possible animal entities. Task 2: Relationships of animal models and exposures vs experimentas ot endpoints vs experiments.
Iris.ai	Any	Task 1: NER of entities selected by the user. Task 2: Entity linking of the identified entities.	Task 1: ML classifier. Algorithim is unknown. Task 2: Uses a knowledge graph to represent the relations of within the entities on the paper or between the entities of the table. The technical implementation is unknown.	Task 1: Embeddings: word embedding. Task 2: Not Applicable.	Full-text paper.	Task 1: Possible entities based on a confidence interval. Task 2: Additional semantics on the extracted entities.

Table 8 Systematic literature review tools analysed based on general features

Tool	Multiple user roles	Multiple user support	Project auditing	Automated full-text retrieval	Automated search	Snowballing	Manual reference importing	Manually inserting full-text	Deduplication	Discrepancy resolving	In-/excluding references	Reference labelling & comments	Flow diagram creation	Protocol	Living/updatable	Free to use
Abstrackr	Sing.	2	Yes	No	None	No	Yes	No	No	Yes	No	Yes	No	No	No	Yes
Colandr	Sing.	2	No	No	None	No	No	No	No	Yes	Yes	Yes	No	Yes	No	Yes
DistillerSR	Mult.	>1	Yes	Yes	PubMed	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
EPPI-Reviewer	Mult.	>1	Yes	No	PubMed	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No
LitSuggest	Sing.	No	No	No	PubMed	No	No	No	No	No	No	No	No	No	Yes	Yes
Nested Knowledge	Mult.	>1	Yes	Yes	PubMed; Europe PMC; DOAJ; ClinicalTrials.gov	No	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Rayyan	Mult.	>1	Yes	No	None	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Yes
RobotAnalyst	Sing.	No	No	No	PubMed	No	Yes	No	No	No	Yes	No	No	No	No	Yes
SWIFT-Active Screener	Mult.	>1	Yes	No	None	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No
SWIFT-Review	Sing.	No	No	No	None	No	No	No	No	No	No	No	No	No	No	Yes
FAST2	Sing.	No	No	No	None	No	No	No	No	No	No	No	No	No	No	Yes
ASReview	Sing.	>1	No	No	None	No	No	No	No	No	Yes	No	No	No	No	Yes
Research Screener	Mult.	>1	No	No	None	No	No	No	Yes	Yes	Yes	No	No	No	No	Yes
pitts.ai	Mult.	>1	No	No	PubMed	No	No	No	No	Yes	Yes	No	No	No	No	No
SysRev.com	Mult.	>1	Yes	No	PubMed	No	No	Yes	No	Yes	Yes	Yes	No	No	No	No
Covidence	Mult.	>1	No	No	None	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No

Table 8 (continued)

Tool	Multiple user roles	Multiple user support	Project auditing	Automated full-text retrieval	Automated search	Snowballing	Manual reference importing	Manually inserting full-text	Deduplication	Discrepancy resolving	In-/excluding references	Reference labelling & comments	Flow diagram creation	Protocol	Living/updatable	Free to use
RobotReviewer / RobotSearch	Sing.	No	No	No	None	No	No	No	No	No	No	No	No	No	No	Yes
Iris.ai	Sing.	No	Yes	No	CORE; PubMed; US Patent Office; CORDIS	No	No	No	No	No	No	No	No	No	No	No
PICO Portal	Mult.	>1	Yes	Yes	None	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes
Dextr	Sing.	No	No	NA	None	NA	NA	NA	NA	No	NA	NA	NA	NA	NA	Yes
ExaCT	Sing.	No	No	NA	None	NA	NA	NA	NA	No	NA	NA	NA	NA	NA	Yes

Acknowledgements We would like to express our gratitude to the developers of the following tools for providing additional information via email or in personal interviews: Covidence, DistillerSR, Nested Knowledge, Pitts.ai, PICOPortal, Rayyan, SWIFT-Active Screener, SWIFT-Reviewer, and SysRev.

