Large Language Models in der systematischen Literaturrecherche – eine Evidenzübersicht

Uses for large language models in systematic literature searching – an overview of the evidence

Irma Klerings[1]

1 Department für Evidenzbasierte Medizin und Evaluation, Universität für Weiterbildung Krems, Österreich

Abstract

Background: The methodology of systematic literature searching requires that the information retrieval process has a high recall and is as transparent and reproducible as possible. The introduction of large language models (LLMs) such as ChatGPT raised the expectations for automation of evidence synthesis processes through artificial intelligence (AI). However, assessments of the usefulness of LLMs for systematic searching are heterogeneous. This narrative review examines the current evidence on the use of LLM tools compared to systematic searches performed by humans (as of August 2025).
Results: The majority of studies focus on two areas of application: the creation of Boolean search strategies by LLMs and the generation of comprehensive literature lists using AI-supported search platforms (Elicit, Consensus). In both cases, AI tools achieved insufficient recall rates compared to traditional systematic search methods. However, AI-supported search platforms were able to identify additional studies that were not found by Boolean search strategies. Few studies investigated the use of LLMs for error detection in database search strategies. AI was able to find errors, but there were problems in creating improved search strategies.
Conclusion: Based on the available evidence, AI-supported methods should at most be used to complement established methods of systematic literature research. On their own, they neither achieve the necessary high recall nor are their results reproducible. However, there are also significant gaps in the evidence. Independent evaluations and critical assessment of AI tools by users remain essential.

Keywords: artificial intelligence, large language models, systematic literature searching

Zusammenfassung

Hintergrund: Die Methodik systematische Literaturrecherche stellt hohe Anforderungen an Transparenz, Reproduzierbarkeit und Vollständigkeit von Literatursuchen. Die Einführung von Large Language Models (LLMs) wie ChatGPT hat der Automatisierung von Evidenzsynthese-Prozessen durch künstliche Intelligenz (KI) neuen Aufschwung gegeben. Allerdings sind Einschätzungen des Nutzens von LLMs für die systematische Suche heterogen. Diese narrative Übersichtsarbeit untersucht die aktuelle Evidenzlage zur Anwendung von LLM-Tools im Vergleich zu von Menschen durchgeführten systematischen Suchen (Stand: August 2025).
Ergebnisse: Die Studienlage konzentriert sich auf zwei Anwendungsbereiche: Erstellung boolescher Suchstrategien durch LLMs und Generierung von umfassenden Literaturlisten mit KI-gestützten Rechercheplattformen (Elicit, Consensus). In beiden Fällen erzielten KI-Tools unzureichende Recall-Raten im Vergleich zu klassischen systematischen

Suchmethoden. Allerdings konnten KI-gestützte Such-Plattformen zusätzliche Studien identifizieren, die von Booleschen Suchstrategien nicht gefunden wurden. Wenige Studien untersuchten die Verwendung von LLMs zur Fehlererkennung in Datenbank-Suchstrategien. KI war in der Lage Fehler zu finden, allerdings gab es Probleme bei der Erstellung von verbesserten Suchstrategien.
Schlussfolgerung: Basierend auf den verfügbaren Evaluationen, sollten KI-gestützte Methoden höchstens komplementär zu etablierten Methoden der systematischen Literaturrecherche angewandt werden. Allein erreichen sie weder den notwendigen hohen Recall, noch sind ihre Ergebnisse reproduzierbar. Die Evidenzlage weist allerdings auch erhebliche Lücken auf. Unabhängige Evaluationen und die kritische Bewertung von KI-Tools durch Anwender*innen bleiben essenziell.

Schlüsselwörter: künstliche Intelligenz, Large Language Models, systematische Literaturrecherche

Einleitung

Die systematische Literaturrecherche bildet eine methodische Grundlage für alle Arten von systematischen Evidenzsynthesen (z.B. Systematic Reviews, Rapid Reviews, Evidence Maps). Im Gegensatz zu anderen Literatursuche-Prozessen stellt sie hohe Anforderungen an Transparenz, Reproduzierbarkeit und Vollständigkeit des Rechercheprozesses [1], [2]. Während Automatisierungs-Ansätze für die Erstellung von Evidenzsynthesen seit langem untersucht und entwickelt werden [3], hat die Verfügbarkeit von Large Language Models (LLMs) wie ChatGPT, Claude oder Gemini dem Thema neuen Aufschwung gegeben: Neue Tools und eine steigende Anzahl von Publikationen über diese Tools versprechen, dass künstliche Intelligenz (KI) den langwierigen Evidenzsynthesen-Prozess schneller und einfacher macht.
Allerdings scheinen nicht alle Bereiche dieses Prozesses in gleichem Maße für LLM-Anwendungen geeignet. Ein Scoping Review [4] mit Literatur bis Anfang 2024 identifizierte zahlreiche Publikationen zur LLM-Anwendung bei der Erstellung von Evidenzsynthesen. In der Übersicht zeigte sich, dass Studienautor*innen den Nutzen von LLMs für Literaturscreening und Datenextraktion als überwiegend vielversprechend oder schlimmstenfalls neutral sahen. In anderen Bereichen, wie Risk of Bias Bewertung und Literatursuche, gab es keinen einheitlichen Trend: Manche Studien bewerteten LLM-Nutzung als vielversprechend, andere als neutral, aber – insbesondere bei der Literatursuche – gab es auch viele negative Bewertungen.
Dieses breite Spektrum an Einschätzungen stellt alle, die systematische Literatursuchen erstellen, vor ein praktisches Problem: Gibt es unter den vielen verfügbaren KI-Tools und Methoden, solche, die für den Suchprozess tatsächlich nützlich sind? Und wenn ja, welche?

