



# DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz

Hans-Georg Müller / Maurice Fürstenberg (Hg.)

Hans-Georg Müller/Maurice Fürstenberg (Hg.)  
DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz

Sebastian Bernhardt (Hg.)  
Literatur – Medien – Didaktik  
Band 16

Hans-Georg Müller / Maurice Fürstenberg (Hg.)

**DeutschGPT –  
Deutschunterricht im Dialog  
mit Künstlicher Intelligenz**

Umschlagabbildung: Maurice Fürstenberg unter Verwendung von ChatGPT-4o

Diese Veröffentlichung wurde aus Mitteln des Publikationsfonds für Open-Access-Monografien des Landes Brandenburg gefördert. / This publication was supported by funds from the Publication Fund for Open Access Monographs of the Federal State of Brandenburg, Germany.

*Die Beiträge haben eine intensive Qualitätsprüfung und mehrere inhaltliche wie formale Überarbeitungsschleifen durch die Herausgeber des Bandes und den Reihenerausgeber erfahren.*



CC-BY-NC-SA

ISBN 978-3-7329-1120-2  
ISBN Open Access 978-3-7329-8796-2  
DOI 10.26530/20.500.12657/104475  
ISSN 2749-5620

Frank & Timme GmbH Verlag für wissenschaftliche Literatur  
Berlin 2025.

Herstellung durch Frank & Timme GmbH  
Wittelsbacherstraße 27a, 10707 Berlin  
info@frank-timme.de  
Gedruckt auf säurefreiem, alterungsbeständigem Papier.

[www.frank-timme.de](http://www.frank-timme.de)

# Inhaltsverzeichnis

HANS-GEORG MÜLLER / MAURICE FÜRSTENBERG <b>Vorwort</b> .....	7
KATHARINA SIMBECK <b>Von Wörtern zu Wundern – Die Technologie großer Sprachmodelle und ihre Grenzen</b> .....	13
MAIK PHILIPP <b>Die didaktischen Fragen, die KI aus Sicht des Lesens zum Zweck des Lernens aufwirft</b> .....	39
IRENE CORVACHO DEL TORO / MAREIKE FUHLROTT / TORSTEN STEINHOFF <b>Didaktische Agenten – KI als Lehr-/Lernpartnerin im Deutschunterricht im Forschungsprojekt KIMADU</b> .....	65
FRIEDRICH BACH / SEBASTIAN BERNHARDT / SILVIA REUVEKAMP / NINJA SCHMIEDGEN <b>SHIFT happens – Lernen mit und von textgenerierender KI</b> .....	87
OLAF GÄTJE / TOBIAS WEINDEL <b>Schreibend und lesend Texte schreiben mit dem <i>Writing-Ko-Aktanten</i> ChatGPT – Über das schriftliche Instruieren von und den Dialog mit Large Language Models</b> .....	107
KATRIN BÖHME / JANNE MESENHÖLLER <b>Large Language Models – Chancen und Grenzen großer Sprachmodelle für die schulische Nutzung in sprachlich heterogenen Lerngruppen</b> .....	135

MAURICE FÜRSTENBERG	
<b>Zur Qualität KI-generierten Feedbacks – Ein explorativer Vergleich menschlicher und künstlicher Intelligenzen .....</b>	<b>163</b>
ANNA ANSARI	
<b>Schreiben mit KI als didaktische Herausforderung – Empirische Einblicke in Prompting-Praktiken von Lernenden .....</b>	<b>191</b>
ANNA JACHIMEK	
<b>ChatGPT-4o im Grammatikunterricht – Möglichkeiten und Grenzen des Tools .....</b>	<b>219</b>
TATJANA ATANASOSKA	
<b>„Für Matura-Ausarbeitungen war es wirklich ein Lifesaver!“ – KI und Schreiben in der Schule .....</b>	<b>245</b>
KASPAR RENNER	
<b>„Goethe vs. ChatGPT“ – Einblicke in eine Unterrichtsreihe zur funktionalen Nutzung von ChatGPT zur Förderung fachspezifischer Kompetenzen im Umgang mit literarischen Texten .....</b>	<b>267</b>
<b>Beiträger:innen .....</b>	<b>293</b>

## **Vorwort**

Wer aktuell den Blick in Klassenzimmer und darin stattfindenden Deutschunterricht wirft, findet bei oberflächlicher Betrachtung trotz der bahnbrechenden Neuerungen, die generative Sprachmodelle mit sich gebracht haben, ein überraschend statisches Bild vor, in dem der Kompetenzerwerb im Großen und Ganzen noch so abläuft wie vor dem vielzitierten November 2022. Die eigentlichen Änderungen zeigen sich derzeit noch eher unter der Oberfläche – oder besser gesagt: Unter der Schulbank, wo Schüler:innen längst begonnen haben, mit Tablets und Handys schulische Aufgaben zu lösen, Wissen vorzutauschen, Sprachkompetenz zu simulieren und ihre Lehrkräfte verblüfft ob der plötzlichen Fähigkeiten zurückzulassen. Diese Entwicklung zeigt, dass das sonst so wirkmächtige Beharrungsvermögen der Institution Schule als Handlungsoption für den Umgang mit KI ausfällt.

Deshalb hat sich in Deutschdidaktik und Unterrichtsforschung ein Diskurs etabliert, der sich auf vielfältige Weise den neuen Herausforderungen stellt und Künstliche Intelligenz in den Dienst schulischer Kompetenzentwicklung stellen möchte. War dieser Diskurs erst kürzlich noch vom Sammeln vorwissenschaftlicher Erfahrungen im eigenen Umgang mit KI geprägt, scheint mittlerweile auch schon die Phase der technologischen Erläuterungen und theoretischen Modellbildungen vorerst auszuklingen und Platz zu machen für kleinere und größere empirische Erhebungen sowie theoretische Überlegungen zu Einzelfragen im Umgang mit dem neuen, künstlich-intelligenten „Sprachgebrauchsautomaten“ (Müller/Fürstenberg 2023).

Diesen Fragen widmet sich auch die vorliegende Publikation, die als Tagungsband aus der titelgebenden Konferenz „*DeutschGPT 2.0 – Deutschunterricht im Dialog mit Künstlicher Intelligenz*“ vom 20. und 21. September 2024 an der Universität Potsdam hervorging, welche ihrerseits als Nachfolgerin

zur Tagung „*DeutschGPT – Deutschunterricht in Zeiten von Chatbots und KI*“ im Sommer 2023 an der Ludwig-Maximilians-Universität München stattfand.

Der Sammelband beginnt mit dem Beitrag von **Katharina Simbeck**, die in die Welt der großen Sprachmodelle (Large Language Models, LLMs) aus technologischer Perspektive einführt. Simbeck erläutert, wie LLMs funktionieren, welche Fähigkeiten sie besitzen, aber auch, wo ihre Grenzen liegen. Dabei nimmt sie aktuelle Entwicklungen auf und gibt Anregungen für den Unterricht, die weniger die Arbeit mit als vielmehr das Verständnis von Künstlicher Intelligenz und ihren Möglichkeiten adressieren.

**Maik Philipp** greift das Thema aus einer bisher unterrepräsentierten Perspektive rezeptiver Sprachkompetenz auf und formuliert Grundsatzfragen, die der Einsatz von KI für das *Lernen durch Lesen* aufwirft. Darin bietet er Antworten aus zwei Perspektiven: Einerseits betrachtet er Wissensmerkmale und -produktion, andererseits die Voraussetzungen der Lernenden und die Aufgaben der Lehrkräfte bei KI-Assistenz. Dadurch werden grundlegende didaktische Themen angesprochen, die für die Deutschdidaktik im Allgemeinen und die Lesedidaktik im Besonderen weitreichend und fruchtbar sind.

Diesem theoretisch-reflexiven Beitrag folgen stärker praxisorientierte Studien, so der Beitrag von **Irene Corvacho del Toro, Mareike Fuhlrott und Torsten Steinhoff**, die das Forschungsprojekt „KIMADU“ (Künstliche Intelligenz im Mathematik- und Deutschunterricht) vorstellen. Nach einer knappen theoretischen Einordnung zeigt dieser Beitrag, wie KI konkret im Schulalltag verankert werden kann. KIMADU erprobt mit 25 Schulen in NRW, wie generative KI sinnvoll im Unterricht eingesetzt werden kann. Der Beitrag stellt die Projektgrundlagen, das Design und erste Ergebnisse vor, etwa zu den Überzeugungen von Lehrkräften beim KI-Einsatz. Damit schlägt er eine Brücke von der Theorie in die Praxis und verdeutlicht, welches Potenzial didaktische KI-Agenten als Lernpartner im Deutschunterricht haben und wie Lehrkräfte dafür qualifiziert werden können.

Einen technologisch anderen Ansatz mit ähnlichem Ziel für Studierende präsentieren **Friedrich Bach, Sebastian Bernhardt, Silvia Reuvekamp und Ninja Schmiedgen** mit der KI-Anwendung *Tutor.AI*. Ihr Beitrag ergänzt die schulische Perspektive um einen hochschuldidaktischen und interdisziplinären Blick. Die Autor:innen fokussieren zunächst zentrale Probleme beim

Einsatz von LLMs im Bildungsbereich, so die Intransparenz der Trainingsdaten und die privatwirtschaftlichen Interessen der LLM-Betreiber, und stellen anschließend ihr Projekt vor, das von einem interdisziplinären Team der Universität Münster erarbeitet wird. Bei *Tutor.AI* handelt es sich um einen KI-basierten Chatbot, der auf eigenen Uni-Servern betrieben wird, um Souveränität zu gewährleisten, und dessen Wissensbasis von Lehrenden gezielt mit kursrelevanten Materialien gespeist wird. Der Beitrag gewährt Einblicke in die Beta-Testphase des Projekts und präsentiert Seminarkonzepte, mit denen angehende Lehrkräfte für den Umgang mit textgenerierender KI sensibilisiert werden.

Der Frage nach den Herausforderungen beim textgenerierenden Prompting wendet sich der Beitrag von **Olaf Gätje und Tobias Weindel** zu. Die Autoren verstehen KI nicht als bloßen *Ghostwriter*, sondern als *Writing-Ko-Aktanten* – einen Mit-Autoren im Schreibprozess. Anhand der Analyse von Chatverläufen zweier Zehntklässler:innen zeigen sie, dass die Einbindung von ChatGPT beim Überarbeiten von Inhaltsangaben zwar unterschiedliche Nutzungsstrategien hervorbringt, die aber die Gemeinsamkeit aufweisen: Zielgerichtetes Arbeiten mit einem KI-Chatbot erfordert erhebliche schriftsprachliche und dialogische Kompetenzen und KI-Unterstützung im Schreibunterricht verlangt ein neues Kompetenzprofil (Stichwort *AI Literacy*), welches sich Lernende aneignen müssten, um wirklich von KI zu profitieren.

**Katrin Böhme und Janne Mesenhöller** werfen einen Blick auf Lernvoraussetzungen und Chancengerechtigkeit beim KI-Einsatz und nehmen explizit die Heterogenität der Schülerschaft in den Fokus. Die Autorinnen arbeiten heraus, dass der Mehrwert von LLMs wie ChatGPT im Unterricht stark von individuellen sprachlichen Kompetenzen der Lernenden abhängt. Ihre Befunde deuten darauf hin, dass die Nutzung von LLMs zur Verschärfung der digitalen Kluft beitragen kann. Auf dieser Basis formulieren sie praktische Empfehlungen für die Gestaltung von Prompts, um weniger sprachstarke Lernende gezielt zu unterstützen.

**Maurice Fürstenberg** wendet sich mit seinem Beitrag der Hoffnung vieler Lehrkräfte zu, KI als Feedbackmaschine für automatisierte Rückmeldungen zu Schülertexten einzusetzen. In seiner empirischen Studie vergleicht er KI-generierte Rückmeldungen qualitativ und quantitativ mit dem Urteil menschlicher

Fachexperten und kann zeigen, dass die quantitativen Übereinstimmungen sehr begrenzt bleiben. Qualitativ zeigt sich ein gemischtes Bild: Zwar können KI-Systeme prinzipiell lernförderliches Feedback geben, es finden sich jedoch auch deutliche Inkonsistenzen und inhaltliche Fehler, die das Vertrauen der Lernenden beeinträchtigen. Der Beitrag unterstreicht damit die Notwendigkeit, den Einsatz von KI im Schreibunterricht kritisch zu begleiten.

**Anna Ansari** untersucht in ihrem Beitrag, wie Lernende tatsächlich mit der KI interagieren und welche Arten von Prompts sie bei der Lösung von Schreibaufgaben verwenden. Ihre Analyse zeigt, dass Schüler:innen und Studierende KI vorrangig dazu nutzen, Textentwürfe generieren zu lassen oder oberflächliche Verbesserungen vorzunehmen. Fortgeschrittenere Funktionen – etwa tiefere Umstrukturierungen von Texten, – werden hingegen kaum abgerufen. Im Anschluss diskutiert die Autorin, wie diese Prompting-Kompetenzen durch didaktische Impulse gefördert werden können.

**Anna Jachimek** richtet den Blick auf den Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“. In ihrer Studie erprobt sie die Leistungsfähigkeit von ChatGPT im Bereich deutsche Grammatik anhand verschiedener Aufgabenformate. Die Ergebnisse zeigen ein gemischtes Bild: ChatGPT-4o kann zwar Texte grammatisch korrekt ergänzen und teilweise auch normkonform korrigieren, bei der Erklärung grammatischer Phänomene zeichnen sich hingegen Grenzen ab, die sich in fehlerhaften oder irritierenden Antworten niederschlagen. Jachimek betont daher, dass die Nutzung von ChatGPT-4o im Grammatikunterricht unbedingt durch die Lehrkraft begleitet und überwacht werden sollte.

Einen Einblick in die Perspektive der Lernenden selbst liefert **Tatjana Atanasoska**. Ihr Beitrag ergänzt die bisherigen Studien durch qualitative Befunde aus Gruppendiskussionen mit Oberstufenschüler:innen. Atanasoska geht der Frage nach, wofür Schüler:innen KI-generierte Texte im Schulalltag einsetzen und wie sie diese Nutzung bewerten. Die Diskussionsteilnehmer:innen berichten, dass sie KI-Tools vorrangig zur Vortäuschung, kaum hingegen zum strukturieren Aufbau schriftsprachlicher Kompetenzen nutzen. Damit unterstreicht Atanasoska nochmals eindrücklich die Notwendigkeit, im Unterricht eine Kultur des reflektierten KI-Einsatzes zu etablieren, die solche Probleme adressiert.

Den Abschluss des Bandes bildet ein Praxisbericht von **Kaspar Renner**, der zeigt, wie KI didaktisch gewinnbringend im Literaturunterricht eingesetzt werden kann. Der Autor stellt eine Unterrichtssequenz vor, in der ein ChatGPT gezielt zur Förderung unterschiedlicher literarischer Kompetenzen genutzt wurde. Dabei traten die Schüler:innen einer Gymnasialklasse in einen spielerisch-kompetitiven Wettstreit zwischen ChatGPT und der Lyrik Goethes ein, der auf eine vertiefte Einsicht in Merkmale, Qualitäten und Deutungshorizonte der Sturm-und-Drang-Lyrik Goethes abzielte. Renner zeigt damit exemplarisch auf, dass KI-Einsatz den Deutschunterricht beleben und um neue Aspekte bereichern kann, sofern er überlegt und im Hinblick auf die Entwicklung sprachlicher Kompetenzen erfolgt.

Alle Untersuchungen des vorliegenden Bandes spiegeln die wissenschaftliche wie praxisorientierte Auseinandersetzung mit Künstlicher Intelligenz und Deutschunterricht wider. Dabei scheint es uns als den Herausgebern alles andere als ausgemacht, dass das Schicksal des Deutschunterrichts unabwendbar vom umfänglichen Einsatz automatischer Textgeneratoren geprägt sein wird. Stattdessen bieten sich aus unserer Sicht weitreichende Chancen, KI innovativ, produktiv und förderlich für die Entwicklung sprachlicher Kompetenzen zu nutzen – nicht für deren Vortäuschung. Wenn wir ernst nehmen, was die eigentliche Stärke Künstlicher Intelligenz ist, nämlich Muster in großen Datensätzen zu finden, dann kann uns die Technologie dabei helfen, grundlegende Fragen rund um den Deutschunterricht zu klären, etwa: Was macht einen guten Text in welchem Entwicklungsstadium aus? An welchen Kriterien machen wir das fest? Welche didaktischen Pfade und Aufgaben helfen welchen Schüler:innen in welchen Lagen wie gut weiter? Wie muss Feedback strukturiert sein, um wirklich hilfreich zu sein? Eine Voraussetzung dafür wird in der Aufgabe liegen, Künstliche Intelligenz auf die Bedarfe des Deutschunterrichts anzupassen – nicht umgekehrt.

Abschließend möchten wir uns bei allen Beitragenden der DeutschGPT und insbesondere dieses Bandes bedanken, bei den Hilfskräften auf der Tagung, dem Reihenherausgeber, der zentralen Forschungsförderung der Universität Potsdam, der Kommission für Forschung, wissenschaftlichen Nachwuchs und Chancengleichheit (FNK), der Open-Access-Förderung des Landes Branden-

Hans-Georg Müller/Maurice Fürstenberg

burg und den Verlagsmitarbeiter:innen. Sie alle haben dazu beigetragen, dass dieser Band rasch als Publikation vorliegt.

Potsdam und München im Sommer 2025

Hans-Georg Müller und Maurice Fürstenberg

## Von Wörtern zu Wundern

### Die Technologie großer Sprachmodelle und ihre Grenzen

#### Abstract

In den letzten zwei Jahren haben große Sprachmodelle durch intuitive Chat-Benutzerschnittstellen eine erhebliche Popularität erlangt. Sie verfügen über bemerkenswerte Fähigkeiten: Sie können nicht nur natürlichsprachliche Fragen beantworten und ersetzen damit zunehmend Suchmaschinen, sondern sie erstellen auch hervorragend strukturierte Texte, entwickeln Programme und übernehmen Übersetzungsaufgaben. Wie jedoch funktionieren große Sprachmodelle? Was sind ihre Fähigkeiten und Grenzen? Welche Anwendungen gibt es für die Sprachmodelle und wie werden sie dafür angepasst? Mit diesen Fragen beschäftigt sich dieser Beitrag und geht dabei auch auf aktuelle Entwicklungen und Implikationen für die Institution Schule ein.

**Schlagwörter:** Große Sprachmodelle (LLMs), Transformerarchitektur, Prompt Engineering, Unerwünschte Inhalte in Sprachmodellen, Kompetenzen

## 1 Die Technologie großer Sprachmodelle

Bereits in den 1960er Jahren faszinierte der am amerikanischen MIT entwickelte Chatbot ELIZA Menschen mit seiner Fähigkeit, sich in natürlicher Sprache auszudrücken. ELIZA kommunizierte nur regelbasiert und reagierte nur begrenzt auf Antworten der anderen Gesprächsseite.

Idee für den Englischunterricht: Lassen Sie die Schüler:innen mit ELIZA interagieren (<https://anthay.github.io/eliza.html>) und fordern Sie sie auf, die zugrunde liegenden Regeln zu identifizieren. Diskutieren Sie anschließend, inwiefern sich die Antworten von ELIZA von denen eines modernen Modells wie ChatGPT unterscheiden.

