Journal of the American Medical Informatics Association, 2025, **32(6)**, 1071–1086
https://doi.org/10.1093/jamia/ocaf063
Advance access publication 7 May 2025
Review

Review

The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review

Dmitry Scherbakov, PhD[1], Nina Hubig, PhD[1,2], Vinita Jansari, PhD[3], Alexander Bakumenko, MSc[3], Leslie A. Lenert, MD, MS*[,1]

[1]Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina (MUSC), Charleston, SC 29403, United States, [2]Interdisciplinary Transformation University, OG 2 A-4040 Linz, Austria, [3]School of Computing, Clemson University, Charleston, SC 29634, United States

*Corresponding author: Leslie A. Lenert, MD, MS, Biomedical Informatics Center, Department of Public Health Sciences, Medical University of South Carolina (MUSC), 22 WestEdge Street, Suite 200, Room WG213, Charleston, SC 29403, United States (lenert@musc.edu)

D. Scherbakov and N. Hubig contributed equally to this work.

Abstract

Objectives: This study aims to summarize the usage of large language models (LLMs) in the process of creating a scientific review by looking at the methodological papers that describe the use of LLMs in review automation and the review papers that mention they were made with the support of LLMs.

Materials and Methods: The search was conducted in June 2024 in PubMed, Scopus, Dimensions, and Google Scholar by human reviewers. Screening and extraction process took place in Covidence with the help of LLM add-on based on the OpenAI GPT-4o model. ChatGPT and Scite.ai were used in cleaning the data, generating the code for figures, and drafting the manuscript.

Results: Of the 3788 articles retrieved, 172 studies were deemed eligible for the final review. ChatGPT and GPT-based LLM emerged as the most dominant architecture for review automation (n = 126, 73.2%). A significant number of review automation projects were found, but only a limited number of papers (n = 26, 15.1%) were actual reviews that acknowledged LLM usage. Most citations focused on the automation of a particular stage of review, such as Searching for publications (n = 60, 34.9%) and Data extraction (n = 54, 31.4%). When comparing the pooled performance of GPT-based and BERT-based models, the former was better in data extraction with a mean precision of 83.0% (SD = 10.4) and a recall of 86.0% (SD = 9.8).

Discussion and Conclusion: Our LLM-assisted systematic review revealed a significant number of research projects related to review automation using LLMs. Despite limitations, such as lower accuracy of extraction for numeric data, we anticipate that LLMs will soon change the way scientific reviews are conducted.

Key words: large language models; review automation; systematic review; scoping review; Covidence.

Introduction

The abundance of scientific information available can be overwhelming, posing a challenge for researchers to navigate relevant data. Consequently, the number of scoping and systematic reviews helping scientists synthesize the evidence has increased significantly over the years. Toh and Lee noted an exponential rise in the number of scoping reviews, with 2665 published in 2020 alone, compared with fewer than 10 reviews published annually before 2009.[1] The same trend is observed in systematic reviews and meta-analyses. For example, in cardiology, over 2400 meta-analyses were published in 2019, quadrupling the number reported in 2012.[2]

The completion of a review requires substantial resources[3]; furthermore, there is often unpredictable uncertainty in the amount of resources required.[4] The time to complete a single systematic review varies, but authors typically give estimates in months and even years.[5] Screening automation platforms, such as Covidence,[6] facilitate systematic and scoping reviews by streamlining established guidelines and checklists, such as the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and Population, Intervention, Comparison, and Outcome (PICO) to ensure transparency and rigor in the review process.[7] The use of such platforms may reduce the time needed to complete reviews by providing tools that automate key tasks, such as removing duplicate references and generating flow charts of the screening process, visual extraction designers, and workflows for several independent reviewers.

Although, for example, Covidence includes features to reduce the time to complete screening, such as key term highlighting and embedded natural language processing (NLP) algorithm,[8] it primarily organizes the significant manual work

Received: February 12, 2025; **Revised:** April 2, 2025; **Editorial Decision:** April 6, 2025; **Accepted:** April 11, 2025

that is still needed from human reviewers, such as screening and extraction. Each of these steps typically involves 2 independent analysts, with a third optional human expert supervising the process and resolving the disagreements.

Even with 2 reviewers resolving disagreements through discussion, as many as 3% of relevant citations are missed, and if only a single reviewer is used (for example, in rapid reviews), as many as 13% of relevant publications can be missed.[9] The relatively weak performance of humans in screening relevant articles has led some investigators to develop natural language processing tools[10–13] to automate screening. A recent statement by the National Institute for Health and Care Excellence (NICE) highlights the potential and accompanying risks of artificial intelligence (AI) in the systematic review process automation.[14] Large language models (LLMs) have recently emerged as some of the most powerful NLP tools across different ranges of tasks,[15–17] which are reviewed in this publication.

While LLMs signify a notable advancement, earlier methodologies rooted in machine learning (ML) and natural language processing (NLP) laid the groundwork for assisted reviewing and annotation.[18] Prior methods to automate systematic reviews primarily focused on text classification and data extraction, aiming to decrease manual review burden and improve review efficiency.[10,19]

In title and abstract screening, ML techniques such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression trained on human-labeled data were prevalent, with tools such as Abstrackr and EPPI-Reviewer demonstrating workload reductions of approximately 40-50% while maintaining high recall (eg, 95% or higher).[20] Some studies indicated potential workload reductions up to 88-98% when using text mining as a second screener.[21] These tools typically prioritized citations by relevance probability, enabling reviewers to efficiently manage screening tasks by assessing highly relevant citations first. Active learning, a key approach, involved the machine learning from reviewer decisions on a subset of citations, strategically selecting the most informative citations for subsequent review, thus enhancing classifier accuracy with less human effort. Crowdsourcing, notably via platforms such as Cochrane Crowd, also contributed to eligibility assessment automation.[22]

For data extraction, earlier approaches relied extensively on rule-based systems, such as MedEx, which extracted specific clinical data from texts with reasonable accuracy.[10,18] Machine learning methods were explored but faced constraints, including limited availability of annotated datasets for training and validation.[19,23] For instance, RobotReviewer achieved automated risk of bias assessment with accuracy close to human performance, effectively identifying supporting textual evidence for bias judgments.[23]

Pre-LLM literature search strategies employed predominantly optimized Boolean queries, with emerging text mining approaches reducing manual screening workloads by approximately 30-70%, though sometimes at the cost of a 5% recall loss.[21] However, early text mining methods required considerable setup time, specialized technical expertise, and extensive collaboration or training for research teams.[21] Machine learning-based review automation systems faced significant dependence on high-quality labeled datasets, necessitating substantial manual annotation efforts, complicated further by limited public availability of clinical data due to privacy regulations like HIPAA.[10,18,22,23] Additionally, these systems exhibited poor transferability, with models specifically tailored to individual systematic reviews, leading to repeated resource-intensive training processes for each new project.[10,21,22] Rule-based systems, despite interpretability benefits, also lacked generalizability due to reliance on handcrafted, domain-specific rules developed collaboratively by experts and physicians.[18] Furthermore, systematic bias, particularly automation bias—where reliance on automated systems could lead to overlooked errors—posed an ongoing challenge to their adoption.[24] Thus, while pre-LLM text mining approaches showed potential for automating systematic review processes, their adoption was hindered by extensive initial investments, labeled data dependencies, limited model portability, systematic biases, and requisite domain-specific expertise.[10,18,21,24] These challenges underscored the advantages offered by advancements such as LLMs.

In summary, previous efforts leveraging NLP and ML in systematic review processes achieved significant efficiencies in screening and extraction stages, though adoption was constrained by concerns about bias, limited annotated datasets, and accuracy in highly specific tasks. LLMs have since qualitatively expanded these capabilities, building upon foundational work done by earlier NLP and ML systems.

In this systematic review, we evaluated the use of LLMs to assist with several components of the review process. The review aims to (1) summarize the current state-of-the-art research projects using LLMs to automate the review process, (2) look at the range of review types and review stages that are being automated, and (3) assess the performance of LLMs used for automation. This review aims to cover both, the review papers created with various degrees of LLM support, and methodological papers describing review automation with LLMs. As detailed below, we used LLM to assist with several key aspects of our review.

Methods

The study's research plan was formulated by the author team and the review was registered in the Open Science Framework (OSF) database.[25] The results are reported using the checklist provided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020, checklist is provided in Supplementary File S2).[26]

We decided that to be included in the review, citations had to be either reviews created with LLMs (and LLM usage was disclosed by authors) or methodological papers centered around the usage of LLMs in the automation of different phases of a systematic review. Only English-language journal publications, including conference abstracts and review publications that used LLMs in their creation, were considered.

Publications were excluded if they:

- Did not use some kind of LLM (eg, ChatGPT, Mistral, GPT-3.5, and BERT);
- Did not describe the automation of any stage of the review process;
- The paper was a review article itself that did not use LLM to conduct the review;
- The full text of the article could not be retrieved or was not published in English.

The initial search was conducted by a human reviewer (D. S.) in June 2024 using title and abstract fields in PubMed,

Table 1. Search strategy.

((“large language models” OR “large language model” OR “LLM” OR “LLMs” OR “ChatGPT” OR “GPT-3” OR “GPT-4” OR “LLaMA” OR “Mistral” OR “Mixtral” OR “BARD” OR “BERT” OR “Claude” OR “PaLM” OR “Gemini” OR “Copilot”) AND (“systematic review*” OR “scoping review*” OR “literature review*” OR “narrative review*” OR “umbrella review*” OR “rapid review*” OR “integrative review*” OR “evidence synthesis” OR “meta-analysis”))

Scopus, and Dimensions[27] databases and with default filters in Google Scholar. All publications were searched from the time of inception. Table 1 presents the search strategy for the databases.

All citations were then uploaded to Covidence. Covidence was used as a review protocol to track the progress of the study. The screening and extraction process took place in Covidence with the help of the LLM plugin for Covidence that our team developed. This plugin is used during the screening and extraction phases. The process of using the LLM plugin for screening and extraction is shown in Figure 1.

The developed add-on works by interacting with the Covidence platform programmatically via an intermediary software solution that was created in Python and R. The solution passes content between Covidence and the LLM OpenAI GPT-4o model provided by Microsoft Azure cloud service.[28] Once the LLM generates the response, a script automates actions in Covidence, such as clicking the Include/Exclude buttons or leaving notes.

The review process involved 3 stages that were automated by the Covidence add-on: abstract screening, full-text screening, and extraction. In each stage, 2 human reviewers were calibrated by screening a sample to refine the inclusion criteria and extraction categories. They then created and tested prompts for the LLM. LLM inference was programmed to run inference 3 times to determine the final decision (eg, “include” or “exclude”) based on the majority vote. Three prompts per phase are detailed in Table S1.

For the screening phases, a human-LLM consensus was reached through the process of using 2 human reviewers who first agreed and reached a human consensus on the subset of 100 abstracts (30 full-texts for the full-text screening phase), and then by comparing the results of their consensus against LLM votes, establishing a new human-LLM consensus (for instance, LLM can reveal false positives or false negatives in human consensus). LLM extraction precision was measured by a single human reviewer, and for categories with low precision (<80%), a manual reviewer was assigned to validate and correct LLM outputs. Benchmarks are provided in Tables S2-S4.

