

(How) can AI-based automation tools assist with systematic searching?

Artificial Intelligence (AI) methods in evidence synthesis: Cochrane Learning Live webinar series

16 January 2025

**Trusted evidence.
Informed decisions.
Better health.**

Dr Maria-Inti Metzendorf, Public Health and Information Scientist, Cochrane Evidence Synthesis Unit Germany/UK, Sub-Unit Düsseldorf
Irma Klerings, Senior Information Specialist, Cochrane Austria



Conflicts of interest

None known.

Overview

- Terminology
- Overview of search process and tools
- Can AI help with ...?
 - identifying relevant references
 - designing search strategies
 - reviewing a search strategy
 - running search strategies, exporting and aggregating
- Environmental impacts
- Conclusions and outlook

Starting polls (one answer)

- a) What is currently your main role in systematic review production?

- b) Do you use AI for searching?

Terminology

Automation tools

- Software/tools aiding with task execution in systematic reviews (SR) without replacing the skills needed for their conduct. [1]

Artificial intelligence (AI)

- Technology that performs tasks “that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours or solving problems.” [2]
- **Machine learning (ML):** an algorithm is trained on large datasets, allowing it to recognize patterns and make assumptions and predictions. Training data helps teach the machine how to respond to similar data in the future. [3]
 - **Supervised learning:** machine learns through data that has been labeled by a human.
 - **Unsupervised learning:** machine processes loads of data provided by a human and discovers patterns on its own.
- **Large language models (LLM):** first learn language (words, phrases and complete sentences), then learn context to predict what comes next in a sentence. LLMs can generate human-like text, rewrite content, summarize content, and converse. [3]

1. O'Connor AM, et al. Large language models, updates, and evaluation of automation tools for systematic reviews: a summary of significant discussions at the eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR). Syst Rev. 2024 Nov 27;13(1):290.

2. CADTH. Development of an Evaluation Instrument on Artificial Intelligence Search Tools for Evidence Synthesis: Health Technologies. Canadian Agency for Drugs and Technologies in Health; 2024 Oct. AI0001.

3. Warner L. Defining AI: a lexicon for librarians and their patrons. Computers in Libraries, January/February 2025.

Search tools that are already helpful

- **Designing the search strategy**
 - Text analysis: PubReMiner, Yale MeSH Analyzer, Voyant Tools [1], VOSviewer, searchbuildR, Medsyntax
 - Search strategy construction: litsearchR
 - Syntax translation: SRA Polyglot Search Translator (TERA) [2]
- **Deduplication**
 - Deduklick [3], SRA Deduplicator (TERA), Covidence, Rayyan [4]
- **Full text retrieval**
 - EndNote “find full text” feature
- **Screening reduction**
 - Screen4Me [5], Covidence, EPPI-Reviewer (custom) classifiers

1. O'Keefe H, Rankin J, Wallace SA, et al. Investigation of text-mining methodologies to aid the construction of search strategies in systematic reviews of diagnostic test accuracy-a case study. 2023.

2. Clark J, McFarlane C, Cleo G, et al. The Impact of Systematic Review Automation Tools on Methodological Quality and Time Taken to Complete Systematic Review Tasks: Case Study. JMIR Med Educ. 2021.

3. Borissov N, Haas Q, Minder B, et al. Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research. 2022 .

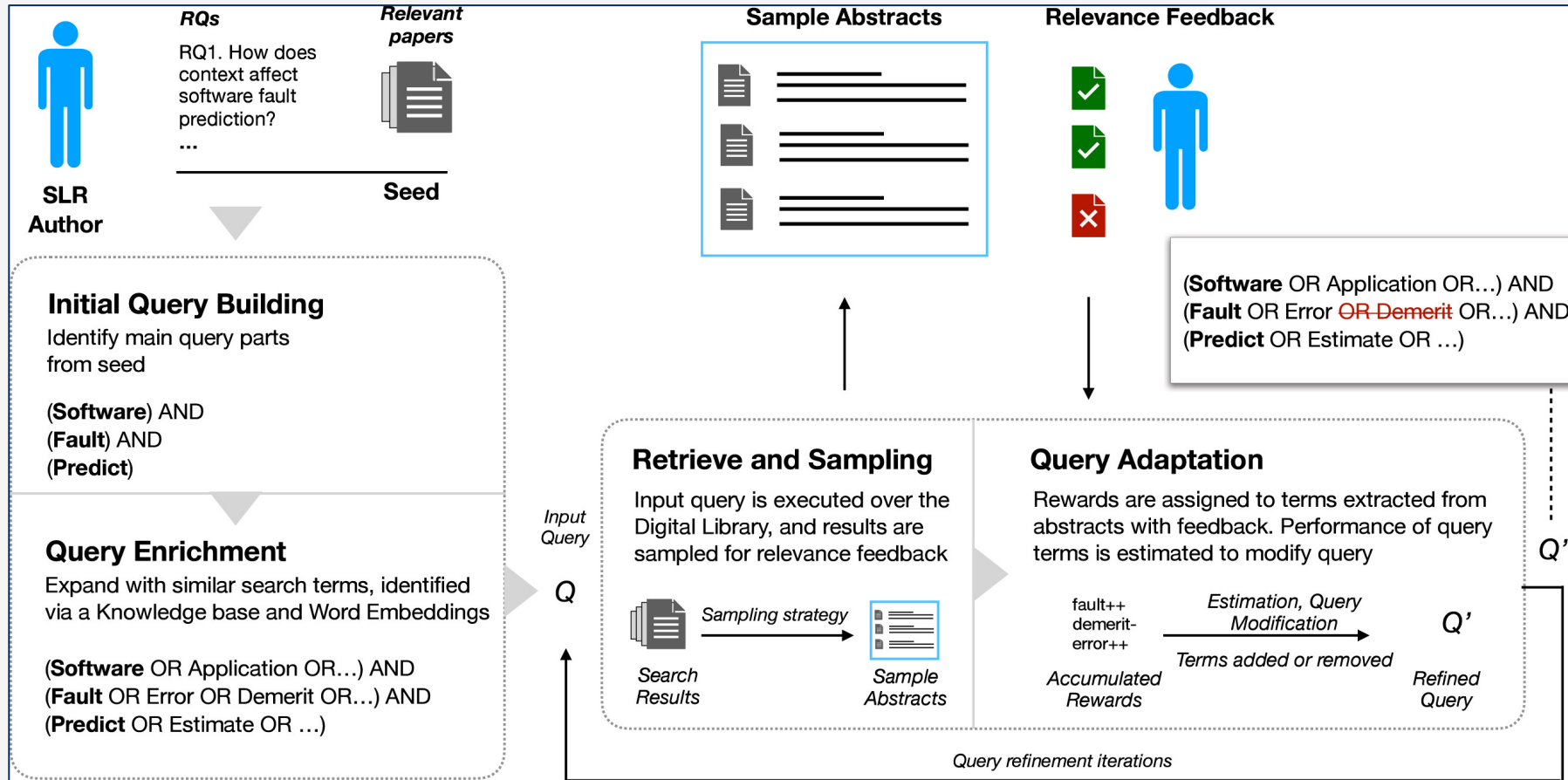
4. Janka H, Metzendorf MI. High precision but variable recall – comparing the performance of five deduplication tools. 2024.