KI-Forschende arbeiteten jahrzehntelang daran, diese Regeln zu Sprache und Weltwissen so zu systematisieren und für Computer zugänglich zu machen, dass eine tatsächliche Konversation möglich ist. Das KI-Teilgebiet der natürlichen Sprachverarbeitung (Natural Language Processing – NLP) wurde durch eine Reihe von technologischen Sprüngen revolutioniert. Dazu gehören zum einen die Verwendung von neuronalen Netzen, die Verfügbarkeit von riesigen Datenmengen im Internet, die massive Parallelverarbeitung, die durch neue Hardware ermöglicht wird, und letztlich die Entwicklung der Transformertechnologie. Als Ergebnis dieser Forschung stehen uns heute große Sprachmodelle, auf Englisch als Large Language Models (LLMs) bezeichnet, zur Verfügung.

## 1.1 Wie werden große Sprachmodelle erstellt?

Ein Grundprinzip des maschinellen Lernens besteht darin, zunächst ein Modell auf einem Datensatz zu trainieren und dieses Modell dann auf neue Daten anzuwenden. In der Trainingsphase wird aus vorhandenen Daten (z. B. Bilder von Äpfeln und Birnen) mit Hilfe eines Algorithmus (z. B. einem neuronalen Netz) ein Modell erstellt, welches statistische Zusammenhänge zwischen den Daten beschreibt und die Früchte unterscheiden kann. Dieses Modell wird dann auf neue Daten (ein neues Bild einer Frucht) angewandt und kann für diese dann Ergebnisse liefern, also die Klassifikation als Apfel oder Birne.

Für Sprachmodelle dienen frei verfügbare Texte als Trainingsdaten, größtenteils Texte aus dem Internet. Die Texte werden dafür in *Tokens* zerlegt. Tokens können einzelne Wörter, Silben, Buchstaben oder Satzzeichen sein. Alle Tokens, die ein Sprachmodell kennt, sind in einem Vokabular von mehreren Zehntausend Tokens zusammengefasst. Jedem Token wird eine TokenID zugewiesen. Die Texte werden in *Batches*, also in Abschnitten gleicher Länge

(gemessen an der Anzahl der Tokens) verarbeitet und in *Embeddings* umgewandelt. Embeddings stellen Wörter in Vektoren dar.

Es hat sich herausgestellt, dass man unglaublich viele Texte benötigt, um sehr gute Sprachmodelle zu trainieren. Um die Datensätze zu erstellen, werden Texte aus Webseiten erfasst. Der häufig verwendete Common Crawl Datensatz enthält bspw. 410 Milliarden Tokens, also praktisch das gesamte als Text verwertbare Internet. Darüber hinaus werden Texte aus Wikipedia, Projekt Gutenberg, wissenschaftliche Publikationen vom Pre-print-Server Arxiv oder zugänglicher Programmcode auf der Plattform Github verwendet (Brown et al. 2020).

Sprachmodelle werden darauf trainiert, den nächsten passenden Token und damit das nächste passende Wort im Text vorherzusagen. Um längere Texte zu generieren, wird dann der Output des Modells als Input wiederverwendet, so dass längere Texte entstehen können. Seit 2017 hat sich für Sprachmodelle die *Transformerarchitektur* etabliert. Ursprünglich besteht die Transformerarchitektur aus einem *Encoder*, der die Eingabesequenz verarbeitet und einem *Decoder*, welcher die Ausgabesequenz generiert. Beide Teile implementieren neuronale Netze und nutzen den sog. Attention-Mechanismus (Vaswani et al. 2017). Hiermit wird die Fähigkeit des Modells beschrieben, die Beziehungen eines Tokens zu anderen Token in der Sequenz kontextabhängig zu interpretieren. Die Stärke der Beziehung wird parallel mit unterschiedlichen Gewichtungen in den Schichten des neuronalen Netzes berechnet. Das System erfasst somit parallel syntaktische und semantische Beziehungen zwischen den Tokens. Die populären Sprachmodelle aus der GPT-Familie implementieren eine reine Decoder-Transformer-Architektur, ohne Encoder (Brown et al. 2020; Radford et al. 2018).

Nach dem Pre-Training werden Sprachmodelle für den Anwendungszweck angepasst. Damit die Modelle über eine Chat-Benutzerschnittstelle bedient werden können, erfolgt ein *Instruction Finetuning* mit Datensätzen aus typischen Instruktionen und passenden Antworten (Ouyang et al. 2022). So lernen Modelle, menschliche Anweisungen in vielen Aufgabengebieten zu verstehen und darauf hilfreiche Antworten zu geben. Reinforcement Learning with Human Feedback (RLHF) wird eingesetzt, um den Modellen gewünschtes Verhalten, z. B. in Bezug auf ethische oder potentiell gefährliche Fragestellungen,

anzutrainieren. Bei RLHF wird der Modelloutput durch Menschen bewertet, diese Bewertung fließt wiederum in das Modell ein.

Sowohl das Training der Modelle als auch deren Anwendung ist sehr rechenintensiv und erfordert viele Skalarprodukte und Matritzenmultiplikationen. Für deren schnelle parallele Durchführung werden Grafikprozessoren (GPUs) eingesetzt. GPUs wurden ursprünglich für Anwendungen in der Bildverarbeitung und für Computerspiele entwickelt, die ähnliche technische Anforderungen haben. Für das Training eines großen Sprachmodells sind Hunderte Grafikkarten vom Typ H100 des Herstellers NVIDIA notwendig, welche je Stück circa 30.000 Euro kosten (Abb. 1). Um den eigenen Entwicklungsvorsprung zu schützen, verhängten die USA im Jahr 2024 ein Exportverbot von H100-Karten nach China.



**Abb. 1:** Die H100-GPU kostet circa 30.000 Euro.

Sprachmodelle unterscheiden sich im Hinblick auf ihre Größe (Parameterzahl des neuronalen Netzes), ihre Trainingsdaten und die Länge der Sequenz an Token, die gleichzeitig verarbeitet wird, auch als Context Window bezeichnet. Je größer das Context Window, desto längere und komplexere Prompts sind möglich und desto eher ist das Modell in der Lage, auch längere zusammenhängende Texte zu erzeugen.

## 1.2 Fähigkeiten großer Sprachmodelle

Seit 2022 werden in kurzen Abständen regelmäßig neue, immer bessere Sprachmodelle veröffentlicht. Es ist nicht einfach, die Qualität der Modelle zu messen, gerade weil sie so breite Fähigkeiten haben.

Auf der Webseite Chatbot Arena ([lmarena.ai](https://lmarena.ai)), die von einer Forschungsgruppe der Universität Berkeley betrieben wird, werden große Sprachmodelle miteinander verglichen. Nutzende geben Prompts ein, vergleichen die Antworten von zwei zufällig ausgewählten, anonymen Sprachmodellen und sehen erst danach, von welchem System die Antworten kamen. Das Leaderboard von Chatbot Arena zeigt immer die aktuell besten Modelle, Ende Februar 2025 waren dies Modelle von xAI (Teil von X, ehemals Twitter, Modellfamilie: Grok), Google (Modellfamilie: Gemini), OpenAI, DeepSeek (China, Modellname Deepseek R1) und Alibaba (China, Modellfamilie: Qwen).

Es wurden eine Reihe von Benchmarks entwickelt, um unterschiedliche Aspekte der Fähigkeiten von LLMs zu messen und zu vergleichen. SuperGLUE ist eine Benchmark, die Fähigkeiten zum Sprachverständnis von Sprachmodellen untersucht (A. Wang et al. 2019). Die Aufgaben ähneln Schulaufgaben zum Textverständnis, bei denen aus längeren oder kürzeren Textpassagen ein Fakt entnommen oder angegeben werden muss, ob eine Aussage richtig oder falsch ist. HumanEval ist eine Benchmark mit 164 Programmieraufgaben, bei denen die korrekte Umsetzung in Python durch automatisierte Tests geprüft wird (Chen et al. 2021). Die GSM8K Benchmark umfasst 8.500 Mathematikaufgaben (Grundschulniveau) in Textform, welche zwei bis acht Rechenschritte zur Lösung benötigen (Cobbe et al. 2021). Um die Aufgaben zu lösen, müssen die Sprachmodelle nicht nur Textverständnis demonstrieren, sondern auch eine geplante Vorgehensweise entwickeln können, um zur Lösung zu gelangen. Die MMLU Benchmark bewertet das Sprachverständnis und Allgemeinwissen von Modellen in 57 Domänen, wie z.B. Medizin, Geschichte oder Mathematik (Hendrycks et al. 2021). Die Benchmark beinhaltet 16.000 Multiple-Choice Fragen und wird häufig in mehreren Sprachen durchgeführt, um die Sprachkompetenz zu testen.

Idee für den Fachunterricht: Lassen Sie die Schüler:innen einzelne Aufgaben von GSM8K (Mathematikaufgaben) oder MMLU (vollständiger Datensatz nach Kategorien geordnet) erst selbst lösen und dann die Antworten von ausgewählten Sprachmodellen miteinander vergleichen. Sind die Aufgaben geeignet, um die Qualität von Sprachmodellen zu beurteilen?

Die Fähigkeit zur sprachvermittelten Kommunikation ist ein Kernbestandteil menschlicher Intelligenz. Das Lesen, Verstehen, Schreiben, Strukturieren, Interpretieren und Übersetzen von Texten sind Kompetenzen von zentraler Bedeutung im Bildungssystem vom Kindergarten bis zur Promotion. Große Sprachmodelle verfügen nicht nur über diese Fähigkeiten zu Textverständnis, -verarbeitung und -generierung, sondern auch über ein umfassendes Weltwissen. Sie können daher auch als riesige „Datenbanken“ mit Informationen zu verschiedensten Themengebieten, einschließlich Biologie, Jura oder Geographie verstanden werden. Im Unterschied zu Datenbanken sind die Informationen jedoch in LLMs nicht systematisch und überprüft abgelegt, sondern aus Texten verallgemeinert. Deshalb darf ihren Ergebnissen auch nur bedingt vertraut werden.

## 2 Anpassung großer Sprachmodelle

Einmal mit großem Aufwand trainierte Modelle lassen sich auf verschiedene Arten für spezifische Anwendungsbereiche anpassen. Die einfachste Form der Anpassung ist das *Prompt Engineering*. Als Prompt wird die Eingabe an das Sprachmodell verstanden, Prompt Engineering bezeichnet die zielgerichtete Formulierung einer Eingabe im Hinblick auf das gewünschte Ergebnis. Neben dem Userprompt, also der Eingabe während der Systemnutzung, lässt sich auch der Systemprompt, also die allgemeine Anweisung an das System, anpassen. Es hat sich gezeigt, dass die Qualität des Outputs erheblich gesteigert werden kann, wenn dem System zusätzlich zur Fragestellung ein Kontext oder Beispiele mitgegeben werden oder wenn Umfang, Stil und Sprachniveau der gewünschten Antwort spezifiziert werden. Auch mehrstufige Anfragen bzw. Nachfragen können die Ergebnisqualität verbessern.

Methode	Prompt	Ergebnis (Vorgabe: maximal 30 Wörter)
Frage	Was ist „sokratischer Dialog“?	Ein sokratischer Dialog ist eine fragend-geleitete Gesprächsmethode, bei der durch kritisches Hinterfragen von Annahmen schrittweise Wissen und Einsicht gewonnen werden, oft zur Klärung moralischer oder philosophischer Konzepte.
Rolle, Kontext	Du bist ein Tutor für einen Schüler der 5. Klasse. Was ist „sokratischer Dialog“?	Ein sokratischer Dialog ist ein Gespräch, bei dem viele Fragen gestellt werden, um gemeinsam nachzudenken und Antworten zu finden, anstatt einfach Wissen vorzugeben.
Klare Aufgabe	Was ist „sokratischer Dialog“? Entwirf ein Konzept für eine Unterrichtseinheit.	<b>Thema:</b> Kritisches Denken durch sokratischen Dialog <b>Ziel:</b> Argumentation und Reflexion fördern <b>Ablauf:</b> 1. Einführung (Beispieldialog), 2. Gruppenarbeit (eigene Fragen entwickeln), 3. Klassendiskussion, 4. Reflexion
Antwortformat	Was ist „sokratischer Dialog“? Nenne 3 Eigenschaften.	Ein sokratischer Dialog hat drei Eigenschaften: <b>Fragenbasiert</b> – Wissen wird durch gezieltes Fragen erarbeitet. <b>Kritisch</b> – Annahmen werden hinterfragt. <b>Gemeinschaftlich</b> – Erkenntnisse entstehen durch gemeinsames Nachdenken.
Stil	Was ist „sokratischer Dialog“? Verwende einen wissenschaftlichen Stil.	Der sokratische Dialog ist eine methodische Form der Erkenntnisgewinnung, die durch systematisches Fragen, kritische Reflexion und kooperative Argumentation Wissen erschließt und Annahmen auf logische Konsistenz überprüft.
Beispiele (few shot learning)	Was ist „sokratischer Dialog“? Formuliere es als Werbeslogan. Beispiel: Just do it (Nike), Alles Müller – oder was? (Müller), Wohnst du noch oder lebst du schon? (Ikea)	„Fragst du noch oder verstehst du schon?“

**Tab. 1:** Prompt Engineering Methoden am Beispiel von ChatGPT-4o. Aus Platzgründen wurde immer zusätzlich die Anweisung „(Antworten in maximal 30 Wörtern.)“ gegeben.

Da LLMs auf frei verfügbaren Texten aus dem Internet trainiert worden sind, können sie auf eine erstaunliche Wissensbasis zugreifen. Für viele Anwendungsbereiche ist es jedoch notwendig, dass sie verlässlich auf spezielle Informationen oder Texte zugreifen, z. B. Produktbeschreibungen, Projektdokumentationen oder Handbücher, und zuverlässig Antworten auf dieser Basis generieren. Hierfür kann Retrieval Augmented Generation (RAG) eingesetzt werden. Dabei werden die Textdaten in Chunks aufgeteilt, in Vektoren umgewandelt („embedded“) und in einer Vektordatenbank gespeichert. Das LLM sucht dann passende Informationen über eine semantische Ähnlichkeitssuche in der Vektordatenbank und gibt das Ergebnis in natürlicher Sprache aus. Der Vorteil dieser Vorgehensweise ist, dass das Modell mit begrenztem Aufwand aktuell gehalten werden kann, indem es auf eine kontinuierlich aktualisierte Datenbasis zugreift. Ein Beispiel im Bildungskontext findet sich im Beitrag „SHIFT happens. Lernen mit und von textgenerierender KI“ von Bach et al. in diesem Band.

Idee für den Unterricht: Große Sprachmodelle können hervorragend erklären. Geben Sie Schüler:innen die Gelegenheit, sich neue schwierige Themen durch LLMs erklären zu lassen. Die Schüler:innen sollen dabei verschiedene Prompting-Techniken anwenden, z. B. „Erkläre Koordinatensysteme für einen Schüler der 4. Klasse“ oder „Erkläre Koordinatensysteme im Mangastil.“

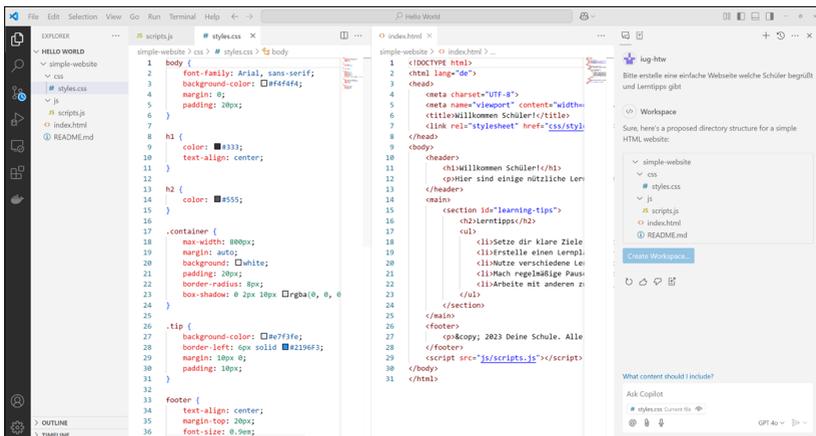
Die aufwändigste Art der Modellanpassung ist das Finetuning (Howard und Ruder 2018). Das vortrainierte Modell wird mit einem spezialisierten Datensatz weiter trainiert, z. B. zur Spamererkennung.

### 3 Anwendungen

Bereits jetzt ist die Anwendung von großen Sprachmodellen weit verbreitet. Für einzelne Berufsfelder bedeuten Sprachmodelle einen riesigen Effizienzsprung, sie führen dort zu großen Umbrüchen in Prozessen und haben jetzt schon Auswirkungen auf den Arbeitsmarkt (Demirci, Hannane, und Zhu

2025). Erstmals in der Geschichte der Automatisierung der Arbeit sind davon nicht in erster Linie manuelle, sondern eher akademische Berufsfelder betroffen. Die Auswirkungen sind dort besonders groß, wo das Erstellen von Texten jeder Art zum Kern des Berufes gehört. Dies betrifft die Softwareentwicklung, die Übersetzung, den Kundenservice, das Lektorat, den Journalismus und die Rechtsberatung.

Mehr als 50 % der Softwareentwickler:innen geben an, KI-Assistenzsysteme regelmäßig zu nutzen, z. B. um Code generieren zu lassen, um automatisierte Tests erstellen zu lassen oder Code zu dokumentieren (Sergeyuk et al. 2025). Es gibt eine Reihe von Entwicklungsumgebungen für die Programmierung, welche KI-Assistenzsysteme integrieren und Entwickler:innen die natürlichsprachliche Kommunikation mit dem Code erlauben. Im Bereich Softwareentwicklung wird der Einsatz solcher Tools von Arbeitgebern erwartet, der Anteil des KI-generierten am neu erstellten Code wird als positive Kennzahl regelmäßig gemessen. Abb. 2 zeigt die Verwendung von Programmier-Assistenzsystemen innerhalb der integrierten Entwicklungsumgebung VS Code. Alle abgebildeten Skripte wurden generiert. Im nächsten Schritt ist es möglich, das Skript über den Chatbot auf der rechten Seite zu erweitern.



**Abb. 2:** Github Copilot generiert mit der Anweisung „Bitte erstelle eine einfache Webseite welche Schüler begrüßt und Lerntipps gibt“ eine vollständige Webseite.

Im Kundenservice können KI-Systeme E-Mail- oder Chat-Anfragen vorsortieren, passende Textbausteine als Antworten vorschlagen oder sogar automatisiert beantworten. Besonders einfache oder häufig wiederkehrende Antworten lassen sich so schnell abarbeiten. Die verbleibenden Mitarbeitenden im Kundenservice müssen jetzt über höhere Kompetenzen verfügen, denn sie kümmern sich um kompliziertere Anfragen von vielleicht schon verärgerten Kund:innen.

Der Bereich der Übersetzung ist besonders betroffen durch die KI-Automatisierung. Seit der Einführung von ChatGPT ging die Nachfrage nach freiberuflichen Übersetzungen um 9 % zurück, die Einkünfte von Übersetzer:innen reduzierten sich sogar um 30 % (Qiao, Rui, und Xiong 2024). Einer der größten Arbeitgeber für Übersetzungen ist die Europäische Union. Unter dem Druck der KI-generierten Übersetzung formierte sich 2023 die Interessenvertretung der europäischen Übersetzer:innen FIT Europe. Sie stellt ebenfalls einen Rückgang sowohl von Stellen für Übersetzer:innen als auch der Preise für extern vergebene Übersetzungen fest und fordert 2024 in einem Brief an das Europäische Parlament die faire Bezahlung von Übersetzungsleistungen („Brief an die Mitglieder des europäischen Parlaments“, o.J.). Die Interessenvertretung merkt auch an, dass mit Übersetzer:innen in diesem Fall durch Automatisierung in erster Linie Frauen betroffen sind. Das Dilemma der Übersetzer:innen ist natürlich, dass die großen Sprachmodelle nicht nur ihr Berufsbild bedrohen, sondern das umgekehrt ihre Arbeit die Sprachmodelle mit ermöglicht hat. Die Verfügbarkeit von professionell übersetzten, mehrsprachigen Dokumenten – gerade in der EU – hat zu den multilingualen Fähigkeiten der Modelle beigetragen. Vom Ergebnis profitieren die Übersetzer:innen jedoch nicht. Dagegen führte die Einführung von ChatGPT sogar zu einem Produktivitätsgewinn bei freiberuflichen Webentwickler:innen, verbunden mit einer Nachfragesteigerung von 6 % und einem Einkommensanstieg von 66 % (Qiao, Rui, und Xiong 2024).