References

Abbasi K (2023) A commitment to act on data sharing. BMJ 382 https://doi.org/10.1136/bmj.p1609

Adam GP, Wallace BC, Trikalinos TA (2022) Semi-automated tools for systematic searches. Meta-Res: Methods Protocols, 17–40 (2022)

Agai E (2020) A new machine-learning powered tool to aid citation screening for evidence synthesis: Picoportal. Advances in evidence synthesis: special issue. Cochrane Database Syst Rev 9(suppl 1), 172

Aguilera Cora E, Lopezosa C, Codina L (2024) Scopus AI beta: functional analysis and cases

Ali Z, Kefalas P, Muhammad K, Ali B, Imran M (2020) Deep learning in citation recommendation models survey. Expert Syst Appl 162:113790

Allot A, Lee K, Chen Q, Luo L, Lu Z (2021) Litsuggest: a web-based system for literature recommendation and curation using machine learning. Nucleic Acids Res 49(W1):352–358

Altmami NI, Menai MEB (2022) Automatic summarization of scientific articles: a survey. J King Saud Univ—Comput Inf Sci 34(4):1011–1028

Al-Zubidy A, Carver JC, Hale DP, Hassler EE (2017) Vision for SLR tooling infrastructure: prioritizing value-added requirements. Inf Softw Technol 91:72–81

Ampatzoglou A, Bibi S, Avgeriou P, Verbeek M, Chatzigeorgiou A (2019) Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Inf Softw Technol 106:201–230

Angioni S, Salatino A, Osborne F, Recupero DR, Motta E (2021) AIDA: a knowledge graph about research dynamics in academia and industry. Quant Sci Stud 2(4):1356–1398. https://doi.org/10.1162/qss_a_00162

Arno A, Elliott J, Wallace B, Turner T, Thomas J (2021) The views of health guideline developers on the use of automation in health evidence synthesis. Syst Rev 10:1–10

Ayling J, Chapman A (2022) Putting AI ethics to work: are the tools fit for purpose? AI Ethics 2(3):405–429

Baas J, Schotten M, Plume A, Côté G, Karimi R (2020) Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quant Sci Stud 1(1):377–386

Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, Ouzzani M, Thayer K, Thomas J, Turner T et al (2018) Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (ICASR). Syst Rev 7:1–7

Beltagy I, Lo K, Cohan A (2019) Scibert: a pretrained language model for scientific text. arXiv preprint arXiv:1903.10676

Bernard N, Balog K (2023) A systematic review of fairness, accountability, transparency and ethics in information retrieval. ACM Comput Surv

Besta M, Blach N, Kubicek A, Gerstenberger R, Gianinazzi L, Gajda J, Lehmann T, Podstawski M, Niewiadomski H, Nyczyk P, Hoefler T (2023) Graph of thoughts: solving elaborate problems with large language models

Birkland TA (2019) An introduction to the policy process: theories, concepts, and models of public policy making. Routledge

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Borah R, Brown AW, Capers PL, Kaiser KA (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):012545

