Recherchieren mit ChatGPT?

Künstliche Intelligenz in der Literaturrecherche

Seit der Veröffentlichung von ChatGPT durch die Firma OpenAI im November 2022 sind Künstliche Intelligenz (KI), Große Sprachmodelle (Large Language Models, LLM) und Chatbots in aller Munde. Ihr Einsatz im Wissenschaftsbetrieb wird breit rezipiert, so auch in den Bibliotheken, wo der Umgang mit Sprachmodellen Auswirkungen auf die Vermittlung von Informationskompetenz im Allgemeinen und auf Recherchefähigkeiten im Speziellen hat und zukünftig haben wird. An die Theken der Bibliotheken, seien es öffentliche oder wissenschaftliche, sind in jüngster Vergangenheit immer häufiger Listen mit unauffindbaren Literaturtiteln getragen worden – Listen, die mit ChatGPT erstellt worden sind. Die Frage, inwieweit sich ein Werkzeug wie ChatGPT überhaupt zur Literaturrecherche nutzen lässt, stieß auf der 112. BiblioCon 2024 in Hamburg auf außerordentlich großes Interesse.¹

Im Rahmen eines Hands-on-Lab waren die Teilnehmenden dazu eingeladen, Eingaben auszuprobieren, um die Antworten und deren Qualität vergleichen zu können. Zum Einsatz kamen ChatGPT 3.5 von OpenAI, der Copilot von Microsofts Suchmaschine Bing, welcher auf der 2023 veröffentlichten Version GPT-4 (nicht zu verwechseln mit der im Mai 2024 erschienenen Version GPT-4o) von OpenAI basiert, sowie Gemini, das Konkurrenzprodukt von Google.² Alle drei Sprachmodelle kamen im Hands-on-Lab in ausreichender Anzahl zum Einsatz, sodass Ergebnisse und Ausgaben sowohl eines Sprachmodells als auch über verschiedene Modelle hinweg verglichen werden konnten. Ziel des Hands-on-Labs war die Sensibilisierung hinsichtlich der Grenzen und Fallstricke, die der Umgang mit Großen Sprachmodellen bereithalten kann. Aus diesem Grund wurde bewusst die ältere GPT-Version 3.5 gewählt, da hier die Fehlerquote um einiges höher ist als bei den neueren Sprachmodellen und sich die Unzulänglichkeiten besser demonstrieren ließen als mit aktuelleren Versionen wie GPT-4o.³ Die höhere Zahl an Fehlern macht es leichter, das notwendige Bewusstsein für die Funktionsweisen von LLM zu schaffen. Dabei darf nicht außer Acht gelassen werden, dass die Entwicklung rasant voranschreitet. Die Unterschiede zwischen der Veröffentlichung von ChatGPT im Winter 2022 und GPT-4o heute sind enorm. Fehler werden weniger und die Modelle (aktuell) immer besser.

Das bedeutet jedoch nicht, dass neuere Versionen perfekt wären, wie die erste Aufgabe des Hands-on-Labs zeigte. Es wurde eine leicht recherchierbare Information gesucht, nämlich zur Anzahl der Standorte der Deutschen Nationalbibliothek (DNB) – also zwei, Leipzig und Frankfurt. Das Auditorium erhielt über die drei angewandten LLM hinweg, auch in den aktuellen Versionen, ein wahres Füllhorn an Antworten: Zu Leipzig und Frankfurt gesellten sich in einigen Fällen Berlin, in einem Fall auch Hamburg und in einem anderen wurde ausgegeben, dass die DNB Standorte in Helsinki und anderen europäischen Städten besäße. In einigen, wenigen Ausgaben wurde das Kürzel der DNB

1 Dieser Beitrag berichtet vom Hands-on-Lab „Recherchieren mit ChatGPT? Künstliche Intelligenz in der Literaturrecherche“ am 05.06.2023 auf der 112. BiblioCon in Hamburg. Für den Foliensatz vgl. Quasendorf, Friedrich: Recherchieren mit ChatGPT? Künstliche Intelligenz in der Literaturrecherche, BiblioCon 2024 [Präsentationsfolien], <https://nbn-resolving.org/urn:nbn:de:0290-opus4-187689>.