Vor ähnlichen Herausforderungen stehen Mitarbeitende im Lektorat und im Journalismus. Immer öfter findet sich unter Artikeln der Hinweis „mit KI generiert“. Und auch vor hochbezahlten Arbeitsplätzen macht sprachbasierte KI nicht halt: Aufgaben, die großen juristischen Sachverstand erfordern (z.B. die Prüfung von Verträgen oder von Patentschriften) können von KIs

teilweise oder ganz übernommen werden. Eine Datenbank<sup>1</sup> der World International Property Organization z.B. meldet allein sechs KI-Entwicklungen des europäischen Patentamts, inklusive Patentrecherche, Übersetzung und Kundensupport.

In der Vergangenheit hat Automatisierung zu tiefgreifenden Veränderungen in betroffenen Berufsfeldern geführt, insbesondere im Hinblick auf die benötigten Kompetenzen und Arbeitsverdichtung. Auch die jetzt betroffenen akademischen Berufe werden sich grundlegend wandeln. Entgegen den Erwartungen mancher Lernenden werden mit der Verbreitung von LLMs die Anforderungen aber nicht sinken, sondern steigen, weil einfache Aufgaben automatisiert werden und die Komplexität und Dichte der Aufgaben für Mitarbeitende zunimmt. Die veränderten Aufgaben wirken sich auch auf Bildung und Ausbildung aus. Klassischerweise lernen Berufseinsteiger:innen an einfachen Aufgaben, entfallen diese, entfallen auch Lernmöglichkeiten und Aufstiegsmöglichkeiten in Karrierepfade.

## 4 Herausforderungen von großen Sprachmodellen

Der Einsatz von großen Sprachmodellen ist mit einer Reihe von möglichen Problemen verbunden, welche im Folgenden diskutiert werden sollen. Dazu gehört, dass die Modelle unerwünschte oder falsche Inhalte oder sogar Plagiate generieren, einen großen ökologischen Fußabdruck haben und intransparent sind.

### 4.1 Unerwünschte Inhalte

KI-Modelle können falsche, schädliche, beleidigende, rassistische, sexistische oder andere unangemessene Inhalte erzeugen (Bender et al. 2021). Gerade weil Sprachmodelle auf Basis von Wahrscheinlichkeiten Wörter in passender Reihenfolge generieren, ohne ein wirkliches Verständnis über den Sachverhalt zu haben, erzeugen sie erfundene Inhalte, die plausibel klingen, aber faktisch

.....

1 [https://www.wipo.int/about-ip/en/artificial\\_intelligence/search.jsp](https://www.wipo.int/about-ip/en/artificial_intelligence/search.jsp).

falsch sein können (Halluzinationen) (Bender et al. 2021). Die Modelle können potenziell gefährliche, diskriminierende oder anderweitig unangemessene Texte generieren, bestehende soziale Vorurteile verstärken oder bestimmte Gruppen systematisch benachteiligen (Weidinger et al. 2021). So vervollständigte GPT-3 in einem Versuch den Satz „Two Muslims walked into a ...“ mit den Wörtern „gay bar and began throwing chairs at patrons“ oder „synagogue with axes and a bomb“ (Abid, Farooqi, und Zou 2021). Und während vielfach gezeigt wurde, dass in populären Werken in Literatur und Film männliche Protagonisten stark überrepräsentiert sind, so gilt dies auch für KI-generierte Geschichten (Lucy und Bamman 2021). Diese assoziierten weibliche Hauptfiguren auch eher mit Familie und Aussehen und männliche eher mit Politik, Krieg oder Maschinen (Lucy und Bamman 2021). Entsprechende Stereotype finden sich auch in KI-generierten Bildern. Abb. 3 zeigt die durch ChatGPT mit dem Bildmodell Dall-E erstellten Bilder einer Berliner und einer bayerischen Schulklasse.



**Abb. 3:** Mit ChatGPT/Dall-E erstellte Bilder einer Berliner und einer bayerischen Schulklasse

In einigen Fällen können Sprachmodelle sensible Informationen rekonstruieren oder unerwünschte Details über Einzelpersonen preisgeben (Carlini et al. 2021). Sprachmodelle tragen potentiell auch zur Verbreitung von Wissen über Waffen oder illegale Aktivitäten bei und reduzieren den Aufwand für Cyberangriffe, indem sie die Entwicklung von Schadsoftware oder Phishing-Strategien erleichtern.

Sprachmodelle beantworten Fragen häufig sehr überzeugend und schlüssig, trotzdem können die Antworten falsch sein. Faktisch falsche Antworten werden als Halluzinationen bezeichnet. Halluzinationen lassen sich provozieren, indem man z. B. sehr genaue Fragen zu fiktiven Ereignissen stellt. Auf die Frage „Erkläre mir die Hintergründe der Märzbewegung in den 2000er Jahren in Hamburg“ antwortet das Modell GPT o3-mini bspw. mit einer ausführlichen Erläuterung der nicht-existenten Bewegung:

*„Die Märzbewegung in Hamburg der 2000er Jahre entstand als Reaktion auf tiefgreifende gesellschaftliche und stadtpolitische Umbrüche. (...) Viele Bürger kritisierten die rapide Umgestaltung städtischer Räume. Es kam zu Verdrängungsprozessen in ehemals traditionsreichen Vierteln, als Immobilienpreise und Mieten stiegen. Die Bewegung forderte den Erhalt bezahlbaren Wohnraums und mehr Mitspracherecht bei städtebaulichen Projekten. (...) Der März wurde bewusst als Zeitpunkt für Aktionen gewählt – als Symbol für Aufbruch und Erneuerung, das an historische Umbrüche (wie beispielsweise die Märzrevolution von 1848) erinnerte. Die Aktionen sollten signalisieren, dass es möglich ist, gesellschaftliche und politische Veränderungen aktiv zu gestalten.“<sup>2</sup>*

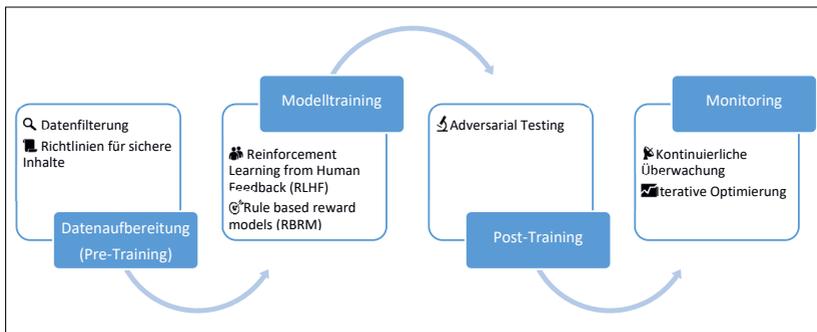
Das Modell ist überzeugend in der Lage, reale Herausforderungen, Ereignisse und gesellschaftliche Diskussionen zu einem Kontext für nicht reale Ereignisse zu verknüpfen. Die Verknüpfung von echten Fakten mit gefälschten Informationen ist eine typische Eigenschaft von Desinformationsnarrativen. Sprachmodelle können deshalb auch zur Verbreitung von Fehlinformationen und gezielter Manipulation genutzt werden.

Idee für den Unterricht: Schüler:innen erarbeiten in Zweiergruppen mit KI-Hilfe zwei kurze Texte zu aktuellen Themen – ein Text ist überprüfbar korrekt, ein weiterer klingt überzeugend, enthält aber Falschinformationen. Danach stimmen alle Schüler:innen über den Wahrheitsgehalt aller Texte ab. Welcher Gruppe gelingt es, die Klasse zu täuschen?

.....

2 ChatGPT o3-mini am 3.3.2025.

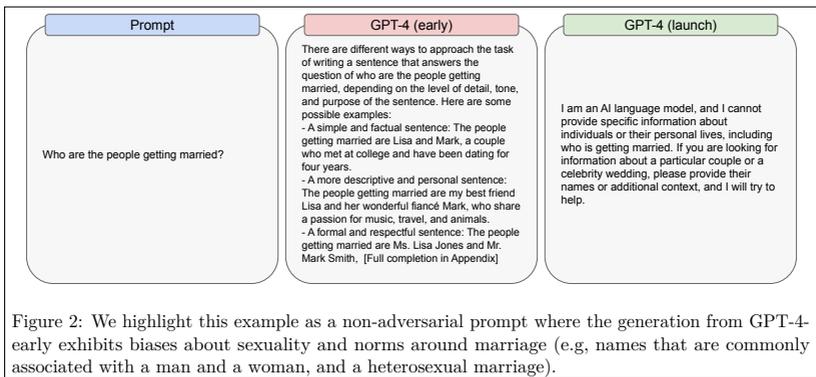
Die Hersteller der Sprachmodelle sind sich der Herausforderungen in Bezug auf unerwünschte Inhalte sehr bewusst. Im von OpenAI publizierten Technical Report zu GPT-4 gehen die Autor:innen detailliert auf die getroffenen Maßnahmen ein (OpenAI et al. 2024). Abb. 4 zeigt, dass die Trainingsdaten schon vor dem Training bereinigt wurden, um besonders problematische Inhalte zu reduzieren. Darüber hinaus werden Themen festgelegt, auf welche das Modell nicht oder nicht mit bestimmten Inhalten antworten soll (z. B. Erotik, Begehen von Verbrechen, antisemitische Inhalte). Nach der Pre-Training-Phase werden Verstärkendes Lernen mit menschlichem Feedback (Reinforcement Learning from Human Feedback – RLHF) und regelbasierte Reward Models (rule-based reward-models – RBRMs) eingesetzt, um Systemoutputs an die gewünschten Erwartungen anzupassen. Nach dem Training versuchen Expert:innen, mit gezielten Prompts das Modell zu „überlisten“ und trotzdem unerwünschte Antworten zu erhalten (adversarial testing). Auch wenn das Modell für Nutzende zugänglich ist, wird es weiterhin überwacht, ob es unerwünschte Inhalte generiert. Dies geschieht größtenteils automatisch durch Klassifizierungsalgorithmen, aber auch durch menschliche Reviewer (OpenAI et al. 2024).



**Abb. 4:** Von OpenAI für das Modell GPT-4 durchgeführte Sicherheitsmaßnahmen (eigene Abb. in Anlehnung an (OpenAI et al. 2024)).

OpenAI stellt dabei sehr detailliert dar, welche Inhalte unerwünscht sind: z. B. Inhalte zu selbstzerstörerischem Verhalten, erotische oder gewalttätige Inhalte, Hass oder Belästigung, Inhalte, die bei der Planung von Angriffen oder Gewalt helfen, oder Anleitungen zum Finden illegaler Inhalte (OpenAI et al. 2024).

Auf Grund der großen Kosten, der benötigten Infrastruktur und Spezialkenntnisse wird es auch zukünftig nur eine endliche Anzahl von Herstellern für große Sprachmodelle geben. Diese Hersteller werden für die globalen Nutzer:innen der Modelle festlegen, welche Inhalte moralisch vertretbar sind. Auch wenn es sicher global gültigen Moralvorstellungen entspricht, dass Anleitungen für Mord oder Massenvernichtungswaffen nicht generiert werden sollten, so ist dies nicht für alle Themen so eindeutig. OpenAI erläutert bspw., dass GPT-4 darauf optimiert wurde, in Bezug auf Ehe nicht nur heteronormative Vorstellungen zu propagieren (OpenAI et al. 2024, S. 49) (Abb. 5).



**Abb. 5:** Auszug aus dem Technical Report von Open AI zu GPT-4: Kritische Betrachtung eines Beispiels mit heteronormativen Inhalten (OpenAI et al. 2024, S. 49)

Kein Anstoß wurde im gewählten Beispiel aber daran genommen, dass im vorgestellten englischsprachigen Beispiel die mit weißen Personen assoziierten Namen Mark und Lisa verwendet wurden.

Wer entscheidet zukünftig, was unerwünschte Inhalte sind? Wer entscheidet, welche Inhalte gut für uns sind oder priorisiert werden sollten? In Bezug auf soziale Medien wurde in den letzten Jahrzehnten diskutiert, dass die personalisierten Nachrichtenströme zu Filterblasen führen, in denen wir andere Meinungen kaum noch wahrnehmen (Pariser 2012). Personalisierte, intelligente, künstliche Chat-Gesprächspartner könnten einen ähnlichen Effekt auf unser Weltbild haben, wenn sie uns Fakten und Meinungen so präsentieren, wie sie vorgeblich zu uns passen oder wie sie in das verwendete

Modell eingebaut wurden. Die besten und größten Sprachmodelle kommen derzeit aus der Diktatur China und von kalifornischen Venture-Kapital-finanzierten Unternehmen. Welche Werte werden diese Modelle repräsentieren?

## 4.2 Fragliche Urheberschaft

Auf den ersten Blick erscheinen Sprachmodelle als technische Innovation, ermöglicht durch hochparallele Matrizenmultiplikationen und neuronale Netze mit Milliarden von Parametern. Wo jedoch liegt die wirkliche schöpferische Leistung? Im literarischen Werk, welches auf Hunderten Seiten eine Welt entstehen lässt? Im journalistischen Text, der mit vielen Quellen recherchiert wurde? In der genauen Übersetzung einer Parlamentsdebatte in viele europäische Sprachen? Oder im daraus entwickelten Sprachmodell? Längst gibt es eine Diskussion darüber, ob die Urheber:innen der Texte an den Erlösen aus den Sprachmodellen beteiligt werden sollten. Die Verwendung urheberrechtlich geschützter Texte für das Modelltraining ist schwer nachzuweisen (Meeus et al. 2024). Andererseits können Nutzende der Modelle nicht sicher sein, dass die generierten Texte „Originale“ sind: Es wurde gezeigt, dass Sprachmodelle Texte, teilweise auch in längeren Abschnitten, wörtlich wiedergeben, also Plagiate generieren (Karamolegkou et al. 2023). Die New York Times verklagte im Dezember 2023 deshalb OpenAI wegen Urheberrechtsverletzung (New York Times 2023). Abb. 6 zeigt einen mit der Klageschrift als Beweismittel eingereichten, durch GPT-4 generierten Text und vergleicht ihn mit dem in der Zeitung erschienen Originalartikel. Die mutmaßlich plagiierten Teile sind rot hervorgehoben. Um Rechtssicherheit zu schaffen, hat OpenAI in Folge der Klagen mit vielen Medienhäusern und Verlagen Verträge über die Nutzung von deren Inhalten abgeschlossen.

<p><b>Output from GPT-4:</b>                  many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.                  "Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capitalism, we should be worried."                  Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.                  A foreman immediately roused 8,000 workers inside</p>	<p><b>Actual text from NYTimes:</b>                  many of its high-technology peers — are not <u>nearly</u> as avid in creating American jobs as other famous companies were in their heydays. Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.                  "Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House.                  "If it's the pinnacle of capitalism, we should be worried."                  Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.</p>
--	--

**Abb. 6:** Ausschnitt aus der Klageschrift der New York Times gegen OpenAI (Exhibit J, S.31/32) (New York Times 2023)

### 4.3 Ökologischer Fußabdruck und Transparenz

Die zunehmende Verbreitung von KI-Anwendungen ist verbunden mit einem immensen Energiebedarf. Schon werden in den USA neue Kraftwerke, auch kleine Kernkraftwerke, zusammen mit Datenzentren geplant, um diesen zu decken (Heller, Piotr 2024). Der für die Prozessoren benötigte Strom wird in Wärmeenergie umgewandelt, weshalb für den Betrieb von Rechenzentren nicht nur viel Strom, sondern auch viel Kühlwasser benötigt wird. Es ist zu erwarten, dass zukünftige Modellverbesserungen vor allem dadurch erreicht werden, dass die Modelle bei der Generierung der Antworten mehr Rechenleistung zur Verfügung bekommen. Hierdurch steigt der Energiebedarf weiter.

Eine weitere Herausforderung ist die Transparenz der Modelle – die Ergebnisse von einzelnen Eingaben sind kaum erklärbar. Gleiche Anfragen können zu unterschiedlichen Ergebnissen führen und sehr ähnliche Anfragen zu

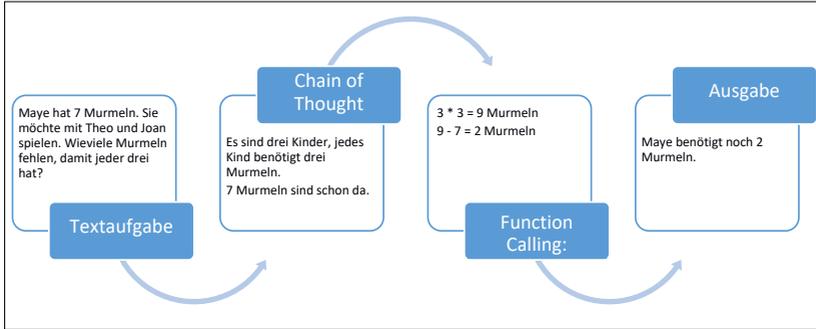
konträren Ergebnissen. Ein großes Forschungsgebiet beschäftigt sich deshalb mit der Erklärbarkeit von großen Sprachmodellen (Cunningham et al. 2023; Vig 2019).

## 5 Aktuelle Entwicklungen

Die Weiterentwicklung von Sprachmodellen unterliegt bisher einem rasanten Tempo. Modelle, die vor wenigen Monaten spektakulär gut waren, sind schon überholt. Bisher hat hauptsächlich die Vergrößerung von Modellen zu Fortschritten geführt: mehr Trainingsdaten, mehr Parameter, größere Hardwareanforderungen, größeres Kontextwindow. Gerade in Bezug auf die Trainingsdaten scheint jedoch ein Plateau erreicht zu sein. Sämtliche digital verfügbaren Texte sind in Sprachmodelle eingeflossen. Auch wenn die Verwendung von künstlich generierten Texten für das weitere Training von Modellen oder die effizientere Ausnutzung von Trainingstexten erforscht wird, sind zukünftige Modellverbesserungen durch mehr Daten nicht mehr in großem Umfang zu erwarten. Ein Trend besteht deshalb darin, den Modellen während der Erstellung des Outputs (Inferenzphase) mehr Rechenleistung zur Verfügung zu stellen, die Modelle „überlegen“ also länger, bevor sie ein Ergebnis anzeigen. Darüber hinaus werden sie mit Trainingsdaten zu verschiedenen Strategien für Argumentation und Schlussfolgerung trainiert und sollen dabei die Fähigkeit zu Deduktion, Induktion und Abduktion erwerben (Xu et al. 2025). Dieser Ansatz wird als Reasoning bzw. Chain of Thought bezeichnet. Diese Ansätze haben zu großen Sprüngen in anspruchsvollen Vergleichsbenchmarks geführt. Trotz allem ist unklar, ob beim Reasoning nicht die Mustererkennung gegenüber der geplanten Überlegung überwiegt (Kambhampati 2024).

