KI trifft auf Literaturrecherche – mit mäßigem Erfolg!

Christoph Holzke, Monika Hotze, Bernhard Mittermaier

Abstract

*Die Nutzung Künstlicher Intelligenz (KI) ist aus unserem Alltag kaum noch wegzudenken. So sind auch – nicht zuletzt – Bibliotheken als zentrale Informationsvermittelnde von dieser Informationstechnologie betroffen und zu einer Stellungnahme aufgefordert. In der, in diesem Beitrag, vorgestellten Studie wurde ein Teilgebiet der KI, die generative KI an einigen Web-Anwendungen zur Informationssuche, im Speziellen zur Literatursuche, untersucht. Beispielhaft wurden fünf KI-gestützte Web-Applikationen in Hinblick auf die Anwendbarkeit einer umfassenden Recherche wissenschaftlicher Literatur (literature review) herangezogen. Für einen Vergleich der Qualität der verschiedenen Angebote wurde ein standardisiertes Recherchekonzept (zunächst entwickelt für die Anwendung ChatGPT *) angewandt. Die Studie zeigt, dass die Ergebnisse einer KI-gestützten Recherche in allen untersuchten Systemen qualitative Fehler aufweisen und die Quantität der ermittelten Literatur nicht den Anforderungen eines literature reviews genügt. Somit wird die klare Notwendigkeit deutlicher Reifung, wenn nicht sogar grundlegender Änderungen des technischen Ansatzes von KI-gestützten Recherchewerkzeugen deutlich.*

The use of artificial intelligence (AI) has become an integral part of our everyday lives. Libraries, as central providers of information, are also affected by this information technology and are being called upon to take a stand. In the study presented in this article, a sub-area of AI, generative AI, was investigated using a number of web applications for information retrieval in particular for literature search.. Five AI-supported web applications were used as examples to assess their suitability for comprehensive search for scientific literature (literature review). A standardized research concept (initially developed for the ChatGPT application) was used to compare the quality of the various offerings. The study shows that the results of AI-supported retrieval in all systems examined contain qualitative errors and that the quantity of literature identified does not meet the requirements of a literature review. This highlights the clear need for significant maturation, if not fundamental changes, in the technical approach of AI-supported search tools.*

* ChatGPT ist ein von OpenAI entwickeltes Sprachmodell
ChatGPT is a language model developed by OpenAI.

Einleitung

Mit dem allgemeinen Hype zur diversen, raumgreifenden Anwendung von Künstlicher Intelligenz (KI) hält diese Technik auch in Bibliotheken Einzug.[1] [2] Dabei ist eine derzeit beliebte Anwendung von KI das sogenannte Dialogsystem Chatbot, ein Kommunikationswerkzeug zwischen Informationsanbietendem und -nutzendem.[3] Wie auch aus anderen Bereichen bekannt, so initiiert auch im Informationsbereich das große Interesse an der Nutzung von KI eine Explosion von Angeboten von Web-Applikationen, natürlich auch – und nicht zuletzt – mit ökonomischem Interesse der Anbietenden. Tatsächlich sprießen so auch im Bereich der KI-unterstützten Literatur- und Informationsrecherche Web-Anwendungen wie Pilze aus dem Boden. Eine Entwicklung der Anwendungen erfolgt dabei in der Regel solange, wie auch für eine gewisse Zeitspanne Fördergelder zur Verfügung stehen.[4] Kritisch anzumerken ist, dass Informationsspezialisten häufig nicht an der Softwareentwicklung beteiligt sind, auch wenn später die Produkte als Werkzeuge der Informations- und Literatursuche vermarktet werden.[5] Das bedeutet aber, dass das Wissen über Dinge wie Referenz-Suchprozesse, Such- und Überprüfungsmethoden und die dazugehörigen wissenschaftlichen Arbeitsabläufe und Anforderungen mit Blick auf die Integrität der Forschung in den Funktionalitäten der KI-gestützten Recherchewerkzeuge unterentwickelt bleiben.

Qualitätsentscheidend ist weiterhin, mit welchen Dokumenten (Informationsgrundlage) das zugrundeliegende Large Language Model (LLM) der Anwendung trainiert wurde. Die meisten KI-Anwendungen können nur auf Informationsquellen zugreifen, welche frei zugänglich im Internet zu finden sind. Volltext von wissenschaftlichen Closed-Access-Verlagen zählen hier nur sehr eingeschränkt oder fast gar nicht dazu. Vielmehr werden im Gro lediglich Metadaten wie Titel, Autor, Ver-

1 Gasparini, A., & Kautonen, H. (2022). Understanding Artificial Intelligence in Research Libraries – Extensive Literature Review. LIBER Quarterly: The Journal of the Association of European Research Libraries, 32(1). (2022) S. 1-36. <https://doi.org/10.53377/lq.10934> [29.04.2025]