5. Noel-Storr A, Dooley G, Affengruber L, Gartlehner G. Citation screening using crowdsourcing and machine learning produced accurate results: Evaluation of Cochrane's modified Screen4Me service. J Clin Epi. 2020

The search process – specific tasks

- 1) Identify relevant (seed) references
- 2) Design search
 - identify main search concepts
 - identify relevant text words (title/abstract), normalize language (truncation, phrases)
 - identify relevant controlled vocabulary (MeSH, Emtree)
 - construct combination of concepts (Boolean/adjacency operators)
 - test search strategy for recall and precision, implement changes
- 3) Translate search syntax to other databases
- 4) Peer review search strategies
- 5) Run search strategies in different databases/interfaces
- 6) Export results from different databases/interfaces and combine into one set
- 7) Deduplicate merged search results
- 8) Document search
- 9) Import results to screening tool
- 10) [Continually update search]

Searching is an iterative process



Study authors' conclusions of the potential of LLMs for systematic searching (Feb 2024)

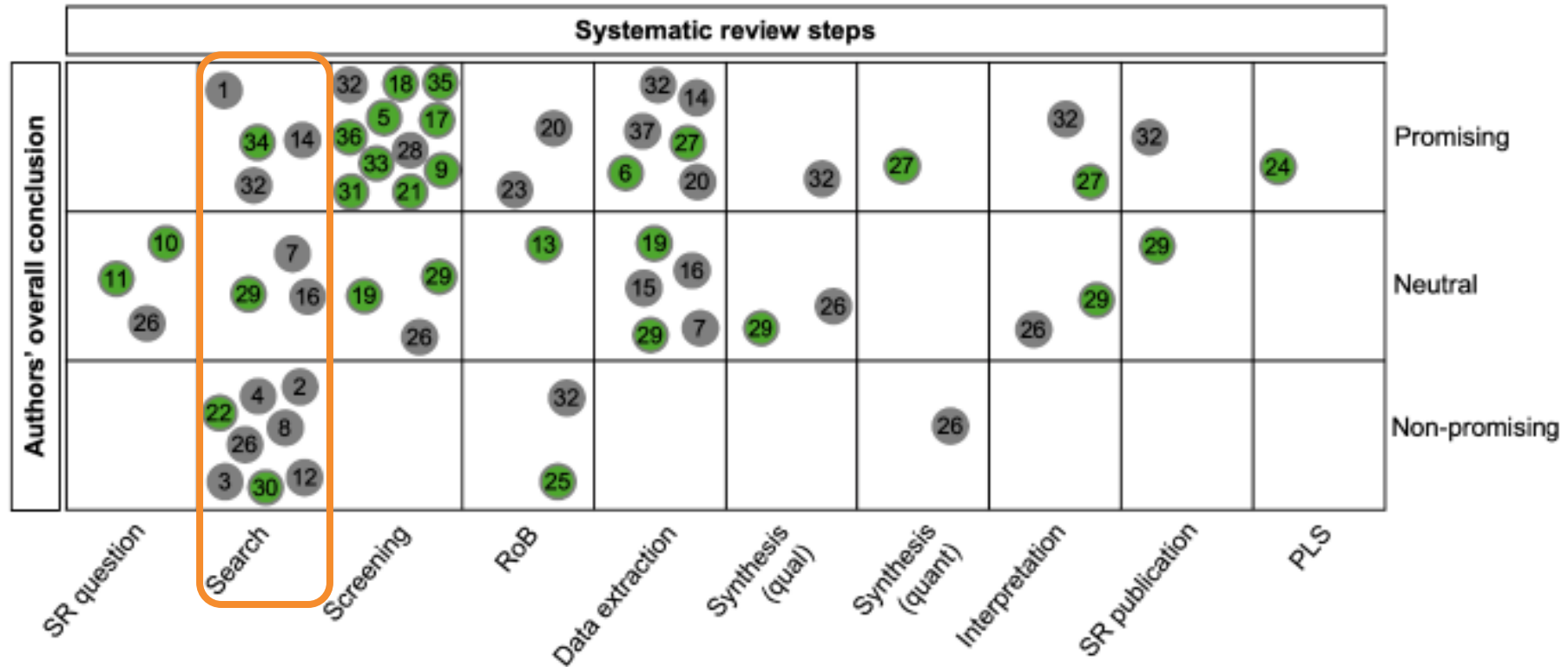


Figure 3: Bubble chart visualizing primary study design (green color: validation studies, grey color: other study designs) and authors' overall categorized conclusion (y-axis) of each SR step (x-axis). Each bubble represents a study with study-ID as listed in Table 1. Studies evaluating several SR steps are represented multiple times accordingly.