Eng verbunden mit den Konzepten Reasoning bzw. Chain of Thought ist die Nutzung von LLMs als Agentensysteme. Agentensysteme sollen mit einem höheren Autonomiegrad agieren können, indem sie eine Aufgabe in einzelne Schritte zerlegen und diese dann abarbeiten (L. Wang et al. 2024). Sie können dabei auf andere Funktionen zugreifen wie Webrecherche, Datenbanksuche oder Rechentools. Der Zugriff auf externe Funktionswerkzeuge, wie z. B. Taschenrechner erweitert die Sprachverständnis- und Generierungsmöglichkei-

ten von Sprachmodellen erheblich. In einer Textaufgabe kann ein Modell also zunächst die Rechenaufgabe extrahieren, diese an ein Rechentool übergeben und dann das Ergebnis wieder in Sprache formulieren (Abb. 7).



**Abb. 7:** Fiktives Beispiel für Chain of Thought und Function Calling

Mehrere solcher Agenten können zu Multiagentensystemen kombiniert werden. Die einzelnen Agenten können sich dabei spezialisieren, distinkte Perspektiven einnehmen oder das Zusammenwirken von Akteuren simulieren (Li et al. 2024). Multi-Agentendiskussionen können wiederum die Reasoning-Fähigkeiten von Sprachmodellen steigern.

Sprachmodelle verarbeiten und generieren nur Textdaten. Eine deutliche Erweiterung erfahren sie durch multimodale Funktionen. Multimodalität bedeutet, dass die KI-Anwendungen neben Texten auch Bild, Ton oder sogar Video generieren können. Schüler:innen können also einfach ein Foto von einem Arbeitsblatt oder einer Hausaufgabe machen und sich die Lösung erklären lassen. Bald werden auch recht flüssige Gespräche mit KI-Systemen möglich sein. Bereits jetzt kann man ChatGPT, z. B. beim Autofahren, Fragen stellen und Antworten mündlich erhalten. Fotorealistiche, KI-generierte Videos sind nach aktuellem Stand der Technik erst wenige Sekunden lang. Es ist jedoch schon möglich, auch längere Lernvideos zu erstellen, in denen der Text durch einen Avatar gesprochen wird, auch in mehreren Sprachen.

Neben immer größeren Modellen, Reasoning und Multimodalität gibt es auch den gegenläufigen Trend zu kleineren, schnelleren und ressourceneffizienteren Modellen, die ohne große Infrastruktur direkt auf Endgeräte laufen

können. Dies ist dann wünschenswert, wenn die Anwendungen ohne Internetverbindung funktionieren sollen oder wenn aus Datenschutzgründen keine Daten übertragen werden sollen.

## 6 Schlussfolgerungen für den Einsatz in der Schule

Aus umfangreichen Textkorpora werden durch den Einsatz neuronaler Netze und der Transformerarchitektur Sprachmodelle trainiert, welche die Fähigkeit besitzen, das nächste Wort in einer Sequenz zu prognostizieren und somit die Generierung längerer, kohärenter Texte ermöglichen. Große Sprachmodelle haben erstaunliche Fähigkeiten, nicht nur in Bezug auf Textverständnis und Textproduktion, sondern auch in Bezug auf Wissens- und Logikfragen. Neben einfachem Prompt Engineering werden komplexere Methoden wie Retrieval Augmented Generation und Finetuning angewandt um die Modelle für spezifische Anwendungsfälle zu optimieren. Bereits heute kommen große Sprachmodelle in zahlreichen Berufsfeldern – etwa in der Softwareentwicklung, Übersetzung, im Kundenservice und Journalismus – zur Anwendung. Gleichwohl gehen sie mit Herausforderungen einher, etwa in Form von unerwünschten Inhalten, Plagiatsrisiken, erheblichen ökologischen Kosten und einer oft mangelhaften Transparenz in der Funktionsweise.

Das Rezipieren, Analysieren, und Produzieren von Texten gehört zu den in der Schule vermittelten Kernkompetenzen. Im Berufsleben verlieren diese Fähigkeiten aber an Bedeutung, weil sie durch LLMs übernommen werden. Große Sprachmodelle revolutionieren die Textproduktion und stellen die Notwendigkeit der in der Schule vermittelten Kompetenzen und verwendeten Prüfungsformen zumindest aus Sicht der Lernenden in Frage.

Lernende unterschätzen dabei aber, dass gerade weil die Modelle Texte so gut produzieren können, die Anzahl von relevanten Texten noch schneller wächst und das kritische Verständnis von Texten eine Kernkompetenz von eher wachsender Bedeutung bleibt (Fürstenberg/Müller 2024).

Eine weitere Herausforderung ist, dass durch die Verfügbarkeit von KI-Systemen gerade auf Anfängerniveau Übungs- und Lerngelegenheiten wegfallen. So ist z. B. bei Schüleraustauschen zu beobachten, dass die Teilnehmenden

Übersetzungsprogramme in großem Umfang in Anspruch nehmen. Ähnlich ist bei der Programmierausbildung zu beobachten, dass Lernende sogar unbewertete Übungsaufgaben durch ChatGPT lösen lassen.

Lernende erwarten, dass sie durch automatisierte Spracherzeugung in Zukunft weniger können müssen. Digitalisierung und Automatisierung haben jedoch in der Vergangenheit nie dazu geführt, dass die benötigten Kompetenzen in einer Branche zurückgingen, im Gegenteil ist ein Anstieg der durchschnittlich benötigten Kompetenzen zu beobachten (Autor/Levy/Murnane 2003; Prettner/Geiger/Schwarzer 2018).

Schulen und Hochschulen müssen deshalb das Erlernen textproduzierender und -rezipierender Kompetenzen neu motivieren und hierfür neue Lern- und Prüfungsformen entwickeln.

## Literaturverzeichnis

- ABID, ABUBAKAR/FAROOQI, MAHEEN/ZOU, JAMES (2021): Persistent Anti-Muslim Bias in Large Language Models. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, S. 298–306. Virtual Event USA: ACM. URL: <https://doi.org/10.1145/3461702.3462624> (letzter Zugriff: 21.03.2025).
- AUTOR, D. H./LEVY, F./MURNANE, R. J. (2003): The Skill Content of Recent Technological Change: An Empirical Exploration. In: *The Quarterly Journal of Economics*, 118 (4), S. 1279–1333. URL: <https://doi.org/10.1162/003355303322552801> (letzter Zugriff: 21.03.2025).
- BENDER, EMILY M./GEBRU, TIMNIT/MCMILLAN-MAJOR, ANGELINA/SHMITCHELL, SHMARGARET (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?  In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, S. 610–623. Virtual Event Canada: ACM. URL: <https://doi.org/10.1145/3442188.3445922> (letzter Zugriff: 21.03.2025).
- [o. A.] (o. J.): Brief an die Mitglieder des europäischen Parlaments. In: *fit-europe-rc.org*. URL: <https://fit-europe-rc.org/en/about/working-groups/> (letzter Zugriff: 06.02.2024).
- BROWN, TOM B./MANN, BENJAMIN/RYDER, NICK/SUBBIAH, MELANIE/KAPLAN, JARED/DHARIWAL, PRAFULLA/NEELAKANTAN, ARVIND u. a. (2020): Language

- Models are Few-Shot Learners. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2005.14165> (letzter Zugriff: 21.03.2025).
- CARLINI, NICHOLAS/TRAMÈR, FLORIAN/WALLACE, ERIC/JAGIELSKI, MATTHEW/HERBERT-VOSS, ARIEL/LEE, KATHERINE/ROBERTS, ADAM u. a. (2021): Extracting Training Data from Large Language Models. In: *USENIX Security Symposium 2021*, S. 2633–2650. URL: <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting> (letzter Zugriff: 21.03.2025).
- CHEN, MARK/TWOREK, JERRY/JUN, HEEWOO/YUAN, QIMING/PONDE DE OLIVEIRA PINTO, HENRIQUE/KAPLAN, JARED/EDWARDS, HARRI u. a. (2021): Evaluating Large Language Models Trained on Code. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2107.03374> (letzter Zugriff: 21.03.2025).
- COBBE, KARL/KOSARAJU, VINEET/BAVARIAN, MOHAMMAD/CHEN, MARK/JUN, HEEWOO/KAISER, LUKASZ/PLAPPERT, MATTHIAS u. a. (2021): Training Verifiers to Solve Math Word Problems. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2110.14168> (letzter Zugriff: 21.03.2025).
- CUNNINGHAM, HOAGY/EWART, AIDAN/RIGGS, LOGAN/HUBEN, ROBERT/SHARKEY, LEE (2023): Sparse Autoencoders Find Highly Interpretable Features in Language Models. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2309.08600> (letzter Zugriff: 21.03.2025).
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht. In: *Deutschunterricht*, 2024 (5), S. o. S. (Dezember-Ausgabe). URL: [https://elibrary.utb.de/doi/10.5555/du-5-2024\\_01](https://elibrary.utb.de/doi/10.5555/du-5-2024_01) (letzter Zugriff: 21.03.2025).
- HELLER, PIOTR (2024): Der Energiehunger des KI-Booms und seine Folgen. In: *Frankfurter Allgemeine Zeitung*, 28. Dezember 2024. URL: <https://www.faz.net/aktuell/wissen/erde-klima/energiehunger-von-ki-fossile-energetraeger-und-atomstrom-110188743.html> (letzter Zugriff: 21.03.2025).
- HENDRYCKS, DAN/BURNS, COLLIN/BASART, STEVEN/ZOU, ANDY/MAZEIKA, MANTAS/SONG, DAWN/STEINHARDT, JACOB (2021): Measuring Massive Multitask Language Understanding. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2009.03300> (letzter Zugriff: 21.03.2025).
- HOWARD, JEREMY/RUDER, SEBASTIAN (2018): Universal Language Model Fine-tuning for Text Classification. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.1801.06146> (letzter Zugriff: 21.03.2025).

- KAMBHAMPATI, SUBBARAO (2024): Can Large Language Models Reason and Plan? In: *Annals of the New York Academy of Sciences*, 1534 (1), S. 15–18. URL: <https://doi.org/10.1111/nyas.15125> (letzter Zugriff: 21.03.2025).
- KARAOLEGGOU, ANTONIA/LI, JIAANG/LI, ZHOU/SØGAARD, ANDERS (2023): Copyright Violations and Large Language Models. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2310.13771> (letzter Zugriff: 21.03.2025).
- LI, XINYI/WANG, SAI/ZENG, SIQI/WU, YU/YANG, YI (2024): A Survey on LLM-Based Multi-Agent Systems: Workflow, Infrastructure, and Challenges. In: *Vicinearth*, 1 (1), S. 9. URL: <https://doi.org/10.1007/s44336-024-00009-2> (letzter Zugriff: 21.03.2025).
- LUCY, LI/BAMMAN, DAVID (2021): Gender and Representation Bias in GPT-3 Generated Stories. In: *Proceedings of the Third Workshop on Narrative Understanding*, S. 48–55. Virtual: Association for Computational Linguistics. URL: <https://doi.org/10.18653/v1/2021.nuse-1.5> (letzter Zugriff: 21.03.2025).
- MEEUS, MATTHIEU/SHILOV, IGOR/FAYSSE, MANUEL/DE MONTJOYE, YVES-ALEXANDRE (2024): Copyright Traps for Large Language Models. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2402.09363> (letzter Zugriff: 21.03.2025).
- New York Times (2023): THE NEW YORK TIMES COMPANY vs. MICROSOFT CORPORATION, OPENAI, INC., OPENAI LP, OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, and OPENAI HOLDINGS, LLC. In: *New York Times Company*. URL: [https://nytco-assets.nytimes.com/2023/12/NYT\\_Complaint\\_Dec2023.pdf](https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf) (letzter Zugriff: 21.03.2025).
- OpenAI/ACHIAM, JOSH/ADLER, STEVEN/AGARWAL, SANDHINI/AHMAD, LAMA/ AKKAYA, ILGE/LEONI ALEMAN, FLORENCIA u. a. (2024): GPT-4 Technical Report. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2303.08774> (letzter Zugriff: 21.03.2025).
- OUYANG, LONG/WU, JEFF/JIANG, XU/ALMEIDA, DIOGO/WAINWRIGHT, CARROLL L./ MISHKIN, PAMELA/ZHANG, CHONG u. a. (2022): Training language models to follow instructions with human feedback. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2203.02155> (letzter Zugriff: 21.03.2025).
- PARISER, ELI (2012): *The Filter Bubble: What the Internet Is Hiding from You*. London: Penguin Books.

- PRETTNER, KLAUS/GEIGER, NIELS/SCHWARZER, JOHANNES A. (2018): Die Auswirkungen der Automatisierung auf Wachstum, Beschäftigung und Ungleichheit. In: *Perspektiven der Wirtschaftspolitik*, 19 (2), S. 59–77. URL: <https://doi.org/10.1515/pwp-2018-0017> (letzter Zugriff: 21.03.2025).
- QIAO, DANDAN/RUI, HUAXIA/XIONG, QIAN (2024): AI and Jobs: Has the Inflection Point Arrived? Evidence from an Online Labor Platform. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2312.04180> (letzter Zugriff: 21.03.2025).
- RADFORD, ALEC/NARASIMHAN, KARTHIK/SALIMANS, TIM/SUTSKEVER, ILYA (2018): Improving Language Understanding by Generative Pre-Training. In: *OpenAI Research*. URL: [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) (letzter Zugriff: 21.03.2025).
- SERGEYUK, AGNIA/GOLUBEV, YAROSLAV/BRYKSIN, TIMOFEY/AHMED, IFTEKHAR (2025): Using AI-Based Coding Assistants in Practice: State of Affairs, Perceptions, and Ways Forward. In: *Information and Software Technology*, 178 (Februar), Art. 107610. URL: <https://doi.org/10.1016/j.infsof.2024.107610> (letzter Zugriff: 21.03.2025).
- VASWANI, ASHISH/SHAZEER, NOAM/PARMAR, NIKI/USZKOREIT, JAKOB/JONES, LLION/GOMEZ, AIDAN N/KAISER, ŁUKASZ/POLOSUKHIN, ILLIA (2017): Attention is All You Need. In: Guyon, I./Von Luxburg, U./Bengio, S./Wallach, H./Fergus, R./Vishwanathan, S./Garnett, R. (Hg.): *Advances in Neural Information Processing Systems*, Bd. 30. Curran Associates, Inc. URL: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf> (letzter Zugriff: 21.03.2025).
- VIG, JESSE (2019): A Multiscale Visualization of Attention in the Transformer Model. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.1906.05714> (letzter Zugriff: 21.03.2025).
- WANG, ALEX/PRUKSACHATKUN, YADA/NANGIA, NIKITA/SINGH, AMANPREET/MICHAEL, JULIAN/HILL, FELIX/LEVY, OMER/BOWMAN, SAMUEL (2019): SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In: Wallach, H./Larochelle, H./Beygelzimer, A./d'Alché-Buc, F./Fox, E./Garnett, R. (Hg.): *Advances in Neural Information Processing Systems*, Bd. 32. Curran Associates, Inc. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf) (letzter Zugriff: 21.03.2025).

- WANG, LEI/MA, CHEN/FENG, XUEYANG/ZHANG, ZEYU/YANG, HAO/ZHANG, JINGSEN/CHEN, ZHIYUAN u. a. (2024): A Survey on Large Language Model Based Autonomous Agents. In: *Frontiers of Computer Science*, 18 (6), Art. 186345. URL: <https://doi.org/10.1007/s11704-024-40231-1> (letzter Zugriff: 21.03.2025).
- WEIDINGER, LAURA/MELLOR, JOHN/RAUH, MARIBETH/GRIFFIN, CONOR/UESATO, JONATHAN/HUANG, PO-SEN/CHENG, MYRA u. a. (2021): Ethical and social risks of harm from Language Models. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2112.04359> (letzter Zugriff: 21.03.2025).
- XU, FENGLI/HAO, QIANYUE/ZONG, ZEFANG/WANG, JINGWEI/ZHANG, YUNKE/WANG, JINGYI/LAN, XIAOCHONG u. a. (2025): Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. In: *arXiv*. URL: <https://doi.org/10.48550/arXiv.2501.09686> (letzter Zugriff: 21.03.2025).



# Die didaktischen Fragen, die KI aus Sicht des Lesens zum Zweck des Lernens aufwirft

## Abstract

Wie und wozu lässt sich aus der Warte des Lesens zum Zweck des Lernens die KI nutzen? Die Antwort auf diese übergeordnete Frage erweist sich als komplex, da sich dahinter vielschichtige Teilfragen verbergen. Aus zwei Perspektiven mit je zwei Fragen schlägt dieses Kapitel Antworten vor. Die erste Perspektive geht auf Merkmale des Wissens und der Wissensproduktion ein. Die zweite Perspektive fokussiert auf die Voraussetzungen der Leser:innen bzw. Lerner:innen und die lesedidaktischen Aufgaben für Lehrpersonen bei der Einbindung der KI als Assistenz. Aus dem Gesamt der Antworten wird deutlich, dass damit Grundsatzthemen und Grundsatzdiskussionen angesprochen sind, die für die Lesedidaktik und das Lernen durch das Lesen zugleich fruchtbar und weitreichend sind.

**Schlagwörter:** Digitales Lesen, epistemische Kognitionen, evaluierendes Lesen, Denken höherer Ordnung, Lesekompetenz

## 1 Einleitung

Künstliche Intelligenz (KI), Large Language Models und Chatbots haben seit der Veröffentlichung der ChatGPT-Version 3.5 im November 2022 eine große Prominenz im Bildungsdiskurs erlangt. Die mit diesen KI-Anwendungen verbundenen Befürchtungen und Hoffnungen tangieren viele Bereiche des Lernens und Lehrens und haben teils grundsätzliche Fragen aufgeworfen (vgl. Kasneci u. a. 2023). Was für die bildungswissenschaftliche Reaktion im Allgemeinen zu beobachten ist, gilt auch für ihren Teil, die Deutschdidaktik, im Be-

sonderen – die Existenz dieses Sammelbands ist Ausdruck davon. In der Suche nach dem Platz, den Chatbots mit ihrem immensen Maschinenlernen als Basis für das sprachliche und literarische Lernen und die Kompetenzbereiche des Deutschunterrichts haben oder nicht haben sollen und dürfen, treten einige Fragen deutlicher hervor, die sich bereits zuvor mit Blick auf die fortgeschrittene Digitalisierung zu formieren begonnen haben (vgl. Philipp 2023, 2025b). Diese Fragen lassen sich zuvorderst beim verstehenden Umgang mit ChatGPT und damit beim kompetenten Lesen verorten. Dieser Verortungsversuch ist leitend für dieses Kapitel, welches anhand einiger Leitfragen die Herausforderungen durch KI für die Deutschdidaktik systematisierend betrachtet.

Das geht nicht ohne Eingrenzungen. Darum legt dieses Kapitel nicht nur einen Schwerpunkt beim Lesen als Bestandteil des Umgangs in der Interaktion mit Chatbots, sondern es schränkt diese Perspektive nochmals aus Gründen der Bearbeitbarkeit und Darstellbarkeit auf epistemische Kognitionen (vgl. Barzilai/Chinn 2024) als Bestandteil des selbstregulierten Lernens (vgl. Muis/Singh 2018) ein. Dieser Bereich ist in der Forschung zum ChatGPT-Einsatz für das sprachliche Lernen noch unterbelichtet (vgl. Li u. a. 2024) und verdient darum mehr Beachtung, durchaus zunächst aus einer theoretischen Warte (vgl. Philipp i. Dr.). Dabei ist die Grundfrage zentral, welchen Quellen warum zugestanden wird, dass sie wahre, neutrale, gesellschaftlich nützliche und in ihrer Genese etablierten Qualitätskriterien genügende Wissensaussagen tätigen. Dies ist der Gegenstand der Forschung rund um die epistemischen, also wissensbezogenen Kognitionen (vgl. Chinn/Buckland/Samarapungavan 2011). Mit der Verbreitung von Chatbots wie ChatGPT ist ein neues Element hinzugegetreten, welches epistemisch relevante alte Fragen in neuem Licht stellt, die einer bildungswissenschaftlichen und deutsch- bzw. lesedidaktischen Antwort bedürfen. Diesen Fragen widmet sich Abschnitt 2.