Der vorliegende Beitrag untersucht den aktuellen Evidenzstand zur KI-gestützten Literatursuche und versucht folgende Fragen zu beantworten:

1. Gibt es Studien, die die Verwendung von KI (besonders LLMs) im systematischen Suchprozess im Vergleich zu etablierten Methoden evaluieren?
2. Lassen sich daraus Empfehlungen für das praktische Vorgehen bei der systematischen Suche ableiten?

Kontext

Um diese Fragen zu beantworten, müssen zuerst einige Grundlagen geklärt werden: der Aufbau und die Anforderungen der systematischen Literaturrecherche, und die Art von KI, um die es im Weiteren geht.

Systematische Literatursuche: Anforderungen und Prozess

Systematische Literatursuchen zielen auf die Identifikation möglichst aller relevanten Studien zu einer Fragestellung ab, der Prozess priorisiert also die größtmögliche Vollständigkeit des Suchergebnisses (hoher Recall). Zudem muss der Prozess transparent dokumentiert und so reproduzierbar wie möglich sein. Diese drei Elemente, Vollständigkeit, Transparenz und Reproduzierbarkeit haben das Ziel, eine Verzerrung der Ergebnisse der Evidenzsynthese aufgrund der verwendeten Literatur zu minimieren.
Der systematische Suchprozess umfasst typischerweise die folgenden Schritte:

- Scoping/Explorative Suchen: Überblick über die Fragestellung und relevante Literatur, Identifikation von relevanten „seed citations“ für weitere Suchschritte, Identifikation von relevanten Informationsquellen (Datenbanken, Journals, Organisationen, etc.) für die weitere Suche,

- Entwicklung der primären Datenbank-Suchstrategie: Konzeptidentifikation, Textwörter, kontrolliertes Vokabular,
- Übersetzung der Suchstrategie auf andere Datenbanken/Suchoberflächen,
- Peer-Review der Suchstrategien,
- Durchführung der Datenbank-Suchen und Export der Suchergebnisse,
- Zusätzliche Suchmethoden: z.B. Citation Searching, Handsuche, Websuche,
- Transparente Dokumentation aller Schritte des Suchprozesses.

Diesem Prozess folgen die Deduplizierung aller Suchergebnisse und die systematische Literaturauswahl (Title/Abstract und Fulltext-Screening).
Für die Anwendung von KI-Tools bei der systematischen Literaturrecherche sind zwei Ansätze denkbar: Entweder werden einzelne Schritte des Prozesses mit KI unterstützt/ersetzt aber der Prozess selbst bleibt unverändert, oder der gesamte Prozess wird durch KI-Nutzung umgestaltet/ersetzt.

Automatisierung, Künstliche Intelligenz, Large Language Models

Künstliche Intelligenz (KI) kann definiert werden als Technologie, die Aufgaben ausführt, für die normalerweise biologische Intelligenz erforderlich wäre (z.B. das Verstehen gesprochener Sprache, das Erlernen von Verhaltensweisen oder das Lösen von Problemen) [5].
Davon lassen sich andere Technologien unterscheiden, die auf mechanische Automatisierung setzen, beispielsweise das automatische Syntax-Mapping von Polyglot Search Translator [6] oder die gewichtete Textanalyse von searchbuildR [7].
Im Kontext der systematischen Literatursuche bezieht sich der Begriff „KI“ primär auf Large Language Models (LLMs). Bei LLMs handelt es sich um eine Form generativer KI, bei der Machine Learning-Algorithmen verwendet werden, um neue Inhalte auf der Grundlage von Mustern zu erstellen, die aus Trainingsdaten erlernt wurden. Konkret dienen LLMs der Texterstellung, sie sind „Chat bots“ [8], [9].
LLMs weisen Charakteristika auf, die potenziell die Erfüllung der Anforderungen an systematische Suchen – insbesondere Transparenz und Reproduzierbarkeit – erschweren [10], [11], [12]:

- Black Box-Problem: Intransparenz der Prozesse, die zum Output führen. Aufgrund ihrer Komplexität ist es selbst den Entwickler*innen kaum möglich die „Entscheidungen“ eines LLMs basierend auf einem Prompt zu erklären oder nachzuvollziehen.
- Bias: Verzerrungen im Output, in den Charakteristika des Trainingsmaterials oder im Modelldesign. LLMs könnten beispielsweise Vorurteile oder Falschinformationen reproduzieren, die in den Trainingsdaten prävalent waren. Andererseits könnten spezifische Funktionen eines KI-Tools das Output verzerren (beispielsweise indem die Trefferzahl der Suche unabhängig von der Fragestellung festgelegt wird).
- Halluzinationen: Generierung faktisch falscher, aber formal plausibler Inhalte. Zum Auftreten von Halluzinationen können verschiedene Faktoren beitragen, die mit den ersten beiden Charakteristika zu tun haben: den verwendeten Trainingsdaten und -methoden sowie den Methoden die das LLM verwendet, um Antworten zu generieren.

Zuletzt sind LLMs auch mit einem erheblichen Ressourcenverbrauch verbunden: Training und Nutzung der Modelle hat einen großen Energie-, Wasser- und Rohstoffbedarf [13], [14], [15].
Diese bekannten Limitationen von LLMs machen es umso wichtiger ihren Nutzen für konkrete Anwendungen zu evaluieren, um ungewollte negative Auswirkungen auf das Endergebnis zu vermeiden.

Quellen und Methodik

Diese narrative Übersichtsarbeit basiert auf:

- Literaturübersichten zur KI-Nutzung in Evidenzsynthesen [16], [17], [18],
- der „Living Evidence Map“ von Farhad Shokraneh [19] (Stand: Ende August 2025),
- Literatur-Surveillance (Semantic Scholar research feed, Embase.com Search Alerts, Stand: Ende August 2025)

Sie inkludiert sowohl publizierte Artikel als auch Preprints. Es wurden nur Studien berücksichtigt, die Performanz der KI-Methode im Vergleich zu menschlicher Arbeit berichteten. Manche Quellen [16], [18] beschränkten sich explizit auf den gesundheitswissenschaftlichen Kontext.

Evidenzübersicht

Zurzeit gibt es Evaluationen zu drei Anwendungsbereichen, bei denen KI-Methoden mit etablierten systematischen Suchen verglichen wurden:

- Die Erstellung von booleschen Suchstrategien,
- die Erstellung von Literaturlisten,
- die Fehlererkennung bei Suchstrategien.

Können LLMs systematische Suchstrategien entwickeln?

Der Großteil der Studien zum Thema KI-Nutzung im systematischen Suchprozess beschäftigt sich mit der Erstellung von Booleschen Datenbank-Suchstrategien, in den meisten Fällen für PubMed [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32].
Bei diesem Vorgehen erstellt ein LLM eine Datenbank-spezifische Suchstrategie basierend auf einem Prompt, das die Forschungsfrage und eventuell zusätzliche Informationen beinhaltet. Dabei bleibt die Transparenz und Reproduzierbarkeit des Suchprozesses an sich gewähr-

leistet, da die Datenbank-Suche an sich ohne KI vonstattengeht.

In der Evidence Map von Adam et al. [16] wurden acht solche Studien ausgewertet und mit der üblichen Performance von Menschen-generierten Suchstrategien verglichen (Median Recall (88%, Range: 65–100%) und Präzision (2%, Range: 1,7–2,2%)). Dabei zeigte sich, dass sogenannte „zero shot prompts" bei denen ein generisches LLM (z.B. ChatGPT, Claude) ohne weiteres Training eine Suchstrategie erstellt, schlecht funktionierten, und zwar unabhängig vom verwendeten Modell und Prompting-Ansatz: Der durchschnittliche Recall lag bei 4–31,6%. Dieser Trend ist auch in neueren Studien [21], [32], die nicht in der Evidence Map berücksichtigt wurden, vorhanden: Durchschnittlicher Recall für zero shot prompts lag zwischen 15% und 43,6%, wobei die Performanz in der Regel große Schwankungsbreiten für dieselben Prompts im selben Modell aufwies.

Vielversprechender waren in der Evidence Map [16] zwei Studien, die „fine-tuned" Modelle verwenden. Dabei wurden existierende LLMs mit tausenden, für die konkrete Aufgabe relevanten, Datensets weiter trainiert. Das Ergebnis war ein hoher Recall bei gleichzeitiger extrem niedriger Präzision: Bei Pourezza et al. [31] hatte der Ansatz mit dem höchsten Recall (96,76%) eine Präzision von 0,03%, bei Adam et al. [23] waren es 91,83% und 0,15%. Allerdings handelt es sich bei diesen Publikationen um Pilotstudien, die noch keine fertigen Anwendungen für andere User*innen zur Verfügung stellen.