Can AI identify relevant references?

- Idea: Retrieve relevant studies based on a research question without complex boolean search strategies
 - Prerequisite: AI needs to retrieve literature from external source (e.g. Retrieval-augmented generation)
 - Potential advantage: No/little search expertise required
 - Potential disadvantage: **No transparency regarding selection criteria or completeness of identified literature**
- ⇒ **Lack of transparency & sensitivity: No replacement for a systematic search strategy, but potentially useful for scoping/exploratory searches**

Example: Elicit & Consensus



<https://elicit.com>



<https://consensus.app/>

- Can be prompted with natural language research question
- Use LLMs to retrieve and summarise papers from Semantic Scholar
- Offer additional filtering and data extraction options (e.g. study design, population, etc.)
- Provide summary of „Top Papers“ **and** a list of results
- **Concerns:** Accuracy of summaries and data extraction, results depend on prompt, retrieved articles not always relevant

Example: Elicit

Prompt: Does the setting - operating room or intensive care unit - affect the rate of complications in external ventricular drain insertion?

Summary

Summary of top 4 papers ▾

Copy

3/4 references not related to question

External ventricular drain (EVD) placement is a common neurosurgical procedure that can be performed in the intensive care unit (ICU) or operating room (OR). A systematic review found inconclusive evidence regarding the impact of setting on EVD complications ([Dawod et al., 2020](#)). A meta-analysis reported absolute risks for EVD-related hemorrhages, infections, and revisions, emphasizing the need for best-practice standards to reduce complications ([Mahto et al., 2022](#)). Regarding EVD weaning, a randomized trial showed no significant difference in shunt implantation rates between rapid and gradual weaning, but gradual weaning led to longer ICU and hospital stays ([Klopfenstein et al., 2004](#)). Another meta-analysis focused on ventriculostomy-related infections (VRI), finding no association with increased mortality or poor neurological outcomes. However, VRI was linked to prolonged EVD placement, extended ICU and hospital stays, and increased internal ventricular shunt placement ([Chadwick et al., 2023](#)).

List

Impact of an Educational Intervention Implanted in a Neurological Intensive Care Unit on Rates of Infection Related to External Ventricular Drains

E. Camacho +6

PLoS ONE

2013 · 49 citations PDF DOI



An educational intervention in a neurological intensive care unit reduced rates of external ventricular drain-related infections.

Reference irrelevant to question

quasi-experimental (before-after intervention) study

Search was limited to SR, MA, RCT

Example:

ResearchRabbit & Semantic Scholar Research Feed

- Information Retrieval based on seed citations
- AI-supported ranking/searching based on data provided by seeds
- Newly identified relevant articles can be added to the seed citations to improve further predictions
- Concern:
 - ResearchRabbit: Results based on citation networks require that studies cite each other
 - Research Feed: finds only papers added within the last 3 months



<https://www.researchrabbit.ai/>



<https://www.semanticscholar.org/>

Can AI design search strategies?

- Idea: AI designs search strategy that can be run in a database (e.g. PubMed)
- Prerequisite: LLM must „know“: requirements for systematic search strategies & database specific syntax
- Potential advantages: fewer errors in search strategies, quicker identification of relevant concepts and synonyms
- Potential disadvantages:
 - Impact of prompt formulation
 - Same prompt leads to different results
 - Hallucinated subject headings (e.g. Mesh)
 - May not work with subscription based databases
 - Recall/sensitivity, precision, specificity may not be adequate of a systematic search strategy

Wang 2023: Test of different prompting strategies for search strategy generation in ChatGPT

	Prompt ID	Prompt
Simple	q1	For a systematic review titled "[review_title]", can you generate a systematic review Boolean query to find all included studies on PubMed for the review topic?
	q2	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. Now you have your information need to conduct research on [review_title]. Please construct a highly effective systematic review Boolean query that can best serve your information need.
Detailed	q3	Imagine you are an expert systematic review information specialist; now you are given a systematic review research topic, with the topic title "[review_title]". Your task is to generate a highly effective systematic review Boolean query to search on PubMed (refer to the professionally made ones); the query needs to be as inclusive as possible so that it can retrieve all the relevant studies that can be included in the research topic; on the other hand, the query needs to retrieve fewer irrelevant studies so that researchers can spend less time judging the retrieved documents.
	q4	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. You are able to take an information need such as: "[example_review_title]" and generate valid pubmed queries such as: "[example_review_query]". Now you have the information need to conduct research on "[review_title]", please generate a highly effective systematic review Boolean query for the information need.
With Example	q5	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the medical literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information need. A professional information specialist will extract PICO elements from information needs in a common practice in constructing a systematic review Boolean query. PICO means Patient/ Problem, Intervention, Comparison and Outcome. PICO is a format for developing a good clinical research question prior to starting one's research. It is a mnemonic used to describe the four elements of a sound clinical foreground question. You are able to take an information need such as: "[example_review_title]" and you generate valid pubmed queries such as: "[example_review_query]". Now you have your information need to conduct research on "[review_title]". First, extract PICO elements from the information needs and construct a highly effective systematic review Boolean query that can best serve your information need.