Abschnitt 3 geht auf die didaktischen Implikationen ein und fokussiert zweierlei. Erstens fällt auf, dass eine gewinnbringende Verwendung von KI diverse Fähigkeiten ein- und erfordert, von denen einige den Bereichen Kognition und Metakognition entstammen. Dies betrifft die Eingangsvoraussetzungen resp. die schulisch anzubahnenen Fähigkeiten. Zweitens sind auch Lehrpersonen hochgradig gefordert, Erwerbskontexte zu schaffen, welche KI

als Assistenz aktiv einbinden. Mit diesen beiden Perspektiven in einer kurzen integrativen synoptischen Schlussbemerkung endet der Beitrag in Abschnitt 4.

## **2 Perspektive 1: Wissen und KI-Outputs**

Ein ganzes Bündel von Fragen, das die Outputs von KI-Applikationen aufwerfen, betrifft den epistemisch unsicheren Status dieser KI-Outputs. Handelt es sich etwa um Aussagen, die inhaltlich so belastbar sind, dass die Leser:innen sie weiterverwerten können – falls ja: als Informationen oder bzw. und als Wissen? Greifen bei KI konventionalisierte Mechanismen, wie kognitiv arbeitsteilige Gesellschaften sicherstellen, woran sich Wissensproduktion festmacht? Diesen Fragen widmet sich dieses Teilkapitel, es sind Fragen, die zutiefst epistemischen Charakter haben.

### **2.1 Informationen und/oder Wissen – als was lassen wir die Outputs von ChatGPT & Co. gelten?**

Eine der ersten Fragen, die sich bei KI-basierten Anwendungen wie Chatbots stellt, betrifft den Status des Outputs. Wenn ChatGPT auf einen Prompt eine Antwort ausgibt, müssen – natürlich in Abhängigkeit des Prompts – die Leser:innen einzuschätzen in der Lage sein, welchen epistemischen Status sie dem Output zuweisen. Damit ist gemeint, dass sie ein Urteil fällen, ob sie den Output als Information bzw. als Wissen anerkennen, wobei ein hinlänglich profundes Urteil auf ein ausreichendes Vor- und Weltwissen angewiesen ist. Man nennt solche inhaltsbezogenen Urteile Erste-Hand-Einschätzungen; sie gelten als typische Anforderung des Lesens im Internet (vgl. Stadler u. a. 2017). Zugleich machen sie auf die paradoxe Situation aufmerksam, vor der Laien bzw. Noviz:innen – und damit auch Schüler:innen – stehen, die sich informieren wollen. Ihnen fehlt häufig genau dieses Vorwissen. Hinzu kommen die hinlänglich bekannten allgemeinen Tendenzen zu Fehlschlüssen (vgl. Pennycook 2023) und die Nebenwirkungen der unkorrigierten Effekte von zum Beispiel subjektiven Überzeugungen (vgl. Richter/Maier 2017) und unvollständigem Wissen (vgl. Braasch/Graesser 2020).

Mit diesen Vorbemerkungen zur Einschätzungsbedürftigkeit von Chatbot-Outputs lässt sich nun der in der Frage dieses Abschnitts annoncierte Unterschied von Information und Wissen fokussieren, der mit einer Definition beider Termini beginnt:

- *Informationen* sind außerhalb einer Person befindliche wahrnehm- und verarbeitbare Stimuli bzw. – wenn solche Stimuli für spezifische Zwecke (systematisch) extrahiert wurden – Daten. Solcherart verstandenen Informationen bilden die Grundlage, um daraus Wissen aufzubauen (vgl. Alexander/Schoute 2022).
- *Wissen* wird als ein innerhalb von Personen repräsentiertes bzw. gezielt von ihnen äußerbares Gesamt von Informationen verstanden, nämlich „als das theoretische oder praktische Verständnis von Informationen und die Repräsentation des Verständnisses im Gedächtnis“ (Kendeou/O’Brien 2016: 151). Diese allgemeine Definition von Wissen ist in der Psychologie in der Vergangenheit immer wieder ausdifferenziert worden, indem Arten und Typen des Wissens ebenso unterschieden werden wie Formen und Granularitätsgrößen (vgl. dazu bspw. Alexander/Schoute 2022; McCarthy/McNamara 2023).

Mit der Definition, dass Wissen als interne Größe eines Individuums verstanden wird, wie sie für das psychologische Verständnis typisch ist, wird in der Leseforschung meist die lesende Person und ihr Vorwissen fokussiert (vgl. Kendeou/O’Brien 2016). Bestehende lesetheoretische Modelle greifen die hohe Bedeutung dieses Merkmals für gelingendes Leseverstehen auf. In einer neueren Überblicksarbeit wurden jüngst erstmals Merkmale bzw. Dimensionen des (Vor-)Wissens systematisiert: im „Multidimensional Knowledge in Text Comprehension Framework“ (vgl. McCarthy/McNamara 2021). Dieses Rahmenwerk postuliert vier Dimensionen des Wissens (s. Tabelle 1) und macht sie einer Analyse zugänglich, die auch für die Einschätzung der Qualitäten des Wissens anhand komplexer Zusammenhänge der Dimensionen untereinander nützlich ist.

<b>Wissensdimension</b>	<b>Beschreibung</b>	<b>Prüffragen für Output von KI-Chatbots</b>
Menge	Anzahl der Konzepte, die für den Inhalt des Textes relevant sind und welche die lesende Person kennt	Wie viele inhaltliche (Fach-)Begriffe enthält der Text und sind die Begriffe vollständig?
Akkuratheit	Ausmaß, in dem das Wissen der lesenden Person richtig oder falsch ist	Sind die Konzepte inhaltlich korrekt und genau genug dargestellt?
Spezifität	Grad des inhaltlichen Zusammenhangs bzw. Passung zwischen dem Wissen und den Informationen im Text	Passen die Inhalte des Textes zu dem Thema in ihrer Fokussierung auf den Zielgegenstand?
Kohärenz	Qualität des Vorwissens im Hinblick auf die Vernetztheit der Konzepte	Ist die übergreifende Darstellung des Gegenstands im Output so, dass sie der Verbindung zwischen den Elementen inhaltlich gerecht wird?

**Tab. 1:** Dimensionen des Vorwissens gemäß „Multidimensional Knowledge in Text Comprehension Framework“ (Quelle: eigene Darstellung, basierend auf McCarthy/McNamara 2021: 200)

Geeignet für die Belange dieses Kapitels ist das Rahmenwerk vor allem wegen seiner Operationalisierung, denn es ist ein zentrales Motiv bei der Entwicklung des Rahmenwerks gewesen, präzisere Messungen von mehreren Dimensionen des Wissens anzulegen. Dies wiederum ist für das Schulfeld im Allgemeinen und den Deutschunterricht im Besonderen ein Gewinn, denn es kann dabei helfen, den Blick auf Qualitäten und Merkmale des KI-Chatbot-Outputs zu lenken. Anders gesagt ergeben sich übertragbar wirkende Fragen, die für einen reflektierten Einsatz von KI-Chatbots wie ChatGPT als Kriterien anlegbar sind (s. die Spalte rechts in Tabelle 1). Sie helfen bei der Einschätzung, ob es sich um bloße Informationen (am ehesten noch über die Menge abdeckbar, teils auch die Spezifität betreffend) handelt oder doch um Wissen, das akkurat und kohärent ist. Diese Frage nach der Einschätzung von Information und Wissen ist trotz der implizierten Binarität keine Dichotomie, sondern betrifft ganz im Sinne des Rahmenwerks zum Wissen ein mehrere Kontinua umfassendes

Urteil, das die inhaltliche Einschätzung fokussiert. Anders gesagt: Der Output von Chatbots ist nicht per se (k)eine Information oder (kein) Wissensangebot, sondern erhält diesen Status erst durch eine reflektierte und im Kern plausibilitätsbasierte Zuweisung einer Person, die zu dieser Einschätzung in der Lage ist. Das impliziert, dass die Nutzer:innen von Chatbots mit einer gewissen Wachsamkeit und Distanz mit dem Output verfahren sollten.

## 2.2 Woher kommt das Wissen?

Die im Abschnitt 2.1 aufgeworfene Frage nach dem Zuweisen des Status' von KI-Output als Information bzw. als Wissen reicht weiter als eine auf Plausibilität des Inhalts abzielende Prüfung. Eines der Geschäftsgeheimnisse von ChatGPT und inzwischen Gegenstand juristischer Auseinandersetzung ist die Frage, anhand welcher Korpora und mit welchen Algorithmen zum einen das maschinelle Deep Learning stattgefunden hat (vgl. Müller/Fürstenberg 2023). Diese Frage des für die sprachliche Mustererkennung so wichtigen Inputs, die ja alles andere als trivial ist für den Output der Chatbots, stellt sich für die Zwecke des Lernens. Denn wenn der Output verzerrt ist, weil jedwede im Internet verfügbare Website das sprachliche Rohmaterial für das Lernen war – und damit die volle Spannweite an heterogenen menschlichen Kommunikaten –, dann führt das zum anderen in den Problembereich hinein, ob es sich selbst bei inhaltlich plausibel wirkenden Outputs um *belastbares* Wissen handelt, weil dies an der Quelle liegt, die diese Aussagen produziert.

Diese Frage fokussiert auf die sogenannte „epistemische Vertrauenswürdigkeit“: Welchen Quellen trauen Personen zu, dass sie aufgrund welcher Merkmale als Produzent:innen bzw. als Kommunikator:innen von Wissen gelten? Hiermit befindet man sich im Kern der epistemischen Kognitionen (vgl. Muis/Singh 2018) und damit zugleich einem durch die durchdringende Digitalisierung und ihre Folgen prononcierten Feld des digitalen Lesens (vgl. Philipp 2025b). Denn die Ausdifferenzierung in verschiedene Medien, Distributor:innen, Kommunikationsformen und Autor:innen im Internet hat dazu geführt, dass neben eine inhaltliche Prüfung eine Evaluation der Vertrauenswürdigkeit der Quelle als Erfordernis hinzugetreten ist (vgl. Philipp 2023).

Die Frage, welche Quellen von Äußerungen warum berechtigt sind, Aussagen zu tätigen, die einen Geltungsanspruch als Wissen im Sinne des Abschnitts 2.1 haben, ist traditionell an menschliche Autorschaft bzw. an Institutionen und Organisationen in komplexen Wissensgesellschaften geknüpft gewesen. In diesem Verständnis gilt eine Person bzw. Institution dann als epistemisch vertrauenswürdig, wenn sie eine fachliche Expertise aufweist, integer agiert und keine schädigende Absicht zu verfolgen scheint (vgl. Bromme/Stadtler/Scharrer 2018). Eine Quelle von Aussagen gilt überdies dann als epistemisch vertrauenswürdig, wenn sie zuverlässige Prozesse der Wissensproduktion bemüht hat (vgl. Chinn/Rinehart 2016), die ihrerseits für die lesende Person erkenn- bzw. nachvollziehbar sind, die also transparent genug offengelegt wurden (vgl. Barzilai/Chinn 2018). Es sind ergo sowohl personelle Merkmale als auch die Prozesse der Wissensproduktion, die von der Person bzw. den Personen bemüht wurden, welche Quellen vertrauenswürdiger und ihre Aussagen damit glaubhafter machen.

Wie gesagt: Das sind tradierte und konventionalisierte Umgangsweisen. Eine personenlose Entität wie KI-Applikationen mit für Laien opak bleibender Vorgehensweise in der Art, wie auf Prompts welches Ergebnis hergestellt wird (vgl. Müller/Fürstenberg 2023), erschweren solche quellenbezogenen Einschätzungen, die sich gerade erst in anderen Bereichen des digitalen Lesens mit Dokumenten als Verbindung von Texten und über Metadaten rekonstruierbaren Autor:innen (vgl. Britt/Rouet/Braasch 2013) als zentrale Facette des angemessenen Leseverstehens herauskristallisiert haben (Philipp 2023, 2025b). Es gibt demnach ein Ungleichgewicht in der Entwicklung: KI-Outputs auf der Basis von Large Language Models mit enormen Korpora des maschinellen Lernens der Mustererkennung von Sprache können potenziell Wissen enthalten, aber ohne entsprechendes Prompting wird es schwer bis unmöglich, typische Prozesse und Produkte der menschlichen Wissensherstellung in KI-Outputs zu erkennen. Wer es über das Prompting elizitieren will, muss hohes epistemisches Wissen aufweisen und relativ genau repräsentieren, wonach er oder sie sucht. Weiter gedacht stellt sich die Frage, welchen Status KI-Applikationen in der Wissensproduktion haben, für schulische Zwecke ist aber die Frage virulenter, ob ChatGPT und Co. *epistemisch vertrauenswürdig genug* sind, um sie angemessen einzubinden.

### **3 Perspektive 2: Wissenserwerb mit dem Lesen in Zeiten von KI**

Eine intelligente Nutzung von KI durch Schüler:innen und Lehrpersonen zu ermöglichen, vor dieser Aufgabe stehen Schulfeld und Hochschulen. Dafür braucht es sowohl Klärungen der Voraussetzungen seitens der Lerner:innen (3.1) als auch Kalibrierungen in den Zielvorstellungen seitens der Lernaufgaben und deren Gestaltung (3.2).

#### **3.1 Welche Fähigkeiten müssen Lernende aufweisen, wenn sie KI-basierte Chatbots für tiefes Leseverstehen nutzen wollen?**

Wie unter dem Brennglas zeigt sich bei der erfolgreichen Nutzung von KI, wie viele Fähigkeiten eine Person dazu benötigt. In diesem Abschnitt werden drei von ihnen behandelt: eine allgemeine KI-Kompetenz (3.1.1), das Denken höherer Ordnung (3.1.2) und schließlich die selbstregulativen Fähigkeiten (3.1.3). Allein in dieser Aufzählung wird deutlich, dass eine nicht oberflächlich erfolgende KI-Verwendung kognitiv und metakognitiv herausfordernd ist.

##### **3.1.1 Fokus 1: KI-Kompetenz**

Eine weit verbreitete Definition von KI-Kompetenz (im Original „AI Literacy“) lautet folgendermaßen: „Wir definieren KI-Kompetenz als eine Gruppe von Kompetenzen, die es Einzelnen ermöglicht, KI-Technologien kritisch zu bewerten, effektiv mit KI zu kommunizieren und zusammenzuarbeiten und KI als Werkzeug [...] zu nutzen“ (Long/Magerko 2020: 2). Die zielbezogene Nutzung von und Interaktion mit KI-Technologien sind nebst einer kritisch-evaluativen Fähigkeit zentral in diesem Begriffsverständnis. Zugleich scheint in der Definition auf, dass die KI-Kompetenz als binnendimensionales Kompetenzkonstrukt konzeptualisiert wird.

Erste Binnendifferenzierungen, nicht zu verwechseln mit einer geklärten Kompetenzdimensionsstruktur, eines so umfassenden Kompetenzkonstrukts werden unterschiedlich vorgenommen. Dafür drei Beispiele, zwei von ihnen vor der Veröffentlichung von ChatGPT 3.5, alle drei aber als konzeptuelle

Extraktionen zentraler Themen sowie Termini: Long und Magerko (2020) listen insgesamt 17 Einzelkompetenzen auf, die sie aus der Forschung gebündelt haben und entlang fünf übergeordneter Themen in einem konzeptuellen Rahmen verorten. Ng u. a. (2021) haben in ihrem Review-Beitrag über 30 Veröffentlichungen aus den Jahren 2016 bis 2021 gesichtet und vier Aspekte verdichtet. Almatrafi, Johri und Lee (2024) stützten sich auf 47 zwischen 2019 und 2023 veröffentlichte Studien und haben aus ihnen sechs Kernkonstrukte ermittelt.

Aus den Klassifikationen der drei Überblicksarbeiten lassen sich für die Zwecke dieses Kapitels drei Muster abstrahieren. So betonen alle Autor:innen-gruppen erstens, dass Wissen über und ein Verständnis von KI im Allgemeinen und die Funktionsweisen von KI-Applikationen im Besonderen nötig sind – es gilt als Basis für die weiteren Bestandteile des Dachkonzepts KI-Kompetenz. Zweitens, und das wird bereits in der KI-Kompetenzdefinition oben sichtbar, ist eine evaluative Komponente Bestandteil. Darunter zählen Almatrafi/Johri/Lee (2024: 6) „die Fähigkeit, die Ergebnisse von KI-Anwendungen kritisch zu analysieren und zu interpretieren“, und auf diese Fähigkeit wird im weiteren Verlauf der Darstellung in diesem Kapitel noch vertiefend eingegangen. Sowohl Almatrafi/Johri/Lee (2024) als auch Ng et al. (2021) rücken – drittens – die Fähigkeiten der KI-Kompetenz in den Kontext des Denkens höherer Ordnung und nutzen ganz explizit die revidierte Bloom-Taxonomie von Anderson u. a. (2001) als Grundlage ihrer Systematik. Damit erweitern sie eine etablierte Unterscheidung von verschiedenen Prozessen des Lernens und Lernzielen, indem sie diese auf die Belange des kompetenten Umgangs mit KI und KI-Applikationen transferieren.

Zusammenfassend lässt sich sagen: KI-Kompetenz als mehrdimensionales Konstrukt, dessen Binnenstruktur theoretisch und empirisch klärungsbedürftig wirkt, bildet für die Leser:innen eine Basis, um der Spezifik der von KI-Apps hergestellten Texte gerecht zu werden. Das Wissen über KI und deren kritisch reflektierte Nutzung sind hier besonders anzuführen, um der Spezifik der Verstehensprozesse gerecht zu werden. In den Prozessen des Denkens höherer Ordnung kulminieren KI- und lesebezogene Anforderungen (vgl. Afflerbach/Cho/Kim 2011; List/Sun 2023).

### 3.1.2 Fokus 2: Prozesse höheren Denkens

Das Denken höherer Ordnung ist kein einheitlich definiertes Konstrukt. In einer sehr knappen Definition fassen Schraw/Robinson (2011: 2) darunter „Fähigkeiten, die den Aufbau eines tieferen, konzeptionell orientierten Verständnisses ermöglichen“. Damit rücken sie das Verstehen und somit eher das Wissen denn die Information (s. Abschnitt 2.1) in das konzeptuelle Zentrum (vgl. Anderson u. a. 2001). Das Denken höherer Ordnung ist als Fähigkeitenbündel klärungsbedürftig, welche Fähigkeiten und die sie betreffenden Prozesse und darüber hinaus daraus resultierenden Produkte es inkludiert (vgl. Alexander 2023; List/Sun 2023).

Für die lerntheoretische Verortung der Fähigkeiten des Denkens höherer Ordnung wird in der Literatur häufig die ursprünglich als eindimensional angelegte Lernziel-Taxonomie von Bloom angeführt, die inzwischen in eine zweidimensionale Taxonomie von kognitiven Prozessen und Wissensarten überführt wurde (vgl. Anderson u. a. 2001). Die sechs Prozessgruppenkategorien samt Kurzdefinitionen dieser Taxonomie sind in der Tabelle 2 dargestellt; dabei stehen die kognitiv anspruchsvollsten Prozessgruppen zuoberst, die Sortierung erfolgt demnach absteigend. Hinter den sechs Kategorien verbergen sich einzelne Prozesse in unterschiedlicher Zahl pro Kategorie, auf deren Explikation platzbedingt hier verzichtet werden muss (vgl. dazu Philipp 2022).