Die verhältnismäßig große Anzahl von Studien zu dem Thema KI-Suchstrategie-Entwicklung zeigt auch Probleme bei der Reproduzierbarkeit der KI-Evaluationen selbst auf. Zwei Studien [23], [27], die versuchten, den Ansatz einer einflussreichen, 2023 veröffentlichten Studie [22] zu replizieren, erreichten schlechtere Ergebnisse als in der Originalpublikation berichtet wurden. In ihrer Antwort auf Staudinger et al. [27] wiesen die ursprünglichen Studienautor*innen [21] darauf hin, das unter anderem query validation („manual removal of incorrectly generated queries") und query refinement („improving an initial query automatically") grundlegende Bestandteile ihrer Evaluation waren. Es zeigt sich also, dass nach dem ersten Output weitere Schritte notwendig waren, um die Generierung ausführbarer Suchstrategien -unabhängig von ihrer Performanz – zu gewährleisten.

Ausgehend von diesen Studien und Übersichtsarbeiten ist momentan Fachexpertise notwendig, um von LLMs erstellte Suchstrategien zu bewerten und zu überarbeiten. Es ist unklar, ob diese Vorgehensweise zu Zeitersparnissen oder einer Verbesserung der Qualität von Suchstrategien im Vergleich zu anderen Methoden führt.

Können LLMs relevante Literatur vollständig identifizieren?

Ein anderer Ansatz ist die Erstellung von umfassenden Literaturlisten durch KI-Tools. Diese Vorgehensweise zielt darauf ab, den gesamten Prozess der systematischen Suche durch einen Prompt-basierten Vorgang zu ersetzen. Ein auf der Recherchefrage basierender Prompt soll dabei alle relevanten Publikationen identifizieren. Da hier das KI-Tool selbst für das Auffinden der relevanten Literatur verwendet wird, haben LLM-spezifische Probleme wie Halluzinationen und variierende Ergebnisse für gleichbleibende Prompts einen direkten Einfluss auf das Ergebnis der Suche.

Dieses Thema wurde weniger ausführlich erforscht, aber es gibt Studien, die ChatGPT [33], [34], [35], [36], Elicit [36], [37], [38], Consensus [36], [39] und andere Anwendungen [33], [34], [35] im Vergleich mit systematischen Suchergebnissen untersuchen. Übersichtsarbeiten von Clark et al. [17] und Adam et al. [16] zeigten, dass für ChatGPT der Recall bei dieser Vorgehensweise bei 4–14% der relevanten Literatur lag. Zusätzlich wurden hohe Halluzinationsraten – fehlerhafte oder nicht existierende Literaturangaben – beobachtet.

Die Verwendung von Elicit und Consensus [36], [37], [38], [39] – KI-gestützten Plattformen für wissenschaftliche Literatur, die Retrieval Augmented Generation nützen – war etwas besser mit durchschnittlichen Recall-Raten von 19,6% bis 43,5%. Eine Studie [37] zeigte allerdings, dass die Verwendung eines gleichbleibenden Prompts in Elicit zu unterschiedlichen Ergebnissen führte. Andererseits fanden zwei weitre Evaluationen von Elicit [37], [38] in den KI-generierten Suchergebnissen eine kleine Anzahl an relevanten Artikeln, die durch die klassischen systematischen Suchen nicht gefunden worden waren. Die Autor*innen kamen zum Schluss, dass KI-gestützte Suchen keine systematischen Suchen ersetzen aber als komplementäre Suchmethoden im systematischen Prozess dienen können.

Die Evaluationsstudien zeigen, dass für die Verwendung von LLMs zur Literatursuche unbedingt eine Anbindung an externe Quellen (Websuche, Literaturdatenbanken) notwendig ist, um Halluzinationen zu vermeiden. Aber selbst KI-Tools, die für wissenschaftliche Literatursuchen entwickelt wurden, erstellen gegenwärtig keine umfassenden Listen relevanter Literatur die mit den Ergebnissen systematischer Suchen vergleichbar sind. Sie könnten allerdings für die Vorbereitung systematischer Suchen (Scoping) oder als zusätzliche Suchmethoden von Nutzen sein.

Können LLMs Suchstrategien verbessern?

Manuale für die systematische Literatursuche [2], [40], [41] empfehlen, dass zumindest die primäre Datenbank-Suchstrategie durch eine zweite Person überprüft wird. Das soll die adäquate Erfassung der Fragestellung und Fehlerfreiheit der Suchstrategie gewährleisten.