Step	Prompt	Example Answer
Step1	Follow my instructions precisely to develop a highly effective Boolean query for a medical systematic review literature search. Do not explain or elaborate. Only respond with exactly what I request. First, Given the following statement and text from a relevant study, please identify 50 terms or phrases that are relevant. The terms you identify should be used to retrieve more relevant studies, so be careful that the terms you choose are not too broad. You are not allowed to have duplicates in your list. statement: "Prevalence of Differentiated Thyroid Cancer in Autopsy Studies Over Six Decades: A Meta-Analysis" Text: Ten occult carcinomas of the thyroid gland were found in 274 unselected autopsies at the Pathology Service, Hospital José Joaquín Aguirre, between December 1980 and March 1983. This is the lowest incidence among the most recent published series. The smallest papillary microcarcinomas presented most frequently as nonencapsulated nonsclerosing tumors without inflammatory infiltrate or fibrosis, suggesting that they may represent the early stages of development. Our data show a relatively high and similar frequency of papillary microcarcinomas in surgical and autopsy series, but do not demonstrate risk factors for clinical evolution.	1. Differentiated thyroid cancer 2. Prevalence 3. Autopsy studies 4. Meta-analysis 5. Occult carcinomas 6. Thyroid gland 7. Incidence 8. Environmental factors 9. Etiology 10. Low risk ... 50. Parameters.
Step 2	For each item in the list you created in step 1, classify it into as of three categories: terms relating to health conditions (A), terms relating to a treatment (B), terms relating to types of study design (C). When an item does not fit one of these categories, mark it as (N/A). Each item needs to be categorised into (A), (B), (C), or (N/A).	1. (A) Differentiated thyroid cancer 2. (N/A) Prevalence 3. (C) Autopsy studies 4. (C) Meta-analysis 5. (A) Occult carcinomas 6. (A) Thyroid gland 7. (N/A) Incidence 8. (N/A) Environmental factors 9. (N/A) Etiology 10. (N/A) Low risk ... 50. (N/A) Parameters.
Step 3	Using the categorised list you created in step 2, create a Boolean query that can be submitted to PubMed which groups together items from each category. For example: ((itemA1[Title/Abstract] OR itemA2[Title/Abstract] or itemA2[Title/Abstract]) AND (itemB1[Title/Abstract] OR itemB2[Title/Abstract] OR itemB3[Title/Abstract]) AND (itemC1[Title/Abstract] OR itemC2[Title/Abstract] OR itemC3[Title/Abstract]))	((differentiated[Title/Abstract] OR thyroid[Title/Abstract] OR carcinoma[Title/Abstract] OR papillary[Title/Abstract] OR microcarcinoma[Title/Abstract]) AND (cancer[Title/Abstract] OR incidence[Title/Abstract] OR etiology[Title/Abstract] OR risk[Title/Abstract] OR gender[Title/Abstract] OR hormonal[Title/Abstract] OR nodular[Title/Abstract] OR goiter[Title/Abstract] OR Hashimoto's[Title/Abstract] OR malignancy[Title/Abstract] OR concomitant[Title/Abstract] OR tumor[Title/Abstract] OR infiltrate[Title/Abstract] OR fibrosis[Title/Abstract] OR development[Title/Abstract] OR frequency[Title/Abstract])) AND (autopsy[Title/Abstract] OR surgical[Title/Abstract] OR material[Title/Abstract] OR series[Title/Abstract] OR specimens[Title/Abstract] OR cases[Title/Abstract])).



Staudinger 2024:
Replication of
Wang 2023

Table 3: Obtained average Recall scores with standard deviation for each model and for each prompt type (query). Statistically significantly better performing results (one sample t-test, $p < 0.05$) compared to Wang et al. [54] are indicated by †. Bolded results are the highest overall, and underlined results are the highest for a given query.

CLEF		Recall						
Baseline		0.832						
	Wang et al. [54]	GPT-3.5-1106	GPT-3.5-0125	GPT-4	Mistral-tiny	Mistral-small	Mistral-local	Zephyr
q1	<u>0.129</u>	0.084 ± 0.145	0.019 ± 0.079	0.072 ± 0.142	0.046 ± 0.119	0.062 ± 0.128	0.037	0.015
q2	<u>0.131</u>	0.067 ± 0.125	0.019 ± 0.081	0.093 ± 0.169	0.026 ± 0.071	0.021 ± 0.063	0.025	0.017
q3	<u>0.118</u>	0.115 ± 0.195	0.026 ± 0.106	0.086 ± 0.147	0.041 ± 0.108	0.063 ± 0.123	0.038	0.007
q4-HQE	<u>0.504</u>	0.139 ± 0.212	0.033 ± 0.131	0.086 ± 0.170	0.063 ± 0.166	0.067 ± 0.150	0.020	0.062
q5-HQE	<u>0.334</u>	0.150 ± 0.212	0.027 ± 0.107	0.091 ± 0.142	0.043 ± 0.129	0.060 ± 0.138	0.053	0.005
Seed		Recall						
Baseline		0.711						
Baseline-edit		0.647						
	Wang et al. [54]	GPT-3.5-1106	GPT-3.5-0125	GPT-4	Mistral-tiny	Mistral-small	Mistral-local	Zephyr
q1	0.053	0.148 ± 0.24†	<u>0.203 ± 0.291†</u>	0.132 ± 0.25†	0.132 ± 0.244	0.190 ± 0.293	0.122	0.036
q2	0.039	0.025 ± 0.108	<u>0.147 ± 0.247†</u>	0.141 ± 0.246†	0.047 ± 0.142	0.057 ± 0.152	0.063	0.087
q3	0.052	0.086 ± 0.206†	<u>0.169 ± 0.255†</u>	0.156 ± 0.244†	0.065 ± 0.168	<u>0.214 ± 0.299</u>	0.095	0.002
q4-HQE	0.129	0.213 ± 0.310†	<u>0.237 ± 0.296†</u>	0.145 ± 0.267	0.091 ± 0.184	0.092 ± 0.228	0.121	0.193
q5-HQE	0.079	0.244 ± 0.311†	<u>0.258 ± 0.326†</u>	0.113 ± 0.224†	0.064 ± 0.155	0.084 ± 0.210	0.092	0.153
q4-RE	0.016	0.174 ± 0.256†	<u>0.202 ± 0.280†</u>	0.088 ± 0.205†	0.080 ± 0.171	0.080 ± 0.212	0.055	0.066
q5-RE	—	0.178 ± 0.281	<u>0.267 ± 0.340</u>	0.104 ± 0.205	0.067 ± 0.160	0.075 ± 0.195	0.064	0.060
guided	<u>0.517</u>	0.035 ± 0.130	0.048 ± 0.109	0.125 ± 0.221	0.017 ± 0.090	0.100 ± 0.206	—	—