The data charting form for extraction was designed by human experts (D.S., V.J., A.B., L.L., and N.H.) and adopted into the LLM prompt to collect the following primary information:

- Author, year, title;
- Country and/or US state of the study;
- What types of reviews were automated;
- Stage of review automated in the research project;
- LLM type used;
- Performance metrics reported by authors during each stage of the review. In particular, accuracy, precision, recall, specificity, and F1 were extracted; if other metrics were used instead, they were grouped under the “Other metrics” category; if no metrics were reported, a “Not mentioned/qualitative” value was assigned.
- Number of samples (full-texts or abstracts) that authors used to compute their performance metrics;
- Brief information on how performance metrics were calculated;
- Brief information on reported timesaving;
- What was the general opinion of the study team on the usage of LLMs in review automation (positive, negative, or mixed) with a citation to support this viewpoint;
- Sources of the funding of the research project (public, private, mixed, or unknown);
- Is the paper an actual review that used LLMs or a methods paper?

Due to the diverse nature of publications and study designs, bias and quality assessment were not performed.

An LLM tool by Google NotebookLM (version from August 2024),[29] along with a manual review (D.S., V.J., and A.B.), was used to cross-check the extraction results for the fields where the precision of extraction was low (<0.8) during the benchmark. ChatGPT (4o model)[30] was used to clean the extraction data: the case was formatted, duplicates that were not identified automatically were removed, and similar entries were renamed to a common name. The data were manually fed into the chat window by a human reviewer (D.S.). Scite.ai (version from August 2024)[31] was used to draft parts of the introduction and discussion sections, whereas ChatGPT was used to draft the abstract and results section of this review by generating R code snippets to produce Figures 3-5. Frequency count was the main method to synthesize the results, and column plots were used to present the frequencies. A map figure was used to synthesize location data. To compare performance metrics, boxplots were used to display the median, interquartile range (IQR), whiskers (within 1.5 × IQR), and outliers (outside 1.5 × IQR) of the most frequently mentioned model types. However, mean and standard deviation were used when comparing and reporting performance metrics. Only studies reporting the metrics we specified in the extraction form were used in the comparison. Studies with other numeric metrics or qualitative metrics were not compared.

ChatGPT was used to draft the text of the results section, which was then corrected by our team where needed. As a result, approximately 40% of the Introduction, 90% of the Results, and 30% of the Discussion section were generated by different types of LLMs. Human experts edited and verified the final LLM-generated draft of the manuscript.

Additionally, we report the time savings and the computational costs in Supplementary File S1. We used our time measurements and reference data from experienced reviewers to calculate the time savings.[32]

Results

Figure 2 outlines the PRISMA article selection process for this study. Initially, 3788 studies were identified across several databases: PubMed (n = 2174), Scopus (n = 1207),

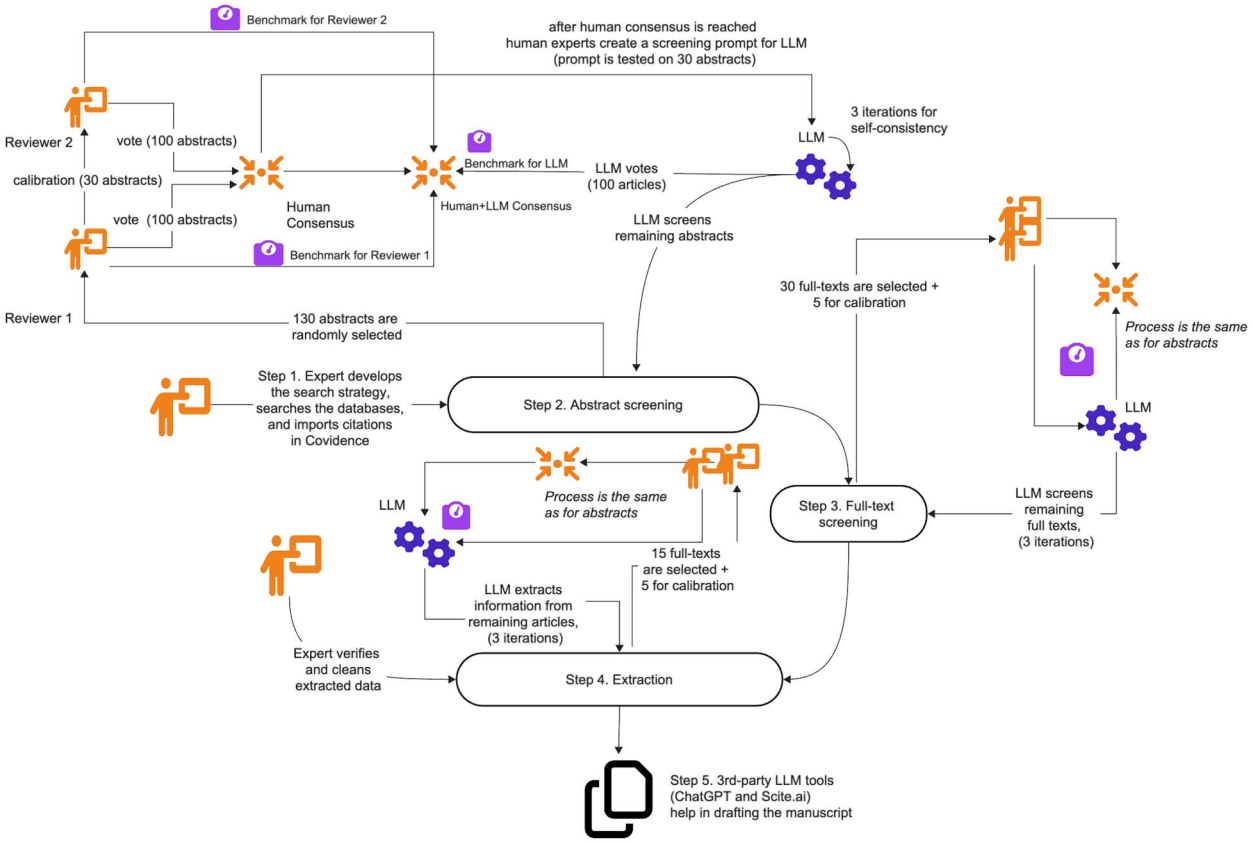

Figure 1. LLM workflow added into Covidence for screening and extraction.

Dimensions (n = 356), and Google Scholar (n = 48), along with 3 additional studies from citation searching. Following the removal of 447 duplicates (1 manually and 446 by Covidence), 3341 studies remained for the screening phase.

During the title and abstract screening process, 3041 studies were excluded, leaving 300 studies for retrieval and full-text eligibility assessment. Out of these 300 studies, 128 were excluded for various reasons, with the most common being "The paper does not describe the automation of any stage of the review process" (n = 88). A total of 172 studies were included in the final review.

Figure 3 shows the geographic distribution of studies across 43 countries. Most citations are from the United States (n = 60, 34.9%), followed by Australia (n = 14, 8.14%), the United Kingdom and China (n = 13, 7.6%), and Germany (n = 11, 6.4%). Other notable contributors include Canada (n = 7, 4.1%) and India (n = 6, 3.5%). Austria, Ireland, Italy, the Netherlands, and South Korea each contributed 4 studies (2.3%), while countries like New Zealand, France, Japan, and others provided 3 (1.7%). The rest contributed 1-2 studies.

In the United States, 47 studies had state-level data. Tennessee, New York, and Massachusetts led with 5 citations each (10.6%), followed by California (n = 4, 8.5%). North Carolina and Ohio contributed 3 studies (6.4%), while several other states provided 2 (4.3%) or 1 (2.1%) citations.

Figure 4A and Table 2 show the types of reviews discussed in automation papers. The most frequently mentioned type is "Systematic Review" (n = 118, 68.6%), followed by "Literature/Narrative Review" (n = 37, 21.5%) and "Meta-Analysis" (n = 19, 11.0%). The remaining categories include "Scoping Review" (n = 8, 4.7%), "Other/Non-specific" (n = 14, 8.1%), and "Rapid Review" (n = 6, 3.5%). "Umbrella Review" has a smaller representation with 2 mentions (1.2%).

Figure 4B and Table 2 illustrate the stages of review discussed in automation papers. The most frequently mentioned stage is "Searching for publications" (n = 60, 34.9%), followed by "Data extraction" (n = 54, 31.4%) and "Evidence synthesis/summarization" (n = 32, 18.6%). Other categories with notable mentions include "Title and abstract screening" (n = 43, 25.0%), "Drafting a publication" (n = 22, 12.8%), "Full-text screening" (n = 14, 8.1%), "Quality and bias assessment" (n = 12, 7.0%), "Publication classification" (n = 10, 5.8%), "Other stages" (n = 6, 3.5%), and "Code and plots generation" (n = 4, 2.3%).

The most frequently mentioned AI model is GPT/ChatGPT, with 126 occurrences (73.3%), showing its widespread use (Figure 5). BERT-based models are also notable with 32 mentions (18.6%). LLaMA/Alpaca models have 8 mentions (4.7%), followed by Google Bard/Gemini with 5 (2.9%) and Claude models with 7 (4.1%). Other models like BART (n = 3, 1.7%) and Mistral (n = 4, 2.3%) are less frequent. Several models, including Bing and XLNet, have 2 mentions each (1.2%), while many others are mentioned just once (0.6%).

Of the 172 citations, 79 (45.9%) reported common metrics like Accuracy, Precision/Recall, and F1, while 36 (20.9%) used less common metrics, such as G-score and Jaccard similarity. The remaining 57 publications (33.1%) relied on