2 Wildgaard, L., Vils, A., & Sandal Johnsen, S. (2023). Reflections on tests of AI-search tools in the academic search process. LIBER Quarterly: The Journal of the Association of European Research Libraries, 33(1), S. 1-34. <https://doi.org/10.53377/lq.13567> [29.04.2025]

3 Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. et al. Large language models in medicine. Nat Med 29, (2023) S. 1930-1940. <https://doi.org/10.1038/s41591-023-02448-8> [29.04.2025]

4 Khalil, H., Ameen, D., & Zarnegar, A. (2022). Tools to support the automation of systematic reviews: a scoping review. Journal of Clinical Epidemiology, 144, S. 22-42. <https://doi.org/10.1016/j.jclinepi.2021.12.005> [05.05.2025]

5 Wildgaard, L., Vils, A., & Sandal Johnsen, S. (2023). Reflections on tests of AI-search tools in the academic search process. LIBER Quarterly: The Journal of the Association of European Research Libraries, 33(1), S. 1-34. <https://doi.org/10.53377/lq.13567> [29.04.2025]

lag plus die Abstracts der Publikationen für das Training der LLMs herangezogen.

Künstliche Intelligenz, auch Artificial Intelligence (AI) genannt (oder auch „Algorithmische Intelligenz“[6]), ist ein Teilgebiet der Informatik, das sich mit der Automatisierung intelligenten Verhaltens, als Teil des Machine Learning befasst. Hieraus können zum Beispiel KI-gestützte Recherche- und Informationssysteme hervorgehen. Diese modernen KI-Verfahren entwickeln ihre Leistungsfähigkeit durch das Trainieren der Systeme mit geeigneten maschinenlesbaren Daten.[7]

Aus der Literatur (z.B. Bach, 2024[8]) ergibt sich, dass im Bibliotheksbereich vornehmlich die sog. generative KI mögliche Anwendungsfelder bereit hält. Als generative KI werden Modelle der Künstlichen Intelligenz bezeichnet, die darauf trainiert sind, neue Inhalte in Form von Text, Audio, Bildern oder Videos zu erzeugen. Hierfür werden Sprachmodelle verwandt. Sogenannte Large Language Models (LLMs), also große Sprachmodelle, sind eine spezielle Form einer KI-Anwendung, mit Grundlage umfangreicher Informationssammlungen. Dabei gilt grundsätzlich: Je umfangreicher die Daten in einem LLM sind („Coverage“) und umso strukturierter, desto umfassender und meist auch diverser kann diese KI arbeiten. Große Sprachmodelle sollten daher grundsätzlich eine bessere, genauere Antwortmöglichkeit auf die Fragen der Nutzenden in das System (also in der Web-Anwendung an den Chatbot) bedeuten. Zudem werden bei dieser Technik Abfragen in menschlicher („natürlicher“) Sprache in das System gegeben, danach in mathematische Algorithmen umgewandelt und die Ergebnisse der generativen KI wieder in, vom Menschen verständlicher Sprache ausgegeben. Hierbei ist wichtig zu wissen, dass die KI ihre Ergebnisse erzeugt, indem sie wahrscheinlichkeitsgetrieben arbeitet. Aus der zugrundeliegenden Informationssammlung errechnet die KI wie wahrscheinlich eine Antwort auf eine, von Nutzenden gestellte Frage ist und gibt diese Antwort als Ergebnis des Systems aus.

Bei der Anwendung der KI-Systeme für eine Literaturrecherche ist daher entscheidend, eine passgenaue Formulierung der Fragestellung (KI-bezogen das sog. Prompting) zu generieren. Die Kompetenz und Kreativität des Nutzenden liegt darin, den Dialog mit der KI über eine präzise formulierte Eingangsfrage zu starten, das ausgegebene Ergebnis zu überprüfen und über mehrere Folgefragen schrittweise das gewünschte Ergebnis zu erarbeiten. Erhält das System unzureichend genaue Anfragen, so kann die KI grundsätzlich auch weniger genau passende Antworten, also ein brauchbares Ergebnis erzeugen und ausgeben.[9]

Davon unabhängig ist aber ein, besonders in Bezug auf die Anwendung einer wissenschaftlichen Literaturrecherche schwerwiegendes Problem der KI-Technik bereits bekannt: Das sogenannte „Halluzinieren“, bei dem die KI falsche Informationen erzeugt (unabhängig von der Qualität der gestellten Frage) oder Referenzen angibt, die real nicht existieren.[10]

Ziel der Untersuchung war es, herauszufinden, inwieweit KI-gestützte Systeme für eine umfassende, qualitativ hochwertige und wissenschaftlich akzeptable Literaturrecherche (literature review) geeignet sind. Methodisch ermöglicht dabei eine standardisierte Abfolge von Prompts, die von der KI erzeugten Ergebnisse zu vergleichen und im Hinblick auf den zu erarbeitenden literature review zu bewerten. Das dabei in dieser Untersuchung angewandte Konzept ist auf eine umfassende Sichtung der relevanten Literatur zu einer Fragestellung ausgerichtet.