Adam et al 2024: Creating a specifically trained model

Training Data						
RecordID	Review_title	Review_question	pubmed	X	CRDAccessionNumber	RecordType
1050	Is there evidence for single limb tra	What are the effects of single limb tra	("obstructive lung disease"[tiab]) OR ("copd"[tiab]) OR ("coad"[tiab]) OR ("cobd"[tiab]) OR ("aebc"[tiab]) OR (chronic obstructive lung disease[tiab]) OR (chronic obstructive		42011001050	Clinical
1132	Inspiratory muscle training to facili	In patients who receive mechanical ve	(([muscu* OR muscul*] AND (train* OR condition*)) AND (inspirat* OR respirat* OR ventilat* OR pulmonary)) OR (([socapn* OR normocapn*] AND hyperpn*) OR (inspirat* ANI		42011001132	Clinical
1143	The role of von Willebrand factor i	Does vWF-containing FVIII-concentrat	(([Hemofil-M[tiab] OR Immunate[tiab] OR Emoclot[tiab] OR Replenate[tiab] OR Alphanate[tiab] OR FVIII-LFB[tiab] OR Factane[tiab] OR "Beriarte P"[tiab] OR "Factor VIII SD"[tia		42011001143	Clinical
1157	Effect of non-steroidal anti-inflamr	To assess the anti-inflammatory effect -	([aceclofenac* OR acemetacin* OR Acetylsalicylic* OR alclofenac* OR alminoprofen* OR Aspirin* OR azapropazone* OR Bendazac* OR benoxaprofen* OR benzydamine* OR E		42011001157	Clinical
1182	Inhaled Cannabis for chronic pain	The objective of our systematic review	(Pain[Mesh:no exp] OR pain[tw]) AND (cannabi*[tw] OR Cannabis[Mesh:no exp] OR marihuana[tw] OR Marijuana[tw] OR Hashish[tw] OR Tetrahydrocannabinol[Mesh:no exp]		42011001182	Clinical
1190	Is the lateral position a safe and use	Lying supine/on back when unconsocio	(([Unconsciousness][Mesh] OR "Glasgow Coma Scale"[Mesh] OR unconscious* OR (loss AND conscious*)) OR (minimal AND conscious*)) OR coma* OR pseudocoma* OR ((per		42012001190	Clinical
1192	A systematic review of first trimest	Are there any markers in the first trim	("sensitivity and specificity" OR "sensitivity and specificity/standards" OR "sensitivity" OR "specificity" OR "screening" OR "false positive" OR "false negative") All Fields) OR "acc		42011001192	Clinical
1241	Systematic review of the prevalenc	What is the prevalence (adult lifetime	("Domestic violence"[Mesh] OR "Partner abuse"[Mesh] OR "Partner violence"[Mesh] OR "Spouse abuse"[Mesh] OR "Battered women"[Mesh] OR ([abus* OR batter* OR viole		42011001241	Clinical
1253	Interventions designed to reduce e	The objective of the present study is to	(Health education OR patient education OR out-of-hours OR walk-in centers OR continuing care points OR fast track areas OR fast track unit OR nurse practitioners OR nurs		42011001253	Clinical