Bornmann L, Mutz R (2015) Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. J Assoc Inf Sci Technol 66(11):2215–2222
Borrego A, Dessi D, Hernández I, Osborne F, Recupero DR, Ruiz D, Buscaldi D, Motta E (2022) Completing scientific facts in knowledge graphs of research concepts. IEEE Access 10:125867–125880
Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, pp 144–152
Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D et al (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the stard initiative
Bozada T Jr, Borden J, Workman J, Del Cid M, Malinowski J, Luechtefeld T (2021) Sysrev: a fair platform for data curation and systematic evidence review. Front Artif Intell 4:685298
Brockmeier AJ, Mu T, Ananiadou S, Goulermas JY (2018) Self-tuned descriptive document clustering using a predictive network. IEEE Trans Knowl Data Eng 30(10):1929–1942
Brody S (2021) Scite. J Med Libr Assoc 109(4):707
Brooke J et al (1996) Sus—a quick and dirty usability scale. Usability Evaluation Ind 189(194):4–7
Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
Brynjolfsson E, Li D, Raymond LR (2023) Generative AI at work. Technical report, National Bureau of Economic Research
Burgard, T., Bittermann, A.: Reducing literature screening workload with machine learning. Zeitschrift für Psychologie (2023)
Buscaldi D, Dessí D, Motta E, Murgia M, Osborne F, Recupero DR (2024) Citation prediction by leveraging transformers and natural language processing heuristics. Inf Process Manag 61(1):103583
Carver JC, Hassler E, Hernandes E, Kraft NA (2013) Identifying barriers to the systematic literature review process. 2013 ACM/IEEE international symposium on empirical software engineering and measurement. IEEE, pp 203–212
Castelvecchi D (2016) Can we open the black box of AI? Nat News 538(7623):20
Chai KE, Lines RL, Gucciardi DF, Ng L (2021) Research screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev 10:1–13
Chen Z, He L, Liu S, Liu H, Yu J, Li Y (2022) The grading of recommendations, assessment, development, and evaluation approach was rarely followed and inconsistently applied in pressure injury prevention guidelines development: a cross-sectional survey. J Tissue Viability 31(3):438–443
Cheng S, Augustin C (2021) Keep a human in the machine and other lessons learned from deploying and maintaining Colandr. Chance 34(3):56–60
Cheng S, Augustin C, Bethel A, Gill DA, Anzaroot S, Brun JL, Dewilde B, Minnich R, Garside R, Masuda YJ, Miller DC, Wilkie DS, Wongbusarakum S, McKinnon MC (2018) Using machine learning to advance synthesis and use of conservation and environmental evidence. Conserv Biol 32
Cierco Jimenez R, Lee T, Rosillo N, Cordova R, Cree IA, Gonzalez A, Indave Ruiz BI (2022) Machine learning computational tools to assist the performance of systematic reviews: a mapping review. BMC Med Res Methodol 22(1):1–14
Cowie K, Rahmatullah A, Hardy N, Holub K, Kallmes K et al (2022) Web-based software tools for systematic literature review in medicine: systematic search and feature analysis. JMIR Med Inform 10(5):33219
Craig D, Rice S (2007) NHS economic evaluation database handbook. Centre for Reviews and Dissemination, York
Dawes, M., Pluye, P., Shea, L., Grad, R., Greenberg, A., Nie, J.-Y.: The identification of clinically important elements within medical journal abstracts: Patient–population–problem, exposure–intervention, comparison, outcome, duration and results (pecodr). Informatics in Primary care 15(1) (2007)
de la Torre-López J, Ramírez A, Romero JR (2023) Artificial intelligence to automate the systematic review of scientific literature. Computing: 1–24
Dell NA, Maynard BR, Murphy AM, Stewart M (2021) Technology for research synthesis: an application of sociotechnical systems theory. J Soc Soc Work Res 12(1):201–222
Dessí D, Osborne F, Reforgiato Recupero D, Buscaldi D, Motta E (2022) Cs-kg: a large-scale knowledge graph of research entities and claims in computer science. In: Sattler U, Hogan A, Keet M, Presutti V, Almeida JPA, Takeda H, Monnin P, Pirrò G, d'Amato C (eds) The Semantic Web—ISWC 2022. Springer, Cham, pp 678–696
Di Nunzio GM, Kanoulas E, Majumder P (2022) Augmented intelligence in technology-assisted review systems (altars 2022): evaluation metrics and protocols for ediscovery and systematic review systems. European conference on information retrieval. Springer, pp 557–560

