

Haimo Stiemer/Evelyn Gius/Dominik Gerstorfer

Künstliche Intelligenz und literaturwissenschaftliche Expertise

Abstract: Based on Heinrich von Kleist's essay „On the Gradual Production of Thoughts During Speech“, this article discusses the extent to which his ideas can be adapted to the current debates on artificial intelligence (AI) and large language models (LLMs). Despite the opacity of LLMs, they are powerful tools. Therefore, the article argues for their assertive integration into the practice of literary studies serving as sparring partners in teaching, as assistants or as co-creative partners in research. First, the article deals with the problems and conditions that arise when the ChatGPT language model is used to retrieve literary knowledge. It then presents and reflects on examples of its use in the teaching of literature at the university level. Finally, the implications of AI for literary studies as a discipline are discussed.

„Wenn du etwas wissen willst und es durch Meditation nicht finden kannst“, schreibt Heinrich von Kleist im Jahr 1805, „so rate ich dir, mein lieber, sinnreicher Freund, mit dem nächsten Bekannten, der dir aufstößt, darüber zu sprechen. Es braucht nicht eben ein scharfdenkender Kopf zu sein, auch meine ich es nicht so, als ob du ihn darum befragen solltest: nein! Vielmehr sollst du es ihm selber allererst erzählen.“¹ Der mit diesen Sätzen beginnende, in der Germanistik viel diskutierte² Essay Kleists trägt den Titel „Über die allmähliche Verfertigung der Gedanken beim Reden“ und resümiert Beispiele, bei denen anspruchsvolle Denkleistungen durch den Akt der Rede befördert beziehungsweise erbracht werden. Die Rede ist dabei eingebunden in eine vermeintlich dialogische Situation und bleibt doch Selbstgespräch wie -aufklärung. Der Erkenntniszuwachs beim Redner ergibt sich also nicht aus der Kommunikation im eigentlichen, den Gedankenaustausch mit den Gesprächspartner:innen suchenden Sinn. Vielmehr versetzt ihn bereits deren Anwesenheit während seines Monologs zu einem bestimmten Moment in den für die

1 Heinrich von Kleist, „Über die allmähliche Verfertigung der Gedanken beim Reden“, in *Sämtliche Werke und Briefe in vier Bänden*, Bd. 3, *Erzählungen/Anekdoten/Gedichte/Schriften*, hg. von Ilse-Marie Barth et al., Frankfurt a. M. 1990, 534–540, 534.

2 Vgl. u. a. Gabriele Kapp, „Die Verortung des Aufsatzes: ‚Über die allmähliche Verfertigung der Gedanken beim Reden‘“, in *„Des Gedankens Senkblei“*. *Studien zur Sprachauffassung Heinrich von Kleists 1799–1806*, hg. von ders., Stuttgart/Weimar 2000, 289–311; Rüdiger Campe, „Verfahren. Kleists Allmähliche Verfertigung der Gedanken beim Reden“, *Sprache und Literatur*, Bd. 43(2), 2012, 2–21.

Erkenntnis notwendigen Erregungszustand.³ Mit Blick auf die ursprünglich französische Redewendung, nach welcher der Appetit beim Essen komme („l'appétit vient en mangeant“), konstatiert Kleist: die Idee kommt beim Sprechen („l'idée vient en parlant“).

Inwiefern kann Kleists „Modell spontaner sprachlicher Performanz“⁴ aber in Bezug auf die gegenwärtigen Debatten um Künstliche Intelligenz (KI) und große Sprachmodelle (Large Language Models, LLMs) aktualisiert werden? Oder genauer gefragt: Welche kognitiven Leistungen sind möglich, wenn der „nächste[...] Bekannte[...]“ und Gesprächspartner ein GPT-basierter Chatbot ist? In der von der Veröffentlichung von ChatGPT durch das Unternehmen OpenAI angeregten Debatte um die Potentiale, aber auch Begrenzungen von Sprachmodellen wurde immer wieder das Bild des stochastischen Papageis bemüht.⁵ Denn die KI generiert in Form neuronaler Netze auf Basis von Trainingsdaten statistisch ähnlichen Output und *imitiert* auf diese Weise das menschliche Sprachverhalten. Die Chat-Funktion des Programms erschafft weiterhin eine dialogische Situation, in der die Nutzer:innen mit Nachfragen oder auch Widerspruch auf die vom KI-Chatbot generierten Antworten reagieren können. Wenn wir abermals mit Kleist sprechen, so ist dieser stochastische Papagei vielleicht „kein scharfdenkender Kopf“, mit ihm können aber durchaus komplexe Konversationen geführt werden. In diesem Beitrag soll deshalb dargelegt und diskutiert werden, welche Anwendungsmöglichkeiten sich für Sprachmodelle in der akademischen Forschung und Lehre anbieten, wobei dies aus einer literaturwissenschaftlichen Perspektive geschehen soll. Hierfür sollen zunächst cursorisch die Bedingungen und Probleme erläutert werden, die sich im akademischen respektive literaturwissenschaftlichen Kontext beim Einsatz von Large Language Models ergeben (1.). Im Weiteren werden wir Anwendungsbeispiele für die literaturwissenschaftliche Lehre präsentieren, die im *fortext lab* an der TU Darmstadt konzipiert wurden (2.), um letztlich, auch als Fazit, die aus unserer Sicht möglichen Auswirkungen der KI auf die Literaturwissenschaft als Disziplin beziehungsweise die mit KI verbundenen Herausforderungen für sie zu skizzieren (3.).