Evaluation data							
no	id	source	title	GS_query	included_studies	query_title	altquery_tikq
1	43	Wang	*Prevalence of Differentiated Thyroid Cancer in Autc	((Thyroid Neoplasms OR Adenocarcinoma Follicu	17159249, 2302665, 6504772, 6514426, 2270	[autopsy OR postmortem OR post mx (autopsy OR autopsies OR autopsy studies OR autopsy series OR autopsy findings C	
2	96	Wang	*Prevalence of Differentiated Thyroid Cancer in Autc	((Thyroid Neoplasms OR Adenocarcinoma Follicu	17159249, 2302665, 6504772, 6514426, 2270	[autopsy OR postmortem OR post mx (autopsy OR postmortem OR post mortem OR cadaver OR corpse OR death OR dea	
3	52	Wang	*Comparing the quantity and quality of randomised	(Anti-Bacterial Agents OR Macrolides OR beta-L	1725148, 12236276, 6276973, 17354160, 455	(antibiotic OR antimicrobial OR anti (antibiotics OR antimicrobials OR antibacterial OR antimicro* OR antifungal OR a	
4	53	Wang	*Comparing the quantity and quality of randomised	(Anti-Bacterial Agents OR Macrolides OR beta-L	1773154, 18431478, 26575060, 3525736, 128	(antibiotic OR antimicrobial OR anti (antibiotics OR antimicrobials OR antibacterial OR antimicro* OR antifungal OR a	
5	8	Wang	*Is there a surgeon's effect on patients,ÃO physical h	(Physicians OR Doctor OR Doctors OR Physician	26511481, 28893748, 30174142, 26169083, 1	(surgeon* OR surgeon* OR surger* C (surgeon* OR surgeon* OR surger* OR operation* OR operations OR surgical OR c	
6	103	Wang	*Increased fluid intake to prevent urinary tract infec	((Urinary Tract Infections OR Urinary Tract Infec	30285042, 28608666, 23076891, 29705442, 2	(urinary tract infection OR urinary tr (urinary tract infection OR urinary tract incontinence OR urinary tract infection OR	
7	112	Wang	*Is there a doctors,ÃO effect on patients,ÃO physical	(Physicians OR Doctor OR Doctors OR Physician	15833750, 29992198, 25921589, 29283899, 2	(doctor* OR physician* OR clinician (doctor* OR physician* OR general practitioner* OR GP OR family physician* OR f	
8	51	Wang	*Comparing the quantity and quality of randomised	(Anti-Bacterial Agents OR Macrolides OR beta-L	21923567, 6369096, 2112167, 3129741, 2865	(antibiotic OR antimicrobial OR anti (antibiotics OR antibacterial OR antimicrobial OR antimycobacterial OR antiviral	
9	17	Wang	*Comparison of immunogenicity and safety of licens	(Japan OR Japanese OR Japan) AND (Encephaliti	24717964, 21150279, 19857447, 11027812, 2	(Japanese encephalitis* OR JE) AND ((Japanese encephalitis OR Japanese encephalitis OR Japanese encephalitis OR	

Wang et al. GPT replication data										
id	title	description	question query	included_studies	wang_q1	wang_q2	wang_q3	wang_q4	this_work1	this_work2
43	Prevalence of Differe	Autopsy studies for lu	((Thyroid Neoplasms	17159249, 2302665, 65	("differentiated thyroid cancer"	("differentiated thyroid cancer" C	("differentiated thyroid cancer"	("Differentiated Thyroid Canc	("differentiated thyroid cancer" OR "p	("differentiated thyr
96	Prevalence of Differe	Incidental Thyroid C	((Thyroid Neoplasms	17159249, 2302665, 65	("differentiated thyroid cancer"	("differentiated thyroid cancer" C	("differentiated thyroid cancer"	("Differentiated Thyroid Canc	("differentiated thyroid cancer" OR "p	("differentiated thyr
52	Comparing the quan	Antibiotics for UTIs	("Anti-Bacterial Age	1725148, 12236276, 62	("randomized controlled trial"	("randomized controlled trial"	("randomized controlled trial"	("randomized controlled trial"	("randomized controlled trial"	("randomized contro
53	Comparing the quan	Antibiotics for ARIs	("Anti-Bacterial Age	1773154, 18431478, 26	("randomized controlled trial"	("randomised controlled trial"	OF ("randomized controlled trial"	("randomized controlled trial"	("randomized controlled trial"	("randomized contro
8	Is there a surgeon's e	Doctors lifestyle imp	(Physicians[Mesh] O	26511481, 28893748, 3	((("surgeon's effect" OR "surgeo	("surgeon's effect" OR "surgeon e	("surgeon's effect" OR "surgeon	("surgeon's effect" OR "surgeon	("surgeon effect"[All Fields] Of	((("surgeon's effect" OR "surgeon effect"
103	Increased fluid intake	UTI self management	((Urinary Tract Inf	30285042, 28608666, 2	((("fluid intake" OR "fluid consum	("fluid intake" OR "fluid consum	("fluid intake" OR "fluid consum	("fluid intake"[All Fields] OR "fl	((("fluid intake" OR "water intake" OR	((("fluid intake" OR "wa
112	Is there a doctors,ÃO	Doctors lifestyle imp	(Physicians[Mesh] O	15833750, 29992198, 2	((("doctor" OR "physician" OR "h	("doctor's effect" OR "physician's	("doctor's effect" OR "physician	("doctors' effect"[All Fields] OI	((("physician-patient relations" OR "doc	((("physician-patient r
51	Comparing the quan	Antibiotics for skin di	("Anti-Bacterial Age	21923567, 6369096, 21	("randomized controlled trial"	("randomised controlled trial"	OF ("randomized controlled trial"	("randomized controlled trial"	("randomized controlled trial"	("randomized contro
17	Comparison of immu	Japanese encephalit	("Japan"[Mesh] OR J	24717964, 21150279, 1	((("Japanese encephalitis" OR "J	("Japanese encephalitis" OR "JE"	((("Japanese encephalitis vaccin	("Japanese encephalitis"[MeS	((("Japanese encephalitis" OR "Japanes	((("Japanese encephal

Figure 2. Data. The top row shows the structure of the data used to train/validate the models. It includes the title and key question in free text along with a manually curated PubMed search query. The middle row shows the structure of the data used to evaluate the models. It shows the title source (Wang/Adam), the free-text title, the query used in the original review (GS_query), the PMIDs of the included studies, and the 2 model-generated queries, using title only and using title + key questions). The bottom row shows the structure of the data used to replicate Wang et al's GPT analysis. It shows the free-text title and key questions, the query used in the original review (query), the PMIDs of the included studies, and the model-generated queries, for each prompt.

Table 1. Summarized results for each model on the evaluation set.

	Sensitivity, %	Precision, %	NNR, N	Word count, N	Failed queries, N (%) ^b
Overall					
Human query ^a	100 (88-100)	0.17 (0.03-0.62)	580 (161-3466)	144 (96-240)	0 (0%)
Mistral-Instruct-7b trained on title and key questions query	86 (51-100)	0.11 (0.03-0.58)	908 (171-3906)	66 (49-83)	2 (3%)
Mistral-Instruct-7b trained on title only query	85 (40-100)	0.08 (0.02-0.49)	1206 (205-5810)	71 (55-83)	1 (2%)

Table 2. Summarized results for each model on the full evaluation set.