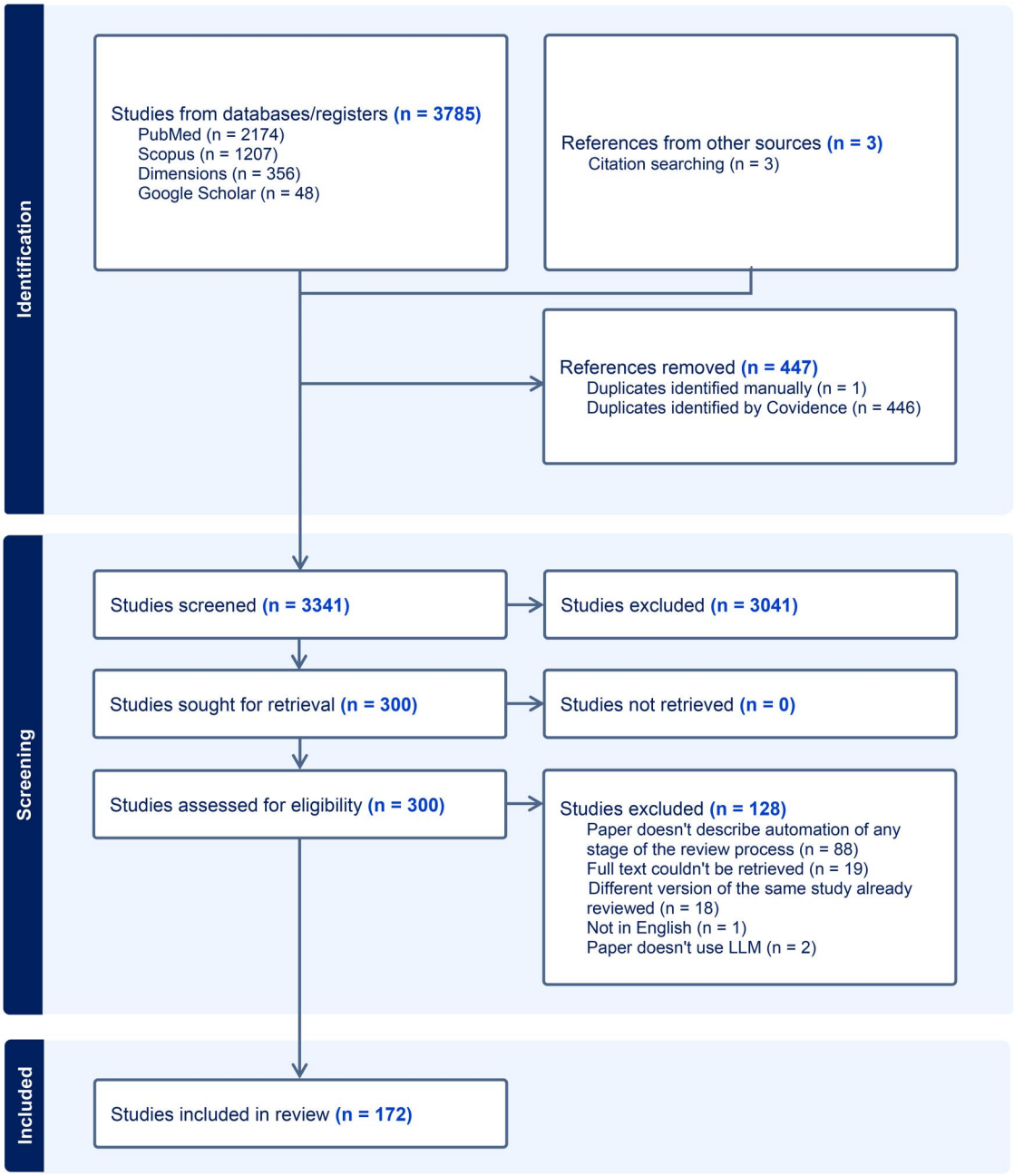

Figure 2. Flow diagram of the systematic review process.

qualitative assessments. Figure 6 shows the performance metrics for GPT- and BERT-based models. Based on the comparison of mean values across reviewed papers, GPT models had lower accuracy in title/abstract screening (M = 77.34, SD = 13.06) compared to BERT models (M = 80.87, SD = 11.81). However, GPT models performed better in data extraction, with precision (M = 83.07, SD = 10.43) and recall (M = 85.99, SD = 9.82), while BERT models had lower precision (M = 61.06, SD = 31.26) and similar recall (M = 80.03, SD = 10.09). In title/abstract screening, BERT models had higher precision (M = 65.6, SD = 17.65) but lower recall (M = 72.93, SD = 23.95) than GPT models (precision M = 63.2, SD = 24.34; recall M = 80.42, SD = 23.31).

The majority of the reviewed publications were papers describing how LLM could be used to automate a certain phase of the review (n = 146, 84.9%). Only 26 (15.1%) papers were actual reviews conducted with some help from LLM tools. Most authors were positive about the usage of LLMs in reviews (n = 120, 69.8%), with 43 citations (25.0%) containing mixed or cautious views on LLM usage. Only 9 (5.2%) study teams had negative experiences with LLM usage. More than half of the studies had public funding reported (n = 97, 56.4%).

Table S5 in presents the complete extraction table with all extracted categories across 172 citations.

Discussion

Our LLM-assisted systematic review revealed a significant number of research projects related to review automation

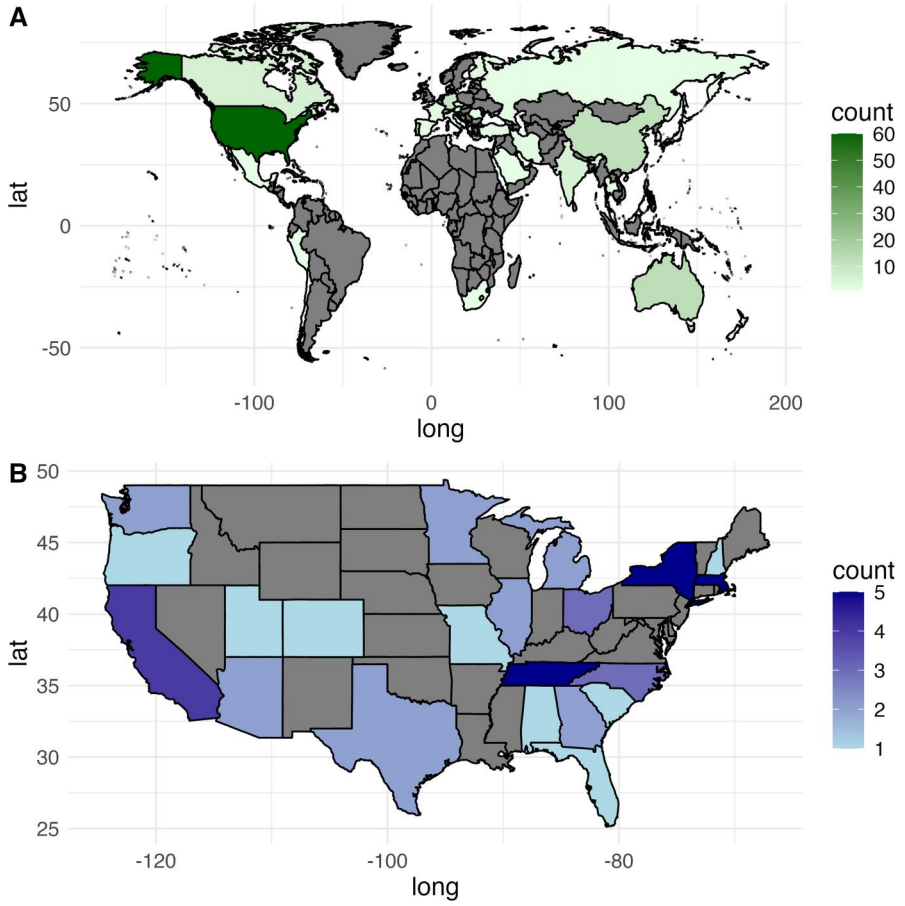

Figure 3. (A) Publications by country of origin. (B) Publications by state in the United States.

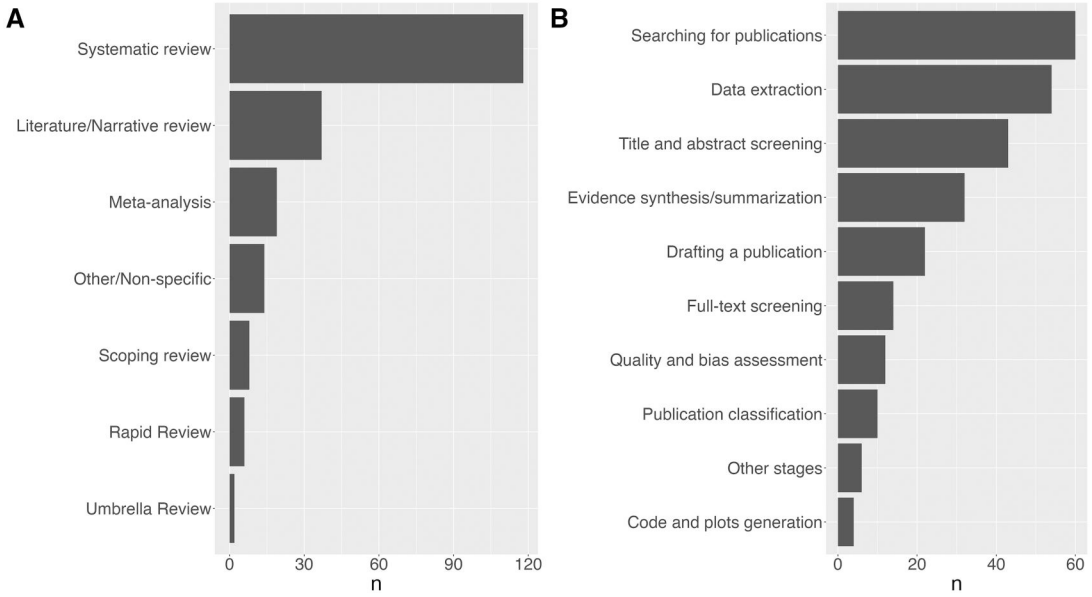

Figure 4. (A) Types of automated review. (B) Which stages of review are automated in the paper.

with LLM. Despite finding a significant number of projects using LLMs to automate some stages of the review process, only a few papers focused on the full cycle of review automation.[53,54] There might be perceived publication barriers; for example, journals have recently started to ask about LLM-generated content, although we do not have information on whether this leads to changes in the reviewing process. A growing number of LLM-generated papers will probably

Table 2. Summary table of reviewed citations.