3 Vgl. Peter Philipp Riedl, „Über die allmähliche Verfertigung der Gedanken beim Reden. An R. v. L.“, in *Kleist-Handbuch. Leben – Werk – Wirkung*, hg. von Ingo Breuer, Stuttgart/Weimar 2013, 150–152, 151.

4 Riedl 2013, 151.

5 Emily M. Bender et al., „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?“, in *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, hg. von Association for Computing Machinery (ACM), New York 2021, 610–623, in [<https://dl.acm.org/doi/10.1145/3442188.3445922>] (Zugriff: 06.10.2023).

1 Die literaturwissenschaftliche Expertise von ChatGPT

Wenn wir der Frage nachgehen, inwiefern ein großes Sprachmodell wie Chat GPT-3 oder -4 literaturwissenschaftliche Expertise bereithält und inwiefern diese abrufbar ist, gilt es vorab zu klären, was unter dieser Expertise zu verstehen ist. In einer sehr grundlegenden Bestimmung kann hierunter ein Set von rezeptiven und produktiven Textumgangsformen verstanden werden, das auf der Basis nicht nur anerkannter fachwissenschaftlicher Methoden und Kategorien, sondern auch literarhistorischen Wissens erkenntnistiftend auf literarische Texte angewendet wird.⁶ Die Begriffsbestimmung wird gleichwohl prekär, wenn sie im Hinblick auf den Methoden- und Theoriepluralismus innerhalb der Literaturwissenschaft präzisiert werden soll. Selbst bei der im Selbstverständnis der klassischen Literaturwissenschaft sicherlich zentralen hermeneutischen Praxis (im weitesten Sinn der Interpretation von Texten) besteht Uneinigkeit darin, welches spezifische Aufmerksamkeitsverhalten und damit welche kon-, ko- oder intratextuellen Aspekte besonders erklärungs wirksam und also interpretationsrelevant sind.⁷ Wenn wir zunächst jedoch von der grundlegenden Auffassung von der Expertise als fachwissenschaftlich fundierten, literarhistorisch geschulten Umgang mit literarischen Texten ausgehen, bleibt zu klären, welche Komponenten dieser Expertise zu welchem Zeitpunkt in der literaturwissenschaftlichen Praxis von großen Sprachmodellen bereitgestellt werden beziehungsweise bei welchen sie sekundieren können. Bei der Verwendung von ChatGPT-3.5 fällt die Antwort bei Prompts, mit denen literarhistorisches Wissen abgefragt wird, vorerst ernüchternd aus. Das über Prompts abgerufene Wissen besteht häufig aus einer Melange aus Fakten und Fiktion, wird allerdings durchaus eloquent präsentiert. Der Papagei kommt hier also an seine Grenzen und *halluziniert* mitunter, wie es in der Diskussion um KI-Chatbots heißt.

6 Vgl. Steffen Martus, „Interpretieren – Lesen – Schreiben. Zur hermeneutischen Praxis aus literaturwissenschaftlicher Perspektive“, in *Hermeneutik unter Verdacht*, hg. von Andreas Kablitz et al., Berlin/Boston 2021, 45–81; Walter Erhart, „Was wollen Philologen wissen? Über Praktiken und Passionen der Literaturwissenschaft“, in *Wissens-Ordnungen. Zu einer historischen Epistemologie der Literatur*, hg. von Nicola Gess/Sandra Janßen, Berlin/Boston 2014, 145–179.

7 Vgl. Klaus Weimar, „Text, Interpretation, Methode. Hermeneutische Klärungen“, in *Wie international ist die Literaturwissenschaft? Methoden- und Theoriediskussion in den Literaturwissenschaften: kulturelle Besonderheiten und interkultureller Austausch am Beispiel des Interpretationsproblems (1950–1990)*, hg. von Lutz Danneberg/Friedrich Vollhardt, Stuttgart/Weimar 1996, 110–122.