Dieste O, Grimán A, Juristo N (2009) Developing search strategies for detecting relevant experiments. Empir Softw Eng 14:513–539
Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C (2014) Content-based citation analysis: the next generation of citation analysis. J Am Soc Inf Sci 65(9):1820–1833
Dunn A, Dagdelen J, Walker N, Lee S, Rosen AS, Ceder G, Persson K, Jain A (2022) Structured information extraction from complex scientific text with fine-tuned large language models. arXiv preprint arXiv:2212.05238 (2022)
Feng L, Chiam YK, Lo SK (2017) Text-mining techniques and tools for systematic literature reviews: a systematic literature review. 2017 24th Asia-Pacific Software Engineering Conference (APSEC). IEEE, pp 41–50
Finstad K (2010) The usability metric for user experience. Interact Comput 22(5):323–327. https://doi.org/10.1016/j.intcom.2010.04.004
Fontaine G, Maheu-Cadotte M-A, Lavallee A, Mailhot T, Lavoie P, Rouleau G, Vinette B, García M-PR, Bourbonnais A (2022) Designing, planning, and conducting systematic reviews and other knowledge syntheses: six key practical recommendations to improve feasibility and efficiency. Worldviews Evid-Based Nurs 19(6):434–441
Garousi V, Felderer M (2017) Experience-based guidelines for effective and efficient data extraction in systematic reviews in software engineering. In: Proceedings of the 21st international conference on evaluation and assessment in software engineering, pp 170–179
Glanville J, Dooley G, Wisniewski S, Foxlee R, Noel-Storr A (2019) Development of a search filter to identify reports of controlled clinical trials within cinahl plus. Health Info Lib J 36(1):73–90
Google (n.d.) https://drive.google.com/file/d/1gsg8s8WGrTETJxL3dL2eqzPowxbwcr37/view
Gough D, Thomas J, Oliver S (2017) An introduction to systematic reviews. SAGE Publications Ltd, London
Grant MJ, Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies. Health Inf Libraries J 26(2):91–108
Groth P, Gibson A, Velterop J (2010) The anatomy of a nanopublication. Inf Serv Use 30(1–2):51–56
Haddaway NR, Callaghan MW, Collins AM, Lamb WF, Minx JC, Thomas J, John D (2020) On the use of computer-assistance to facilitate systematic mapping. Campbell Syst Rev 16(4):1129
Hannousse A (2021) Searching relevant papers for software engineering secondary studies: semantic scholar coverage and identification role. IET Softw 15(1):126–146
Harrison H, Griffin SJ, Kuhn I, Usher-Smith JA (2020) Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. BMC Med Res Methodol 20:1–12
Hassler E, Carver JC, Hale D, Al-Zubidy A (2016) Identification of SLR tool needs-results of a community workshop. Inf Softw Technol 70:122–129
Hassler E, Carver JC, Kraft NA, Hale D (2014) Outcomes of a community workshop to identify and rank barriers to the systematic literature review process. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering, pp 1–10
Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intell Syst Their Appl 13(4):18–28
Higgins J (2011) Cochrane handbook for systematic reviews of interventions. version 5.1. 0 [updated march 2011]. The Cochrane collaboration. www. cochrane-handbook. org
Higgins JP, Altman DG (2008) Assessing risk of bias in included studies. Cochrane handbook for systematic reviews of interventions: Cochrane book series, pp 187–241
Hitzler P (2021) A review of the semantic web field. Commun ACM 64(2):76–83
Hope T, Downey D, Weld DS, Etzioni O, Horvitz E (2023) A computational inflection for scientific discovery. Commun ACM 66(8):62–73
Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, Holmgren S, Pelch KE, Walker V, Rooney AA et al (2016) Swift-review: a text-mining workbench for systematic review. Syst Rev 5:1–16
Howard BE, Phillips J, Tandon A, Maharana A, Elmore R, Mav D, Sedykh A, Thayer K, Merrick BA, Walker V et al (2020) Swift-active screener: accelerated document screening through active learning and integrated recall estimation. Environ Int 138:105623
Iansiti M, Lakhani KR (2020) Competing in the age of AI: strategy and leadership when algorithms and networks run the world. Harvard Business Press, Boston
Jaradeh MY, Oelen A, Farfar KE et al (2019) Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th international conference on knowledge capture, pp 243–246
Jesso ST, Kelliher A, Sanghavi H, Martin T, Parker SH (2022) Inclusion of clinicians in the development and evaluation of clinical artificial intelligence tools: a systematic literature review. Front Psychol 13

Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of hallucination in natural language generation. ACM Comput Surv 55(12):1–38
Justitia A, Wang H-C (2022) Automatic related work section in scientific article: Research trends and future directions. 2022 International Seminar on Intelligent Technology and Its Applications (ISITIA). IEEE, pp 108–114
Kebede MM, Le Cornet C, Fortner RT (2023) In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. Res Synth Methods 14(2):156–172
Keele S, et al (2007) Guidelines for performing systematic literature reviews in software engineering. Technical report, ver. 2.3 EBSE technical report. EBSE
Khalil H, Ameen D, Zarnegar A (2022) Tools to support the automation of systematic reviews: a scoping review. J Clin Epidemiol 144:22–42
Kim SN, Martinez D, Cavedon L, Yencken L (2011) Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics 12:1–10
Kiritchenko S, De Bruijn B, Carini S, Martin J, Sim I (2010) Exact: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inform Decis Mak 10:1–17
Kohl C, McIntosh EJ, Unger S, Haddaway NR, Kecke S, Schiemann J, Wilhelm R (2018) Online tools supporting the conduct and reporting of systematic reviews and systematic maps: a case study on cadima and review of existing tools. Environ Evid 7(1):1–17
Krichen M (2023) Convolutional neural networks: a survey. Computers 12(8):151
Kung J (2023) Elicit (product review). J Can Health Libr Assoc/Journal de l'Association des bibliothèques de la santé du Canada 44(1)
Kusa W, Lipani A, Knoth P, Hanbury A (2023) An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. Intell Syst Appl 18:200193
Kusa W, Knoth P, Hanbury A (2022) Evaluation of automated citation screening in systematic literature reviews with work saved over sampling: an analysis. In: 1st workshop on augmented intelligence for technology-assisted reviews systems: evaluation metrics and protocols for ediscovery and systematic review systems, pp 1–7
La Toile Q (2004) Database of abstracts of reviews of effects (dare). Douleurs 5(2)
Larsson S, Heintz F (2020) Transparency in artificial intelligence. Internet Policy Rev 9(2)
Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. In: HCI and usability for education and work: 4th symposium of the workgroup human-computer interaction and usability engineering of the Austrian Computer Society, USAB 2008, Graz, 20–21 Nov 2008. Proceedings, vol. 4. Springer, pp 63–76
Lawrence J, Reed C (2020) Argument mining: a survey. Comput Linguist 45(4):765–818
Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: International Conference on Machine Learning. PMLR, pp 1188–1196
Lewis JR (2018) The system usability scale: past, present, and future. Int J Hum–Comput Interaction 34(7):577–590
Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 33:9459–9474
Li Q, Li J, Sheng J, Cui S, Wu J, Hei Y, Peng H, Guo S, Wang L, Beheshti A et al (2022) A survey on deep learning event extraction: approaches and applications. IEEE Trans Neural Netw Learn Syst
Linardatos P, Papastefanopoulos V, Kotsiantis S (2020) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):18
Li X, Ouyang J (2022) Automatic related work generation: a meta study. arXiv preprint arXiv:2201.01880
Lipscomb CE (2000) Medical subject headings (mesh). Bull Med Libr Assoc 88(3):265
Liu J, Timsina P, El-Gayar O (2018) A comparative analysis of semi-supervised learning: the case of article selection for medical systematic reviews. Inf Syst Front 20:195–207
Liu P, Gao W, Dong W, Huang S, Zhang Y (2022) Open information extraction from 2007 to 2022—a survey. arXiv preprint arXiv:2208.08690
Long J (2023) Large language model guided tree-of-thought
Lortie CJ, Braun J, Filazzola A, Miguel F (2020) A checklist for choosing between r packages in ecology and evolution. Ecol Evol 10(3):1098–1105
Machine learning functionality in EPPI-reviewer. https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf

Marshall IJ, Kuiper J, Banner E, Wallace BC (2017) Automating biomedical evidence synthesis: RobotReviewer. In: Proceedings of the conference: association for computational linguistics, meeting, vol. 2017. NIH Public Access, p. 7

Marshall C, Kitchenham B, Brereton P (2018) Tool features to support systematic reviews in software engineering—a cross domain study. e-Inf Softw Eng J 12(1):79–115

Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC (2018) Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Research synthesis methods 9(4):602–614

Marshall C, Brereton P (2015) Systematic review toolbox: a catalogue of tools to support systematic reviews. In: Proceedings of the 19th international conference on evaluation and assessment in software engineering, pp 1–6

Marshall C, Brereton P, Kitchenham B (2014) Tools to support systematic reviews in software engineering: a feature analysis. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering, pp 1–10

Marshall C, Brereton P, Kitchenham B (2015) Tools to support systematic reviews in software engineering: a cross-domain survey using semi-structured interviews. In: Proceedings of the 19th international conference on evaluation and assessment in software engineering, pp 1–6

Meloni A, Angioni S, Salatino A, Osborne F, Recupero DR, Motta E (2023) Integrating conversational agents and knowledge graphs within the scholarly domain. IEEE Access 11:22468–22489

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 26

Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, Agirre E, Heintz I, Roth D (2023) Recent advances in natural language processing via large pre-trained language models: a survey. ACM Comput Surv 56(2):1–40

Minion JT, Egunsola O, Mastikhina L, Farkas B, Hofmeister M, Flanagan J, Salmon C, Clement F (2021) Pico portal. J Can Health Libraries Assoc 42(3):181

Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S (2014) Reducing systematic review workload through certainty-based screening. J Biomed Inform 51:242–253

Moher D, Schulz KF, Altman DG (2001) The consort statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Lancet 357(9263):1191–1194

Moher D, Tsertsvadze A, Tricco AC, Eccles M, Grimshaw J, Sampson M, Barrowman N (2007) A systematic review identified few methods and strategies describing when and how to update systematic reviews. J Clin Epidemiol 60(11):10951

Moher D, Liberati A, Tetzlaff J, Altman DG et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the Prisma statement. Ann Intern Med 151(4):264–269

Mourão E, Kalinowski M, Murta L, Mendes E, Wohlin C (2017) Investigating the use of a hybrid search strategy for systematic reviews. 2017 ACM/IEEE international symposium on empirical software engineering and measurement (ESEM). IEEE, pp 193–198

Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Percie du Sert N, Simonsohn U, Wagenmakers E-J, Ware JJ, Ioannidis J (2017) A manifesto for reproducible science. Nat Hum Behav 1(1):1–9

Munn Z, Tufanaru C, Aromataris E (2014) JBI's systematic reviews: data extraction and synthesis. AJN Am J Nurs 114(7):49–54

Myllyaho L, Raatikainen M, Männistö T, Mikkonen T, Nurminen JK (2021) Systematic literature review of validation methods for AI systems. J Syst Softw 181:111050