Citations	Review stage	M/R	LLM	Review
Agarwal, 2024[33]; Anghelescu, 2023[34]; Bersenev, 2024[35]; Dossantos, 2023[36]; Jenko, 2024[37]; Khlaif, 2023[38]; Li, 2024[39]; Livberber, 2023[40]; Lozano, 2024[41]; Najafali, 2023[42]; Semrl, 2023[43]; Teperikidis, 2024[44]; Wang, 2024[45]; Wu, 2023[46]; Yun, 2023[47]; Zhao, 2024[48]	Drafting	M	GPT	Sys., N, Umbrella, Other, M
Yun, 2023[47]	Drafting	M	Other	Sys.
Huang, 2023[49]; Lamovšek, 2023[50]; Liu, 2023[51]; Pedroso-Roussado, 2023[52]; Schopow, 2023[53]; Teperikidis, 2023[54]	Drafting	R	GPT	Sys., Umbrella, Other, N
Ahmed, 2023[55]; Marshalova, 2023[56]; Mutinda, 2022[57]; Panayi, 2023[58]; Scells, 2023[59]; Shinde, 2022[60]; Wang, 2022[61]; Whitton, 2023[62]; Yazi, 2021[63]	Extraction	M	BERT	Other, Scop., Sys., M, N
Oami, 2024[64]; Prasad, 2024[65]; Ye, 2024[66]	Extraction	M	Bard/Gemini	Sys., N
Gartlehner, 2024[67]; Oami, 2024[64]	Extraction	M	Claude	Sys.
Ahmed, 2023[55]; Aronson, 2023[68]; Flaherty, 2024[69]; Gue, 2024[70]; Kartchner, 2023[71]; Khraisha, 2024[72]; Kılıç, 2023[73]; Lozano, 2024[41]; Mahmoudi, 2024[74]; Mahuli, 2023[75]; Miao, 2023[76]; Oami, 2024[64]; Prasad, 2024[65]; Reason, 2024[77]; Schmidt, 2024[78]; Serajeh, 2024[79]; Shah-Mohammadi, 2024[80]; Susnjak, 2023[81]; Susnjak, 2024[82]; Tang, 2024[83]; Tao, 2024[84]; Teperikidis, 2024[44]; Tovar, 2023[85]; Uittenhove, 2024[86]; Urrutia, 2023[87]; Wang, 2024[88]; Yun, 2024[89]; Zamani, 2024[90]; Zhao, 2024[48]	Extraction	M	GPT	Sys., M, Other, N, Umbrella
Ghosh, 2024[91]; Serajeh, 2024[79]; Tovar, 2023[85]; Yun, 2024[89]	Extraction	M	Llama	M, Other, N, Sys.
Susnjak, 2024[82]; Tsai, 2024[92]; Yun, 2024[89]	Extraction	M	Mistral	Sys., M
Hossain, 2024[93]; Jain, 2024[94]; Sami, 2024[95]	Extraction	M	Non-specific	Sys., N
Grokhowsky, 2023[96]; Yun, 2024[89]	Extraction	M	Other	M, R
Sun, 2024[97]; White, 2023[98]	Extraction	R	Claude	Sys.
Janes, 2022[99]; Liu, 2023[51]; Noe-Steinmüller, 2024[100]; Pattyn, 2023[101]; Schopow, 2023[53]; Teperikidis, 2023[54]	Extraction	R	GPT	Sys., Umbrella, Other
Beheshti, 2023[102]; Sun, 2024[97]	Extraction	R	Other	Sys.
Ambalavanan, 2020[103]; Martenot, 2022[104]	Full-text	M	BERT	M, Sys., N
Ye, 2024[66]	Full-text	M	Bard/Gemini	N, Sys.
Aronson, 2023[68]; Khraisha, 2024[72]; Lozano, 2024[41]; Susnjak, 2023[81]	Full-text	M	GPT	Sys., Other
Tsai, 2024[92]	Full-text	M	Mistral	Sys.
Hossain, 2024[93]; Sami, 2024[95]	Full-text	M	Non-specific	Sys.
Guo, 2023[105]	Full-text	M	Other	Sys.
Liu, 2023[51]; Schopow, 2023[53]; Teperikidis, 2023[54]	Full-text	R	GPT	Sys., Umbrella, Other
Scells, 2023[59]	Other	M	BERT	Sys.
Atkinson, 2023[106]; Demir, 2024[107]; Giunti, 2024[108]; Kılıç, 2023[73]; Najafali, 2023[42]; Qureshi, 2023[109]; Whang, 2024[110]; Zhao, 2024[48]	Other	M	GPT	Other, M, Sys., N
Abd-Alrazaq, 2024[111]	Other	R	BERT	Other
Khadhraoui, 2022[112]; Liang, 2023[113]; Likhareva, 2024[114]	Publication classification	M	BERT	Sys.
Alshami, 2023[115]; Guler, 2023[116]; Lam, 2024[117]	Publication classification	M	GPT	N, Sys.
Grokhowsky, 2023[96]; Platt, 2023[118]; Raja, 2024[119]	Publication classification	M	Other	R, Sys.
Twinomurinzi, 2023[120]	Publication classification	R	GPT	Scop.
Wang, 2022[121]	Quality/bias assessment	M	BERT	Sys.
Lai, 2024[122]; Woelfle, 2024[123]	Quality/bias assessment	M	Claude	M, Sys.
Barsby, 2024[124]; Chern, 2023[125]; Hasan, 2024[126]; Lai, 2024[122]; Mahuli, 2023[75]; Pitre, 2023[127]; Roberts, 2023[128]; Srivastava, 2023[129]; Teperikidis, 2024[44]; Treviño-Juarez, 2024[130]; Woelfle, 2024[123]	Quality/bias assessment	M	GPT	M, Sys., N, Umbrella

(continued)

Table 2. (continued)

Citations	Review stage	M/R	LLM	Review
Woelfle, 2024[123]	Quality/bias assessment	M	Mistral	M, Sys.
Alchokr, 2022[131]; Lu, 2021[132]; Tang, 2023[133]	Searching	M	BERT	N, Scop., Sys.
Aiumtrakul, 2023[134]; Chelli, 2024[135]	Searching	M	Bard/Gemini	Sys.
Agarwal, 2024[33]; Aiumtrakul, 2023[134]; Anghelescu, 2023[34]; Antu, 2023[136]; Chelli, 2024[135]; Choueka, 2024[137]; Demir, 2024[107]; Díaz, 2023[138]; Dossantos, 2023[36]; Flaherty, 2024[69]; Goldfarb, 2024[139]; Gupta, 2023[140]; Gupta, 2023[141]; Gwon, 2024[142]; Herbst, 2023[143]; Jafari, 2024[144]; Kim, 2024[145]; Kılıç, 2023[73]; Li, 2024[146]; Liu, 2023[147]; Lozano, 2024[41]; Maniaci, 2024[148]; Najafali, 2023[42]; Qureshi, 2023[109]; Roy, 2024[149]; Ruksakulpiwat, 2024[150]; Sanii, 2024[151]; Semrl, 2023[43]; Singh, 2023[152]; Spillias, 2023[153]; Suppadungsuk, 2023[154]; Susnjak, 2023[81]; Teperikidis, 2024[44]; Tovar, 2023[87]; Wang, 2023[155]; Yan, 2024[156]; Zamani, 2024[90]; Zhao, 2024[48]; Zhu, 2023[157]; Zimmermann, 2024[158]	Searching	M	GPT	Sys., N, Other, Scop., R, M, Umbrella
Tovar, 2023[85]	Searching	M	Llama	N
Tsai, 2024[92]	Searching	M	Mistral	Sys.
Hossain, 2024[93]; Jain, 2024[94]; Sami, 2024[95]	Searching	M	Non-specific	Sys., N
Aiumtrakul, 2023[134]; Guo, 2023[105]; Gwon, 2024[142]; Sanii, 2024[151]; Zhu, 2023[157]	Searching	M	Other	Sys., N
Anghelescu, 2023[159]; Cambaz, 2024[160]; Haltaufderheide, 2024[161]; Liu, 2023[51]; Pattyn, 2023[101]; Ruksakulpiwat, 2023[162]; Sallam, 2023[163]; Schopow, 2023[53]; Srivastava, 2023[164]; Teperikidis, 2023[54]; Zhao, 2024[165]	Searching	R	GPT	R, Sys., Other, Scop., Umbrella, M, N
Beheshti, 2023[102]; Cambaz, 2024[160]	Searching	R	Other	Sys.
Lan, 2024[166]; Lu, 2021[132]; Shinde, 2022[60]; Teslyuk, 2020[167]	Synthesis	M	BERT	Sys., N
Anghelescu, 2023[34]; Antu, 2023[136]; Atkinson, 2023[106]; Aydın, 2022[168]; Blasingame, 2024[169]; Chaker, 2024[170]; Dossantos, 2023[36]; Jenko, 2024[37]; Kim, 2024[145]; LamHoai, 2023[171]; Li, 2024[146]; Lozano, 2024[41]; Qureshi, 2023[109]; Susnjak, 2023[81]; Susnjak, 2024[82]; Tang, 2023[172]; Teperikidis, 2024[44]; Wang, 2024[45]; Yan, 2023[173]; Zhao, 2024[48]	Synthesis	M	GPT	Sys., M, Other, N, Umbrella
Susnjak, 2024[82]	Synthesis	M	Mistral	Sys.
Yu, 2022[174]	Synthesis	M	Other	Sys.
Lamovšek, 2023[50]; Li, 2024[175]; Liu, 2023[51]; Noe-Steinmüller, 2024[100]; Pedroso-Roussado, 2023[52]; Rajjoub, 2024[176]; Temsah, 2023[177]	Synthesis	R	GPT	N, Sys., Other
Ambalavanan, 2020[178]; Aum, 2021[179]; Edwards, 2024[180]; Hasny, 2023[181]; Kats, 2023[182], Mao, 2024[183]; Martenot, 2022[104]; Ng, 2023[184]; Qin, 2021[185]; Wang, 2022[186]	Title/abstract	M	BERT	Sys., Other, N
Ye, 2024[66]	Title/abstract	M	Bard/Gemini	N, Sys.
Castillo-Segura, 2023[187]	Title/abstract	M	Claude	Sys.
Akinseloyin, 2023[188]; Ali, 2024[189]; Cai, 2023[190]; Castillo-Segura, 2023[187]; Guo, 2024[191]; Huotala, 2024[192]; Issaiy, 2024[193]; Kataoka, 2023[194]; Khraisha, 2024[72]; Kılıç, 2023[73]; Li, 2024[195]; Lozano, 2024[41]; Robinson, 2023[196]; Susnjak; 2023[81]; Syriani, 2023[197]; Teperikidis, 2024[44]; Tran, 2024[198]; Urrutia, 2023[87]; Wang, 2023[199]; Wang, 2024[200]; Wilkins, 2023[201]; Yang, 2024[202]; Zamani, 2024[90]	Title/abstract	M	GPT	Sys., Scop., R, M, Other, Umbrella

(continued)

Table 2. (continued)

Citations	Review stage	M/R	LLM	Review
Li, 2024[195]; Robinson, 2023[196]; Wang, 2023[199]; Wang, 2024[200]	Title/abstract	M	Llama	M, Sys.
Tsai, 2024[92]	Title/abstract	M	Mistral	Sys.
Hossain, 2024[93]; Sami, 2024[95]	Title/abstract	M	Non-specific	Sys.
Castillo-Segura, 2023[187]; Li, 2024[195]; Robinson, 2023[196]	Title/abstract	M	Other	M, Sys.
Buchlak, 2022[203]; Buchlak, 2022[204]	Title/abstract	R	BERT	Sys.
White, 2023[98]	Title/abstract	R	Claude	Sys.
Liu, 2023[51]; Schopow, 2023[53]; Teperikidis, 2023[54]	Title/abstract	R	GPT	Sys., Umbrella, Other
Buchlak, 2022[203]; Buchlak, 2022[204]	Title/abstract	R	Other	Sys.

R, Review paper; M, Methods paper; Title/abstract , Title or abstract screening; Full-text, Full-text screening; Extraction, Data extraction; Searching, Publication searching; Synthesis, Evidence Synthesis; Sys., Systematic review; Scop., Scoping review; M, Meta-analysis; N, Narrative/literature review; GPT, GPT/ChatGPT—based models; BERT, BERT-based models; Llama, Llama/Alpaca-based models.

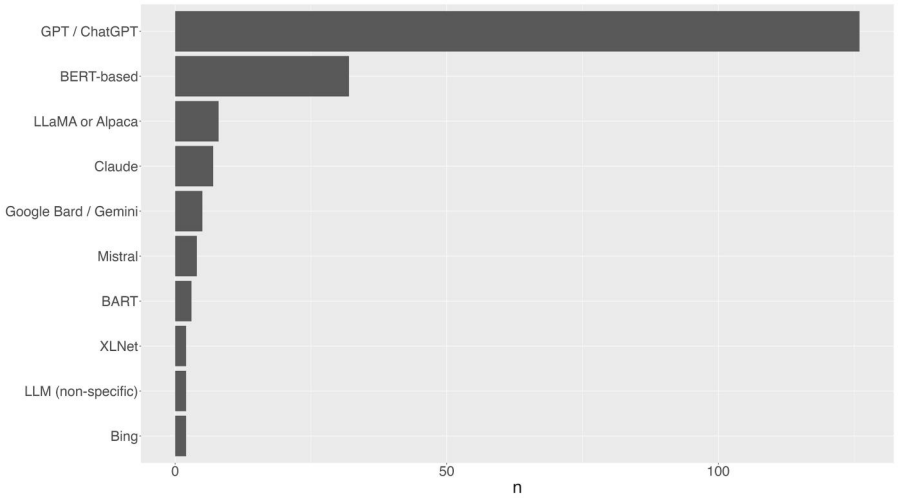

Figure 5. LLM types proposed for automation (models mentioned in 2 or more studies shown).

eventually change how the review is conducted (reviewers might be assisted by LLMs, or the review paper format could eventually be replaced by online real-time information retrieval).