Dies geschieht nach unseren bisherigen Erfahrungen, erstens, bei jenen Themenbereichen, die im literaturwissenschaftlichen Diskurs einen eher randständigen Status besitzen, bei Themen also, für die vermutlich wenig oder gar keine Trainingstexte dem Sprachmodell zugrunde liegen. Anzunehmen ist, dass dies auch für nicht, nur schwach kanonisierte oder dekanonisierte literarische Texte gilt. Folgen wir dieser Vermutung, dann stiege mit zunehmender Kanonizität die Wahrscheinlichkeit, dass die Ausgaben des Sprachmodells zumindest weitestgehend korrekt sind, wenn es zum Beispiel um Handlungszusammenfassungen von literarischen Texten geht. Mit dieser Annahme steht dann wiederum die Frage im Raum, welche Parameter für die Kanonizität oder besser Relevanz bei der Bildung des Trainingstextkorpus relevant waren. Für die englischsprachige Literatur liegt inzwischen eine Studie vor, die anhand der Ausgaben von ChatGPT-4 rekonstruiert hat, welche literarischen Werke von diesem Modell memoriert⁸ wurden, und damit, welche Werke sein implizites Wissen über Literatur prägen.⁹ Es kam heraus, dass sowohl vor 1928 veröffentlichte, urheberrechtsfreie, als auch (urheberrechtlich geschützte) zeitgenössische Science Fiction- und Fantasy-Romane memoriert wurden, wie zum Beispiel „Alice in Wonderland“, die Harry Potter-Romane, die „Lord of the Rings“-Trilogie, „The Hunger Games“ von Suzanne Collins oder „Pride and Prejudice“ von Jane Austen. Der Grad der Memorisierung hängt vermutlich, so die Autor:innen der Studie, mit der Häufigkeit zusammen, mit der Passagen aus diesen Werken im Internet erschienen sind. Der Fokus auf vor 1928 veröffentlichte, kanonisierte Texte und zeitgenössische Science Fiction- und Fantasy-Literatur hat in der Konsequenz zur Folge, dass das Sprachmodell zu Texten anderer Genres deutlich schlechtere Antworten generiert. Auch wenn eine vergleichbare Studie zur deutschsprachigen Literatur noch nicht vorliegt, muss dieser Bias immer mitgedacht werden: Das Modell liefert bessere Aussagen über populäre Gegenwartstitel als über den Long Tail der Literatur. Weniger populäre Texte werden im Zweifelsfall mit wenig sinnvollen oder auch falschen Informationen beschrieben, die das Modell anhand der populären Titel generalisiert. Eine weitere Beobachtung erschwert das Zutrauen in die ChatGPT-Ausgaben zu literaturwissenschaftlichem Wissen.

⁸ Zu beachten ist, dass das Memorieren im Kontext des maschinellen Lernens ein unerwünschter Effekt ist: Das Modell passt sich zu stark an den Trainingsdatensatz an und generalisiert in der Konsequenz schlechter in Bezug auf neue Daten. Das hier gemeinte Memorieren unterscheidet sich also von dem menschlichen Einprägen bzw. Memorieren. Vgl. Adrian de Wynter et al., „An Evaluation on Large Language Model Outputs: Discourse and Memorization“, *Natural Language Processing Journal*, Bd. 4, 2023, in [https://doi.org/10.1016/j.nlp.2023.100024] (Zugriff: 06.10.2023).

⁹ Kent K. Chang et al., „Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4“, 2023, in [https://doi.org/10.48550/arXiv.2305.00118] (Zugriff: 18.08.2023).

Der halluzinatorische Gehalt der Antworten scheint, zweitens, ebenso in jenen Bereichen groß, in denen die Fachterminologie nah der Alltagssprache ist oder die fachsprachlichen Termini in der Alltagssprache mit einer anderen Bedeutung verwendet werden. Auszüge aus einem Chat mit dem Sprachmodell zum Konzept der narrativen Ebenen vom 21. Februar 2023 verdeutlichen das:

Prompt: Erläutere das Konzept der narrativen Ebenen oder Erzählebenen.

ChatGPT: Das Konzept der narrativen Ebenen oder Erzählebenen bezieht sich auf die verschiedenen Ebenen oder Schichten einer Erzählung, die dazu beitragen, eine komplexe und tiefgründige Geschichte zu schaffen. Die Idee ist, dass eine Erzählung aus mehr als nur einer einfachen Handlung besteht, und dass es mehrere Ebenen gibt, auf denen eine Geschichte erzählt werden kann.