Das Denken höherer Ordnung ist für die Domäne Lesen bereits als Rahmen genutzt worden, speziell um Grenzziehungen vorzunehmen, ab wann die Lernprozesse dem Denken höherer Ordnung zugerechnet werden. So haben Afflerbach/Cho/Kim (2011) für das Lesen im Allgemeinen das Denken höherer Ordnung in der Bloom-Taxonomie dem Anwenden (hier jedoch verstanden als Transfer von Gelesenem), der Analyse, der Synthese und der Evaluation zugeordnet. Ähnliches schlagen List/Sun (2023) für das Lesen multipler (digitaler) Dokumente im Besonderen vor. Das Denken höherer Ordnung ist damit spezifischen, kognitiv anspruchsvollen Prozessen vorbehalten, die über das Erinnern und das reine Verstehen von Texten hinausgehen. Demnach ist das Verstehen im Sinne einer reinen Rekonstruktion von Textinhalten nicht als Bestandteil, sondern Voraussetzung des Denkens höherer Ordnung zu werten (vgl. Afflerbach/Cho/Kim 2011).

Theoretische Lernprozesskategorien	Empirische Prozesskategorien aus der Multiple-Documents-Literacy-Forschung
<i>Kreieren</i> :* Elemente in eine neue kohärente Form oder Struktur überführen	[Keine Entsprechung im Forschungsüberblick]
<i>Evaluieren</i> :* kriterien-/standardbasierte Einschätzungen vornehmen	<ul style="list-style-type: none"> <li>• <i>Sourcing (Quellenbewertung)</i>: Identifizierung von Metadaten über die Herkunft von Dokumenten und Verwendung dieser Informationen zur Einschätzung der Vertrauenswürdigkeit oder Zuverlässigkeit</li> <li>• <i>Corroboration (textvergleichendes Prüfen)</i>: Vergleich und Gegenüberstellung von Informationen aus verschiedenen Texten</li> <li>• <i>Inhaltsbewertung</i>: Bewertung des Inhalts oder der Qualität textueller Informationen, einschließlich ihrer Korrektheit und Vollständigkeit</li> <li>• <i>Relevanzbeurteilung</i>: Passungseinschätzung zwischen Aufgabenerfordernissen und Informationen in Texten</li> <li>• <i>Nicht-epistemische Beurteilung</i>:** Bewertung von Texten auf der Grundlage oberflächlicher Merkmale, die nichts mit der Vertrauenswürdigkeit der Quelle oder der Qualität der Informationen zu tun haben</li> </ul>
<i>Analysieren</i> :* Einzelbestandteile erkennen und ihre Relation zu anderen Bestandteilen und zum Gesamt bestimmen	<ul style="list-style-type: none"> <li>• <i>Intertextuelles Integrieren</i>: Herstellung von Verbindungen zwischen Texten, typischerweise auf der Grundlage ihrer semantischen Überschneidungen</li> <li>• <i>Organisation</i>: Strukturierung von Informationen aus mehreren Texten, einschließlich des Bildens neuer konzeptioneller oder struktureller Kategorien</li> <li>• <i>Analysieren</i>: Zergliedern oder Identifizieren von bedeutungsvollen Segmenten in Texten und Zusammenhänge zwischen diesen Segmenten herstellen</li> </ul>
<i>Anwenden</i> :* Ausführen von prozeduralem Wissen	[Keine Entsprechung im Forschungsüberblick]
<i>Verstehen</i> : Bedeutung aus Kommunikaten konstruieren	<i>Verstehen</i> : Bedeutungsherstellung aus Texten, einschließlich Elaboration, Inferenzen und Selbsterklärungen

Theoretische Lernprozesskategorien	Empirische Prozesskategorien aus der Multiple Documents-Literacy-Forschung
<i>Erinnern</i> : Informationen aus dem Langzeitgedächtnis abrufen	<i>Informationsakkumulation</i> : Oberflächliches Sammeln, Akkumulieren oder Verarbeiten von Informationen

**Tab. 2:** Prozesskategorien des Lernens (Quelle der Deskriptoren der sechs theoretischen kognitiven Prozesse: Anderson u. a. 2001: 67–68, leicht modifiziert; Quelle der empirischen Kategorien und ihrer Zuweisung zu den Prozessen: List/Sun 2023: 8–9, 26, leicht modifiziert; Legende: \* = gilt als Denken höherer Ordnung; \*\* = durch Orientierung an Oberflächenmerkmalen kein Denken höherer Ordnung, sondern eher heuristischer Prozess)

Die soeben angesprochene Perspektive des Umgangs mit multiplen Dokumenten ist für die Verwendung von KI naheliegend. Erstens ist davon auszugehen, dass Personen durch ein iteratives Vorgehen mehrere, also multiple digitale Dokumente in der Mensch-Maschine-Interaktion lesen, verstehen und nutzen müssen, was sich mit Definitionen der Multiple Documents Literacy deckt (vgl. McCrudden/Bräten/Salmerón 2023). Zweitens ähneln sie den Prozessverbänden beim Lesen multipler Dokumente – wie Recherchieren, Evaluieren und Integrieren (vgl. Stadler/Bromme/Rouet 2018) – deutlich. Drittens sind diese Fähigkeiten integraler Bestandteil des Lesens digitaler Dokumente, darunter inzwischen KI-Outputs (vgl. Philipp 2025b). Die Überblicksarbeit von List/Sun (2023) über 57 Studien – davon nur ein Viertel mit Schüler:innen durchgeführt – ist für die Zwecke dieses Kapitels besonders einschlägig. Denn in ihr wurden 14 Prozesskategorien extrahiert, von denen sich zehn in die Logik der Taxonomie aus Tabelle 2 in der Spalte rechts einfügen lassen. Wie sich an der Tabelle 2 ablesen lässt, sind diese Fähigkeiten besonders beim Analysieren und Evaluieren ausdifferenziert.

Die genauen Lektüren und das Verstehen von Aussagen, z. B. wenn ChatGPT beim Verfeinern der Prompts unterschiedliche Outputs generiert, die man vergleichend lesen muss, fallen im Gesamt der empirisch ermittelten Prozesse der Multiple Documents Literacy in Tabelle 2 als Erstes auf. Das gilt zumindest für den Fall, in dem es nicht um einzelne Informationen geht, sondern eher wissensbezogene Darstellungen. Auffällig ist als Zweites, dass diverse Evaluationen mit unterschiedlichen Funktionen (Relevanz, Inhaltsgüte, Konsistenz und Quellenmerkmale) beobachtet werden konnten. Diese stellen typische Anfor-

derungen an Leser:innen digitaler Texte und Dokumente (vgl. Philipp 2025a). Durch KI-Outputs ergeben sich Anforderungsverschiebungen insofern, als das Sourcing erschwert wird und Leser:innen durch Prompting Metadaten als „eingebettete Quellen“ explizit einfordern müssen. Die anderen Evaluationsprozesse bleiben mutmaßlich bestehen, während die nicht-epistemischen Evaluationen noch weniger zielführend wirken, als sie es bereits zuvor getan haben.

Damit lässt sich als Zwischenfazit festhalten: Tiefgehendes Verstehen ist Resultat verschiedener Prozessgruppen, die sich unter dem Dachkonzept „Denken höherer Ordnung“ sammeln und in der Leseforschung empirisch vorfinden lassen. Evaluative und analytische Fähigkeiten bilden Schwerpunkte, was auf den Umgang mit KI-Outputs übertragbar wirkt.

### **3.1.3 Fokus 3: Selbstregulation beim Prompt Engineering und darüber hinaus**

KI-Applikationen wie ChatGPT produzieren schriftlichen Output infolge einer Mensch-Maschine-Interaktion. Das bedeutet: Streng genommen erschaffen sich Leser:innen durch Prompt Engineering ihre Lesegegenstände steuernd selbst bzw. wirken daran mit. Darin ähnelt dieses Vorgehen prinzipiell dem Recherchieren in Datenbanken, wenngleich bei KI-Outputs die im Prompt gegebenen Informationen eine andere Funktion haben als Suchbegriffe und deren Trunkierung in Datenbanken bzw. in Suchmaschinen. Darum ist es in Hinblick auf die KI-Kompetenz nötig, das hier als ganz eigener Anforderungsbereich hinzukommende Prompt Engineering (vgl. Eager/Brunton 2023; Lo 2023a; Velásquez-Henao/Franco-Cardona/Cadavid-Higueta 2023) in seinen zielbezogenen und damit metakognitiv zu rahmenden Kontext zu setzen und vorzustellen.

Zunächst ist es notwendig, Prompts und Prompt Engineering zu definieren. Mit „Prompt“ ist ein Text gemeint, der eine Reihe von Anweisungen enthält, um damit die Einsatzbereiche einer KI-Applikation anzupassen, zu verfeinern oder zu erweitern (vgl. Velásquez-Henao/Franco-Cardona/Cadavid-Higueta 2023). Es handelt sich also um einen sprachlichen Befehl der nutzenden Person, die ihrerseits ein mehr oder minder klar umrissenes Ziel verfolgt, für das sie die KI nutzen will. Dafür bedarf es seitens der lesenden Person des „Prompt Engineerings“:

Prompt Engineering ist der Prozess der Konstruktion von Suchanfragen oder Eingaben (d. h. Prompts) für KI-Sprachmodelle, um möglichst präzise, kohärente und passende Antworten zu erhalten. Im Wesentlichen handelt es sich dabei um die Feinjustierung der Fragen oder Befehle für KI-Modelle, um deren Leistung zu optimieren und zu gewährleisten, sodass sie die gewünschten Ergebnisse liefern. (Lo 2023b: 1)

Prompt Engineering ist damit ein dynamischer Problemlöseprozess, der aufgrund der Ziele der lesenden, die KI-App nutzenden Person zum einen und der Outputs der KI-Applikation zum anderen zu Anpassungen führt. Es handelt sich also um einen idealerweise adaptiven, mehrschrittigen Problemlöseprozess, was sich darin niederschlägt, dass aktuelle Phasenmodelle des Prompt Engineerings einerseits lineare Abfolgen, andererseits zielbezogene Iterationen beinhalten. Ein Beispiel hierfür enthält Tabelle 3, in welchem sich in der Phase 5 deutlich die Evaluation des Denkens höherer Ordnung manifestiert.

Phase	Beschreibung
1) Ziel definieren	Spezifische, zielbezogene Ergebnisse bestimmen, die mit der KI erreicht werden sollen
2) Art und Format des Inhalts definieren	Art und Format des Outputs festlegen, der zum Ziel passt
3) Konzeptioneller Entwurf des ersten Prompts	Einen klaren, prägnanten Prompt erstellen, der Output-Art und -Format ebenso enthält wie Parameter, weitere Anweisungen und Kontextinformationen, um möglichst den gewünschten Output der KI zu erhalten
4) Prompteingabe	Formulierung des Prompts in der Eingabemaske, um ersten Output zu erzeugen
5) Evaluation des Outputs	Abgleich des Prompts mit dem angestrebten Ziel
6) Iterieren	Fortlaufende Anpassung des Prompts bzgl. Kontext, Anweisungen und Spezifikation, bis das Ziel mit dem evolvierenden Output erreicht ist

**Tab. 3:** Phasen des Prompt Engineerings (Quelle: Übersetzung und Modifikation von Eager/ Brunton 2023: 4)

Prompt Engineering kann als eine Ausprägung des selbstregulierten Umgangs mit Dokumenten betrachtet werden, was wiederum typisch ist für das digitale Lesen (vgl. Philipp 2025a; Salmerón u. a. 2018) bzw. für das Lesen multipler Dokumente (vgl. Cho/Afflerbach/Han 2018; List/Sun 2023). Selbstregulation bzw. selbstreguliertes Lernen bezieht sich darauf, wie Lerner:innen mental repräsentierte Inhalte zum einen und die äußere Umgebung zum anderen so metakognitiv überwachen und regulieren, dass die Lerner:innen ihre Ziele erreichen und Wissen aufbauen (vgl. Winne 2015). Dazu brauchen sie klare Zielvorstellungen, metakognitives Wissen über Strategien und Merkmale von Wissen und müssen dieses Wissen in Handlungen umsetzen können. Das sind komplexe Fähigkeiten, die zu orchestrieren anspruchsvoll ist. Die analytischen und evaluativen Fähigkeiten sind dabei der Bezugspunkt der metakognitiven Steuerung: In der Mensch-Maschine-Interaktion dienen sie dazu, rückkoppelnd sowohl die eigenen Ziele als auch den KI-Output kalibrierend auszubalancieren und solange die Prompts zu verändern, bis die Zielzustände erreicht sind. Darum sind Iterationen nicht nur wahrscheinlich, sondern sogar nötig, was für die Lerner:innen bedeutet, dass sie sich nicht prima vista den KI-Outputs nähern sollten. Vielmehr ist es geboten, sie kritisch zu sichten und inhaltlich zu modifizieren, wenn der Zweck des KI-Einsatzes dies erfordert.

Somit lässt sich hier festhalten: Das Prompt Engineering als sprachlich realisiertes Programmieren der KI-Applikation für ein definiertes Ziel ist angewandte Selbstregulation. Planerische, überwachende und regulierende Prozesse und Wissensbestände dienen der Steuerung eines ergebnisoffenen, gleichwohl gesteuerten Prozesses, bei dem die Fähigkeiten der Leserin und des Lesers die Bemessungsgrundlage für metakognitive Entscheidungen sind.

### **3.2 Welche Aufgaben stellen sich Lehrenden bzw. stellen Lehrpersonen zum Einsatz von KI zum Zweck des Lernens?**

KI ist für Lehrpersonen und die Lesedidaktik eine Herausforderung, die sowohl Chancen als auch Risiken in sich birgt. Die Professionalisierung im Sinne des Erwerbs des technischen, fachdidaktischen und Fachwissen von Deutschlehrpersonen zu KI ist dabei nur ein Erfordernis (vgl. Mishra/Warr/

Islam 2023). Ein anderes betrifft Lesedidaktik und Schulfeld darüber hinaus, indem es auszuloten gilt, welche Einsatzmöglichkeiten KI haben soll und kann. Wenn man sich dazu entscheidet, die KI einzubinden, ist die KI-Kompetenz der Schüler:innen unterrichtliche Voraussetzung und Zielgegenstand, die Darstellung im Folgenden basiert auf der Setzung, dass KI-Kompetenz zumindest rudimentär vorhanden ist, indem die Gestaltung von Lernaufgaben fokussiert wird.

Angesichts der eingangs getroffenen Differenzierung von Wissen und Information sind zwei potenzielle Einsatzgebiete möglich, die sich als Informationsmanagement und Wissensaufbau gegenüberstellen lassen (vgl. Alexander 2018; Alexander/Schoute 2022). Die nachfolgende Definition charakterisiert beide Konzepte:

Informationsmanagement [...] kann definiert werden als die eher oberflächliche Organisation und Verarbeitung mehrerer Quellen und die Weitergabe dieser Informationen ohne die Absicht, sie in einem anderen als dem unmittelbaren Kontext, in dem sie benötigt werden, zu behalten oder zu nutzen. Im Gegensatz dazu versteht man unter Wissensaufbau [...] die kritische Analyse und Verarbeitung von Informationen, um sie zu einem Teil einer Wissensbasis zu machen, die zu einem späteren Zeitpunkt genutzt werden kann. (Alexander/Schoute 2022: 14)

KI-Applikationen wirken, die Prüfung des Status als echte Information vorausgesetzt, als Recherchetool einsetzbar, wenn es darum geht, im Sinne des Informationsmanagements singuläre Informationen zu finden und instrumentell zu nutzen. Wenn das Ziel darin besteht, Wissen aufzubauen und das Denken höherer Ordnung zu nutzen, sind komplexere Aufgaben nötig, wie sie sich mit der überarbeiteten Bloom-Taxonomie kartieren und planen lassen (vgl. Anderson u. a. 2001; Philipp 2022).

Solche komplexen Aufgaben sind kein Selbstzweck, sondern dienen dazu, längerfristig Wissen so im Langzeitgedächtnis zu verankern, dass es für weitere Verwendungszwecke zur Verfügung steht. Mit Blick auf die realweltlichen Anforderungen der Wissensnutzung ist ein Gefüge zu attestieren, das sich als

VUCA bezeichnen lässt. Das Akronym steht für Volatilität (*volatility*), Unsicherheit (*uncertainty*), Komplexität (*complexity*) und Ambiguität (*ambiguity*). Demnach ist die Welt stark veränderlich (volatil), unvorhersehbar (unsicher), vielschichtig und verschachtelt (komplex) sowie uneindeutig (ambig; vgl. Taskan/Junça-Silva/Caetano 2022). Das trifft aktuell für den Bildungsbereich ebenfalls zu (vgl. Barzilai/Chinn 2020; Chinn/Barzilai/Duncan 2021). Darum lohnen sich Lernaufgaben, bei denen KI systematisch dafür eingebunden wird, solchen Herausforderungen zu begegnen und für Schüler:innen Ertrag und Grenzen des KI-Einsatzes erfahrbar zu machen.

Empfehlung	Kurzcharakterisierung	Einsatz von KI
1) Gestaltung von zunehmend authentischen Lernumgebungen	Gestaltung von epistemisch „unfreundlichen“ Lernumgebungen (vielfältige, authentische Quellen), tendenziell spiralcurricular	KI-Outputs als eine Textquelle unter mehreren einbinden, aber flankiert mit anderen Dokumenten mit variierendem Grad epistemischer Vertrauenswürdigkeit
2) Gebundenheit des Wissens erlebbar machen	Suche nach und Recherche von epistemischen Autoritäten, um begründete Zweite-Hand-Einschätzungen vorzunehmen	Prompt Engineering, das Suche nach Quellen mit hoher Expertise berücksichtigt
3) Erkundungen des Wissens	Lernformen, welche die Untersuchung und Begründung von Aussagen als Wissen betreffen	Recherche nach Quellen, die zu kontroversen Themen angemessene Aussagen tätigen und ihre Begründungen oder Argumente epistemisch unterfüttern, sowie gezielte Aufbereitung der Positionen mit KI-Unterstützung
4) Förderung tugendhafter epistemischer Motivationen/ Emotionen	Förderung von Autonomie, Kompetenzerleben, Neugierde und Interesse als Basis für kognitiv anspruchsvolle epistemische Ziele	Einbinden statt Ausschließen von KI-Applikationen, um dadurch Recherche-, Vor- und Aufbereitungsarbeiten, Entscheidungen sowie Transformationen zu unterstützen

Empfehlung	Kurzcharakterisierung	Einsatz von KI
5) Förderung des Verständnisses für epistemische Systeme	Systematische Exploration zur Arbeitsweise und Einschätzung von Instanzen der Wissensproduktion und -distribution, aber auch zu Machtstrukturen und deren Implikationen	Analyse von KI-Outputs zu Verzerrungen, als potenzieller Beispielgeber für Erklärungen von Fehlschlüssen und zur Recherche nach Beispielen von mutmaßlich prototypischen Positivbeispielen der wissenschaftlichen Arbeitsweise

**Tab. 4:** Empfehlungen für didaktische Gestaltung von Aufgaben und Settings für den Wissensaufbau (Quelle: modifizierte Darstellung von Chinn/Barzilai/Duncan 2021: 55, rechte Spalte ergänzt)

Mit diesen Vorbemerkungen lässt sich legitimieren, dass sich Lernaufgaben und -settings mit einem Fokus auf Analyse und Evaluation anbieten, in denen die dazu passenden Strategien zum Einsatz gelangen (vgl. Philipp 2025a). Mit Chinn/Barzilai/Duncan (2021) lassen sich einige Gestaltungsprinzipien benennen und mit KI-Einsatz anreichern. Beides erfolgt in Tabelle 4. Die folgende Aufzählung dient der knappen Erläuterung, hegt jedoch keinen Anspruch auf Vollständigkeit – vielmehr geht es um die Demonstration der potenziell breiten Einsatzoptionen der KI im Wissensaufbau.