	Sensitivity, %	Precision, %	NNR, N
Wang prompts 1-4, GPT-4			
Prompt 1	0 (0-5.55)	1.61 (0.50-3.70)	62 (27-202)
Prompt 2	0 (0-9.11)	3.85 (0.50-1.00)	26 (1-201)
Prompt 3	0 (0-2.55)	4.76 (0.61-33.3)	21 (3-163)
Prompt 4	0 (0-2.0)	1.15 (0.27-3.70)	87 (27-368)
Updated prompts, GPT-4			
New prompt 1	12.5 (0-45.4)	1.61 (0.66-4.35)	62 (23-151)
New prompt 2 GPT-4	5.9 (0-27.6)	3.45 (0.99-8.33)	29 (12-101)
Fine-tuned model			
Mistral-Instruct-7b trained on title and key questions	86 (51-100)	0.11 (0.03-0.58)	908 (171-3906)
Mistral-Instruct-7b trained on title only	85 (40-100)	0.08 (0.02-0.49)	1206 (205-5810)

Results are median (worst performance-best performance).
Abbreviations: N = number or count; NNR = number needed to read.

Can AI design search strategies?

AI-designed searches can be run in PubMed but, they

- Either have low sensitivity (= miss relevant studies) or low precision (= higher screening workload) or both
- Need to be heavily revised to be fit for purpose: requires expertise

⇒ **May be useful as a starting point for search strategy development, but not as unique/unaudited method**

Can AI "peer" review a search strategy?

Apparently little/no research on this topic(?)

Our tests with ChatGTP 4 were not promising:

- Identified some errors in search strategies, but not all of them
- Suggested incorrect or inappropriate revisions

Prompt: Please identify errors in this Ovid MEDLINE search strategy, such as missing lines, spelling mistakes and wrong search filters used.

Answer: The provided Ovid MEDLINE search strategy has several errors, including spelling mistakes, missing lines, and incorrect search filters. Here is a detailed identification of the errors:

Line 7: The "or/1-5" line should be "or/1-6" to include all previous lines related to Insulin Lispro.

→ CORRECT

Line 8: "Insulin Asport/" should be corrected to "Insulin Aspart/".

→ CORRECT

Line 14: "insulin glulisine.mp." should ideally use the MeSH term if available, i.e., "Insulin Glulisine/".

→ Partially CORRECT: The MeSH term does not exist (yet).

Line 22: "Insulin/aa" is incorrect. This is not a standard field tag. If the aim is to use the subheading for analogs & derivatives, it should be specified properly or corrected.

→ INCORRECT: This is correct Ovid syntax.

Line 31: There is a misplaced quotation mark. "type1" should be "type 1".

→ Partially CORRECT: It is not a misplaced quotation mark, but a missing blank, which has been used in purpose to capture possible errors.

Line 32: The wildcard is incorrect. It should be "insulin* depend*" or "insulin-dependent" without the question mark.

→ INCORRECT: This is correct Ovid syntax.

Line 44: "exp humans/ not animals/" should be "exp animals/ not humans/" to correctly exclude animal studies.

→ CORRECT

Line 50: "34 and 50" should be "34 and 49".

→ CORRECT

Line 51: Date range selection is missing. It should include the proper date range filter.

→ INCORRECT: This is correct Ovid syntax.

Prompt: Please identify errors in this PubMed search strategy, such as missing lines, wrong Boolean operators, spelling mistakes, and wrong search filters used. The research question is: Is there an association of autistic spectrum disorder and measles vaccines?

Search number	Query	Filters	Results
10	#5 AND #8	Systematic Re	8
9	#5 AND #8		224
8	#6 OR #7		81,834
7	("Autistic Disorder"[Mesh]) AND "Autism Spectrum Disorder"[Mesh]		27,681
6	autistic disorder OR autism		81,834
5	#4 AND#1		130,176
4	#2 OR #3		285,141
3	vaccination[Title/Abstract] OR vaccine[Title/Abstract]		284,214
2	("Vaccination"[Mesh]) AND "Measles Vaccine"[Majr]		2,540
1	"Measles"[Mesh] OR measles[tiab]		30,821

Positive:

- Recognised a problem and suggested a solution.
- Identified spelling error and missing field tags.

7. Boolean Operator Logic Issues

- Query 5: #4 AND#1 implies combining measles-related terms (Line 1) with vaccine terms (Line 4). However, Line 4 includes vaccine terms broadly without specific focus on measles
- This could lead to many irrelevant results. Refine Line 4 to focus on measles vaccines explicitly.

This prompt: <https://chatgpt.com/share/6783f9cd-ad1c-8007-9cba-a252115ce04e>

Alternative prompt: <https://chatgpt.com/share/67878ec5-2a80-8007-9f4b-9ccc4897166d>

Alternative prompt: <https://chatgpt.com/share/678791e3-bad8-8007-8c9f-b52aaa27c29c>

Negative:

- Added new logic error
- Produced orphan line
- Added inadequate concept

Corrected Example Search Strategy:

1. "Measles"[Mesh] OR measles[Title/Abstract] OR "Measles Vaccine"[Mesh]
2. "Vaccination"[Mesh] OR vaccine[Title/Abstract]
3. vaccination[Title/Abstract] OR vaccine[Title/Abstract]
4. #1 OR #2
5. autistic disorder[Title/Abstract] OR autism[Title/Abstract]
6. "Autism Spectrum Disorder"[Mesh]
7. #5 OR #6
8. #4 AND #7
9. #8 AND association[Title/Abstract] OR relationship[Title/Abstract]
10. #9 AND systematic[sb]

Can AI run search strategies, export and aggregate?