Die erste Ebene einer Erzählung ist die Handlungsebene, auf der die grundlegende Geschichte erzählt wird. Das ist die Ebene, auf der die Charaktere handeln und die Ereignisse stattfinden. Es ist die Ebene, auf der die meisten Menschen eine Geschichte erleben und sie verstehen.

Die zweite Ebene ist die Reflexionsebene. Hierbei handelt es sich um eine Ebene, die sich auf die Gedanken, Überlegungen und inneren Konflikte der Charaktere konzentriert. Diese Ebene gibt dem Leser Einblick in die Psyche der Charaktere und hilft dabei, ihre Entscheidungen und Handlungen zu verstehen.

Die dritte Ebene ist die Metaebene. Hierbei handelt es sich um eine Ebene, auf der die Erzählung selbst reflektiert und hinterfragt wird. Auf dieser Ebene kann die Erzählung selbst zum Thema werden, indem sie die Frage aufwirft, was es bedeutet, eine Geschichte zu erzählen und warum wir als Leser oder Zuschauer uns für sie interessieren.

Das in der Narratologie einschlägige Konzept der narrativen Ebene bezieht sich auf Segmente einer Erzählung, die anhand von Sprecherwechseln oder in einigen Fällen auch von ontologischen Wechseln bestimmt werden. Neben Rahmen- und Binnenerzählungen, die jeweils eigene narrative Ebenen darstellen, ist jede Erzählung innerhalb einer Erzählung potentiell eine narrative Ebene, sofern sie von einer anderen Erzählinstanz (typischerweise einer Figur der Erzählung) hervorgebracht wird, ebenso wie jene Teile der Erzählung narrative Ebenen darstellen, welche sich durch einen Wechsel der fiktionalen Welt von anderen Teilen der Erzählung unterscheiden. Wenn zum Beispiel in einer Erzählung eine Figur selbst eine Erzählung eröffnet, einer anderen Figur also etwas erzählt, dann wird diese Erzählung in der Erzählung als eine zweite Ebene betrachtet. Auf der zweiten Ebene kann erneut durch eine weitere Figur eine Erzählung stattfinden und damit eine dritte Ebene entstehen, die somit zweifach eingebettet ist.¹⁰ Mit anderen Worten:

10 Matías Martínez/Michael Scheffel, *Einführung in die Erzähltheorie*, München ¹¹2020, 79 ff.; zum

Die erste Ebene ist auf keinen Fall eine Handlungs-, die zweite keine Reflexionsebene. Die Antwort von ChatGPT kontaminiert das narratologische Konzept mit anderen, eher dem Alltagswissen entstammenden Kriterien zur Benennung von Ebenen. Mit der vom Chatbot als „Metaebene“ ausgegebenen dritten Ebene wird hingegen durchaus ein Terminus bemüht, der dem der *Metadiegesis* aus dem Begriffsrepertoire der Narratologie im Bereich der narrativen Ebenen ähnelt. Nur bezeichnet dieser wiederum einen eigenständigen Erzählakt auf der zweiten Ebene. Es gibt in den Ausgaben des Chatbots folglich Anklänge von Expertise und es gibt diverse Halluzinationen beziehungsweise Falschaussagen, die mit schwadronierenden Ausführungen verknüpft werden, zum Beispiel wenn von einer „komplexe[n] und tiefgründige[n] Geschichte“ die Rede ist.

Problematisch ist die Verwendung des Chatbots für die Bereitstellung literaturwissenschaftlichen Wissens also in mehrfacher Hinsicht. Neben dem halluzinatorischen Gehalt und der Kontamination der Fachterminologie mit alltagssprachlichen Vorstellungen erweist sich das Black-Box-Problem, die epistemische Opazität des Programms, als schwierig für die wissenschaftliche Praxis. Zunächst kann nur näherungsweise und mit großen Mühen ermittelt werden, auf welchen Texten beziehungsweise Quellen die modellgenerierten Texte basieren. Denn OpenAI, aber auch andere kommerzielle Anbieter großer Sprachmodelle nehmen keine detaillierte Offenlegung der verwendeten Trainingsdaten vor. In der oben bereits erwähnten Studie zum englischsprachigen ChatGPT-Korpus wurde dessen literarisches Wissen letzten Endes ausschließlich anhand der Auslasstexte des Modells, also über den Output, rekonstruiert. Außerdem benötigen die LLMs sehr große Mengen an Trainingsdaten. Diese werden häufig unkuratiert aus dem Internet (Seiten wie Reddit oder Wikipedia) entnommen und enthalten dementsprechend alle möglichen Probleme des Internets, von Verschwörungstheorien über Hassrede bis hin zu falschen Informationen. Diese werden, sofern sie in größerem Umfang im Korpus enthalten sind, vom Modell als erwünscht gelernt und entsprechend im Output reproduziert. Neben dem Problem des für die Nutzer:innen unkenntlichen Quellenmaterials kommt hinzu, dass die Algorithmen und Modelle, die aus einem Input den entsprechenden Output erzeugen, ebenfalls nicht transparent sind, das heißt es sind nicht nur das Quellenmaterial, sondern auch die Prozesse der Verarbeitung nicht nachvollziehbar. Während das Trainingsmaterial im Prinzip einfach offengelegt und damit transparent gemacht werden könnte, lässt sich die Frage der Verarbeitungsprozesse hingegen nicht so leicht lösen, da diese aufgrund der algorithmischen Struktur von Deep Learning per se nicht für Men-