Napoleão BM, Petrillo F, Hallé S (2021) Automated support for searching and selecting evidence in software engineering: a cross-domain systematic mapping. 2021 47th Euromicro conference on Software Engineering and Advanced Applications (SEAA). IEEE, pp 45–53

Nasar Z, Jaffry SW, Malik MK (2021) Named entity recognition and relation extraction: state-of-the-art. ACM Comput Surveys 54(1):1–39

National Academies of Sciences Engineering Medicine (2019) Reproducibility and replicability in science. The National Academies Press, Washington, DC (2019). https://doi.org/10.17226/25303

Ng, J.Y., Maduranayagam, S.G., Lokker, C., Iorio, A., R., Haynes, B., Moher, D. (2023) Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: a protocol for a cross-sectional survey. In: medRxiv

Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 14

Nicholson JM, Mordaunt M, Lopez P, Uppala A, Rosati D, Rodrigues NP, Grabitz P, Rife SC (2021) Scite: asmart citation index that displays the context of citations and classifies their intent using deep learning. Quant Sci Stud 2(3):882–898

Noel-Storr A, Dooley G, Affengruber L, Gartlehner G (2021) Citation screening using crowdsourcing and machine learning produced accurate results: evaluation of Cochrane's modified screen4me service. J Clin Epidemiol 130:23–31

Nowak A, Kunstman P (2019) Team ep at tac 2018: automating data extraction in systematic reviews of environmental agents. arXiv preprint arXiv:1901.02081

Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal M-E, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E et al (2020) Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdisc Rev: Data Mining Knowl Discov 10(3):1356

O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Wolfe MS (2018) Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of international collaboration for the automation of systematic reviews (ICASR). Syst Rev 7:1–5

O'Connor AM, Tsafnat G, Gilbert SB, Thayer KA, Shemilt I, Thomas J, Glasziou P, Wolfe MS (2019) Still moving toward automation of the systematic review process: a summary of discussions at the third meeting of the international collaboration for automation of systematic reviews (ICASR). Syst Rev 8:1–5

O'Connor AM, Tsafnat G, Thomas J, Glasziou P, Gilbert SB, Hutton B (2019) A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? Syst Rev 8(1):1–8

O'Connor AM, Glasziou P, Taylor M, Thomas J, Spijker R, Wolfe MS (2020) A focus on cross-purpose tools, automated recognition of study design in multiple disciplines, and evaluation of automation tools: a summary of significant discussions at the fourth meeting of the international collaboration for automation of systematic reviews (icasr). Syst Rev 9(1):1–6

O'Connor D, Green S, Higgins JP (2008) Defining the review question and developing criteria for including studies. In: Cochrane handbook for systematic reviews of interventions: Cochrane book series, pp 81–94

National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Board on Research Data and Information; Committee on Toward an Open Science Enterprise (2018) Open science by design: realizing a vision for 21st century research. National Academies Press, Washington

Onan A, Korukoğlu S, Bulut H (2016) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57:232–247

OpenAI: GPT-4 technical report (2023)

Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A (2016) Rayyan-a web and mobile app for systematic reviews. Syst Rev 5:1–10

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE et al (2021) The Prisma 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 88:105906

Patel A, Jain S (2021) Present and future of semantic web technologies: a research statement. Int J Comput Appl 43(5):413–422

Peng C, Xia F, Naseriparsa M, Osborne F (2023) Knowledge graphs: opportunities and challenges. Artif Intell Rev: 1–32

Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543 (2014)

Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. ArXiv abs/1802.05365

Petticrew M, Roberts H (2008) Systematic reviews in the social sciences: a practical guide. John Wiley & Sons, Glasgow

Pranckutė R (2021) Web of science (WoS) and Scopus: the Titans of bibliographic information in today's academic world. Publications 9(1) (2021). https://doi.org/10.3390/publications9010012

Pride D, Cancellieri M, Knoth P (2023) Core-GPT: combining open access research and large language models for credible, trustworthy question answering. In: International conference on theory and practice of digital libraries. Springer, pp. 146–159

Project EPHP (1998) Quality assessment tool for quantitative studies. McMaster University, Hamilton, National Collaborating Centre for Methods and Tools

Przybyła P, Brockmeier AJ, Kontonatsios G, Le Pogam M-A, McNaught J, von Elm E, Nolan K, Ananiadou S (2018) Prioritising references for systematic reviews with RobotAnalyst: a user study. Res Synth Methods 9(3):470–488

Pullin AS, Stewart GB (2006) Guidelines for systematic review in conservation and environmental management. Conserv Biol 20(6):1647–1656