The strength of the present review includes the large-scale (over 3000 abstracts screened and 172 full-text publications eligible for extraction) automation of different stages of review, including drafting the manuscript sections and plot generation. Only a few citations focused on the automation of the full cycle of review, while most focused only on specific areas like extraction or screening, including our previous systematic review where GPT-3.5 was used with LDA-based topic modeling for validation of findings made by human reviewers.[100] In contrast, the LLM-based method that we applied in this work demonstrated its direct applicability by facilitating the automation of the abstract and full-text screening, data extraction, as well as knowledge synthesis stages, with the discussed constraints. Furthermore, our method is domain-agnostic; thus, it can be integrated into large-scale review projects across different domains. The implications of such automation include reducing human workload and improving the overall efficiency of systematic reviews. Furthermore, such tools in their more mature form will require less expertise from human reviewers, which could contribute to the democratization of the systematic and scoping review process, with the potential to add features related to meta-analysis into the process.

GPT-based LLMs were the most dominant type of LLMs and the ones that seemed to yield remarkable results in the data extraction, arguably the most complex and time-consuming stage of any review. The relatively low result in searching for publications can be attributed to high hallucination rates, when these models are prompted to generate scientific citations on the research topic, this was especially noticeable in earlier versions of GPT-based models.

At this moment, there are few restrictions on the type of information users can load into ChatGPT, and published papers are unlikely to contain any sensitive information, making ChatGPT, with its high-performing model, developed desktop application and API, an obvious choice. Usage of these expensive models shows overall a significant reduction in cost to complete a review.[205] At the same time, smaller models, such as BERT, Llama, or Mistral, can be run and fine-tuned locally at much lower costs; we expect to see more automation projects with this LLM in the future.[206]

Limitations

We used calibrated LLMs as reviewers in this project. Some extraction categories, such as performance metrics, had

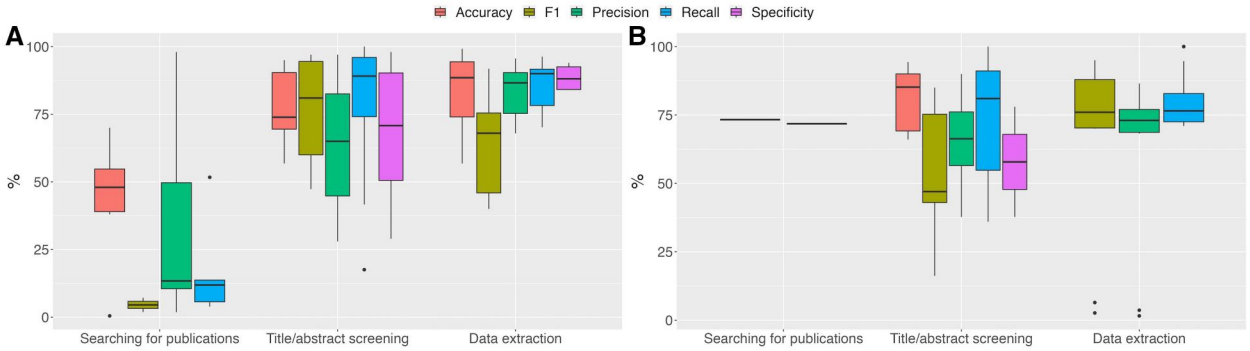

Figure 6. Performance metrics reported for the 3 most common automated stages for (A) GPT-based models and (B) BERT-based models. Boxplots display median value, while comparison is provided using means.

relatively lower accuracy. Therefore, the results of this extraction category should be taken with caution. Nevertheless, in this review, LLMs achieved remarkable results in accuracy, making it possible to delegate time-consuming phases of review to LLMs. Studies generally recommend a single-reviewer approach in some cases, such as rapid reviews.[207] However, we believe that the LLM approach could substitute human reviewers, and human effort should be redirected to supervision of the review process. Future research should focus on improving LLM performance metrics, particularly precision and recall in lower-accuracy extraction categories. Additionally, integrating and evaluating different LLMs, possibly in combination with other AI models, should be explored to enhance performance. The short- and long-term impacts of these integrations on review quality, along with ethical considerations, must also be assessed to maintain research credibility and trust.

It is important to mention that we relied on the disclosure of LLM usage by the authors of reviewed publications, and this study did not use any type of automatic LLM usage detection; thus, we could have missed publications, especially potential reviews, that could have been created with LLM support.

Conclusion

The use of LLMs in review automation is rapidly growing, with expected radical changes in scientific evidence synthesis. LLMs are likely to significantly reduce the time needed for reviews while producing similar or higher-quality data in greater quantities than manual reviews do. Research shows it is becoming increasingly difficult to distinguish between LLM-generated and human-written texts,[208] and the presence of LLM-generated text in scientific publications is growing exponentially.[209] To promote transparency and proper acknowledgment, researchers are encouraged to openly disclose their use of LLMs in academic papers, providing information on the prompts employed and the sections of text affected.[210]

Despite early successes, few systematic reviews using LLMs were identified in our review. Although still in its early stages, AI-assisted reviews are already yielding impressive results, with growing interest as researchers develop semi-automated pipelines. However, generating trustworthy and useful AI-driven reviews still presents both technological and ethical challenges, particularly for quantitative meta-analyses comparing treatment effects. However, the conduct of more simple systematic reviews, such as scoping reviews, appears to be well within the capabilities of current or near-future AI methods.

Acknowledgments

Screening and extraction process took place in Covidence with the help of LLM add-on based on the OpenAI GPT-4o model as described in the Methods section. An LLM tool by Google NotebookLM (version from August 2024) was used to cross-check the extraction results for the fields where the precision of extraction was low during the benchmark. ChatGPT (4o model) was used to clean the extraction data (case formatting, removal of duplicates, and standardization of names). Scite.ai (version from August 2024) was used to draft parts of the introduction and discussion sections, whereas ChatGPT was used to draft the abstract and results section of this review by generating R code snippets to produce Figures 3-5.

Author contributions

Dmitry Scherbakov (Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Nina C. Hubig (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing—review & editing), Vinita Jansari (Conceptualization, Formal analysis, Investigation, Methodology, Validation), Alexander Bakumenko (Conceptualization, Formal analysis, Investigation, Validation, Writing—review & editing), and Leslie Andrew Lenert (Conceptualization, Methodology, Project administration, Resources, Supervision, Writing—review & editing)

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This publication was supported, in part, by the National Center for Advancing Translational Sciences of the National Institutes of Health under Grant Number UL1 TR001450. D.S. was

supported by grant T15 LM013977, Biomedical Informatics and Data Science for Health Equity Research (SC BIDS4Health). This publication was supported in part by a Smart-state Chair endowment. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data underlying this article are available in the article, supplementary materials, and in OSF registration record (https://doi.org/10.17605/OSF.IO/EJKSY).

References

1. Toh TS, Lee JH. Statistical note: using scoping and systematic reviews. *Pediatr Crit Care Med*. 2021;22:572-575.
2. Abushouk AI, Yunusa I, Elmehrath AO, et al. Quality assessment of published systematic reviews in high impact cardiology journals: revisiting the evidence pyramid. *Front Cardiovasc Med* 2021;8:671569.
3. Acar IH, Avcılar G, Yazıcı G, Bostancı S. The roles of adolescents' emotional problems and social media addiction on their self-esteem. *Curr Psychol*. 2020;41:6838-6847. https://doi.org/10.1007/s12144-020-01174-5
4. Borah R, Brown AW, Capers PL, et al. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*. 2017;7:e012545.
5. Munn Z, Peters MDJ, Stern C, et al. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Med Res Methodol*. 2018;18:143-147.
6. Kellermeyer L, Harnke B, Knight S. Covidence and Rayyan. *JMLA*. 2018;106:580.
7. Chan JL, Murphy KA, Sarna JR. Myoclonus and cerebellar ataxia associated with COVID-19: a case report and systematic review. *J Neurol*. 2021;268:3517-3548.
8. Machine learning—the game changer for trustworthy evidence. Accessed August, 14, 2024. https://www.covidence.org/blog/machine-learning-the-game-changer-for-trustworthy-evidence/
9. Gartlehner G, Affengruber L, Titscher V, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. *J Clin Epidemiol*. 2020;121:20-28.
10. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev*. 2019;8:163-110.
11. Rasheed Z, Waseem M, Systä K, Abrahamsson P. 2024. Large language model evaluation via multi ai agents: preliminary results. arXiv, preprint arXiv: 2404.01023, preprint: not peer reviewed.
12. Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G, eds. Zero-shot generative large language models for systematic review screening automation. In: *European Conference on Information Retrieval*. Springer; 2024:403-420.
13. Zaki M, Namireddy SR, Pittie T, et al. Natural language processing-guided meta-analysis and structure factor database extraction from glass literature. *J Non-Crystalline Solids: X*. 2022;15:100103.
14. National Institute for Health and Care Excellence. Use of AI in evidence generation: NICE position statement. Accessed April 24, 2025. https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation–nice-position-statement?utm_medium=social&utm_source=linkedin&utm_campaign=aiposition
15. Liu Z. ChatGPT—a new milestone in the field of education. *ACE*. 2024;35:129-133.
16. Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. *Int J Gen Med*. 2024;17:817-826.
17. Tlili A, Shehata B, Adarkwah MA, et al. What if the devil is my guardian angel: ChatGPT as a case study of using Chatbots in education. *Smart Learn Environ*. 2023;10:15.
18. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49.
19. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. *Syst Rev*. 2015;4:78-16.
20. Tsou AY, Treadwell JR, Erinoff E, et al. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. *Syst Rev*. 2020;9:73-14.
21. O'Mara-Eves A, Thomas J, McNaught J, et al. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Syst Rev*. 2015;4:1-22.
22. Thomas J, Noel-Storr A, Marshall I, Living Systematic Review Network, et al. Living systematic reviews: 2. Combining human and machine effort. *J Clin Epidemiol*. 2017;91:31-37.
23. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc*. 2016;23:193-201.
24. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc*. 2012;19:121-127.
25. Scherbakov D. Large language models in scoping and systematic reviews automation: an automated systematic review [protocol registration]. Accessed April 24, 2025. https://doi.org/10.17605/OSF.IO/EJKSY
26. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
27. *Dimensions (database)*. 2025. Accessed April 24, 2025. https://app.dimensions.ai/
28. Microsoft. *Azure OpenAI Service* [cited 2025 Mar 20]. Accessed April 24, 2025. https://azure.microsoft.com/en-us/products/ai-services/openai-service
29. Google. *NotebookLM* [cited 2025 Mar 20]. Accessed April 24, 2025. https://notebooklm.google.com
30. OpenAI. *Hello GPT-4o*. 2025 [cited 2025 Mar 20]. Accessed April 24, 2025. https://openai.com/index/hello-gpt-4o/
31. scite. *AI for Research*. [cited 2025 Mar 20]. Accessed April 24, 2025. https://scite.ai/
32. Haddaway NR, Westgate MJ. Predicting the time needed for environmental systematic reviews and systematic maps. *Conserv Biol*. 2019;33:434-443.
33. Agarwal S, Laradji IH, Charlin L, Pal C. 2024. LitLLM: a toolkit for scientific literature review. arXiv, preprint arXiv, not peer reviewed.
34. Anghelescu A, Ciobanu I, Munteanu C, et al. ChatGPT: "To be or not to be" . . . in academic research. The human mind's analytical rigor and capacity to discriminate between AI bots' truths and hallucinations. *Balneo PRM Res J* 2023;14:614.
35. Bersenev D, Yachie-Kinoshita A, Palaniappan SK. Replicating a high-impact scientific publication using systems of large language models. bioRxiv 2024.04.08.588614, 2024.
36. Dossantos J, An J, Javan R. Eyes on AI: ChatGPT's transformative potential impact on ophthalmology. *Cureus*. 2023;15:e40765.
37. Jenko N, Ariyaratne S, Jeys L, et al. An evaluation of AI generated literature reviews in musculoskeletal radiology. *Surgeon*. 2024;22:194-197.