Konzept des illokutionären und ontologischen Ebenenwechsels: vgl. Marie-Laure Ryan, *Possible Worlds, Artificial Intelligence, and Narrative Theory*, Bloomington 1991, 175–177.

schen nachvollziehbar sind.¹¹ KI-Modelle liefern also nur Ergebnisse, aber keine anhand des Verarbeitungsweges nachvollziehbaren Rechtfertigungen für diese Ergebnisse; zum anderen ist nicht ersichtlich, was aus welchen Trainingsdaten gelernt wurde.

Wenn, wie soeben dargelegt, die Bereitstellung literaturwissenschaftlicher Erkenntnisse und Wissensbestände durch Sprachmodelle problembehaftet ist, so steht zur Diskussion, wie ChatGPT und Co. auf andere Weise produktiv mit dieser Disziplin zusammengeführt werden können. In einer Metastudie von Caleb Ziems et al. vom April 2023 wurden 13 Sprachmodelle und ihr Nutzen für die Sozialwissenschaften untersucht, hierbei wurden explizit auch text- beziehungsweise literaturwissenschaftliche Forschungsfragen berücksichtigt.¹² Die Forscher:innen kommen zu dem Schluss, dass zum Beispiel ChatGPT gut für diverse Zero-Shot-Klassifikationsaufgaben, also Klassifikationsleistungen ohne vorheriges Einlesen von Trainingsdaten, geeignet ist. Beispielhaft führen sie die Erkennung von Einstellungen beziehungsweise Meinungen in eingegebenen Texten an. Welche Meinung nimmt beispielsweise ein bestimmter Text gegenüber Atheismus oder der Legalisierung der Abtreibung ein? Ebenso gut schneiden die LLMs bei der Erkennung von Fehlinformationen oder Emotionen ab, was die Autor:innen der Studie darauf zurückführen, dass es für jede dieser Aufgaben entweder eine objektive, evidenzbasierte Grundlage gibt (wie im Fall von Fake News) oder ein an umgangssprachlichen Definitionen ausgerichtetes Schema. Der damit aufgerufene Anwendungsbereich erscheint insbesondere für die computationale Literaturwissenschaft¹³ interessant, in welcher manuelle wie automatisierte Annotationsaufgaben von zentraler Bedeutung sind. Einschränkend ist jedoch zu berücksichtigen, dass laut der erwähnten Studie die Sprachmodelle bei komplexeren Taxonomien schlechter abschneiden, also bei Aufgaben, bei denen die Annotationsrichtlinien durch Expertenwissen geprägt sind. Der Einsatz von Sprachmodellen im Forschungskontext bietet sich also lediglich bei kleineren, einfach strukturierten An-

11 Inzwischen wird dazu unter den Schlagwörtern der Interpretable AI bzw. Whiteboxing geforscht. Vgl. u. a. Ajay Thampi, *Interpretable AI. Building Explainable Machine Learning Systems*, New York 2022.

12 Caleb Ziems et al., „Can Large Language Models Transform Computational Social Science?“, 2023, in [<https://doi.org/10.48550/arXiv.2305.03514>] (Zugriff: 18.08.2023).

13 Die computationale bzw. computergestützte Literaturwissenschaft (computational literary studies, CLS) ist ein interdisziplinäres Teilgebiet der Digital Humanities, welches Methoden aus der Informatik resp. des Natural Language Processing (NLP), der computationellen Linguistik und der Literaturwissenschaft für die Analyse literarischer Texte bzw. Korpora vereint. Im deutschsprachigen Bereich erhielt dieses Forschungsfeld eine größere Aufmerksamkeit durch das von der Deutschen Forschungsgemeinschaft (DFG) seit dem Jahr 2020 geförderte Schwerpunktprogramm SPP 2207 „Computational Literary Studies“.

notationsaufgaben an, für welche keine Trainingsdaten mehr generiert werden müssen. Die Frage nach dem Nutzen der LLMs für die literaturwissenschaftliche Fachexpertise und Praxis bleibt dennoch prekär – wenn sie nicht von einer anderen Perspektive her angegangen wird. In einer solchen Perspektive können in der literaturwissenschaftlichen Praxis und Lehre die soeben genannten Defizite wie Fehler in den ChatGPT-Texten produktiv genutzt werden, quasi wie bei einem sokratischen Gespräch oder eben im Sinn von Kleists Überlegungen bezüglich der Ideenfindung beim Reden.