Quesenbery W (2014) The five dimensions of usability. Content and complexity. Routledge, New York, pp 93–114

Reimers N, Gurevych I (2019) Sentence-Bert: sentence embeddings using Siamese Bert-networks. arXiv preprint arXiv:1908.10084
Renda A, Arroyo J, Fanni R, Laurer M, Sipiczki A, Yeung T, Maridis G, Fernandes M, Endrodi G, Milio S et al (2021) Study to support an impact assessment of regulatory requirements for artificial intelligence in Europe. Brussels, European Commission
Rife SC, Rosati D, Nicholson JM (2021) scite: The next generation of citations. Learn Pub 34
Robinson A, Thorne W, Wu BP, Pandor A, Essat M, Stevenson M, Song X (2023) Bio-sieve: exploring instruction tuning large language models for systematic review automation. arXiv preprint arXiv:2308.06610
Robledo S, Grisales Aguirre AM, Hughes M, Eggers F (2023) "hasta la vista, baby''-will machine learning terminate human literature reviews in entrepreneurship? J Small Bus Manag 61(3):1314–1343
Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS (1996) Evidence based medicine: what it is and what it isn't. British Medical Journal Publishing Group
Salatino AA, Osborne F, Birukou A, Motta E (2019) Improving editorial workflow and metadata quality at springer nature. The Semantic Web—ISWC 2019. Springer, Cham, pp 507–525
Salatino A, Osborne F, Motta E (2022) Cso classifier 3.0: a scalable unsupervised method for classifying documents in terms of research topics. Int J Digit Libr: 1–20
Sanderson K (2023) AI science search engines are exploding in number—are they any good? Nature 616(7958):639–640
Schall MC Jr, Cullen L, Pennathur P, Chen H, Burrell K, Matthews G (2017) Usability evaluation and implementation of a health information technology dashboard of evidence-based quality indicators. CIN: Comput Inf Nurs 35(6):281–288
Schmidt L, Mohamed S, Meader N, Bacardit J, Craig D (2023) Automated data extraction of unstructured grey literature in health research: a mapping review of the current research literature. medRxiv, 2023–06
Schmidt L, Mutlu ANF, Elmore R, Olorisade BK, Thomas J, Higgins JPT (2023) Previously titled: data extraction methods for systematic review (semi)automation: a living systematic review. https://api.semanticscholar.org/CorpusID:235752381
Schmidt L, Olorisade BK, McGuinness LA, Thomas J, Higgins JP (2021) Data extraction methods for systematic review (semi) automation: a living systematic review. F1000Research 10
Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z (2021) Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol 138:80–94
Shah CA, Phadnis PN (2022) Text summarization using extractive and abstractive techniques. Int J Sci Res Comput Sci Eng Inf Technol
Shemilt I, Khan N, Park S, Thomas J (2016) Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. System Rev 5:1–13
Shojania KG, Sampson M, Ansari MT, Ji J, Doucette S, Moher D (2007) How quickly do systematic reviews go out of date? A survival analysis. Ann Intern Med 147(4):224–233
Sikka P, Mago V (2020) A survey on text simplification. arXiv preprint arXiv:2008.08612
Stahl BC, Antoniou J, Bhalla N, Brooks L, Jansen P, Lindqvist B, Kirichenko A, Marchal S, Rodrigues R, Santiago N, Warso Z, Wright D (2023) A systematic review of artificial intelligence impact assessments. Artif Intell Rev 56(11):12799–12831. https://doi.org/10.1007/s10462-023-10420-8
Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB et al (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA 283(15):2008–2012
Sumbul G, de Wall A, Kreuziger T, Marcelino F, Costa H, Benevides P, Caetano M, Demir B, Markl V (2021) Bigearthnet-mm: a large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. IEEE Geosci Remote Sens Mag 9(3):174–180. https://doi.org/10.1109/MGRS.2021.3089174
Sybrandt J, Tyagin I, Shtutman M, Safro I (2020) Agatha: automatic graph mining and transformer based hypothesis generation approach. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 2757–2764
Tagawa Y, Taniguchi M, Miura Y, Taniguchi T, Ohkuma T, Yamamoto T, Nemoto K (2019) Relation prediction for unseen-entities using entity-word graphs. In: Proceedings of the thirteenth workshop on graph-based methods for natural language processing (TextGraphs-13), pp 11–16
Team E (2007) Information resources group (IRG) workshop: pushing the frontiers of HTA information management. Evid Based Libr Inf Practice
Thomas J (2013) Diffusion of innovation in systematic review methodology: why is study selection not yet assisted by automation. OA Evid-Based Med 1(2):1–6