Nur zwei Studien untersuchten die Verwendung von LLMs zu diesem Zweck: Hill et al. [42] verwendeten einen random error generator, um Tippfehler und Verknüpfungsfehler in sechs Suchstrategien einzufügen. Dann wurde überprüft, ob LLMs diese Fehler finden und eine verbes-

serte Suchstrategie vorschlagen konnten. Die Fehlererkennungsrate war dabei mit 75-93% teilweise vergleichbar mit der von menschlichen Reviewern (93%). Allerdings konnten LLMs nicht zuverlässig verbesserte Suchstrategien produzieren: Die Anzahl der durchführbaren Suchstrategien lag bei 0/6 (Gemini), 1/6 (Claude) und 2/6 (ChatGPT). Gitman et al. [43] entfernten von 16 Suchstrategien ein gesamtes Konzept (alle Suchbegriffe für „observational studies“ oder „drug harms“). ChatGPT konnte dieses Fehlen erkennen und eine Liste von relevant erscheinenden Suchbegriffen vorschlagen. Es wurden allerdings keine verbesserten Suchstrategien erstellt und mit dem Recall der Original-Suchen verglichen.
Basierend auf diesen Ergebnissen könnten LLMs für die Identifikation von groben Fehlern und das Vorschlagen von Suchbegriffen nützlich sein. Allerdings ist unklar, inwiefern die künstlich generierten Fehler mit „echten“ fehlerhaften Suchstrategien vergleichbar sind.

Fazit und Ausblick

Die Evidenzlage zur Verwendung von KI im systematischen Suchprozess weist große Lücken auf. Einzig die Erstellung von PubMed-Suchstrategien durch generische LLMs ist inzwischen gut untersucht. Sie zeigt sich bis jetzt aber wenig erfolgreich, wenn man sie mit dem Recall vergleicht, den durchschnittliche von Menschen erstellte Suchstrategien erreichen. Zu anderen Ansätzen wie der Verwendung von fine-tuned Modellen, der Suche über KI-gestützte Rechercheplattformen und der Überprüfung von Suchstrategien gibt es wenig Studien. Zu weiteren möglichen Anwendungen – beispielsweise der systematischen Suche nach grauer Literatur – wurden keine passenden Evaluationen identifiziert.
Abgesehen vom unbefriedigenden Recall, stellt die Variabilität der Ergebnisse ein ungelöstes Problem dar: Die wiederholte Ausführung gleichbleibender Prompts führte zu unterschiedlichen Ergebnissen. Wenn KI-Tools für das eigentliche Information Retrieval (nicht nur für die Erstellung einer Datenbank-Suchstrategie) verwendet werden, untergräbt das die fundamentale Anforderung von größtmöglicher Reproduzierbarkeit und Transparenz, die wir an systematische Suchen stellen.
Daher sollten KI-gestützte Methoden zum gegenwärtigen Zeitpunkt nur komplementär zu den etablierten Methoden der systematischen Recherche angewandt werden: KI-gestützte Rechercheplattformen können beispielsweise für explorative Scoping-Suchen nützlich sein, generische LLMs können eine zusätzliche Quelle für Freitext-Suchbegriffe darstellen.
Diese Einschätzung steht im Kontrast mit den Erwartungen, die von KI-Anbietern im Evidenzsynthesen-Bereich geschürt werden. So verspricht beispielsweise otto-SR: „Systematic reviews in hours, not months. otto-SR performs end-to-end evidence synthesis from thousands of citations with better-than-human performance.“ [44] Der zugehörige Preprint [44] zeigt aber, dass es sich dabei nur um die Schritte Abstract Screening, Fulltext Screening und Data Extraction handelt. Auch Elicit verspricht: „With AI and language models, Elicit can help you save up to 80% of the time it takes to run systematic reviews, without compromising on accuracy. The Systematic Reviews workflow guides you step by step through search, title & abstract screening, and full-text data extraction, also providing a research report to summarize the most relevant papers in your review at the end of the process.“ [45]. Elicits eigene Evaluation [46] bezieht sich allerdings ebenfalls nur auf Screening und Data Extraction.
Systematische Literaturauswahl und Datenextraktion sind auch die Schritte der Evidenzsynthese, bei denen LLM-Anwendungen unabhängig von den verwendeten Tools und Modellen vielversprechende Performance liefern [16]. Für den Prozess der systematischen Literaturrecherche ist das bis jetzt nicht der Fall. Aber neue Modelle oder Tools, beispielsweise die Veröffentlichung von nutzerfreundlichen fine-tuned Modellen für die Suchstrategie-Erstellung, könnten das ändern. Aus diesem Grund sind weiterhin gut gemachte unabhängige Evaluationen von KI-Anwendungen unerlässlich. Es ist aber auch wichtig, dass die Anwender*innen notwendige Kenntnisse haben, um den Nutzen der KI-Tools für ihre Zwecke einzuschätzen.
Initiativen wie die joint Artificial Intelligence Methods Group von Cochrane, der Campbell Collaboration, JBI und der Collaboration for Environmental Evidence (CEE) [44] und die Responsible AI in Evidence Synthesis (RAISE) [47] guidance können hierzu einen Beitrag leisten. RAISE richtet sich an Ersteller*innen von Evidenzsynthesen, aber auch KI-Tool Entwickler*innen und Methodenforscher*innen. Die guidance empfiehlt Ersteller*innen von Evidenzsynthesen unter anderem KI-Evaluationen kritisch zu lesen und nur Tools zu verwenden, die nachgewiesenermaßen geeignet für die gewünschte Aufgabe sind [48]. Zusätzlich werden Metriken beschrieben, die für solche Evaluationen relevant sind [10], sowie Fragen/Überlegungen, die die Einschätzung von KI-Tools leiten sollten [49]. Es ist zu hoffen, dass es dadurch auch für potenzielle Anwender*innen einfacher wird den potentiellen Nutzen von KI-Tools für die systematische Literatursuche oder anderen Evidenzsynthese Schritte einzuschätzen.