2 ChatGPT in der literaturwissenschaftlichen Lehre

Um die Vorzüge von ChatGPT in der literaturwissenschaftlichen Lehre zu verdeutlichen, sollen Beispiele herangezogen werden, die von uns für Lehrveranstaltungen an der TU Darmstadt konzipiert wurden. Grundsätzliches Ziel bei der Verwendung des Sprachmodells in der Lehre ist es dabei, den Studierenden nicht allein Kompetenzen im Umgang mit KI-Textgeneratoren zu vermitteln, sondern ebenso ihre literaturwissenschaftliche Expertise zu stärken.¹⁴ Der bereits erwähnte Chat, in welchem das Sprachmodell nach den narrativen Ebenen befragt wurde, ließe sich im Seminar demnach mit der Aufgabe an die Studierenden präsentieren, jene Stellen im ChatGPT-Text zu identifizieren, die entweder falsch, teilweise richtig oder ungenau sind. Auf diese Weise kann einerseits das erzähltheoretisch richtige Verständnis von Erzählebenen mobilisiert werden, andererseits können die Studierenden im Fortgang durch verbesserte, reaktive Prompts versuchen, den Chatbot einen erzähltheoretisch fundierten Beispieltext mit drei Erzählebenen generieren zu lassen, wobei dieses *Prompt Engineering* im Seminar ebenfalls reflektiert werden soll. Ein weiteres Beispiel zielt in eine ähnliche Richtung und betrifft das Konzept der Handlung beziehungsweise des Plots von Erzählungen. Die Studierenden werden gebeten, eine Handlungszusammenfassung von Kleists „Das Erdbeben von Chili“ zu verfassen. Verbunden ist diese Aufgabe lediglich mit der Auflage, hierfür nicht mehr als 20 Hauptsätze zu verwenden und für jeden neuen Satz einen Zeilenwechsel vorzunehmen. Die Zusammenfassungen der Studierenden werden dann mit jener des Sprachmodells verglichen.

¹⁴ Vgl. Henner Gimpel et al., „Unlocking the Power of Generative AI Models and Systems such as GPT-4 and ChatGPT for Higher Education: A Guide for Students and Lecturers. Discussion Paper“, 2023, in [<https://t1p.de/kana7>] (Zugriff: 24.10.2023).

Prompt: Schreibe eine möglichst einfache Zusammenfassung der Erzählung „Das Erdbeben in Chili“ von Heinrich von Kleist in 20 Sätzen, möglichst als Hauptsätze ohne Nebensätze. Fasse dabei die wichtigsten Geschehnisse des Textes zusammen.

ChatGPT:

1. „Das Erdbeben in Chile“ ist eine Erzählung von Heinrich von Kleist.
2. Es geht um ein verheerendes Erdbeben in Santiago de Chile.
3. Viele Menschen sterben bei dem Beben.
4. Die Protagonisten Jeronimo und Doña Blanca sind ein Paar und werden durch das Beben getrennt.
5. Sie suchen verzweifelt nach einander.
6. Sie erleben viele Abenteuer, während sie nach einander suchen.
7. Sie finden sich schließlich wieder.
8. Sie beschließen zusammen zu fliehen, um dem Chaos zu entkommen.
9. Sie begegnen vielen anderen Menschen, die das gleiche Schicksal erlitten haben.
10. Sie versuchen, ihr Leben wieder aufzubauen.
11. Sie beschließen, sich in einer anderen Stadt niederzulassen.
12. Sie beginnen ein neues Leben zusammen.
13. Das Erdbeben hat ihr Leben verändert.
14. Sie müssen alles von vorne aufbauen.
15. Sie finden jedoch Trost in ihrer Liebe zueinander.
16. Sie beschließen, ihr Leben gemeinsam weiterzuführen.
17. Sie sind bereit für die Zukunft.
18. Die Erzählung endet damit, dass das Paar ein neues Leben beginnt.