Methode der Analyse von KI-gestützten Recherchewerkzeugen

Für die Analyse verschiedener KI-Tools bezüglich einer umfassenden Literaturrecherche wurde ein standardisiertes Abfragekonzept erstellt. Grundlage für dieses angepasste Konzept war das Recherchekonzept des YouTube-Tutorial-Kanals „AI and Tech for Education“. [11] Dieser Kanal befasst sich seit längerem mit der Nutzung verschiedener moderner DV/IT-Technologien zur Unterstützung wissenschaftlichen Arbeitens. Das zunächst für ChatGPT[12] entwickelte Recherchekonzept wurde von uns auf die Anwendung in einer umfassenden wissenschaftlichen Literaturrecherche angepasst.

6 Definition von Thorsten Koch, Zuse Institut Berlin, TU Berlin 2021 siehe dazu: <https://www.zib.de/de/research/mai/aim> und <https://www.digis-berlin.de/tags/algorithmische-intelligenz/> [29.04.2025]

7 Seeliger et al. Zum erfolgversprechenden Einsatz von KI in Bibliotheken Teil 1. b.i.t.online 24 2/ (2021), S. 173-178. <https://www.b-i-t-online.de/heft/2021-02-fachbeitrag-seeliger.pdf> [06.05.2025]

8 Bach, Nicolas (2024): KI in Bibliotheken vor und nach dem Aufkommen von ChatGPT: Eine kritische Diskursanalyse. preprint, <https://zenodo.org/records/11153610> [29.04.2025]

9 vgl. Wildgaard, L., Vils, A., & Sandal Johnsen, S. (2023). Reflections on tests of AI-search tools in the academic search process. LIBER Quarterly: The Journal of the Association of European Research Libraries, 33(1), S. 1-34. <https://doi.org/10.53377/lq.13567> [29.04.2025]

10 Ziwei Ji et al. (2023) Survey of Hallucination in Natural Language Generation. ACM Comput. Surv. 55, 12, Article No 248, S. 1-38. <https://doi.org/10.1145/3571730> [29.04.2025]

11 The BEST ChatGPT Prompts for Research and Literature Review. AI and Tech for Education. <https://www.youtube.com/watch?v=wlkr0gMhtNA> [30.04.2025]

12 ChatGPT ist ein von OpenAI entwickeltes Sprachmodell

Um in einem systematischen Ansatz zeigen zu können, welche unterschiedlichen Ergebnisse die verschiedenen KI-Tools liefern (qualitativ und quantitativ), wurden bei jeder getesteten Anwendung die identischen Fragen in Folge zu dem ausgewählten Thema gestellt. Dabei beinhaltete das Recherchekonzept vorformulierte Fragen, in die dann das gewählte wissenschaftliche Thema eingefügt wurde.[13] Dieses standardisierte Vorgehen der Befragung ermöglichte, dass wir die Ergebnisse bewerten und vergleichen konnten und Schwachstellen der einzelnen KI-Tools eindeutig zuzuordnen waren. Hierbei wurde sowohl die Qualität als auch die Quantität der Recherche-Ergebnisse ermittelt und analysiert.

Unser standardisiertes Recherchekonzept umfasste 17 Fragen (Prompts) an die jeweilige KI-Anwendung. Diese waren teilweise aufeinander aufbauend gestaltet. Es wurde dabei die Frage verfolgt, wie gut die KI bei Verwendung von Prompt-Folgen die Informationen aus den Vorfragen bei nachfolgenden Fragen berücksichtigt. Die Prompts waren so ausgerichtet, dass das System letztendlich mögliche Forschungsfragen zum gewünschten Thema entwickelte, sowie final eine umfassende Literaturrecherche durchführte.

Die 17 Fragen haben sich dabei an folgenden fünf Entwicklungsschritten eines wissenschaftlichen Promptings[14] orientiert:

1. *Starten Sie jeweils einen neuen Chat für ein neues Thema.*
2. *Bauen Sie auf den jeweils zuvor bereitgestellten Informationen auf und werden Sie konkret.*
3. *Weisen Sie der KI eine Rolle zu (z.B.: „Antworte als ein wissenschaftlicher Mitarbeitender auf die Frage…").*
4. *Fügen Sie dem Chat weitere spezifische Informationen hinzu.*
5. *Weisen Sie die KI an, „alle oben genannten Punkte" zu verwenden, um eine ausführliche, umfassende Literaturübersicht zu erstellen (alternativ können Sie auch einen bestimmten Teil der „oben genannten Antworten" verwenden).*

Der Literaturrecherche lag das Thema „terpene emissions by plants" zugrunde. Dieses Topic umfasst ein umfangreiches Themenfeld, das verschiedenste Teilaspekte aus den Wissenschaftsbereichen Biologie (Botanik und Ökologie), Agrarwissenschaften, Pharmazie und zudem der Ökonomie beinhaltet.