Anmerkung

Den Ausgangspunkt dieses Beitrags bilden ein Webinar [50] und ein Workshop [51], den die Autorin mit Dr. Maria-Inti Metzendorf (Public Health and Information Scientist, Cochrane Planetary Health Thematic Group) abgehalten hat.

ORCID der Autorin

Irma Klerings: 0000-0001-6644-9845

Interessenkonflikte

Die Autorin erklärt, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel hat.

Literatur

1. Gusenbauer M, Haddaway NR. What every researcher should know about searching - clarified concepts, search advice, and an agenda to improve finding in academia. Res Synth Methods. 2021 Mar;12(2):136-147. DOI: 10.1002/jrsm.1457
2. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Chapter 4: Searching for and selecting studies. In: Higgins J, Thomas J, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.5 (updated 2024 Sep). Cochrane; 2024. Available from: https://training.cochrane.org/handbook/current/chapter-04
3. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev. 2014 Jul;3:74. DOI: 10.1186/2046-4053-3-74
4. Lieberum JL, Toews M, Metzendorf MI, Heilmeyer F, Siemens W, Haverkamp C, Böhringer D, Meerpohl JJ, Eisele-Metzger A. Large language models for conducting systematic reviews: on the rise, but not yet ready for use – a scoping review. J Clin Epidemiol. 2025 May;181:111746. DOI: 10.1016/j.jclinepi.2025.111746
5. Canada's Drug Agency (CDA-AMC). Development of an Evaluation Instrument on Artificial Intelligence Search Tools for Evidence Synthesis. Canadian Journal of Health Technologies. 2024;4(10). DOI: 10.51731/cjht.2024.1004
6. Clark JM, Sanders S, Carter M, Honeyman D, Cleo G, Auld Y, et al. Improving the translation of search strategies using the Polyglot Search Translator: a randomized controlled trial. J Med Libr Assoc. 2020;108(2):195-207. DOI: 10.5195/jmla.2020.834
7. Kapp C, Fujita-Rohwerder N, Lilienthal J, Sieben W, Waffenschmidt S, Hausner E. The searchbuildR shiny app: A new implementation of the objective approach for search strategy development in systematic reviews. Cochrane Evid Synth Methods. 2024 Jun;2(6):e12078. DOI: 10.1002/cesm.12078
8. Warner L. DEFINING AI: A Lexicon for Librarians and Their Patrons. Computers in Libraries. 2025;45(1):16-8.
9. Meenn. Differences between LLM, Deep learning, Machine learning, and AI. Medium; 2024 Sep 30. Available from: https://medium.com/@meenn396/differences-between-llm-deep-learning-machine-learning-and-ai-3c7eb1c87ef8
10. Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall IJ, Hajji R, et al. Responsible AI in Evidence Synthesis (RAISE) 2: building and evaluating AI evidence synthesis tools. 2025 Jun 3. Available from: https://osf.io/fwaud/
11. Ayyamperumal SG, Ge L. Current state of LLM Risks and AI Guardrails. arXiv. 2024 Jun 14. DOI: 10.48550/arXiv.2406.12934
12. Siebert J. Halluzinationen von generativer KI und großen Sprachmodellen (LLMs). Fraunhofer-Institut für Experimentelles Software Engineering IESE; 2024. Available from: https://www.iese.fraunhofer.de/blog/halluzinationen-generative-ki-llm/
13. Jegham N, Abdelatti M, Koh CY, Elmoubarki L, Hendawi A. How hungry is AI? Benchmarking energy, water, and carbon footprint of LLM inference [Preprint]. arXiv. 2025. DOI: 10.48550/arXiv.2505.09598
14. Emberson L, Rahman R. The power required to train frontier AI models is doubling annually. Available from: https://epoch.ai/data-insights/power-usage-trend
15. Berthelot A, Caron E, Jay M, Lefèvre L. Estimating the environmental impact of Generative-AI services using an LCA-based methodology. Procedia CIRP. 2024;122:707-12. DOI: 10.1016/j.procir.2024.01.098
16. Adam G, Davies M, George J, Caputo EL, Htun JM, Coppola E, et al. Machine Learning Tools To (Semi-)Automate Evidence Synthesis: A Rapid Review and Evidence Map. Available from: https://effectivehealthcare.ahrq.gov/products/machine-learning-tools/white-paper
17. Clark J, Barton B, Albarqouni L, Byambasuren O, Jowsey T, Keogh J, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. Research Synthesis Methods. 2025;16(4):601-19. DOI: 10.1017/rsm.2025.16