Auch hier, bei einem in der Germanistik kanonischen Text, finden sich diverse Ungenauigkeiten und Halluzinationen. Dies beginnt mit den Figuren – eine Dona Blanca gibt es in Kleists Text überhaupt nicht; die Liebesbeziehung besteht zwischen Jeronimo und Josephe, die von ChatGPT überhaupt nicht erwähnt wird. Auch werden die Protagonist:innen nicht erst durch das Erdbeben getrennt, sondern durch ihre Inhaftierung in Folge ihrer vom Vater von Josephe nicht gebilligten Verbindung. Frappierend ist, dass ChatGPT der Erzählung ein positives Ende zuschreibt. Im eigentlichen Kleist-Text wird das Paar in einem Tumult vor einer Kirche brutal ermordet, während sie bei ChatGPT am Ende, nachdem sie mehrere Abenteuer überstanden haben, ein neues Leben beginnen.

Im Seminar kann nun das Wissen der Studierenden über den Text mit dem der ChatGPT-Ausgabe kontrastiert und die Relevanz der einzelnen Handlungsabschnitte, Figuren und Ereignisse für das Handlungsgerüst sowie die Strukturierung des Textes diskutiert werden. Was fehlt in den Zusammenfassungen von ChatGPT und in ausgewählten Texten von den Studierenden? In einem zweiten Schritt, verbunden mit der Relevanzfrage, können dann das literaturwissenschaftliche Konzept der Handlung wie die Methodik bei der Erstellung von Handlungszusammenfassungen reflektiert, diskutiert und evaluiert werden. Überprüft und gefördert wird dergestalt die Fachexpertise, ein wissenschaftlich gestütztes Handlungskon-

zept zu haben, es definieren, erläutern *und* anwenden zu können. Allerdings können literaturwissenschaftliche Konzepte mit einem Sprachmodell im Seminar nicht nur eingeübt, sondern auch problematisiert werden. Möglich wäre zum Beispiel, den Begriff der Epoche zu problematisieren und ChatGPT nach einer begründeten Epochenzuordnung von einzelnen Textauszügen zu fragen, ohne den Autor:innenamen anzugeben. Wenn die erste und zweite Strophe eines Gedichts von Novalis dann – wie geschehen – dem jungen Schiller der Sturm-und-Drang-Zeit zugewiesen wird, können im Seminar die Konvergenz von ästhetischen Programmen, aber auch die in ihrer schematischen Orientierungsfunktion nicht selten prekären Epochenkonstruktionen in der Literaturwissenschaft thematisiert werden.

Die offensive Integration von textgenerativer Software in die literaturwissenschaftliche Lehre eröffnet also viele Möglichkeiten. Mit ihr können die rezeptiven wie produktiven Textumgangsformen der Studierenden, ihre Fachexpertise und ihr kritischer Umgang mit Forschungspositionen und -routinen gefördert werden. Fachunabhängiges, übergreifendes Ziel wäre dabei zugleich, die Mündigkeit im Umgang mit großen Sprachmodellen zu befördern und Kompetenzen im Bereich *Prompt Engineering* zu vermitteln, also Wege aufzuzeigen, wie Prompts gestaltet und verfeinert werden können, um bestmöglichen Output zu erzeugen.

3 Fazit: Von der Output- zur Input-Optimierung

Dem Vorbild der dialogischen Kommunikation folgend, können die Übungen mit den Studierenden zum *Prompt Design* auch bewusst tentativ angelegt sein. Das Sprachmodell figuriert dann als Gesprächspartner im Kleistschen Sinn, dessen Äußerungen als Artikulation eines *kollektiven Wissens* im Wechselspiel von In- und Output, in der Abfolge von Frage, Antwort und Nachfrage hilft, den Prompt weniger in seiner Form zu verbessern als in seinem Inhalt zu verändern. Kleists eingangs zitiertes Essay auf die KI-Textgeneratoren zu übertragen, bedeutet demnach, nicht auf die Output-, sondern die Input-Optimierung zu orientieren. Oder, mit Kleists Worten:

Aber weil ich doch irgendeine dunkle Vorstellung habe, die mit dem, was ich suche, von fern her in einiger Verbindung steht, so prägt, wenn ich nur dreist damit den Anfang mache, das Gemüt, während die Rede fortschreitet, in der Notwendigkeit, dem Anfang nun auch ein Ende zu finden, jene verworrene Vorstellung zur völligen Deutlichkeit aus, dergestalt, daß die Erkenntnis zu meinem Erstaunen mit der Periode fertig ist. Ich mische unartikulierte Töne ein, ziehe die Verbindungswörter in die Länge, gebrauche wohl eine Apposition, wo sie nicht nötig

wäre, und bediene mich anderer, die Rede ausdehnender, Kunstgriffe, zur Fabrikation meiner Idee auf der Werkstätte der Vernunft, die gehörige Zeit zu gewinnen.¹⁵