38. Khlaif ZN, Mousa A, Hattab MK, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Med Educ*. 2023;9:e47049.
39. Li X, Ouyang J. 2024. Explaining relationships among research papers. arXiv, preprint arXiv, not peer reviewed.
40. Livberber T. Toward non-human-centered design: designing an academic article with ChatGPT. *EPI*. 2023;32
41. Lozano A, Fleming SL, Chiang C-C, et al. Clinfo.ai: an open-source retrieval-augmented large language model system for answering medical questions using scientific literature. *Pac Symp Biocomput*. 2024;29:8-23.
42. Najafali D, Camacho JM, Reiche E, et al. Truth or lies? The pitfalls and limitations of ChatGPT in systematic review creation. *Aesthet Surg J*. 2023;43:NP654-NP655.
43. Semrl N, Feigl S, Taumberger N, et al. AI language models in human reproduction research: exploring ChatGPT's potential to assist academic writing. *Hum Reprod*. 2023;38:2281-2288.
44. Teperikidis E, Boulmpou A, Papadopoulos C. Prompting ChatGPT to perform an umbrella review. *Acta Cardiol*. 2024;79:403-404.
45. Wang J, Huang C, Yan S, Xie W, He D. When young scholars cooperate with LLMs in academic tasks: the influence of individual differences and task complexities. *Int J Hum–Comput Interact* 2024:41:4624-4639.
46. Wu CL, Cho B, Gabriel R, et al. Addition of dexamethasone to prolong peripheral nerve blocks: a ChatGPT-created narrative review. *Reg Anesth Pain Med*. 2023;49:777-781
47. Yun HS, Marshall IJ, Trikalinos TA, Wallace BC. 2023. Appraising the potential uses and harms of LLMs for medical systematic reviews. arXiv, preprint arXiv, not peer reviewed.
48. Zhao S, Chen S, Zhou J, et al. Potential to transform words to watts with large language models in battery research. *Cell Rep Phys Sci*. 2024;5:101844.
49. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res*. 2023;13:1148-1154.
50. Lamovšek N. Analysis of research on artificial intelligence in public administration. *CEPAR*. 2023;21:77-96.
51. Liu H, Peng Y, Weng C. How good is ChatGPT for medication evidence synthesis? *Stud Health Technol Inform*. 2023;302:1062-1066.
52. Pedroso-Roussado C. Investigating the limitations of fashion research methods in applying a sustainable design practice: a systematic review. Preprints.org 202310.0250.v2, 2023.
53. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. *JMIR Med Inform*. 2023;11:e48933.
54. Teperikidis E, Boulmpou A, Potoupni V, et al. Does the long-term administration of proton pump inhibitors increase the risk of adverse cardiovascular outcomes? A ChatGPT powered umbrella review. *Acta Cardiol*. 2023;78:980-988.
55. Ahmed U. Reimagining open data ecosystems: a practical approach using AI, CI, and Knowledge Graphs. In: *BIR Workshops*. 2023:235-249.
56. Marshalova A, Bruches E, Batura T. 2023. Automatic aspect extraction from scientific texts. arXiv, preprint arXiv, not peer reviewed.
57. Mutinda FW, Liew K, Yada S, et al. Automatic data extraction to support meta-analysis statistical analysis: a case study on breast cancer. *BMC Med Inform Decis Mak*. 2022;22:158.
58. Panayi A, Ward K, Benhadji-Schaff A, et al. Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. *Syst Rev*. 2023;12:187.
59. Scells H, Schlatt F, Potthast M. Smooth operators for effective systematic review queries. 2023. Accessed April 24, 2025. https://doi.org/10.1145/3539618.3591768
60. Shinde K, Roy T, Ghosal. An T. Extractive-abstractive approach for multi-document summarization of scientific articles for literature review. 2022. Accessed April 24, 2025. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174492822&partnerID=40&md5=657f586349a915d084d9edd9031ecc23
61. Wang Q, Liao J, Lapata M, et al. PICO entity extraction for preclinical animal literature. *Syst Rev*. 2022;11:209.
62. Whitton J, Hunter A. Automated tabulation of clinical trial results: a joint entity and relation extraction approach with transformer-based language representations. *Artif Intell Med*. 2023;144:102661.
63. Yazi FS, Vong WT, Raman V, Then PHH, Lunia MJ. Towards automated detection of contradictory research claims in medical literature using deep learning approach. 2021. Accessed April 24, 2025. https://doi.org/10.1109/CAMP51653.2021.9498061
64. Oami T, Okada Y, Nakada T-A Accuracy and reliability of data extraction for systematic reviews using large language models: a protocol for a prospective study. medRxiv 2024.05.22.24307740, 2024.
65. Prasad D, Pimpude M, Alankar A. 2024. Towards development of automated knowledge maps and databases for materials engineering using large language models. arXiv, preprint arXiv, not peer reviewed.
66. Ye A, Maiti A, Schmidt M, et al. A hybrid semi-automated workflow for systematic and literature review processes with large language model analysis. *Future Internet*. 2024;16:167.
67. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods*. 2024;15:576-589.
68. Aronson SJ, Machini K, Shin J, et al. 2023. Preparing to integrate generative pretrained transformer series 4 models into genetic variant assessment workflows: assessing performance, drift, and nondeterminism characteristics relative to classifying functional evidence in literature. arXiv, preprint arXiv, not peer reviewed.
69. Flaherty HB, Yurch J. Beyond plagiarism: ChatGPT as the Vanguard of technological revolution in research and citation. *Res Social Work Pract*. 2024;34:483-486.
70. Gue CCY, Rahim NDA, Rojas-Carabali W, et al. Evaluating the OpenAI's GPT-3.5 Turbo's performance in extracting information from scientific articles on diabetic retinopathy. *Syst Rev*. 2024;13:135.
71. Kartchner D, Al-Hussaini I, Kronick O, Ramalingam S, Mitchell C. Zero-shot information extraction for clinical meta-analysis using large language models. 2023. Accessed April 24, 2025. https://par.nsf.gov/servlets/purl/10502330
72. Khraisha Q, Put S, Kappenberg J, et al. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods*. 2024;15:616-626.
73. Kılıç DK, Vasegaard AE, Desoeuvres A, Nielsen P. 2023. A semi-automated solution approach recommender for a given use case: a case study for AI/ML in oncology via Scopus and OpenAI. arXiv, preprint arXiv, not peer reviewed.
74. Mahmoudi H, Chang D, Lee H, Ghaffarzadegan N, Jalali MS. A critical assessment of large language models for systematic reviews: utilizing ChatGPT for complex data extraction. *SSRN J* 2024.
75. Mahuli SA, Rai A, Mahuli AV, et al. Application ChatGPT in conducting systematic reviews and meta-analyses. *Br Dent J*. 2023;235:90-92.
76. Miao H, Yu X, Wu H. Mining topic structure of AI algorithmic literature. 2023. Accessed April 24, 2025. https://doi.org/10.1109/IEIR59294.2023.10391253
77. Reason T, Benbow E, Langham J, et al. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models. *Pharmacoecon Open*. 2024;8:205-220.
78. Schmidt L, Hair K, Graziozi S, et al. 2024. Exploring the use of a large language model for data extraction in systematic reviews: a rapid feasibility study. arXiv, preprint arXiv, not peer reviewed.