Thomas J, McDonald S, Noel-Storr A, Shemilt I, Elliott J, Mavergames C, Marshall IJ (2021) Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for cochrane reviews. J Clin Epidemiol 133:140–151
Thomas J, Brunton J, Graziosi S (2010) Eppi-reviewer 4.0: software for research synthesis. EPPI-Centre Software. Social Science Research Unit, Institute of Education, London
Tranfield D, Denyer D, Smart P (2003) Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Br J Manag 14(3):207–222
Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E (2014) Systematic review automation technologies. System Rev 3:1–15
Van Altena A, Spijker R, Olabarriaga S (2019) Usage of automation tools in systematic reviews. Res Synth Methods 10(1):72–82
Van De Schoot R, De Bruin J, Schram R, Zahedi P, De Boer J, Weijdema F, Kramer B, Huijts M, Hoogerwerf M, Ferdinands G et al (2021) An open source machine learning framework for efficient and transparent systematic reviews. Nat Mach Intell 3(2):125–133
van den Bulk LM, Bouzembrak Y, Gavai A, Liu N, van den Heuvel LJ, Marvin HJ (2022) Automatic classification of literature in systematic reviews on food safety using machine learning. Curr Res Food Sci 5:84–95
Van der Mierden S, Tsaioun K, Bleich A, Leenaars CH et al (2019) Software tools for literature screening in systematic reviews in biomedical research. Altex 36(3):508–517
Van Noorden R (2023) Chatgpt-like AIs are coming to major science search engines. Nature 620(7973):258–258
Visser M, Van Eck NJ, Waltman L (2021) Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, Crossref, and Microsoft academic. Quant Sci Stud 2(1):20–41
Vladika J, Matthes F (2023) Scientific fact-checking: a survey of resources and approaches. arXiv preprint arXiv:2305.16859
Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP (2007) The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. Lancet 370(9596):1453–1457
Wagner G, Lukyanenko R, Paré G (2022) Artificial intelligence and the conduct of literature reviews. J Inf Technol 37(2):209–226
Walker VR, Schmitt CP, Wolfe MS, Nowak AJ, Kulesza K, Williams AR, Shin R, Cohen J, Burch D, Stout MD et al (2022) Evaluation of a semi-automated data extraction tool for public health literature-based reviews: Dextr. Environ Int 159:107025
Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA (2012) Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, pp 819–824
Wang LL, DeYoung J, Wallace B (2022) Overview of mslr2022: a shared task on multi-document summarization for literature reviews. In: Proceedings of the third workshop on scholarly document processing
Wang S, Zhou W, Jiang C (2020) A survey of word embeddings based on deep learning. Computing 102:717–740
Wang R, Cheng R, Ford D, Zimmermann T (2023) Investigating and designing for trust in AI-powered code generation tools
Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Quart
Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2023) Chain-of-thought prompting elicits reasoning in large language models
Wells GA, Shea B, O'Connell D, Peterson J, Welch V, Losos M, Tugwell P et al (2000) The Newcastle–Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses
Wijkstra M, Lek T, Kuhn T, Welbers K, Steijaert M (2021) Living literature reviews. arXiv preprint arXiv:2111.00824
Wohlin C (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering, pp. 1–10
Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer
Wohlin C, Kalinowski M, Felizardo KR, Mendes E (2022) Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. Inf Softw Technol 147:106908

Wu R, Stauber V, Botev V, Elosua J, Brede A, Ritola M, Marinov K (2018) Scithon™ - an evaluation framework for assessing research productivity tools. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Paris

Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Chen, E.: Large language models for generative information extraction: a survey. arXiv preprint arXiv:2312.17617 (2023)

Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K (2023) Tree of thoughts: deliberate problem solving with large language models

Young J (2005) Research, policy and practice: why developing countries are different. J Int Dev 17(6):727–734

Yu Z, Menzies T (2019) Fast2: an intelligent assistant for finding relevant papers. Expert Syst Appl 120:57–71

Yu Z, Kraft NA, Menzies T (2018) Finding better active learners for faster literature reviews. Empir Softw Eng 23:3161–3186

Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. Int J Mach Learn Cybern 1:43–52

Zhou, Y., Zhang, H., Huang, X., Yang, S., Babar, M.A., Tang, H.: Quality assessment of systematic reviews in software engineering: a tertiary study. In: Proceedings of the 19th international conference on evaluation and assessment in software engineering, pp. 1–14 (2015)

Zuckarelli J (2023) packagefinder: comfortable search for r packages on CRAN directly from the R console. CRAN