Im Gespräch probiert sich der Redner, tastet sich mit Kunstgriffen nach vorn, mit Vorstellungen, die nur von fern her mit dem noch nicht fest definierten Ziel in Verbindung stehen und die an Deutlichkeit erst im Gespräch gewinnen. Die Problemlösung obliegt nicht dem Gesprächspartner, dessen Antworten oder Reaktionen daher auch nicht der Präzision bedürfen. Was Kleist hier beschreibt, kann durchaus mit dem konnektionistischen Verfahren der LLMs gleichgesetzt werden, bei welchem Wissensbestandteile nach ihrem wahrscheinlichen Näheverhältnis zusammengefügt werden. Der Kompetenzüberschuss aber liegt immer bei den Nutzer:innen des Sprachmodells, die für die „Fabrikation der Idee“ im Gespräch, im Chat mit dem Sprachmodell ihre ungefähren Vorstellungen überprüfen, anpassen und bestenfalls präzisieren. Geraten sie dabei in eine Sackgasse, sollten sie sich einen neuen Gesprächspartner suchen, also einen neuen Chat beginnen, denn das Programm merkt sich nur den Verlauf der aktuellen Unterhaltung und kann sich im nächsten Chat nicht wieder drauf beziehen.

Zusammenfassend lässt sich festhalten, dass es sich bei Chatbots trotz ihrer elaborierten sprachlichen Form nicht um Expertensysteme beziehungsweise Experten handelt. Vielmehr sind sie als sehr mächtige, aber opake statistische Werkzeuge anzusehen, deren Integration in die literaturwissenschaftliche Praxis auf vielfältige Weise sinnvoll sein kann. Während der Output neue und interessante Beobachtungen enthalten kann, ist dessen kritische Beurteilung und unabhängige Überprüfung nach den Standards der Disziplin nötig. Die Sprachmodelle können somit als akademischer Sparringspartner in der Lehre eingesetzt werden und das hochschuldidaktische Instrumentarium erweitern. Sie können in der Forschung eine Assistenzfunktion übernehmen, zum Beispiel für Zero-Shot-Klassifikationsaufgaben, und sie können ebenso eine ko-kreative Funktion bei der Entwicklung von Forschungsfragen oder der Überprüfung von Forschungsergebnissen bezüglich ihrer Plausibilität wahrnehmen. Diese Ko-Kreativität konstituiert dabei einen Resonanz- und Reflexionsraum und führt also nicht zu einer Veräußerung der eigentlichen wissenschaftlichen Tätigkeit.

Für die Geisteswissenschaften im Allgemeinen und die Literaturwissenschaft im Besonderen stellen die KI-Textgeneratoren, insofern ihre Anwendung kritisch, reflektiert erfolgt, keine Gefährdung der akademischen Integrität dar. In Bezug auf die Hochschullehre wurde in den letzten Monaten hingegen mehrfach die Befürchtung geäußert, dass Studierende ihre Fähigkeiten zur Recherche, zum kriti-

15 Kleist 1990, 535.

schen Denken und zur schriftlichen Kommunikation verlernen könnten. Richtig ist, dass die Assistenzfunktion von Sprachmodellen Plagiate oder Urheberrechtsverletzungen erleichtert und vor allem, je nach Einsatz der KI-Textgeneratoren, die geforderte Eigenständigkeit wissenschaftlichen Arbeitens unterlaufen kann. Wissenschaftliches Fehlverhalten aber hat es auch davor gegeben. Wichtig ist, den Studierenden den kritischen Umgang mit den generierten Texten und die sinnvolle Verwendung der KI-Tools aufzuzeigen. Ihre offensive Einbeziehung in die Hochschullehre erscheint uns daher nicht nur wünschenswert, sondern geboten. Nicht zu vergessen ist darüber hinaus, dass der weit vorangeschrittene digitale Transformationsprozess bereits unser Verständnis von Kommunikation und Text verändert hat. Der Umgang mit KI-generierten Texten und Bildern wird somit eine Schlüsselkompetenz der nahen Zukunft sein und die Literaturwissenschaft als gesellschaftlichen Ort, an dem zu Texten, Textformen und deren Rezeption geforscht wird, in besonderer Weise fordern. Die „Algorithmisierung der Kultur“ (Gerhard Lauer) verlangt nach philologischer Kompetenz respektive Expertise. Eine Literaturwissenschaft, die sich dieser Herausforderung stellt, braucht sich um ihre gesellschaftliche Legitimation nicht zu sorgen.