2 Siehe <https://chat.openai.com/>; <https://copilot.microsoft.com/>; <https://gemini.google.com/>; jeweils Stand: 29.07.2024

3 Hinzu kommt, dass im Juni 2024 in der kostenfreien Fassung von ChatGPT die Version GPT-4o auf 15 Eingaben beschränkt war, ChatGPT 3.5 hingegen fast unbegrenzt genutzt werden konnte.

nicht korrekt aufgelöst und statt der Deutschen Nationalbibliothek einer niederländischen Bank zugeordnet. Während die falsche Anzahl von Standorten vor allem bei ChatGPT 3.5 auftrat, traten die letztgenannten falschen Ausgaben in Bings Copiloten auf, trotz des aktuelleren Sprachmodells und der damit einhergehenden erwartbaren höheren Zuverlässigkeit. Dass solche Falschinformationen vereinzelt auch von den zuverlässigeren LLM ausgegeben werden würden, war zu erwarten, dass es allerdings in solcher Deutlichkeit geschehen würde und die Antworten derart weit auseinandergingen, überraschte.

Während eine Deutsche Nationalbibliothek mit Standorten in ganz Europa recht problemlos als Falschinformation identifiziert werden kann, ist der Wahrheitsgehalt mancher Ausgaben ohne zusätzliche Recherche häufig gar nicht zu erkennen. Nach der Adresse des vermeintlichen DNB-Standortes in Berlin gefragt, gab ChatGPT 3.5 nicht aus, dass es einen solchen Standort nicht gäbe – wie das etwa Gemini und der Copilot mit weitaus größerer Zuverlässigkeit taten –, sondern: „Der Standort der Deutschen Nationalbibliothek in Berlin befindet sich in der Nähe des Potsdamer Platzes. Genauer gesagt befindet er sich in der Potsdamer Straße 33.“⁴⁴ – dabei handelt es sich um die Anschrift der Staatsbibliothek zu Berlin.

Die schwankende Qualität und Zuverlässigkeit, die LLM in der Informationsabfrage aktuell bieten, zeigte sich auch in der Frage nach der Leitung verschiedener Bibliotheken. Aus dem Auditorium wurde gemeldet, dass das genutzte Sprachmodell zwar den aktuellen Direktor der Universitätsbibliothek Wien korrekt ausgab, statt der langjährigen Vorgängerin allerdings zwei völlig unbekannte Namen nannte. Im Gegensatz zum Copiloten und zu Gemini, wo die Ausgaben in höherem Maße korrekt waren, erwies sich ChatGPT 3.5 in Fragen der Bibliotheksleitung als besonders unzuverlässig. Zwar konnte die Leitung der DNB im Jahre 2020 korrekt genannt werden, wer im selben Jahr jedoch die Universitäts- und Landesbibliothek (ULB) Sachsen-Anhalt leitete, war nicht zu ermitteln: „Im Jahr 2020 wurde die Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB Sachsen-Anhalt) von Herrn Dr. Ulrich Bernius geleitet.“⁴⁵ – jedoch mit der identischen Eingabe unmittelbar an die hier zitierte Ausgabe anschließend: „Im Jahr 2020 wurde die Universitäts- und Landesbibliothek Sachsen-Anhalt (ULB Sachsen-Anhalt) von Herrn Dr. Thomas Rahn geleitet.“⁴⁶ Keiner der beiden genannten Namen stimmt.

Die Fokussierung auf die Falschinformationen, die in den oben genannten Beispielen auftraten, diene als Präludium zum Kern des Hands-on-Labs. Wenn schon die Abfrage von Einzelinformationen unzuverlässig ist, wie sieht es dann mit der Abfrage von Literaturangaben aus, die aus einer Kombination von Informationen (Titel, Erscheinungsjahr, Verlag usw.) bestehen?

Gefragt wurde nach drei Literaturtiteln zur Geschichte des Bibliothekswesens in Deutschland. Angegeben werden sollten neben dem Titel des Buches die Verfasserangabe, das Erscheinungsjahr, der Verlag und die ISBN. Der erste Titel, den ChatGPT 3.5 ausgab, war ein Buch mit dem Titel „Geschichte der deutschen Bibliothek“ von Wolfgang Schmitz, das 1998 bei Harrassowitz mit der

4 Prompt: Wo befindet sich der Standort der DNB in Berlin?, ChatGPT 3.5, 05.06.2024, <https://chat.openai.com>.

5 Prompt: Wer leitete 2020 die ULB Sachsen-Anhalt?, ChatGPT 3.5, 05.06.2024, <https://chat.openai.com>.

6 Prompt: Wer leitete 2020 die ULB Sachsen-Anhalt?, ChatGPT 3.5, 05.06.2024, <https://chat.openai.com>.

ISBN 978-3447039231 erschienen sein soll.⁷ Auf dem ersten Blick sieht die Angabe vertrauenswürdig aus: Der Titel passt zum gesuchten Themenfeld, Harrassowitz ist ein bekannter Verlag und die ISBN, mit 978 beginnend, hat die korrekte Anzahl an Ziffern. Jedoch: Weder der Katalog der DNB noch der Karlsruher Virtuelle Katalog (KVK) kennen ein Buch mit diesem Titel von einem Autor dieses Namens. Mehr noch, die ISBN selbst ist ungültig und führt ins Nichts. Zum gleichen Ergebnis kamen Recherchen zu den anderen beiden Titeln, die ChatGPT 3.5 ausgegeben hat.