Beispielhaft wurden vier verschiedene KI-gestützte Web-Anwendungen auf die Möglichkeiten einer umfassenden Recherche nach wissenschaftlicher Literatur untersucht. Die untersuchten Systeme waren ChatGPT (LLM: GPT-4o - Omni), Consensus (250 ResNet18-Modelle), PerplexityPro (GPT-4 - Omni, Sonnet, Opus, Sonar Large, LLama 3.1, pplx-7b, pplx-70b) und der lizenzpflichtige Research Assistent des Web of Science (basiert auf ChatGPT 3.5 und greift auf die Daten der Web of Science (WoS) Core Collection der Jahre 1900 bis in die Gegenwart zurück; Clarivate™). Zudem wurde die Helmholtz Inferenz Testinstanz Blablador (FZJ-HIFIS) befragt, welche als LLM Evaluation Server die Auswahl und Anwendung verschiedener LLMs ermöglicht. Für die Studie wurde folgendes LLM verwendet: Llama3 405 on WestAI with 4b quantization.

Untersucht wurden die Anwendungen auf folgende Aspekte hin: 1. Umfang der Quellenstärke, also der Informationssammlung auf der das LLM trainiert ist (coverage), 2. Behandelt die KI die Anfrage im Sinne der Fragestellung? und die Fähigkeit des Systems, Bezüge auf vorangegangen gestellte Fragen herzustellen (Verständigkeit), 3. die Ausführlichkeit und die Differenzierung der Ergebnisse sowie die Richtigkeit der vom System ausgegebenen Ergebnisse, 4. die Richtigkeit von Zitaten (Ist die Zuordnung von Zitat zu Referenz korrekt?, existiert die Referenz?), 5. die Zitatbelastbarkeit (Zitierung einer Primär- oder Sekundärliteratur?), sowie die technischen Fragen inwieweit 6. die Ergebnisse in andere Werkzeuge übertragen und weiterverwendet werden konnten (Übertragbarkeit in Textverarbeitungsprogramme) und 7. die Übertragbarkeit der Referenzen in ein Literaturverwaltungsprogramm (LVP).

Ergebnisse

Die Analyse erbrachte, dass grundsätzlich alle untersuchten Systeme zwar Inhalte der Ergebnisse zu gestellten Prompts größtenteils richtig wiedergaben, diese allerdings grundsätzlich sehr allgemein gehalten waren. Zudem zeigten sich sehr oft Probleme mit der Zuordnung von Referenzen zu, in der Ergebnisausgabe, zitierten Inhalten. Nicht selten war die Bindung von Zitat zu Referenz fehlerhaft. Eine Übertragung von Referenzen in ein LVP war nur für den lizenzpflichtigen Research Assistent des Web of Science (WoS) möglich. Auch die Weiterverarbeitung der vom System ausgegebenen Ergebnisse in ein Textverarbeitungsprogramm wie Word war nur unzureichend

13 Holzke, Christoph (2025) retrival questions for AI-systems sytematic validation for literature review, Datenpublikation. Jülich DATA. <https://doi.org/10.26165/JUELICH-DATA/QCFNHM> [06.05.2025]

14 The BEST ChatGPT Prompts for Research and Literature Review. AI and Tech for Education. <https://www.youtube.com/watch?v=wlkr0gMhtNA> [30.04.2025]

gegeben. Ausnahme war hier die Anwendung PerplexityPro. Tabelle 1 zeigt die Zusammenfassung der Untersuchungen.[15]

Das in dieser Untersuchung angewandte Recherchekonzept ist auf eine umfassende Erfassung der relevanten Literatur zu einer Fragestellung ausgerichtet. Diesem Anspruch konnte keines der getesteten Systeme ausreichend genügen. Die verwendete KI zeigte in allen untersuchten Systemen eine deutliche Tendenz der Verkürzung und der Verallgemeinerung der ausgegebenen Ergebnisse. Somit waren die erzielten Informationen aus wissenschaftlicher Sicht eher als erste Übersicht und nicht als tiefgreifende, umfassende Literaturrecherche einzustufen.