18. Musleh A, Alryalat SA. Artificial Intelligence and Large Language Model Powered Literature Review Services. High Yield Medical Reviews. 2025;3(1). DOI: 10.59707/hymrPSEY7778
19. Shokraneh F. Living Evidence Map for Automation of Systematic Reviews (LEMASyR). Available from: https://nested-knowledge.com/nest/21035
20. De Cassai A, Dost B, Karapinar YE, Beldagli M, Yalin MSO, Turunc E, Turan EI, Sella N. Evaluating the utility of large language models in generating search strings for systematic reviews in anesthesiology: a comparative analysis of top-ranked journals. Reg Anesth Pain Med. 2025 Jan:rapm-2024-106231. DOI: 10.1136/rapm-2024-106231
21. Wang S, Scells H, Koopman B, Zuccon G. Reassessing Large Language Model Boolean Query Generation for Systematic Reviews [Preprint]. arXiv. 2025. DOI: 10.48550/arXiv.2505.07155
22. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2023 Jul 23-27; Taipei, Taiwan. Association for Computing Machinery; 2023. p. 1426-36. DOI: 10.1145/3539618.3591703
23. Adam GP, DeYoung J, Paul A, Saldanha IJ, Balk EM, Trikalinos TA, Wallace BC. Literature search sandbox: a large language model that generates search queries for systematic reviews. JAMIA Open. 2024 Oct;7(3):ooae098. DOI: 10.1093/jamiaopen/ooae098
24. Chen XS, Feng Y. Exploring the use of generative artificial intelligence in systematic searching: A comparative case study of a human librarian, ChatGPT-4 and ChatGPT-4 Turbo. IFLA Journal. 2024;51(1):03400352241263532. DOI: 10.1177/03400352241263532
25. Gosak L, Štiglic G, Pruinelli L, Vrbnjak D. PICOT questions and search strategies formulation: A novel approach using artificial intelligence automation. J Nurs Scholarsh. 2025 Jan;57(1):5-16. DOI: 10.1111/jnu.13036
26. Reed RB, Barnett DJ. Assessing the Quality of Biomedical Boolean Search Strings Generated by Prompted and Unprompted Models Using ChatGPT: A Pilot Study. Med Ref Serv Q. 2025;44(1):31-40. DOI: 10.1080/02763869.2024.2440848
27. Staudinger M, Kusa W, Piroi F, Lipani A, Hanbury A, editors. A Reproducibility and Generalizability Study of Large Language Models for Query Generation [Preprint]. arXiv. 2024. DOI: 10.48550/arXiv.2411.14914
28. Budau L, Ensan F. Fully Automated Scholarly Search for Biomedical Systematic Literature Reviews. IEEE Access. 2024;12:83764-73. DOI: 10.1109/access.2024.3405529
29. Bourgeois JP, Ellingson H. Ability of ChatGPT to Generate Systematic Review Search Strategies Compared to a Published Search Strategy. Med Ref Serv Q. 2025;44(3):279-291. DOI: 10.1080/02763869.2025.2537075

30. Boyle A, Huo B, Sylla P, Calabrese E, Kumar S, Slater BJ, Walsh DS, Vosburg RW. Large language model-generated clinical practice guideline for appendicitis. Surg Endosc. 2025 Jun;39(6):3539-3551. DOI: 10.1007/s00464-025-11723-3

31. Pourreza M, Ensan F. Towards semantic-driven boolean query formalization for biomedical systematic literature reviews. Int J Med Inform. 2023 Feb;170:104928. DOI: 10.1016/j.ijmedinf.2022.104928

32. Featherstone R, Walter M, MacDougall D, Morenz E, Bailey S, Butcher R, et al. A Comparative Analysis of Artificial Intelligence Search Tools for Evidence Synthesis [Preprint]. Authorea. 2025. DOI: 10.22541/au.174897559.99564896/v1

33. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, Raynier JL, Clowez G, Boileau P, Ruetsch-Chelli C. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. J Med Internet Res. 2024 May;26:e53164. DOI: 10.2196/53164

34. Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, Shim SR. The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. JMIR Med Inform. 2024 May;12:e51187. DOI: 10.2196/51187

35. Sanii RY, Kasto JK, Wines WB, Mahylis JM, Muh SJ. Utility of Artificial Intelligence in Orthopedic Surgery Literature Review: A Comparative Pilot Study. Orthopedics. 2024;47(3):e125-e130. DOI: 10.3928/01477447-20231220-02

36. Seth I, Lim B, Xie Y, Ross RJ, Cuomo R, Rozen WM. Artificial intelligence versus human researcher performance for systematic literature searches: a study focusing on the surgical management of base of thumb arthritis. Plastic and Aesthetic Research. 2025;12:1.