Je aktueller das LLM, desto besser die Ergebnisse. Von den drei angewandten Sprachmodellen war der Copilot dank seiner Grundlage von GPT-4 am treffsichersten. Hier meldete das Auditorium am häufigsten, dass die genannten Titel auffindbar und korrekt angegeben waren – obgleich der Copilot häufig nicht alle geforderten Angaben lieferte und sich auf Verfasserangabe und Titel beschränkte. In Anbetracht der Tatsache, dass viele der Titel existierten und recherchierbar waren, dürfte das aber verschmerzbar sein. Als besonders hilfreich wurde gesehen, dass der Copilot zu seinem Ergebnis auch die Quellen angab und direkt auf die Verlagsseite oder den Buchhandel verlinkte.

Die Ergebnisse sind jedoch alles andere als perfekt, denn auch der Copilot gab Literatur an, die schlichtweg nicht existiert. Er hatte damit aber immer noch eine höhere Präzision als Gemini, das sich als unzuverlässig herausstellte. Bei ChatGPT 3.5 gab es allenfalls Zufallstreffer, hier muss bei Literaturangaben generell mit einer Fehlinformation gerechnet werden. Mehr noch, anders als bei den anderen beiden Sprachmodellen gibt ChatGPT 3.5 keine Quellen oder weiterführenden Hinweise zu den ausgegebenen Titeln an.

Genau diese Angabe von Quellen und Verweisen kann verhängnisvoll sein, da diese falsche Sicherheit suggerieren: Wenn die Angabe eine Fußnote hat, wird sie schon stimmen. Nutzerinnen und Nutzer werden durch die bloße Anwesenheit von Verweisen dazu verleitet, die Ausgabe des LLM ungeprüft zu übernehmen.

Der Vergleich der tatsächlich existierenden Titel machte ein weiteres Problem deutlich. Abgesehen vom Copiloten, der bei mehreren Personen auf die gleiche Literatur verwies, waren die Angaben breit gestreut. In manchen Fällen war keiner der drei genannten Literaturtitel jünger als zwanzig Jahre. Kriterien der Relevanz und Aktualität spielten bei der Generierung der Ausgaben keine Rolle. Die Auswahl erschien daher völlig willkürlich.

Das große Problem, das beim Einsatz von Sprachmodellen in der Literaturrecherche erkannt werden konnte, ist die schwankende Qualität der Ausgaben. Wie die Beispiele gezeigt haben, waren die Ausgaben nicht allesamt unbrauchbar, aber unzuverlässig. Zu den durchaus richtigen Angaben gesellten sich über alle genutzten LLM hinweg immer wieder Fehlinformationen, die ohne gründliche Kenntnisse der Materie oder zusätzliche Recherche als solche nicht zu erkennen waren. Für die Literaturrecherche sind solche Fehler selbstredend fatal: Wer Literatur sucht, will sich darauf verlassen können, existierende Ergebnisse und keine ausgedachten Titel zu erhalten, die lediglich

7 Prompt: Nenne mir drei Titel zur Geschichte des Bibliothekswesens in Deutschland. Nenne Titel, Autor, Erscheinungsjahr, Verlag und ISBN., ChatGPT 3.5, 05.06.2024, <https://chat.openai.com>.

plausibel klingen.⁸ Falschinformationen werden nicht erwartet und sind nicht ersichtlich, sodass eine weiterführende Überprüfung vielen gar nicht erst in den Sinn kommt.

Um die Einsatzmöglichkeiten von LLM in der Literaturrecherche identifizieren zu können, war es nötig, einen Blick auf die elementaren Funktionsweisen zu werfen, denn die Erwartungshaltung, was Sprachmodelle leisten sollen, weicht von ihrer tatsächlichen Leistungsfähigkeit und ihren Stärken ab. Es gilt: „Sprachmodelle sind keine Wissensmodelle“⁹. Alle Sprachmodelle wurden auf gewaltigen Textmengen trainiert und sind in der Lage, verständliche Texte auszugeben. Diese Texte werden Wort für Wort generiert. Das nächste Wort wird über eine statistische Wahrscheinlichkeit ermittelt, indem das am besten passende Wort ausgewählt wird – basierend auf den Trainingsdaten. Die einzelnen Wörter besitzen für Sprachmodelle jedoch keinen Inhalt. Ihre Bedeutung ergibt sich nur aus dem Zusammenhang, in dem sie im Trainingskorpus vorhanden sind. Auf dieser Grundlage kann eine Faktenprüfung nicht erfolgen, sodass offensichtliche und weniger offensichtliche Fehlinformationen ausgegeben werden.