79. Serajeh NT, Mohammadi I, Fuccella V, De Rosa M. 2024. LLMs in HCI data work: bridging the gap between information retrieval and responsible research practices. arXiv, preprint arXiv, not peer reviewed.
80. Shah-Mohammadi F, Finkelstein J. Large language model-based architecture for automatic outcome data extraction to support meta-analysis. 2024. Accessed April 24, 2025. https://doi.org/10.1109/CCWC60891.2024.10427829
81. Susnjak T. 2023. PRISMA-DFLLM: an extension of PRISMA for systematic literature reviews using domain-specific finetuned large language models. arXiv, preprint arXiv: 2306.14905, not peer reviewed.
82. Susnjak T, Hwang P, Reyes NH, et al. 2024. Automating research synthesis with domain-specific large language model fine-tuning. arXiv, preprint arXiv: 2404.08680, not peer reviewed.
83. Tang Y, Xiao Z, Li X, et al. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: a comparative study of GPT-3.5 and GPT-4. medRxiv 2024.03.20.24304572, 2024.
84. Tao K, Osman ZA, Tzou PL, et al. GPT-4 performance on querying scientific publications: reproducibility, accuracy, and impact of an instruction sheet. *BMC Med Res Methodol*. 2024;24:139.
85. Tovar DA. 2023. AI literature review suite. arXiv, preprint arXiv, not peer reviewed.
86. Uittenhove K, Martinelli P, Roquet A. 2024. Large language models in psychology: application in the context of a systematic literature review. PsyArXiv, preprint PsyArXiv, not peer reviewed.
87. Urrutia F, Buc C, Barriere V. 2023. Deep natural language feature learning for interpretable prediction. arXiv, preprint arXiv, not peer reviewed.
88. Wang X, Luo G. 2024. MetaMate: large language model to the rescue of automated data extraction for educational systematic reviews and meta-analyses. EdArXiv, preprint EdArXiv, not peer reviewed.
89. Yun HS, Pogrebitskiy D, Marshall IJ, Wallace BC. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. arXiv, preprint arXiv, not peer reviewed.
90. Zamani S, Sinha R. Generative AI—the end of systematic reviews in PhD projects? *ACM Inroads*. 2024;15:48-50.
91. Ghosh M, Mukherjee S, Ganguly A, et al. AlpaPICO: extraction of PICO frames from clinical trial documents using LLMs. *Methods*. 2024;226:78-88.
92. Tsai H-C, Huang Y-F, Kuo C-W. Comparative analysis of automatic literature review using Mistral large language model and human reviewers. *Research Square*, 2024. Accessed April 24, 2025. https://doi.org/10.21203/rs.3.rs-4022248/v1
93. Hossain MM. Using ChatGPT and other forms of generative AI in systematic reviews: challenges and opportunities. *J Med Imaging Radiat Sci*. 2024;55:11-12.
94. Jain S, Kumar A, Roy T, Shinde K, Vignesh G, Tondulkar R. SciSpace literature review: harnessing AI for effortless scientific discovery. 2024.
95. Sami AM, Rasheed Z, Kemell K-K, et al. 2024. System for systematic literature review using multiple AI agents: concept and an empirical evaluation. arXiv, preprint arXiv, not peer reviewed.
96. Grokhowsky N. Reducing knowledge synthesis workload time using a text-mining algorithm for research location and subtopic extraction from geographically dependent research publications. *Research Square*, 2023. Accessed April 24, 2025. https://doi.org/10.21203/rs.3.rs-3129370/v1
97. Sun Z, Zhang R, Doi SA, et al. How good are large language models for automated data extraction from randomized trials? medRxiv 2024.02. 20.24303083, 2024.
98. White M. Sample size in quantitative instrument-based studies published in Scopus up to 2022: an artificial intelligence aided systematic review. *Acta Psychol (Amst)*. 2023;241:104095.
99. Janes A, Li X, Lenarduzzi V. Open tracing tools: overview and critical comparison. *J Syst Softw*. 2023;204:111793.
100. Noe-Steinmüller N, Scherbakov D, Zhuravlyova A, et al. Defining suffering in pain: a systematic review on pain-related suffering using natural language processing. *Pain*. 2024;165:1434-1449.
101. Pattyn F. Preliminary structured literature review results using ChatGPT: towards a pragmatic framework for product managers at software startups. In: *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE; 2023:367-370.
102. Beheshti M, Amoozad Mahdiraji H, Rocha-Lona L. Transitioning drivers from linear to circular economic models: evidence of entrepreneurship in emerging nations. *Manage Decis*. 2024;62:2714-2736.
103. Ambalavanan AK, Devarakonda M. 2020. Cascade neural ensemble for identifying scientifically sound articles. arXiv, preprint arXiv, not peer reviewed.
104. Martenot V, Masdeu V, Cupe J, et al. LiSA: an assisted literature search pipeline for detecting serious adverse drug events with deep learning. *BMC Med Inform Decis Mak*. 2022;22:338.
105. Guo F, Luo Y, Yang L, Zhang Y. SciMine: an efficient systematic prioritization model based on richer semantic information. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2023:205-215.
106. Atkinson CF. ChatGPT and computational-based research: benefits, drawbacks, and machine learning applications. *Discov Artif Intell*. 2023;3:article 42.
107. Demir GB, Süküt Y, Duran GS, et al. Enhancing systematic reviews in orthodontics: a comparative examination of GPT-3.5 and GPT-4 for generating PICO-based queries with tailored prompts and configurations. *Eur J Orthod*. 2024;46:cjae011.
108. Giunti G, Doherty CP. Cocreating an automated mHealth apps systematic review process with generative AI: design science research approach. *JMIR Med Educ*. 2024;10:e48949.
109. Qureshi R, Shaughnessy D, Gill KAR, et al. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. 2023;12:72.
110. Whang Y. ChatGPT for editors: enhancing efficiency and effectiveness. *Sci Editing*. 2024;11:84-90.
111. Abd-Alrazaq A, Nashwan AJ, Shah Z, et al. Machine learning-based approach for identifying research gaps: COVID-19 as a case study. *JMIR Form Res*. 2024;8:e49411.
112. Khadhraoui M, Bellaaj H, Ammar MB, et al. Survey of BERT-base models for scientific text classification: COVID-19 case study. *Appl Sci*. 2022;12:2891.
113. Liang F, Hou F, Farshidi S, Jansen S. Sentiment analysis for software quality assessment. In: *CEUR Workshop Proceedings*, Vol. 3567. CEUR WS; 2023:17-24.
114. Likhareva D, Sankaran H, Thiyagarajan S. 2024. Empowering interdisciplinary research with BERT-based models: an approach through SciBERT-CNN with topic modeling. arXiv, preprint arXiv, not peer reviewed.
115. Alshami A, Elsayed M, Ali E, et al. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. *Systems*. 2023;11:351.
116. Guler N, Kirshner S, Vidgen R. Artificial intelligence research in business and management: a literature review leveraging machine learning and large language models. *SSRN J*. 2023;
117. Lam MS, Teoh J, Landay JA, Heer J, Bernstein MS. Concept induction: analyzing unstructured text with high-level concepts using LLooM. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 2024:1-28.
118. Platt M, Platt D. Effectiveness of generative artificial intelligence for scientific content analysis. 2023. Accessed April 24, 2025. https://doi.org/10.1109/AICT59525.2023.10313167

119. Raja H, Munawar A, Mylonas N, et al. Automated category and trend analysis of scientific articles on ophthalmology using large language models: development and usability study. *JMIR Form Res*. 2024;8:e52462.
120. Twinomurinzi H, Gumbo S. ChatGPT in scholarly discourse: sentiments and an inflection point. 2023. Accessed April 24, 2025. https://doi.org/10.1007/978-3-031-39652-6_17
121. Wang Q, Liao J, Lapata M, et al. Risk of bias assessment in preclinical literature using natural language processing. *Res Synth Methods*. 2022;13:368-380.
122. Lai H, Ge L, Sun M, et al. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Netw Open*. 2024;7:e2412687.
123. Woelfle T, Hirt J, Janiaud P, Kappos L, Ioannidis JPA, Hemkens LG. Benchmarking human-AI collaboration for common evidence appraisal tools. medRxiv 2024.04.21.24306137, 2024.
124. Barsby J, Hume S, Lemmey HA, Cutteridge J, Lee R, Bera KD. Pilot study on large language models for risk-of-bias assessments in systematic reviews: A(I) new type of bias? *BMJ Evid Based Med*. 2025;30:71-74.
125. Chern IC, Chern S, Chen S, et al. 2023. FacTool: factuality detection in generative AI—a tool augmented framework for multi-task and multi-domain scenarios. arXiv, preprint arXiv, not peer reviewed.
126. Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med*. 2024;29:394-398.
127. Pitre T, Jassal T, Talukdar JR, Shahab M, Ling M, Zeraatkar D. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: a methods study. medRxiv 2023.11.19.23298727, 2024.
128. Roberts RH, Ali SR, Hutchings HA, et al. Comparative study of ChatGPT and human evaluators on the assessment of medical literature according to recognised reporting standards. *BMJ Health Care Inform*. 2023;30:e100830.
129. Srivastava M. A day in the life of ChatGPT as an academic reviewer: investigating the potential of large language model for scientific literature review. OSF Preprints, 2023.
130. Trevino-Juarez AS. Assessing risk of bias using ChatGPT-4 and Cochrane ROB2 Tool. *Med Sci Educ*. 2024;34:691-694.
131. Alchokr R, Borkar M, M, Thotadarya S, Saake G, Leich T. Supporting systematic literature reviews using deep-learning-based language models. 2022. Accessed April 24, 2025. https://doi.org/10.1145/3528588.3528658
132. Lu ZH, Wang JX, Li X. Revealing opinions for COVID-19 questions using a context retriever, opinion aggregator, and question-answering model: model development study. *J Med Internet Res*. 2021;23:e22860.
133. Tang F-SK-B, Bukowski M, Schmitz-Rode T, et al. Guidance for clinical evaluation under the medical device regulation through automated scoping searches. *Appl Sci*. 2023;13:7639.
134. Aiumtrakul N, Thongprayoon C, Suppadungsuk S, et al. Navigating the landscape of personalized medicine: the relevance of ChatGPT, BingChat, and Bard AI in Nephrology literature searches. *J Pers Med*. 2023;13:1457.
135. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. *J Med Internet Res*. 2024;26:e53164.
136. Antu SA, Chen H, Richards CK. Using LLM (large language model) to improve efficiency in literature review for undergraduate research. In: LLM@ AIED, Tokyo, Japan. CEUR-WS.org; 2023:8-16.
137. Choueka D, Tabakin AL, Shalom DF. ChatGPT in urogynecology research: novel or not? *Urogynecology (Phila)*. 2022:10-97.
138. Díaz O, Garmendia X, Contell JP, Pereira J. Inquiry frameworks for research question scoping in DSR: a realization for ChatGPT. 2023. Accessed April 24, 2025. https://doi.org/10.1007/978-3-031-32808-4_19
139. Goldfarb N, Tal N, Cohen I-C, et al. Barriers and suggested solutions to nursing participation in research: a systematic review with NLP Tools (Preprint). JMIR Preprints, 2024.
140. Gupta R, Herzog I, Weisberger J, et al. Utilization of ChatGPT for plastic surgery research: friend or foe? *J Plast Reconstr Aesthet Surg*. 2023;80:145-147.
141. Gupta R, Park JB, Bisht C, et al. Expanding cosmetic plastic surgery research with ChatGPT. *Aesthet Surg J*. 2023;43:930-937.
142. Gwon YN, Kim JH, Chung HS, et al. The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation. *JMIR Med Inform*. 2024;12:e51187.
143. Herbst P, Baars H. Accelerating literature screening for systematic literature reviews with large language models—development, application, and first evaluation of a solution. 2023. Accessed April 24, 2025. https://ceur-ws.org/Vol-3630/LWDA2023-paper4.pdf
144. Jafari SMA. 2024. Streamlining the selection phase of systematic literature reviews (SLRs) using AI-enabled GPT-4 assistant API. arXiv, preprint arXiv, not peer reviewed.
145. Kim J, Lee J-S, Kim H, Lee T. 2024. Systematic review on healthcare systems engineering utilizing ChatGPT. arXiv, preprint arXiv, not peer reviewed.
146. Li Y, Zhao J, Li M, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. *J Am Med Inform Assoc*. 2024;31:2030-2039.
147. Liu Y, Chen S, Cheng H, et al. 2023. CoQuest: exploring research question co-creation with an LLM-based agent. arXiv, preprint arXiv, not peer reviewed.
148. Maniaci A, Saibene AM, Calvo-Henriquez C, et al. Is generative pre-trained transformer artificial intelligence (Chat-GPT) a reliable tool for guidelines synthesis? A preliminary evaluation for biologic CRSwNP therapy. *Eur Arch Otorhinolaryngol*. 2024;281:2167-2173.
149. Roy K, Khandelwal V, Vera V, Surana H, Heckman H, Sheth A. GEAR-Up: generative AI and external knowledge-based retrieval upgrading scholarly article searches for systematic reviews. 2024. Accessed April 24, 2025. https://doi.org/10.1609/aaai.v38i21.30577
150. Ruksakulpiwat S, Phianhasin L, Benjasirisan C, et al. Assessing the efficacy of ChatGPT versus human researchers in identifying relevant studies on mHealth interventions for improving medication adherence in patients with ischemic stroke when conducting systematic reviews: comparative analysis. *JMIR Mhealth Uhealth*. 2024;12:e51526.
151. Sanii RY, Kasto JK, Wines WB, et al. Utility of artificial intelligence in orthopedic surgery literature review: a comparative pilot study. *Orthopedics*. 2024;47:e125-e130.
152. Singh S, Watson S. ChatGPT as a tool for conducting literature review for dry eye disease. *Clin Exp Ophthalmol*. 2023;51:731-732.
153. Spillias S, et al. Human-AI collaboration to identify literature for evidence synthesis. *Research Square*, 2023. Accessed April 24, 2025. https://doi.org/10.21203/rs.3.rs-3099291/v1
154. Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. *J Clin Med*. 2023;12:5550.
155. Wang S, Scells H, Koopman B, Zuccon G. 2023. Can ChatGPT write a good Boolean query for systematic review literature search? arXiv, preprint arXiv, not peer reviewed.
156. Yan C, Grabowska ME, Dickson AL, et al. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. *NPJ Digit Med*. 2024;7:46.
157. Zhu K, Feng X, Feng X, Wu Y, Qin B. 2023. Hierarchical catalogue generation for literature review: a benchmark. arXiv, preprint arXiv, not peer reviewed.