Name KI-Tool	Quellenstärke (coverage)	Verständigkeit der KI	Ausführlichkeit	Richtigkeit der Inhalte	Richtigkeit des Zitats	Zitatbelastbarkeit (Primär/Sekundärliteratur)	Übertragbarkeit in Word	Referenzen in LVP* übertragbar	Bemerkung
ChatGPT (GPT-4o (Omni))	+++	++	+	++	---	---	++	---	**1
PerplexityPro (GPT-4 Omni, Sonnet, Opus, Sonar Large, LLama 3.1, pplx-7b, pplx-70b)	+++	++	+	++	----	---	+++	---	**2
Consensus 1.0 (250 ResNet18-Modelle)	++	+	+	++	++	+++	++	--	**3
WoS Research assistant (herstellergebunden)	+++	+	+	++	+	++	-	++	**4
HAICU Blablador (Llama3 405 on WestAI with 4b quantization)	+++	++	--	-	---	---	++	---	**5
						*LVP=Literaturverwaltungsprogramm			
	Legende: ----... 0 ...++++ (unbrauchbar... 0 ... gut brauchbar)								
**1 Halluzinationen, fast kein Zitat richtig (2/3 falsch), keine Verlinkung zur Quelle									
**2 kein Zitat richtig, Verlinkung zur Quelle									
**3 Probleme mit Phrasen-Abfragen (" "), indirekte Verlinkung zur Quelle (semantic scholar)									
**4 viele Referenzen, welche nicht im Text zitiert werden; Referenzen zu Zitaten im Text können nicht exklusiv in LVP* übertragen werden (Auswahl aus Gesamtreferenz-Liste zum Prompt (meist hunderte!) notwendig), keine Verlinkung zur Quelle, keine aufeinander aufbauenden Fragen möglich (Bezug zur Fragen-History), keine Fragen-Konzepterstellung möglich									
**5 Testportal mit begrenztem Rechenvolumen (1 MB) pro Nutzendem; größtenteils keine ausreichenden Referenzangaben (nur Au et al., Jahr); fast alle Referenzen existieren nicht (Halluzinationen)									
								Stand: 11.09.2024	

Tab. 1: Qualität verschiedener KI-Tools bezüglich einer umfassenden Recherche nach wissenschaftlicher Literatur (literature review)

15 Holzke, Christoph, (2025) Investigation into the quality of AI-supported web applications based on a literature review – documentation of the research results obtained in the investigation for the individual web applications, Datenpublikation. Jülich DATA <https://doi.org/10.26165/JUELICH-DATA/A4KURE> [06.05.2025]

Den gleichen Effekt sieht man ebenfalls bei der Verwendung der genutzten Referenzen: Keine der KI-gestützten Anwendungen kam für dieses hier genutzte – aufwendige und ausführliche – Recherchekonzept zu „Terpen-Emissionen von Pflanzen" über 10-15 zitierte Referenzen hinaus (sicher eine systemisch vorgegebene Einschränkung der jeweiligen Anwendungen). Alleine für die verschiedenen Facetten (Botanik, Ökologie, Pharmazie und Ökonomie) sind allerdings deutlich mehr relevante Referenzen verfügbar (zum heutigen Zeitpunkt mindestens in der Menge eines Faktors 10 der hier ausgegebenen Ergebnisse, eher deutlich mehr) und gehören eindeutig zu einem umfassenden und tiefgehenden literature review dieses Themas. Auch dies also für eine wissenschaftliche Literaturrecherche in der Praxis eindeutig unzureichend.

Als besonders problematisch zeigte sich der Aspekt der Richtigkeit und Verlässlichkeit der Beziehung Zitat/Referenz der vom System ausgegebenen Ergebnisse. Hier traten bei fast allen KI-gestützten Systemen große Unsicherheiten auf. Dabei waren viele Ergebnisse falsch zitiert oder aber Referenzen frei erfunden (Halluzinieren der KI). Spitzenreiter mit leider 100% falscher Zitate war hier PerplexityPro. Auch ChatGPT, als etabliertes System, zeigte eine hohe Fehlerquote von 64%, ebenso der WoS Research Assistent mit immerhin noch 50% Fehlerquote (hier wurde die Hälfte der aufgeführten Referenzen gar nicht im Ergebnistext zitiert). Am besten schnitt für diese Fragestellung Consensus AI mit null Fehlern in der Zitierung/Referenzierung ab.

Zudem zeigte sich eine schlechte oder gar nicht mögliche Übertragung von Rechercheergebnissen der KI-Tools in andere Werkzeuge wie Office-Anwendungen oder LVPs für alle untersuchten Systeme (ausgenommen hier WoS Research Assistent, welcher die bestehenden Möglichkeiten des WoS nutzt).