37. Bernard N, Sagawa Y Jr, Bier N, Lihoreau T, Pazart L, Tannou T. Using artificial intelligence for systematic review: the example of elicit. BMC Med Res Methodol. 2025 Mar;25(1):75. DOI: 10.1186/s12874-025-02528-y

38. Lau O, Golder S. Comparison of Elicit AI and Traditional Literature Searching in Evidence Syntheses Using Four Case Studies. Cochrane Evid Synth Methods. 2025 Nov;3(6):e70050. DOI: 10.1002/cesm.70050

39. Tosi D. Comparing Generative AI Literature Reviews Versus Human-Led Systematic Literature Reviews: A Case Study on Big Data Research. IEEE Access. 2025;13:56210-9. DOI: 10.1109/access.2025.3554504

40. The JBI Information Science Methodology Group, Ross-White A, Lieggi M, Palacio FGL, Solomons T, Swab M, et al. 2.4 Search Methodology for JBI Evidence Syntheses. In: JBI Manual for Evidence Synthesis. 2024. DOI: 10.46658/JBIMES-24-01

41. European network for Health Technology Assessment (EUnetHTA). Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness. Version 2.0. 2020.

42. Hill J, Brini S, Morrison K, Tran A, Towson G. Can artificial intelligence perform accurate peer review of literature search strategies? A proof of concept study. In: Global Evidence Summit; 2024 September 10-13; Prague, Czech Republic. Cochrane; 2024.

43. Gitman V, Maxwell C, Gamble JM. Enhancing search strategies for systematic reviews on drug Harms: An evaluation of the utility of ChatGPT in error detection and keyword generation. Comput Biol Med. 2025 Jul;193:110464. DOI: 10.1016/j.compbiomed.2025.110464

44. Cao C, Arora R, Cento P, Manta K, Farahani E, Cecere M, et al. Automation of Systematic Reviews with Large Language Models [Preprint]. medRxiv. 2025. DOI: 10.1101/2025.06.13.25329541

45. Elicit. Systematic Reviews in Elicit. [Accessed 2025 Oct 26]. Available from: https://support.elicit.com/en/articles/7927169

46. Fortier-Dubois É. How we evaluated Elicit Systematic Review. Elicit; 2025 Mar 18. Available from: https://blog.elicit.com/how-we-evaluated-elicit-systematic-review/

47. Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall IJ, Hajji R, et al. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations. Version 2. [updated 2025 Jun 3]. Available from: https://osf.io/fwaud/

48. Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall IJ, Hajji R, et al. Responsible AI in Evidence Synthesis (RAISE) 1: Recommendations for practice. 2025 Jun 3. Available from: https://osf.io/cqa82

49. Thomas J, Flemyng E, Noel-Storr A, Moy W, Marshall IJ, Hajji R, et al. Responsible AI in Evidence Synthesis (RAISE): 3: selecting and using AI evidence synthesis tools. version 2. [updated 2025 Jun 3]. Available from: https://osf.io/fwaud/files/5xjpk

50. Metzendorf MI, Klerings I. (How) can AI-based automation tools assist with systematic searching? [Webinar]. Cochrane; 2025. Available from: https://www.cochrane.org/events/how-can-ai-based-automation-tools-assist-systematic-searching

51. Klerings I, Metzendorf MI. (Wie) kann KI bei der systematischen Literatursuche helfen?. In: Die EbM der Zukunft – packen wir's an! 26. Jahrestagung des Netzwerks Evidenzbasierte Medizin. Freiburg, 26.-28.03.2025. Düsseldorf: German Medical Science GMS Publishing House; 2025. Doc25ebmWS-03-01. DOI: 10.3205/25ebm118

Korrespondenzadresse:
Irma Klerings
Department für Evidenzbasierte Medizin und Evaluation, Universität für Weiterbildung Krems,
Dr.-Karl-Dorrek-Straße 30, 3500 Krems, Österreich
irma.klerings@donau-uni.ac.at

Bitte zitieren als
Klerings I. Large Language Models in der systematischen Literaturrecherche – eine Evidenzübersicht. GMS Med Bibl Inf. 2025;25(2):Doc25.
DOI: 10.3205/mbi000638, URN: urn:nbn:de:0183-mbi0006389

Artikel online frei zugänglich unter
https://doi.org/10.3205/mbi000638

Veröffentlicht: *19.12.2025*