158. Zimmermann R, Staab M, Nasseri M, Brandtner P. Leveraging large language models for literature review tasks—a case study using ChatGPT. 2024. Accessed April 24, 2025. https://doi.org/10.1007/978-3-031-48858-0_25
159. Anghelescu A, Firan FC, Onose G, et al. PRISMA systematic literature review, including with Meta-Analysis vs Chatbot/GPT (AI) regarding current scientific data on the main effects of the calf blood deproteinized hemoderivative medicine (Actovegin) in ischemic stroke. *Biomedicines*. 2023;11:1623
160. Cambaz D, Zhang X. Use of AI-driven code generation models in teaching and learning programming: a systematic literature review. 2024. Accessed April 24, 2025. https://doi.org/10.1145/3626252.3630958
161. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ Digit Med*. 2024;7:183.
162. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidiscip Healthc*. 2023;16:1513-1520.
163. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11:887.
164. Srivastava A. 2023. A rapid scoping review and conceptual analysis of the educational metaverse in the Global South: socio-technical perspectives. arXiv, preprint arXiv, not peer reviewed.
165. Zhao P, Zhang X, Cao J, Cheng MM, Yang J, Li X. 2024. A literature review of literature reviews in pattern analysis and machine intelligence. arXiv, preprint arXiv, not peer reviewed.
166. Lan M, Cheng M, Hoang L, et al. Automatic categorization of self-acknowledged limitations in randomized controlled trial publications. *J Biomed Inform*. 2024;152:104628.
167. Teslyuk A. The concept of system for automated scientific literature reviews generation. 2020. Accessed April 24, 2025. https://doi.org/10.1007/978-3-030-50420-5_32
168. Aydın Ö, Karaarslan E. OpenAI ChatGPT generated literature review: digital twin in healthcare. In: Aydın Ö, ed. *Emerging Computer Technologies*. Vol. 2. Elsevier; 2022:22-31.
169. Blasingame MN, Koonce TY, Williams AM, et al. Evaluating a large language model's ability to answer clinicians' requests for evidence summaries. medRxiv 2024.05.01.24306691, 2024.
170. Chaker SC, Hung Y-C, Saad M, et al. Easing the burden on caregivers – applications of artificial intelligence for physicians and caregivers of children with cleft lip and palate. *Cleft Palate Craniofac J*. 2024. doi: 10.1177/10556656231223596.
171. Lam Hoai XL, Simonart T. Comparing meta-analyses with ChatGPT in the evaluation of the effectiveness and tolerance of systemic therapies in moderate-to-severe plaque psoriasis. *J Clin Med*. 2023;12:5410.
172. Tang L, Sun Z, Idnay B, et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med*. 2023;6:158.
173. Yan C, Grabowska ME, Dickson AL, et al. Leveraging generative AI to prioritize drug repurposing candidates: validating identified candidates for Alzheimer's disease in real-world clinical datasets. medRxiv 2023.07.07.23292388, 2023.
174. Yu B. Evaluating pre-trained language models on multi-document summarization for literature reviews. In: *Proceedings of the Third Workshop on Scholarly Document Processing*. 2022:188-192.
175. Li Y, Chen L, Liu A, Yu K, Wen L. 2024. ChatCite: LLM agent with human workflow guidance for comparative literature summary. arXiv, preprint arXiv, not peer reviewed.
176. Rajjoub R, Arroyave JS, Zaidat B, et al. ChatGPT and its role in the decision-making for the diagnosis and treatment of lumbar spinal stenosis: a comparative analysis and narrative review. *Global Spine J*. 2024;14:998-1017.
177. Temsah O, Khan SA, Chaiah Y, et al. Overview of early ChatGPT's presence in medical literature: insights from a hybrid literature review by ChatGPT and human experts. *Cureus*. 2023;15:e37281.
178. Ambalavanan AK, Devarakonda MV. Using the contextual language model BERT for multi-criteria classification of scientific articles. *J Biomed Inform*. 2020;112:103578.
179. Aum S, Choe S. srBERT: automatic article classification model for systematic review using BERT. *Syst Rev*. 2021;10:285.
180. Edwards KM, Song B, Porciello J, et al. ADVISE: accelerating the creation of evidence syntheses for global development using natural language processing-supported human-AI collaboration. *J Mech Desig*. 2024;146:1-15.
181. Hasny M, Vasile AP, Gianni M, et al. BERT for complex systematic review screening to support the future of medical research. 2023. Accessed April 24, 2025. https://doi.org/10.1007/978-3-031-34344-5_21
182. Kats T, van der Putten P, Scholtes J. 2023. Relevance feedback strategies for recall-oriented neural information retrieval. arXiv, preprint arXiv, not peer reviewed.
183. Mao X, Koopman B, Zuccon G. 2024. A reproducibility study of goldilocks: just-right tuning of BERT for TAR. arXiv, preprint arXiv, not peer reviewed.
184. Ng SH-X, Teow KL, Ang GY, et al. Semi-automating abstract screening with a natural language model pretrained on biomedical literature. *Syst Rev*. 2023;12:172.
185. Qin X, Liu J, Wang Y, et al. Natural language processing was effective in assisting rapid title and abstract screening when updating systematic reviews. *J Clin Epidemiol*. 2021;133:121-129.
186. Wang S, Scells H, Koopman B, et al. 2022. Neural rankers for effective screening prioritisation in medical systematic review literature search. arXiv, preprint arXiv, not peer reviewed.
187. Castillo-Segura P, Alario-Hoyos C, Kloos CD, Fernandez Panadero C. Leveraging the potential of generative AI to accelerate systematic literature reviews: an example in the area of educational technology. 2023. Accessed April 24, 2025. https://doi.org/10.1109/WEEF-GEDC59520.2023.10344098
188. Akinseloyin O, Jiang X, Palade V. A novel question-answering framework for automated abstract screening using large language models. medRxiv 2023.12.17.23300102, 2024.
189. Ali F. 2024. Can machine learning help accelerate article screening for systematic reviews? Yes, when article separability in embedding space is high. EdArXiv, preprint EdArXiv, not peer reviewed.
190. Cai X, Geng Y, Du Y, et al. Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. medRxiv 2023.09.06.23295072, 2023.
191. Guo E, Gupta M, Deng J, et al. Automated paper screening for clinical reviews using large language models: data analysis study. *J Med Internet Res*. 2024;26:e48996.
192. Huotala A, Kuutila M, Ralph P, Mäntylä M. 2024. The promise and challenges of using LLMs to accelerate the screening process of systematic reviews. arXiv, preprint arXiv, not peer reviewed.
193. Issaiy M, Ghanaati H, Kolahi S, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. *BMC Med Res Methodol*. 2024;24:78.
194. Kataoka Y, So R, Banno M, et al. Development of meta-prompts for large language models to screen titles and abstracts for diagnostic test accuracy reviews. medRxiv 2023.10.31.23297818, 2023.
195. Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Syst Rev*. 2024;13:219.
196. Robinson A, Thorne W, Wu BP, et al. 2023. Bio-SIEVE: exploring instruction tuning large language models for systematic review automation. arXiv, preprint arXiv, not peer reviewed. Accessed April 24, 2025. https://doi.org/10.48550/arxiv.2308.06610

197. Syriani E, David I, Kumar G. 2023. Assessing the ability of ChatGPT to screen articles for systematic reviews. arXiv, preprint arXiv, not peer reviewed. Accessed April 24, 2025. https://doi.org/10.48550/arxiv.2307.06464
198. Tran V-T, Gartlehner G, Yaacoub S, et al. Sensitivity and specificity of using GPT-3.5 Turbo models for title and abstract screening in systematic reviews and meta-analyses. *Ann Intern Med*. 2024;177:791-799.
199. Wang S, Scells H, Koopman B, Potthast M, Zuccon G. Generating natural language queries for more effective systematic review screening prioritisation. In: *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 2023:73-83.
200. Wang S, Scells H, Zhuang S, Potthast M, Koopman B, Zuccon G. Zero-shot generative large language models for systematic review screening automation. In: *European Conference on Information Retrieval*. Springer; 2024:403-420.
201. Wilkins D. 2023. Automated title and abstract screening for scoping reviews using the GPT-4 large language model. arXiv, preprint arXiv, not peer reviewed.
202. Yang J, Walker KC, Bekar-Cesaretli AA, et al. Automating biomedical literature review for rapid drug discovery: leveraging GPT-4 to expedite pandemic response. *Int J Med Inform*. 2024;189:105500.
203. Buchlak QD, Esmaili N, Bennett C, et al. Natural language processing applications in the clinical neurosciences: a machine learning augmented systematic review. *Acta Neurochir Suppl*. 2022;134:277-289.
204. Buchlak QD, Clair J, Esmaili N, et al. Clinical outcomes associated with robotic and computer-navigated total knee arthroplasty: a machine learning-augmented systematic review. *Eur J Orthop Surg Traumatol*. 2022;32:915-931.
205. Cao C, Sang J, Arora R, et al. Development of prompt templates for large language model–driven screening in systematic reviews. *Ann Intern Med*. 2025;178:389-401.
206. Agapiou A, Lysandrou V. Interacting with the artificial intelligence (AI) language model ChatGPT: a synopsis of Earth observation and remote sensing in archaeology. *Heritage*. 2023;6:4072-4085.
207. Waffenschmidt S, Knelangen M, Sieben W, et al. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. *BMC Med Res Methodol*. 2019;19:132.
208. Orenstrakh MS, Karnalim O, Suarez CA, Liut M. Detecting LLM-generated text in computing education: comparative study for ChatGPT cases. In: *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE; 2024:121-126.
209. Liang W, Zhang Y, Wu Z, et al. 2024. Mapping the increasing use of LLMS in scientific papers. arXiv, preprint arXiv: 2404.01268, not peer reviewed.
210. Hosseini M, Resnik DB, Holmes KL. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Res Ethics*. 2023;19:449-465.

Journal of the American Medical Informatics Association, 2025, 32, 1071–1086
https://doi.org/10.1093/jamia/ocaf063
Review