Wahrscheinlichkeiten sind essentiell für die Textgenerierung und erklären die oben genannten Beispiele: Es gibt keinen Standort der DNB in Berlin, aber da die DNB eine Einrichtung von nationaler Bedeutung ist, Einrichtungen von nationalem Rang am wahrscheinlichsten in der Hauptstadt Berlin zu finden sind und mit der Staatsbibliothek zu Berlin zudem eine bedeutende deutsche Bibliothek in Berlin existiert, ist die Wahrscheinlichkeit hoch, dass sich ein Standort der DNB an der Potsdamer Straße befinden müsste. Ebenso bei den Bibliotheksleitungen: Ist der korrekte Name im Trainingskorpus nicht enthalten, wird ein Name ausgegeben, der am wahrscheinlichsten passt. Das zeigt auch das Beispiel der beiden vermeintlichen Direktoren der ULB Sachsen-Anhalt: Sie wurden beide mit einem Dokortitel versehen, weil es wahrscheinlich ist, dass jemand, der eine wissenschaftliche Bibliothek leitet, auch promoviert ist.

Dieses Phänomen, das auch als Halluzinieren bezeichnet wird, tritt entsprechend auch bei der Literaturrecherche auf: Die Titel werden generiert. Es ist wahrscheinlich, dass ein Buch zur Bibliotheksgeschichte einen generischen Titel wie „Geschichte der deutschen Bibliothek“ trägt und es ist wahrscheinlich, dass die ISBN eines solchen Buches so aussehen würde, wie sie angegeben worden ist. GPT-4, wie es der Copilot nutzt, hat den Vorteil, dass vor der Textgenerierung eine Internetsuche durchgeführt wird, für die die Eingabe in Suchbegriffe umgewandelt wird, was sich im laufenden Betrieb beobachten lässt. Die Ausgaben sind daher wesentlich präziser als bei ChatGPT 3.5, doch auch diese Rechercheergebnisse werden nur als Text behandelt, sodass aktuellere Versionen wie GPT-4 oder GPT-4o zwar ungleich höhere Trefferquoten haben als ältere Sprachmodelle, aber – wie die Meldungen aus dem Auditorium gezeigt haben – keineswegs perfekt sind.

Eine direkte Literaturrecherche über Sprachmodelle ist also nicht möglich. Doch über eine indirekte Literaturrecherche lassen sich die Stärken der LLM nutzen: Man sucht nicht nach der Literatur zu

⁸ Man stelle sich vor, ein Discovery-System würde Literatur herbeiphantasieren.

⁹ Glatzmeier, Armin: Was sind Sprachmodelle, was können sie und welche Probleme ergeben sich daraus? Ein Beitrag zur Leistungsfähigkeit und zu den Grenzen aktueller KI-Sprachmodelle, in: BuB – Forum Bibliothek und Information 75 (7), 2023, S. 348.

einem Thema, sondern nach Suchbegriffen oder Schlagwörtern, um so in den Bibliothekskatalogen die passende Literatur zu finden. Im Workshop wurde diese Vorgehensweise anhand von Schlagwörtern zur Geschichte der Französischen Revolution demonstriert. Hier lieferte auch ChatGPT 3.5 völlig zufriedenstellende Ergebnisse. Die durchnummerierte Liste mit den zwanzig Begriffen enthielt unter anderem Ereignisse, Institutionen, Personen und Personengruppen, etwa „Sturm auf die Bastille“, „Nationalversammlung“, „Ludwig XVI.“, „Jakobiner“ oder – völlig redundant – „Französische Revolution“.¹⁰ Der Copilot, so zeigten die Wortmeldungen, gab zwei verschiedene Arten von Listen aus. Entweder handelte es sich um eine reine Aufzählung oder die einzelnen Begriffe wurden zusätzlich kurz erläutert.¹¹ Eine erweiterte Form dessen zeigte Gemini. Hier wurden zwar deutlich mehr als die geforderten zwanzig Begriffe aufgeführt, dafür wurden sie in „Ideen“, „Personen“ oder „Ereignisse“ kategorisiert.¹²

Für all jene, die ohne Vorkenntnisse an ein Thema herangehen, sind Erläuterungen, wie sie vom Copiloten in manchen Fällen – bedauerlicherweise nicht immer – ausgegeben wurden, und Kategorisierungen der Begriffe, wie bei Gemini zu beobachten, ungemein hilfreich. Sie helfen, sich über den eigenen Bedarf klar zu werden, und ermöglichen verfeinerte Suchanfragen durch die Kombination der ausgegebenen Schlagwörter. Mit der entsprechenden Eingabe lassen sich in allen drei Sprachmodellen sowohl die Erläuterungen als auch die Kategorisierungen realisieren.