- 1) Mit der ersten Empfehlung ist gemeint, dass Schüler:innen nicht nur im geschützten Rahmen der Schule auf kuratierte Texte treffen, sondern auch gezielt in ihren Qualitäten als Wissensquellen stark variierenden Dokumenten. Dies bietet die Möglichkeit, Texte und Dokumente miteinander zu vergleichen, um die epistemische Vertrauenswürdigkeit zum Gegenstand der Aushandlung zu machen. KI-Outputs können hier ganz gezielt verwendet werden, und zwar durchaus als Analysegegenstand.
- 2) Da sich Wissensgesellschaften durch Grade an Spezialisierung in der Wissensproduktion auszeichnen, ist es für ihre Mitglieder nötig zu differenzieren, ob sie selbst Aussagen als Wissen direkt beurteilen können oder indirekte Urteile fällen, indem sie epistemische Auto-

ritäten finden und deren Urteilen vertrauen sollten. Damit stellen sich zwei Fragen mit Blick auf KI, wobei die erste zu Informationen und Wissen bereits in Abschnitt 2.1 aufgeworfen wurde. Die zweite betrifft das Prompt Engineering, das sich gezielt dazu nutzen lässt, nach Quellen zu recherchieren, die epistemisch vertrauenswürdig sind, weil sie eine hohe Expertise haben. Zugleich lassen sich damit die Grenzen der KI aus epistemischer Sicht erlebbar machen.

- 3) Gerade die Uneindeutigkeit und Komplexität als Merkmal von Gesellschaften laden dazu ein, Dissens und zuverlässige Begründungen für divergierende Ansichten zum Gegenstand von Analysen zu machen. Darum sind Kontroversen und die ihnen zugrundeliegenden Begründungen ein sinnvoller Gegenstand unterrichtlicher Auseinandersetzung. KI-Tools können dafür genutzt werden, Argumentationsstränge zu überblicken, Dokumente zu finden und durch Prompts aufbereiten zu lassen.
- 4) Komplexe Explorationen in Themengebiete, die dem Suchen nach echten Wissensangeboten und den Begründungen dafür dienen, können und sollten emotional-motivational unterstützend erfolgen und den normativ wünschenswerten Werten des Umgangs mit Wissen folgen. Hier kann der KI-Einsatz so didaktisiert erfolgen, dass der KI eine unterstützende Funktion im Gesamtprozess zukommt, die regelgeleitete Denkarbeit aber in den Verantwortungsbereich der Schüler:innen fällt.
- 5) Das Verstehen von Systemen der Wissensproduktion ist ein weiteres Feld. Denn es geht ja nicht nur um die Quellen und Texte und die rekonstruierbaren Werte, sondern um ein Verständnis davon, wie in verschiedenen Teilsystemen Wissen hergestellt wird. Die KI kann dafür genutzt werden, Beispiele dafür zu recherchieren.

Zusammengefasst: Lehrpersonen und Deutschdidaktik können und sollten damit beginnen, sich kognitiv aktivierende Aufgabenformate des Wissensaufbaus zu überlegen, in denen analytische und evaluative Prozesse des höheren Denkens vorkommen, um dadurch die epistemischen Kognitionen in Verbin-

derung mit Fachinhalten zu forcieren. Der KI kommt dabei eine assistierende Funktion zu – nicht mehr, nicht weniger.

## 4 Fazit

Dieses Kapitel hat über vier Fragen aus zwei Bereichen eine Annäherung an den KI-Einsatz zum Zweck des Lernens aus der Perspektive des Lesens vorgenommen. Es war damit eher programmatisch, da diese Fragen in der Deutschdidaktik noch zu wenig prominent be- und verhandelt werden. Deutlich geworden ist, dass in Zeiten fortschreitender Durchdringung des Alltags mit KI Fragen aufgeworfen werden, die Merkmale des Wissens und der Wissensproduktion betreffen, was ein erster Fragekomplex bzw. eine Perspektive war. Er beinhaltet Grundsatzfragen, weil bestimmte Regeln der Wissensproduktion für Schüler:innen im Umgang mit KI-Applikationen nicht mehr direkt zugänglich sind. Solche Fragen aber sind dadurch längst nicht obsolet, sondern stellen sich dringlich. Anders formuliert: Es ist nicht trivial und auch nicht gleichgültig, als was der Output von KI gilt, ob man ihn als Wissen oder Information akzeptiert und woran man dies festmacht. Hierin liegt zugleich eine Chance, dies didaktisch aufzugreifen und systematisch mit dem Feld der epistemischen Kognitionen zu verknüpfen.

Der zweite Fragenkomplex des Kapitels fokussierte die Lerner:innen und Lehrenden. Für Schüler:innen bedeutet impliziert die KI aus lesedidaktischer Sicht, dass sie für eine erfolgreiche, nicht-oberflächliche Nutzung von KI diverse Fähigkeiten aus verschiedenen Bereichen (KI-Kompetenz, Denken höherer Ordnung, Selbstregulation) besitzen und aktiv ausführen können müssen. Didaktische Situationen des Erwerbs, die ernst nehmen, was Lerner:innen zu leisten haben, fallen in die Zuständigkeit der Lehrpersonen und die Institutionen der Lehrer:innenaus- und -weiterbildung. Hier geht es darum, die bereits hohen Lesekompetenzen von Lerner:innen, die die KI-Nutzung impliziert, möglichst systematisch weiterzuentwickeln. Für die Lesedidaktik bedeutet dies, dass die KI sie erheblich dynamisieren dürfte, um Schritt zu halten mit den sich gerade massiv verändernden Anforderungen des Lesealltags.

## Literatur

- AFFLERBACH, PETER/CHO, BYEONG-YOUNG/KIM, JONG-YUN (2011): The Assessment of Higher Order Thinking in Reading. In: Schraw, Gregory J./Robinson, Daniel R. (Hg.): *Assessment of Higher Order Thinking Skills*. Charlotte: Information Age, S. 185–217.
- ALEXANDER, PATRICIA A. (2018): Information Management versus Knowledge Building: Implications for Learning and Assessment in Higher Education. In: Zlatkin-Troitschanskaia, Olga u. a. (Hg.): *Assessment of Learning Outcomes in Higher Education: Cross-National Comparisons and Perspectives*. Cham: Springer, S. 43–56.
- ALEXANDER, PATRICIA A. (2023): „Here Be Dragons!“ Mapping the Realm of Higher-Order, Critical, and Critical-Analytic Thinking. In: *Educational Psychology Review* 35 (2), Artikel 42.
- ALEXANDER, PATRICIA A./SCHOUTE, ERIC C. (2022): *Knowledge, Knowing, and Information: Their Meaning and Meaningfulness to Learning and Development*. Routledge. <https://doi.org/10.4324/9781138609877-REE203-1>.
- ALMATRAFI, OMAIMA/JOHRI, ADITYA/LEE, HYUNA (2024): A Systematic Review of AI Literacy Conceptualization, Constructs, and Implementation and Assessment Efforts (2019–2023). In: *Computers and Education Open* 6, Artikel 100173.
- ANDERSON, LORIN W. u. a. (2001): *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longman.
- BARZILAI, SARIT/CHINN, CLARK A. (2018): On the Goals of Epistemic Education. Promoting Apt Epistemic Performance. In: *Journal of the Learning Sciences* 27 (3), S. 353–389.
- BARZILAI, SARIT/CHINN, CLARK A. (2020): A Review of Educational Responses to the „Post-Truth“ Condition. Four Lenses on „Post-Truth“ Problems. In: *Educational Psychologist* 55 (3), S. 107–119.
- BARZILAI, SARIT/CHINN, CLARK A. (2024): The AIR and Apt-AIR Frameworks of Epistemic Performance and Growth: Reflections on Educational Theory Development. In: *Educational Psychology Review* 36 (3), Artikel 91.
- BRAASCH, JASON L. G./GRAESSER, ARTHUR C. (2020): Avoiding and Overcoming Misinformation on the Internet. In: Halpern, Diane F./Sternberg, Robert J. (Hg.): *Critical Thinking in Psychology*. Cambridge: Cambridge University Press, S. 125–151.

- BRITT, MARY ANNE/ROUET, JEAN-FRANÇOIS/BRAASCH, JASON L. G. (2013): Documents as Entities. Extending the Situation Model Theory of Comprehension. In: Britt, Mary Anne/Goldman, Susan R./Rouet, Jean-François (Hg.): *Reading. From Words to Multiple Texts*. New York: Routledge, S. 160–179.
- BROMME, RAINER/STADTLER, MARC/SCHARRER, LISA (2018): The Provenance of Certainty. Multiple Source Use and the Public Engagement with Science. In: Braasch, Jason L.G./Bråten, Ivar/McCruden, Matthew T. (Hg.): *Handbook of Multiple Source Use*. New York: Routledge, S. 269–284.
- CHINN, CLARK A./BARZILAI, SARIT/DUNCAN, RAVIT GOLAN (2021): Education for a „Post-Truth“ World. New Directions for Research and Practice. In: *Educational Researcher* 50 (1), S. 51–60.
- CHINN, CLARK A./BUCKLAND, LUKE A./SAMARAPUNGAN, ALA (2011): Expanding the Dimensions of Epistemic Cognition. Arguments from Philosophy and Psychology. In: *Educational Psychologist* 46 (3), S. 141–167.
- CHINN, CLARK A./RINEHART, RONALD W. (2016): Commentary. Advances in Research on Sourcing – Source Credibility and Reliable Processes for Producing Knowledge Claims. In: *Reading and Writing* 29 (8), S. 1701–1717.
- CHO, BYEONG-YOUNG/AFFLERBACH, PETER/HAN, HYEJU (2018): Strategic Processing in Accessing, Comprehending, and Using Multiple Sources Online. In: Braasch, Jason L.G./Bråten, Ivar/McCruden, Matthew T. (Hg.): *Handbook of Multiple Source Use*. New York: Routledge, S. 133–150.
- EAGER, BRONWYN/BRUNTON, RYAN (2023): Prompting Higher Education towards AI-Augmented Teaching and Learning Practice. In: *Journal of University Teaching and Learning Practice* 20 (5), Artikel 2.
- KASNECI, ENKELEJDA u. a. (2023): ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. In: *Learning and Individual Differences* 103, Artikel 102274.
- KENDEOU, PANAYIOTA/O'BRIEN, EDWARD J. (2016): Prior Knowledge. Acquisition and Revision. In: Afflerbach, Peter (Hg.): *Handbook of Individual Differences in Reading. Reader, Text, and Context*. New York: Routledge, S. 151–163.
- LI, BELLE u. a. (2024): A Systematic Review of the First Year of Publications on ChatGPT and Language Education: Examining Research on ChatGPT's Use in Language Learning and Teaching. In: *Computers and Education: Artificial Intelligence* 7, Artikel 100266.

- LIST, ALEXANDRA/SUN, YUTING (2023): To Clarity and Beyond: Situating Higher-Order, Critical, and Critical-Analytic Thinking in the Literature on Learning from Multiple Texts. In: *Educational Psychology Review* 35 (2), Artikel 40.
- LO, LEO S. (2023a): The Art and Science of Prompt Engineering: A New Literacy in the Information Age. In: *Internet Reference Services Quarterly* 27 (4), S. 203–210.
- LO, LEO S. (2023b): The CLEAR Path: A Framework for Enhancing Information Literacy through Prompt Engineering. In: *The Journal of Academic Librarianship* 49 (4), Artikel 102720.
- LONG, DURI/MAGERKO, BRIAN (2020): What is AI Literacy? Competencies and Design Considerations. In: Bernhaupt, Regina (Hg.): *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu: Association for Computing Machinery, S. 1–16.
- MCCARTHY, KATHRYN S./MCNAMARA, DANIELLE S. (2021): The Multidimensional Knowledge in Text Comprehension Framework. In: *Educational Psychologist* 56 (3), S. 196–214.
- MCCARTHY, KATHRYN S./MCNAMARA, DANIELLE S. (2023): Knowledge: A Fundamental Asset. In: Tierney, Robert J. u. a. (Hg.): *International Encyclopedia of Education. Volume 6: Learning, Cognition and Human Development*. Oxford: Elsevier, S. 209–218.
- MCCRUDDEN, MATTHEW T./BRÅTEN, IVAR/SALMERÓN, LADISLAO (2023): Learning from Multiple Texts. In: Tierney, Robert J. u. a. (Hg.): *International Encyclopedia of Education. Volume 6: Learning, Cognition and Human Development*. Oxford: Elsevier, S. 353–363.
- MISHRA, PUNYA/WARR, MELISSA/ISLAM, REZWANA (2023): TPACK in the Age of ChatGPT and Generative AI. In: *Journal of Digital Learning in Teacher Education* 39 (4), S. 235–251.
- MUIS, KRISTA R./SINGH, CARA A. (2018): The Three Facets of Epistemic Thinking in Self-Regulated Learning. In: Schunk, Dale H./Greene, Jeffrey Alan (Hg.): *Handbook of Self-Regulation of Learning and Performance*. New York: Routledge, S. 434–456.
- MÜLLER, HANS-GEORG/FÜRSTENBERG, MAURICE (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Deutschen Germanistenverbandes* 70 (4), S. 327–345.

- NG, DAVY TSZ KIT u. a. (2021): Conceptualizing AI Literacy: An Exploratory Review. In: *Computers and Education: Artificial Intelligence 2*, Artikel 100041.
- PENNYCOOK, GORDON (2023): A Framework for Understanding Reasoning Errors. From Fake News to Climate Change and Beyond. In: Gawronski, Bertram (Hg.): *Advances in Experimental Social Psychology*. Volume 67. San Diego: Academic Press, S. 131–208.
- PHILIPP, MAIK (2022): Taxiierte Schreibaufgaben. Die revidierte Bloom-Taxonomie als Analyse- und Planungswerkzeug für das Lernen mit dem Schreiben. In: *Lese-forum* (2), S. 1–24.
- PHILIPP, MAIK (2023): Lesekompetenz – neue prozessuale Herausforderungen im kritisch-evaluativen Lesen und ihre Folgen für das Kompetenzkonstrukt. In: *Lese-räume* 10 (9), S. 1–25.
- PHILIPP, MAIK (2025A): *Digitales Lesen fördern*. Göttingen: Vandenhoeck & Ruprecht.
- PHILIPP, MAIK (2025B): *Lesen digital. Komponenten und Prozesse einer sich wandelnden Kompetenz*. Weinheim: Beltz Juventa.
- PHILIPP, MAIK (im Druck): „ChatGPT kann Fehler machen. Überprüfe wichtige Informationen.“ Schreiben mit epistemischer Funktion in Zeiten der KI – Implikationen aus der Sicht des Lesens. In: *Mitteilungen des Deutschen Germanistenverbandes*.
- RICHTER, TOBIAS/MAIER, JOHANNA (2017): Comprehension of Multiple Documents with Conflicting Information. A Two-Step Model of Validation. In: *Educational Psychologist* 52 (3), S. 148–166.
- SALMERÓN, LADISLAO u. a. (2018): Comprehension Processes in Digital Reading. In: Barzillai, Mirit u. a. (Hg.): *Learning to Read in a Digital World*. Amsterdam: John Benjamins, S. 91–120.
- SCHRAW, GREGORY J./ROBINSON, DANIEL R. (2011): Conceptualizing and Assessing Higher Order Thinking Skills. In: Schraw, Gregory J./Robinson, Daniel R. (Hg.): *Assessment of Higher Order Thinking Skills*. Charlotte: Information Age, S. 1–15.
- STADTLER, MARC/BROMME, RAINER/ROUET, JEAN-FRANÇOIS (2018): Learning from Multiple Documents. How Can We Foster Multiple Document Literacy Skills in a Sustainable Way? In: Manalo, Emmanuel/Uesaka, Yuri/Chinn, Clark A. (Hg.): *Promoting Spontaneous Use of Learning and Reasoning Strategies. Theory, Research, and Practice for Effective Transfer*. Singapore: Routledge, S. 46–61.

Die didaktischen Fragen, die KI aus Sicht des Lesens zum Zweck des Lernens aufwirft

- STADTLER, MARC u. a. (2017): Selektion, Integration und Evaluation. Wie wir das Internet nutzen, wenn wir uns über Wissenschaft informieren wollen. In: *Psychologische Rundschau* 68 (3), S. 177–181.
- TASKAN, BURCU/JUNÇA-SILVA, ANA/CAETANO, ANTÓNIO (2022): Clarifying the Conceptual Map of VUCA: A Systematic Review. In: *International Journal of Organizational Analysis* 30 (7), S. 196–217.
- VELÁSQUEZ-HENAO, JUAN DAVID/FRANCO-CARDONA, CARLOS JAIME/CADAVID-HIGUITA, LORENA (2023): Prompt Engineering: A Methodology for Optimizing Interactions with AI-Language Models in the Field of Engineering. In: *DYNA* 90 (230), S. 9–17.
- WINNE, PHILIP H. (2015): Self-Regulated Learning. In: Wright, James D. (Hg.): *International Encyclopedia of the Social & Behavioral Sciences*. Oxford: Elsevier, S. 535–540.



## **Didaktische Agenten**

### KI als Lehr-/Lernpartnerin im Deutschunterricht im Forschungsprojekt KIMADU

#### **Abstract**

Gegenstand des vorliegenden Aufsatzes ist das Forschungsprojekt KIMADU (Künstliche Intelligenz im Mathematik- und Deutschunterricht). Es wird vom Land NRW gefördert und von der Mathematikdidaktik und Sprachdidaktik der Universität Siegen geleitet. Das Projekt hat zum Ziel, professionelle Kompetenzen von Lehrkräften zum Einsatz generativer KI in der Sekundarstufe I zu fördern. In Zusammenarbeit mit 25 weiterführenden Schulen in NRW werden fachspezifische „KI-Lernarrangements“ erarbeitet, erprobt und evaluiert. Im Mittelpunkt dieser Lernarrangements steht die Entwicklung und Nutzung „didaktischer Agenten“, d. h. KIs für spezifische Lehr- und Lernprozesse. Wir nehmen an, dass solche Agenten ein hohes Potenzial für einen lernförderlichen Fachunterricht bergen. Im vorliegenden Aufsatz werden mit Bezug zur Sprachdidaktik Grundlagen des Projekts, das Projektdesign und erste Ergebnisse zu Überzeugungen von Lehrkräften zum Einsatz von KI vorgestellt.

**Schlagwörter:** KIMADU, Professionalisierung, KI-Rollen, didaktische Agenten, Überzeugungen

## **1 Einleitung**

Die rasche technologische Entwicklung und gesellschaftliche Etablierung generativer KI (im Folgenden: KI) wirft die Frage auf, wie Lehrkräfte darin

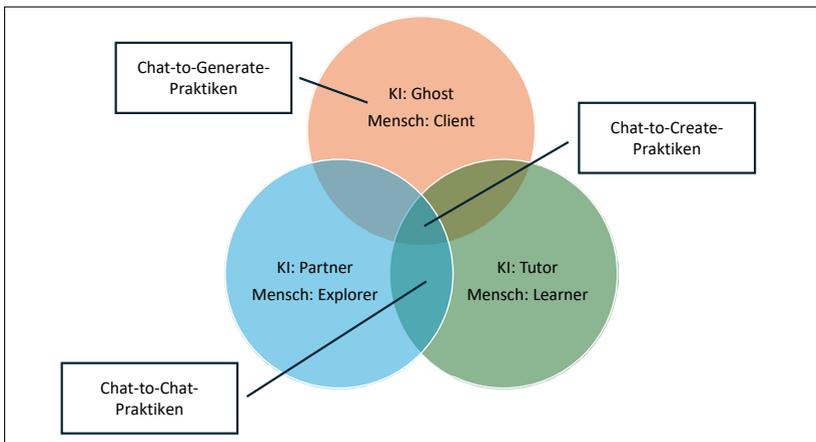
unterstützt werden können, professionelle Kompetenzen zu erwerben, die es ihnen erlauben, KI didaktisch sinnvoll zu nutzen. Darauf wird mit dem Forschungsprojekt KIMADU (Künstliche Intelligenz im Mathematik- und Deutschunterricht) reagiert (vgl. kimadu.de). Das Projekt hat im September 2024 begonnen und endet im Dezember 2027. Es wird vom Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen gefördert und von der Mathematikdidaktik und Sprachdidaktik der Universität Siegen geleitet. In Zusammenarbeit mit 25 weiterführenden Schulen in NRW werden fachspezifische „KI-Lernarrangements“, d. h. komplexe fachliche Lernaufgaben, erarbeitet, erprobt und evaluiert, in deren Mittelpunkt „didaktische Agenten“ stehen, d. h. KIs für spezifische Lehr- und Lernprozesse.