Neben den möglichen Potenzialen von KI in Bibliotheken[16] wurde im Rahmen der hier beschriebenen Recherche zur Nutzung von generativer KI zur umfassenden Recherche wissenschaftlicher Literatur besonders deutlich, dass der Einsatz von entsprechenden Tools zurzeit noch nicht vollständig lösbare Herausforderungen mit sich bringt. Dabei decken sich unsere Ergebnisse mit den Ergebnissen anderer Studien zu diversen anderen KI-gestützten Literaturrecherche-Anwendungen.[17] [18] [19]

Diskussion

Beispielhaft wurden in der vorliegenden Studie verschiedene KI-Tools zur Fähigkeit einer umfassenden Recherche wissenschaftlicher Literatur untersucht. Anders als in anderen Studien[20] lag der Schwerpunkt der Studie in der Untersuchung der Richtigkeit und Nachvollziehbarkeit der KI-generierten Ergebnisse und weniger in der Praktikabilität der untersuchten Systeme. Trotzdem wurde hier auch kurz auf den Aspekt der Nutzbarkeit der untersuchten KI-Tools bezüglich des wissenschaftlichen Workflows (Möglichkeit der Ergebnis- und Referenzübertragung in andere Office-Werkzeuge) eingegangen.

Ein entscheidender Punkt, der als Vorteil der zusätzlichen Verwendung von KI in klassischen Ansätzen der Informationsversorgung hervorgehoben wird, ist die Zeitersparnis beim Retrieval. Der Zeitfaktor ist daher etwas, auf das Entwickler von KI-Recherchesystemen und Nutzende großen Wert legen.[21] Dabei wird davon ausgegangen, dass die erreichte Zeitersparnis der KI-Anwendung die Effizienz der Informationssuche verbessern, und somit Zeit gewonnen wird, um sich auf andere Aktivitäten im Forschungszyklus zu konzentrieren.[22] So also das Versprechen. Was aber, wenn un-

16 Bach, Nicolas (2024): KI in Bibliotheken vor und nach dem Aufkommen von ChatGPT: Eine kritische Diskursanalyse. preprint, <https://zenodo.org/records/11153610> [29.04.2025]

17 Yan, Lixiang et al. (2023) Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55, S. 90-112. <https://doi.org/10.1111/bjet.13370> [06.05.2025]

18 Marshall, I. J., & Wallace, B. C. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews, 8(1), 163. S. 1-10. <https://doi.org/10.1186/s13643-019-1074-9> [06.05.2025]

19 Sallam M. (2023) ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel), 11(6):887. S. 1-20. <https://doi.org/10.3390/healthcare11060887> [06.05.2025]

20 Zhang et al. (2022) Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. Systematic Reviews 11:11. S. 1-7. <https://doi.org/10.1186/s13643-021-01881-5> [06.05.2025]; Beller et al. (2018) Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic Reviews 7:77. S. 1-7. <https://doi.org/10.1186/s13643-018-0740-7> [06.05.2025]; Marshall, I. J., & Wallace, B. C. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Systematic Reviews, 8(1), 163. S. 1-10. <https://doi.org/10.1186/s13643-019-1074-9> [06.05.2025]; Tamara Heck et al. (2025) Quality Criteria for AI-based Research Assistants (AIRAs). Dataset of a study submitted as poster presentation, to be presented at ISI 2025, 18. Internationales Symposium für Informationswissenschaft, Chemnitz Germany, 18.–20. March 2025, <https://isi2025.informationswissenschaft.org/> [06.05.2025], <https://doi.org/10.17605/OSF.IO/KWR6N> [06.05.2025], Datenpublikation unter <https://osf.io/kwr6n/files/osfstorage> [06.05.2025].

21 Zhang et al. (2022) Automation of literature screening using machine learning in medical evidence synthesis: a diagnostic test accuracy systematic review protocol. Systematic Reviews 11:11. S. 1-7. <https://doi.org/10.1186/s13643-021-01881-5> [06.05.2025]

22 Beller et al. (2018) Making progress with the automation of systematic reviews: principles of the International Collaboration for the Automation of Systematic Reviews (ICASR). Systematic Reviews 7:77. S. 1-7. <https://doi.org/10.1186/s13643-018-0740-7> [06.05.2025]; Marshall, I. J., & Wallace, B. C. (2019) Toward systematic review automation: a practical guide to using machine learning tools in research synthesis.

ter dem Zeitgewinn die Qualität des Informationsgewinns leidet oder gar inakzeptabel wird? Hier kommt der Aspekt zum Tragen, dass heutige KI-gestützte Webanwendungen der Informationssuche weitestgehend nicht auf die Volltexte wissenschaftlicher Closed-Access-Publikationen zugreifen können, sondern als Grundlage der Suche lediglich Titel und Abstracts dienen. Es zeigt sich in der Anwendung, dass KI-gestützte Recherchesysteme auch aufgrund der sehr stark eingeschränkten Zugriffsmöglichkeit auf Volltexte qualitativ und quantitativ in den ausgegebenen Ergebnissen zunächst in allgemeineren Aussagen verharren. Selbst ein sehr gutes, also exaktes Prompting stellt also nicht tiefergehendere Ergebnisse in Aussicht. Wir haben in dieser Studie gezeigt, dass diesbezüglich noch sehr großer Entwicklungsbedarf besteht. Dieser Entwicklungsbedarf besteht auch, denkt man an die Probleme wie Halluzinieren der KI oder fehlerhafte Zitierungen, welche regelmäßig bei den KI-Tools auftreten.