In der Abschlussrunde des Hands-on-Labs wurde auf das KI-Werkzeug Perplexity¹³ aufmerksam gemacht, das ebenfalls auf der Technologie von ChatGPT basiert. Perplexity stützt sich zur Genese der Ausgaben stark auf Quellen, die wiederum verlinkt werden. Dadurch halluziniert die KI merklich seltener und gibt häufiger als andere Sprachmodelle an, keine Informationen liefern zu können, was die Ausgaben zuverlässiger macht.¹⁴ Die Zuverlässigkeit von Perplexity wurde durch Wortmeldungen unterstrichen, wonach die Angaben zur Anzahl der DNB-Standorte wie auch zum Generaldirektor der DNB korrekt seien.

Ergänzend zur indirekten Literaturrecherche wurde darauf hingewiesen, dass sich Große Sprachmodelle nicht nur zur Ausgabe von Suchbegriffen für die weiterführende Recherche nutzen ließen, sondern auch für die Erstellung des gesamten Suchstrings – auch das eine enorme Arbeitserleichterung, insbesondere für komplexe Suchabfragen.

Das Hands-on-Lab hat anschaulich gezeigt, dass die direkte Informationsgewinnung und Literaturrecherche über Große Sprachmodelle zu unzuverlässig ist, um sie empfehlen zu können. Das gilt aktuell (Stand Sommer 2024) auch für modernere Sprachmodelle. Von weitaus mehr Erfolg gekrönt

10 Prompt: Nenne mir 20 Schlagworte für die Literaturrecherche zur Geschichte der Französischen Revolution., ChatGPT 3.5, 05.06.2024, <https://chat.openai.com>.

11 Prompt: Nenne mir 20 Schlagworte für die Literaturrecherche zur Geschichte der Französischen Revolution., Copilot, 05.06.2024, <https://copilot.microsoft.com/>.

12 Prompt: Nenne mir 20 Schlagworte für die Literaturrecherche zur Geschichte der Französischen Revolution., Gemini, 05.06.2024, <https://gemini.google.com/>.

13 <https://www.perplexity.ai/> Stand: 31.08.2024.

14 Falsche Angaben können auch hier vorkommen.

ist hingegen die indirekte Literaturrecherche, indem die Stärken der Sprachmodelle genutzt und Begriffe zum Thema für die weitere Suche ermittelt werden.

Diese Befunde sind freilich eine Momentaufnahme. Die eingangs genannte rasante Entwicklung lässt kaum erkennen, wo sie in fünf oder gar zehn Jahren stehen wird. Dass das Thema auch in Zukunft relevant sein wird, dürfte unbestreitbar sein. Fragen aus dem Auditorium, ob es für den Umgang mit KI in der Recherche Handreichungen für den schulischen Unterricht gebe, demonstrieren, dass im Moment ein hoher Bedarf in allen Bibliotheken vorhanden ist, wie mit KI umzugehen ist – ein Bedarf, der nicht so bald wieder verschwinden wird, weder in wissenschaftlichen noch in öffentlichen Bibliotheken. Sobald KI Einzug in die bibliothekarischen Suchsysteme hält, dürfte sich ändern, wie nach Literatur recherchiert wird.¹⁵ Die Schulungen zu Recherchefähigkeiten werden sich entsprechend ausrichten müssen.

Friedrich Quaasdorf, Deutsche Nationalbibliothek, Leipzig, ORCID: <https://orcid.org/0009-0005-3316-6925>

Zitierfähiger Link (DOI): <https://doi.org/10.5282/o-bib/6060>

Dieses Werk steht unter der [Lizenz Creative Commons Namensnennung 4.0 International](#).

15 An dieser Stelle sei auf Retrieval-Augmented Generation (RAG) verwiesen, in dem das klassische Retrieval mit Sprachmodellen kombiniert wird. Zhao, Penghai; Zhang, Hailin; Yu, Qinhan u.a.: Retrieval-Augmented Generation for AI-Generated Content: A Survey, in: arXiv, 21.06.2024. <https://arxiv.org/abs/2402.19473>, Stand: 02.07.2024.