- Time-intensive tasks such as updating searches or running them on a regular basis can currently only be partially automated.
- AI is of "no great help" with supporting "mechanical" steps of a search (run, export, import)
- For running, exporting, aggregating, importing searches we would need:
 - scripts that can connect to databases via APIs (not openly available for most database)
 - tools that would include functionalities of reference management software
- Subscription-based databases play an important role and restrict the possibility of automation.

Environmental impacts of AI

- LLMs require substantial computational power for training and usage
- Required hardware systems and related energy consumption lead to a large carbon footprint, water usage, and significant impacts on mining [1,2].
- **Compute and energy costs:** powering search with generative AI uses at least **4 to 5 times more computing power** than standard search [3].
- “**Google** says that its ML workloads represented about 15% of its total energy use over the last three years.” [4]
- “**A large cloud compute provider** estimates that between 7-10% total spend on compute infrastructure supports AI applications, with 3-4.5% used for training ML models and 4-4.5% spent using these models.” [4]
- We must recognize these environmental impacts and promote a responsible and sustainable use of LLMs for specific tasks in evidence synthesis and searching.



1. Berthelot A, Caron E, Jay M, et al. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. Proc CIRP 2024;122:707-12.
2. Coulson A. The environmental impact of AI in the lab: a double-edged sword? Biotechniques. 2024;76(8):353-356.
3. www.insiderintelligence.com/content/ai-search-s-high-costs-could-vicious-cycle-big-tech-eyes-profitability, 13 Feb 2023.
4. OECD. Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint, OECD Digital Economy Papers, No. 341, 2022.

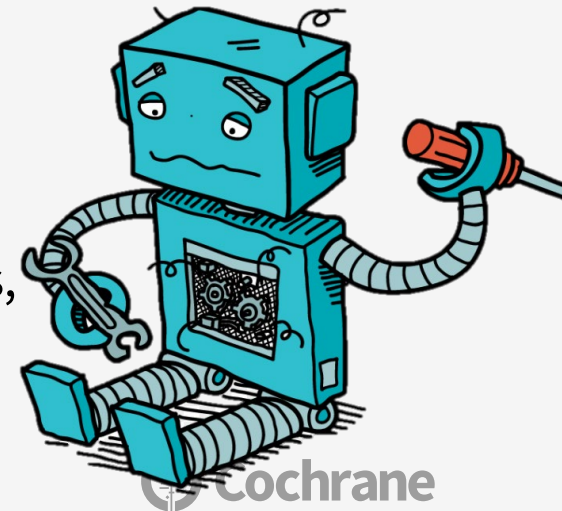
Conclusions based on literature & tests in January 2025

Observations

- AI/LLM-supported academic search engines: Useful for exploratory searches, but no alternative to full systematic search
- AI/LLM-designed search strategies: Inadequate balance of sensitivity and precision
- AI/LLMs may identify some errors in a search strategy but cannot replace true peer review
- Task-specific tools may work better than general purpose chatbots
- LLMs have a significant environmental impact, should only be used where they work best

Conclusions

- Potentially helpful as a starting point, but expertise needed to review/revise/contextualise outputs
- AI development for SR tasks should ideally be multidisciplinary (computer scientists, information and library scientist, systematic reviewers) to ensure robust and adequate tools
- Cannot replace information specialists (yet?)



Outlook: AI could lead to new systematic search approaches

- Design even more sensitive search strategies, if screening can be substantially and robustly reduced by AI according to inclusion/exclusion criteria
- We might have different search interfaces for databases
 - typing in PICO questions
 - prompts: „Please, develop a PubMed search strategy that identifies these 12 relevant references: PMID 1234567, PMID 4567899, PMID 23789879...“
- EPPI-Reviewer for updating existing reviews: combining machine learning classifiers and OpenAlex citation searching [1,2]
- improve interdisciplinary information seeking [3]

1 EPPI-Centre. "Machine learning functionality in EPPI-Reviewer." from https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf.

2. EPPI-Centre. (2022). "OpenAlex in EPPI-Reviewer." from <https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3754>.

3. Zheng C, Zhang Y, Huang Z, Shi C, Xu M, Ma X. DiscipLink: Unfolding Interdisciplinary Information Seeking Process via Human-AI Co-Exploration. ArXiv. 2024;abs/2408.00447.

Ending poll (several answers)

For which steps of the systematic searching process do you consider AI to be most useful?

Thank you... and key references

- CADTH. Development of an Evaluation Instrument on Artificial Intelligence Search Tools for Evidence Synthesis: Health Technologies [Internet]. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health; 2024 Oct. Report No.: AI0001. PMID: 39652682.
- Lieberum JL, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use – a scoping review. [preprint]. 2024 December. doi: 10.1101/2024.12.19.24319326.
- O'Connor AM, et al. Large language models, updates, and evaluation of automation tools for systematic reviews: a summary of significant discussions at the eighth meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Syst Rev*. 2024 Nov 27;13(1):290. doi: 10.1186/s13643-024-02666-2.
- OECD. Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint, OECD Digital Economy Papers, No. 341, 2022.
- Parisi V., Sutton A. The role of ChatGPT in developing systematic literature searches: an evidence summary. *Journal of EAHIL*. 2024, 20(2), 30-34. doi: 10.32384/jeahil20623
- Siemens W, et al. Opportunities, challenges and risks of using artificial intelligence for evidence synthesis. *BMJ Evid Based Med*. 2025 Jan 9;. doi: 10.1136/bmjebm-2024-113320.