Zu den weiteren Hindernissen einer adäquaten Nutzung gehört die immer wieder sichtbar werdende Unfähigkeit der KI-Tools, die Nuancen des menschlichen Urteils und der Expertenmeinung bei der Bewertung der Relevanz nachzubilden.[23] Auch Wildgaard et al. (2023) finden in ihren Untersuchungen, dass die Ergebnisse von Expertenanalysen von KI-generiertem Output und zudem der quantitativen Bewertung dieser Inhalte darauf hin deuten, dass die KI-Suchwerkzeuge Arbeiten von geringerer wissenschaftlicher Qualität liefern.[24] Auch bei unseren Untersuchungen wurde festgestellt, dass erzielte Ergebnisse der KI-Tools tendenziell oberflächlicher Natur sind (maximal in der Qualität eines Abstracts) und wichtige Ergebnisse und Aussagen in Publikationen von der KI nicht erkannt werden.

Ein Hinweis, woher die Probleme der KI-Systeme u.a. kommen könnten, hat die Entwicklung des chinesischen KI-Models DeepSeek gegeben. Hier erzielt die KI deutlich bessere Ergebnisse, indem die Datenbasis des Systems geclustert ist und für eine Anfrage nur in, für diese Frage relevanten, Clustern gearbeitet wird. Die Ergebnisse lassen vermuten, dass es die Funktionalität der KI überfordert auf eine zu große Datenbasis auf einmal zugreifen zu müssen. Genauere, bessere Ergebnisse werden offenbar in passenden Daten-Teilmengen erzeugt. Auch mathematisch ist es nachvollziehbar, dass die Nutzung weniger Daten die Diversität der möglichen Antwortmöglichkeiten reduziert.

Wie bereits zuvor für zwei andere KI-Tools gezeigt[25], genügt aber die Architektur der weitaus meisten KI-gestützten Systeme am Markt leider noch nicht den Ansprüchen einer hochwertigen und umfassenden wissenschaftlichen Literaturrecherche. Als Folge ist der Informationssuchende bei der Verwendung einer KI-gestützten Literaturrecherche für umfassende Ergebnisse, wie in der vorliegenden Studie durchgeführt, überdurchschnittlich stark gefordert, die Qualität der Ergebnisse zu prüfen und zitierte Referenzen exakt und sehr kritisch nachzuprüfen (was bei klassischen Retrieval-Werkzeugen wie Scopus, WoS, etc. mit meist besserer Qualität und mit einer deutlich besseren Verlässlichkeit vonstattengeht).

Aus den hier gezeigten Ergebnissen und ebenso im Vergleich mit vorangegangenen Untersuchungen (vgl. z.B. Wildgaard 2023) ist zu fordern, dass KI-Anwendungen wie generative KI-Tools noch weiter reifen müssen, bevor sie verlässlich für eine gesicherte Recherche wissenschaftlicher Literatur herangezogen werden können.

Aus den Ergebnissen der vorliegenden Analyse lassen sich somit folgende Fragen ableiten:

Systematic Reviews, 8(1), 163. S. 1-10. <https://doi.org/10.1186/s13643-019-1074-9> [06.05.2025]

23 Arno, A., Thomas, J., Wallace, B., Marshall, I. J., McKenzie, J. E., & Elliott, J. H. (2022). Accuracy and efficiency of machine learning-assisted risk-of-bias assessments in "real-world" systematic reviews a noninferiority randomized controlled trial. Annals of Internal Medicine, 175(7), S. 1001-1009. <https://doi.org/10.7326/M22-0092> [07.05.2025]

24 und 25 Wildgaard, L., Vils, A., & Sandal Johnsen, S. (2023). Reflections on tests of AI-search tools in the academic search process. LIBER Quarterly: The Journal of the Association of European Research Libraries, 33(1), S. 1-34. <https://doi.org/10.53377/lq.13567> [29.04.2025]

1. *Wie müssen sich Werkzeuge der generativen KI für wissenschaftliche Literaturrecherchen entwickeln, um verlässlich und brauchbar zu sein?*
2. *Welche zusätzlichen Kriterien und Parameter müssen für eine gut funktionierende KI-gestützte Literaturrecherche in Zukunft erfüllt sein?*

Hier sind zu nennen:

- Halluzinierungseffekte sind inakzeptabel; diese müssen ausgeschlossen werden.
- Ungebrochene und richtige Zuordnung von Zitatstellen zu Metadaten des Zitats (richtig zugeordneter Autor, Titel, Journal, Verlag etc.) sind unerlässlich.
- Weiterentwicklung der LLMs in Bezug auf die Fähigkeit verlässliche Ergebnisse für eine umfassende Literaturrecherche zu erzielen.
- Ausführlichkeit, Tiefe und Qualität der Ergebnisse müssen für einen wissenschaftlichen Anspruch an die Recherche deutlich erhöht werden. Hierfür scheint der Rückgriff auf Volltexte von Publikationen unerlässlich.
- Übertragbarkeit (auch Dokumentation) der Ergebnisse in den wissenschaftlichen Workflows (Übertragung in Literaturverwaltungsprogramme, Office-Anwendungen, etc.) ist zu fordern.

Erst wenn diese Ansprüche erfüllt werden, können auch Bibliotheken den Gebrauch von KI-gestützten Recherchewerkzeugen ihren Nutzenden empfehlen und diese schulen. Eine einfache Lizenzierung eines angebotenen KI-Werkzeugs ist als möglicher Service der Bibliothek nicht ausreichend.

Hierzu sagt Frank Seeliger[26]:

„Bibliotheken ist spartenübergreifend gemein, für eine gewisse Form und Qualität von veröffentlichten Informationen als Wissensspeicher zu stehen und für den Zugang zu ihnen zu sorgen. Beides – Information und den Bibliothekskunden – zusammenzubringen mit dem Anspruch „wie es gesucht wird, so gefunden", gibt Bibliotheken ihre Aufgabe unabhängig jeglicher „Vermittlungs"-Technologie vor. Die besonderen Anforderungen entstehen aus Randbedingungen wie der Diversität an Nutzerinnen und Nutzern. Es gibt nicht eine Nutzerin oder den Nutzer einer Spezial-, Schul-, Stadt-, Hochschul- oder Forschungsbibliothek. Dementsprechend vielfältig müssen Strategien der medialen Verknüpfung mit ihren Interessenten ausfallen. Lösungen von der Stange (ready-made oder off-the-shelf) werden diesen funktionalen Anforderungen nicht gerecht. Der Markenkern von Informationseinrichtungen liegt neben dem Besitz von Medien in der Herstellung des Zugangs über geeignete Nachweisinstrumente und Recherchetools zum gewünschten Werk. Kann hierbei KI Abhilfe schaffen oder sogar Neues?"

Die Ergebnisse der vorliegenden Studie lassen zum jetzigen Zeitpunkt massive Zweifel an der Qualität und Nutzbarkeit einer KI-gestützten Literaturrecherche zu. Rechercheergebnisse werden mit der KI schneller, aber nicht effizienter und schon gar nicht genauer erzielt. Für ein belastbares Ergebnis ist eine umfangreiche, zeitlich aufwändige Nacharbeit bzgl. Prüfung und Einschätzung der Qualität der KI-generierten Ergebnisse nach der Literaturrecherche unabdingbar. Klassische Recherchemethoden weisen diesen Bias – zumindest auf technischer Ebene – nicht auf, zudem die hierbei verwendeten Literaturquellen meist umfangreich validiert sind. Für eine ausführliche Literaturrecherche wissenschaftlicher Literatur sollten daher zurzeit weiterhin etablierte Literaturdatenbanken wie OpenAlex (OurResearch), Web of Science (Clarivate), Scopus (Elsevier) oder auch Google Scholar (Google) etc. verwendet werden, da hier fokussierte, mehr-oder-minder qualitätsgeprüfte wissenschaftliche Informationen angeboten werden.

Dr. Christoph Holzke
arbeitet in der Zentralbibliothek des Forschungszentrums Jülich als Fachinformationsmanager. Arbeitsschwerpunkte sind fachliche Begleitung von Systementwicklungen, Fachreferat, wissenschaftliche Dienste.
c.holzke@fz-juelich.de

Monika Hotze
leitet das Team Informationsvermittlung in der Zentralbibliothek des Forschungszentrums Jülich. Arbeitsschwerpunkte sind Entwicklung und Pflege wissenschaftlicher Dienste, Fachreferat.
m.hotze@fz-juelich.de

Dr. Bernhard Mittermaier
leitet die Zentralbibliothek des Forschungszentrums Jülich. Arbeitsschwerpunkte sind Open Access einschließlich des Open-Access-Monitor, Lizenzverhandlungen u.a. im Projekt DEAL und Bibliometrie.
b.mittermaier@fz-juelich.de

26 Seeliger et al. Zum erfolgversprechenden Einsatz von KI in Bibliotheken Teil 2. b.i.t.online 24 3 (2021) S. 290-299. <https://www.b-i-t-online.de/heft/2021-03/fachbeitrag-seeliger.pdf> [06.05.2025]