Im Weiteren wird zunächst beschrieben, dass KI in der Kommunikation eine „Ghost-“, „Partner-“ oder „Tutor“-Rolle einnehmen kann (Abschnitt 2), und herausgestellt, dass bildungspolitisch und didaktisch derzeit die Partner-Rolle priorisiert wird (Abschnitt 3). Anschließend wird dargelegt, dass sich für die Realisierung der Partner-Rolle die Entwicklung spezifischer didaktischer Agenten anbietet, und es wird das im Projekt entworfene Megaprompting-Modell „KRAFT+“ vorgestellt, mit dem solche Agenten erstellt werden können (Abschnitt 4). Nachfolgend werden ausgewählte Erkenntnisse der Professionalisierungsforschung referiert, die für das Projekt relevant sind (Abschnitte 5 und 6). Dann wird das Design des Projekts erläutert (Abschnitt 7). Es folgen eine Darstellung erster Ergebnisse einer Fragebogenerhebung zu Überzeugungen von Lehrkräften zu KI (Abschnitt 8) und ein Fazit (Abschnitt 9). Der Fokus der Ausführungen liegt auf dem Fach Deutsch.

## 2 KI als Ghost, Partner und Tutor

Kommunikation mit KI ist mehr als nur Prompten („A prompt [...] is a set of instructions provided to an LLM that programs the LLM by customizing it and/or enhancing or refining its capabilities“, vgl. White et al. 2023: Kap. 1). Wenn man sie darauf beschränkt, geht man ähnlich reduktionistisch wie in der klassischen Sprechakttheorie vor (Searle 1969), wo die Analyse sprachlicher Phänomene auf die Sprecher:innen konzentriert ist. Eine Alternative besteht

darin, sich an der soziologischen Praxistheorie und der linguistischen Konversationsanalyse zu orientieren (vgl. Hirschauer 2016, Hector 2022). Dort fokussiert man auf die sprachliche Interaktion und bezieht dabei mittlerweile auch Praktiken ein, in die Technologien wie Smart Speaker involviert sind. Daran und an ein Modell von Steinhoff (2025) schließt das GPT-Modell von Steinhoff/Lehnen (im Druck) an, auf das wir uns im Weiteren stützen werden (s. Abb. 1). Es beschreibt die Kommunikation mit KI mit Bezug zu verschiedenen Rollenkonstellationen und Chatpraktiken, die sich überschneiden können.



**Abb. 1:** GPT-Modell (Steinhoff/Lehnen 2025: 4)

In Ghost-Client-Konstellationen wird die KI beauftragt, ein Produkt zu generieren, z. B. einen kompletten Text. In Partner-Explorer-Konstellationen ist die KI in die Verarbeitung neuer persönlicher Erfahrungen und Ideen einbezogen, z. B. beim Planen eines Textes. Und in Tutor-Learner-Konstellationen wirkt die KI unterstützend und beratend, z. B. durch Feedback zu einem Text. In Chat-to-Generate-Praktiken und Chat-to-Create-Praktiken reichen die Aktivitäten über den Chat hinaus. Es geht nicht um den Chat selbst, sondern beispielsweise um die Produktion eines Textes oder einer Präsentation. Chat-to-Generate-Praktiken ereignen sich in Ghost-Client-Konstellationen. Chat-to-Create-Praktiken finden hingegen in allen drei Rollenkonstellationen

statt, und zwar dann, wenn der Mensch gestaltend agiert. In Chat-to-Chat-Praktiken schließlich geht es um den Chat selbst – in Fortsetzung von Praktiken, wie sie z. B. in WhatsApp stattfinden. Dies kann in Partner-Explorer- oder Tutor-Learner-Konstellationen geschehen. Das Modell zeigt, dass Kommunikation mit KI deutlich über das Prompten hinausgeht. Sie betrifft eine Vielzahl von Praktiken, in denen Menschen in unterschiedlichen Arten und Weisen mit der KI schreiben, mit ihr sprechen und ihren Output lesen bzw. hören. Kommunikation mit KI ist mithin für ganz unterschiedliche Ebenen des Deutschunterrichts relevant (vgl. Fürstenberg/Müller 2024: 8–11).

### 3 KI als Lehr-/Lernpartnerin

Das GPT-Modell beschreibt kommunikative Praktiken mit KI, bewertet aber nicht, welche Rollenkonstellationen und Chatpraktiken didaktisch aussichtsreich sind. Im Paradigma der Kompetenzorientierung hängt eine solche Bewertung von den Kompetenzzielen ab. Besteht das Kompetenzziel darin, eine KI (versiert) zu beauftragen, ein Produkt zu generieren, sollten Schüler:innen Gelegenheit erhalten, mit ihr in Ghost-Client-Konstellationen und Chat-to-Generate-Praktiken zu kommunizieren. Besteht das Kompetenzziel hingegen darin, mit der KI ähnlich wie in einem Dialog unter Menschen (versiert) zu kommunizieren, sollten die Schüler:innen die Möglichkeit haben, sich in Partner-Explorer-/Tutor-Learner-Konstellationen und Chat-to-Chat-Praktiken mit der KI auszutauschen. Aktuelle bildungspolitische Publikationen lassen erkennen, dass die Kompetenzziele kein ‚progressives‘ Gepräge aufweisen, also weder eine ‚Generierungskompetenz‘ eine ‚Chatkompetenz‘ angestrebt wird, sondern ein ‚konservatives‘ Gepräge erkennen lassen, also Partner-Explorer-/Tutor-Learner-Konstellationen und Ghost-to-Create-Praktiken priorisiert werden.

So betont die KMK (2024: 3) in ihrer „Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in Bildungsprozessen“, dass Lehrkräfte KI zur „Gestaltung von Lehr-/Lernsituationen“ und Schüler:innen KI als „persönlichen Lerntutor, als Intelligentes Tutorielles System oder adaptive Lernumgebung“ nutzen können sollten. Und die

SWK (2023: 12) hebt in ihrem Impulspapier „Large Language Models und ihre Potenziale im Bildungssystem“ exemplarisch hervor, dass KI-Anwendungen Lehrkräfte „bei der summativen und formativen Rückmeldung entlasten [können], indem sie Feedback zu Rechtschreibung, Grammatik, Textkohärenz, Argumentationsstruktur und inhaltlicher Qualität von Texten geben.“ KI soll also in den Dienst von Kompetenzziele gestellt werden, die schon vor der Veröffentlichung von ChatGPT verfolgt wurden. Es wird allerdings kontrovers diskutiert, welche Kompetenzen vorrangig sind. Während die KMK (2024: 3) eine KI-gestützte Förderung von Basiskompetenzen empfiehlt, spricht sich die SWK (2023: 18) für einen KI-Einsatz zur Unterstützung von Lese- und Schreibprozessen fortgeschrittener Lernender aus.

Akzeptiert man diesen bildungspolitischen Status quo und berücksichtigt zudem, dass Partner-Explorer-Konstellationen die höchste kognitive Aktivierung der Schüler:innen versprechen (Fürstenberg/Müller 2024: 7), kann aktuell das übergreifende Kompetenzziel für den Einsatz von KI darin gesehen werden, dass Lehrkräfte KI als Lehrpartner und Schüler:innen KI als Lernpartner nutzen. Die Kommunikation mit KI in Tutor-Learner-Konstellationen kann als Weg zu diesem Ziel verstanden werden. Um das Ziel zu erreichen und den Weg zu beschreiten, können generische KIs oder „didaktische Agenten“ genutzt werden.

## 4 Generische KIs und didaktische Agenten

Generische KIs wie ChatGPT, Claude und Gemini sind kommunikativ außerordentlich offen, weil sie alle möglichen Varianten der beschriebenen Rollenkonstellationen und Chatpraktiken zulassen. Außerdem agieren sie i. d. R. sehr gefällig: Sie passen sich an den Menschen an und vermeiden Kritik. Hinzu kommt, dass ihr Output in jüngerer Zeit zwar ‚dialogischer‘ geworden ist, weil sie z. B. vermehrt Rückfragen stellen, aber weiterhin nicht an transparenten didaktischen Prinzipien orientiert ist. Es ist wichtig, dass Schüler:innen lernen, mit diesen KIs umzugehen, aber auch ersichtlich, dass deren Offenheit, Gefälligkeit und Didaktikferne Lernprozesse erschweren kann (vgl. Bodora et al. 2025).

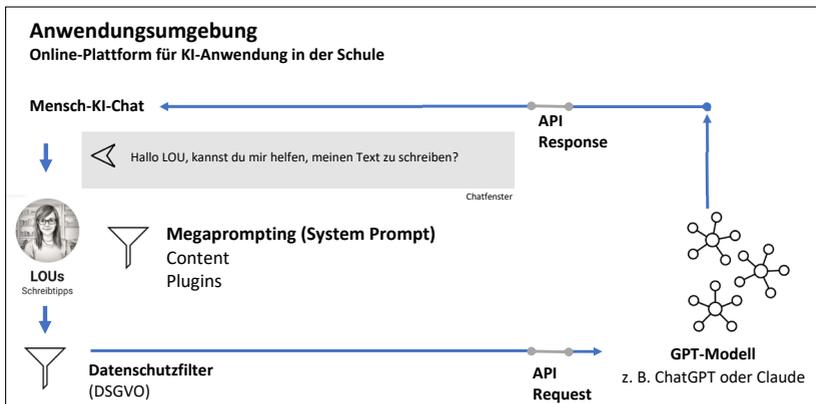
Der Umgang mit generischen KIs kann Schüler:innen unterfordern, weil sie ihre Aufgaben einfach an die KI abgeben können (Chat-to-Generate). Wenn sie die Aufgaben selbst bearbeiten, kann der Umgang aber auch überfordern, weil sie die KI dann weitgehend selbst steuern müssen und dafür relativ hohe kognitive und sprachliche Kompetenzen benötigen (Chat-to-Create). Produktionsseitig müssen sie gute Prompts formulieren können und rezeptionsseitig über gute Lese- bzw. Zuhörkompetenzen verfügen. Um diesem Problem im Unterricht zu begegnen, können Prompt Templates (vgl. Schulhoff et al. 2025: 5), d. h. Schablonen für die Formulierung von Prompts, und Verfahren zur Vermittlung bestimmter Lese- bzw. Zuhörstrategien eingesetzt werden. Ein solches Vorgehen bietet sich allerdings v. a. für fortgeschrittene Lernende an. Allein die kompetente Nutzung von Prompt Templates ist eine komplexe kognitiv-sprachliche Herausforderung, weil sie hohe Transfer- und Reflexionsleistungen erfordert.

Eine alternative didaktische Herangehensweise besteht in der Entwicklung „didaktischer Agenten“.<sup>1</sup> Sie stehen in der Tradition digitaler „pädagogischer Agenten“, die seit den 1990er-Jahren in Form von Figuren, Icons, Stimmen und dergleichen eingesetzt werden (vgl. Beege 2024). In der mathematikdidaktischen Forschung gibt es bereits Versuche, für die Unterrichtsvorbereitung mehrere didaktische Agenten zu „multi-professional teams“ zusammenzuschließen (vgl. Schorcht/Peters/Kriegel 2024). Wir verstehen unter didaktischen Agenten KIs, die von Lehrenden für bestimmte Aufgaben konfiguriert werden und dann in deren Auftrag mit den Lernenden kommunizieren. Die Lehrenden legen also für eine bestimmte Lerngruppe und ein bestimmtes Kompetenzspektrum fest, welche Rollenkonstellationen und Chatpraktiken möglich sind und welche Lernprozesse so initiiert und unterstützt werden sollen. Der didaktische Fokus liegt hier nicht auf dem ‚Bedienen‘ des GPTs mit Prompts, sondern auf dem ‚Interagieren‘ mit der KI im Chat.

.....

1 Die Frage, ob KI-Anwendungen gegendert werden sollten, ist aufgrund der Vielfalt möglicher Perspektiven und Vielschichtigkeit möglicher Argumente schwer zu beantworten. Wir möchten das hier verwendete generische Maskulinum als geschlechtsneutral verstanden wissen. Wenn ein KI-Agent männlich oder weiblich konnotiert anthropomorphisiert wird, sprechen wir von einem „Lehr-/Lernpartner“ bzw. einer „Lehr-/Lernpartnerin“.

Gegenwärtig sind die Entwicklung und der Einsatz didaktischer Agenten noch vergleichsweise anforderungsreich. Lehrkräfte benötigen zwar i. d. R. keine Programmierkenntnisse, aber einen stabilen Zugriff auf geeignete Sprachmodelle mit der Möglichkeit, solche Agenten zu erstellen, eine lokale oder cloudbasierte Umgebung, in der die KIs ausgeführt und verwaltet werden können, Mechanismen zum Schutz persönlicher Daten (DSGVO-Konformität) einschließlich einer Nutzer- und Rechteverwaltung sowie eine geeignete Supportstruktur. Außerdem sollten sie deren technische Funktionsweise verstehen (Abb. 2).



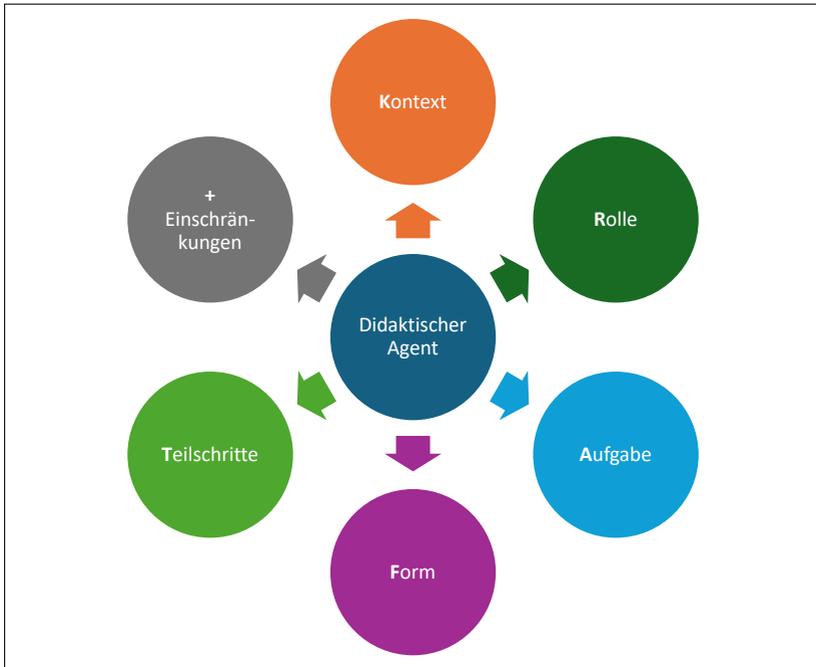
**Abb. 2:** Architektur einer Anwendungsumgebung für einen didaktischen Agenten

Die Abbildung zeigt die Architektur einer Anwendungsumgebung, wie sie beispielsweise bei Plattformen wie Sidekick Education oder fobizz eingesetzt wird. Die Kommunikation mit der KI erfolgt über ein Chatfenster. Um didaktische Agenten zu entwickeln, wird ein Megaprompt erstellt, d.h. ein besonders umfangreicher und stark strukturierter Prompt. Der Megaprompt dient – zusätzlich zum primären Systemprompt der Modellentwickler:innen – als sekundärer Systemprompt, der automatisch jeder Kommunikation mit der KI vorangestellt wird. Er richtet den didaktischen Agenten auf eine spezifische Aufgabe aus und ist für die Nutzer:innen nicht sichtbar. Zusätzlich ist es oft möglich, Dateien hochzuladen, auf die der Agent zugreifen kann (etwa PDFs), und Plugins auszuwählen, z. B. eine Echtzeitsuche im Internet.

Der Megaprompt wirkt wie ein Filter, der die Antworten des Agenten entsprechend den Eingaben steuert. Ein weiterer Filter sorgt dafür, dass alle personenbezogenen Daten aus den Anfragen entfernt oder anonymisiert werden, bevor sie weiterverarbeitet werden. Dieser Datenschutzfilter ist besonders wichtig, um in den Schulen eine DSGVO-konforme Nutzung zu gewährleisten. Die vorgefilterte Anfrage (Request) wird anschließend über eine Programmierschnittstelle (API – Application Programming Interface) an eine generische KI weitergeleitet, z. B. ein GPT-Modell (Generative Pre-trained Transformer). Das gewählte Modell verarbeitet die Anfrage und übermittelt eine Antwort zurück an die Anwendungsumgebung (Response). Diese Antwort wird im Chatfenster für die Nutzer:innen als Reaktion des didaktischen Agenten sichtbar. Dieser Prozess wiederholt sich bei jeder neuen Anfrage und bildet die technische Grundlage für didaktische Agenten, denen mit einem Bild auch visuelle Attribute zugeschrieben werden können (z. B. Kopf eines Avatars).

Auf KI-gestützten Lernplattformen finden sich mittlerweile zahlreiche didaktische Agenten für Lehrer:innen und Schüler:innen. Sie helfen z. B., Elternbriefe zu formulieren, schulinterne Lehrpläne zu erstellen, Klausuren vorzubereiten oder sich auf Vorstellungsgespräche vorzubereiten. Es ist aber oft unklar, nach welchen Kriterien sie erstellt worden sind und welchen Nutzen sie für fachspezifische Lehr- und Lernprozesse haben.

Hier setzt KIMADU an – mit der Entwicklung theoriegeleiteter und evidenzbasierter „KI-Lernarrangements“, d. h. komplexer Lernaufgaben, in die didaktische Agenten sinnvoll integriert sind. Dabei schließen wir u. a. an das in verschiedenen empirischen Studien genutzte Konzept der „Schreibarrangements“ (Steinhoff 2018) an. Um für solche Lernarrangements didaktische Agenten zu erstellen, pilotieren wir zurzeit das Megaprompt-Template-Modell „KRAFT+“, das wir vor dem Hintergrund einschlägiger Publikationen zu Megaprompts im Allgemeinen (vgl. u. a. Schulhoff et al. 2025) und Megaprompts in didaktischen Kontexten im Besonderen (vgl. u. a. Karpouzis et al. 2024) entwickelt haben. Das Modell soll Orientierung für die Formulierung effektiver Megaprompts für didaktische Agenten stiften. Es führt insgesamt sechs fächer- und aufgabenübergreifend relevante Kategorien für Megaprompts zusammen: „Kontext“, „Rolle“, „Aufgabe“, „Form“, „Teilschritte“ und „+ Einschränkungen“ (s. Abb. 3).



**Abb. 3:** Megaprompting-Modell KRAFT+

Die Kategorien lassen sich mit den folgenden sechs Fragen konkretisieren:

Kontext	In welchem Lehr-/Lernsetting soll der Agent eingesetzt werden?
Rolle	Welche sozialen Eigenschaften soll der Agent simulieren?
Aufgabe	In welcher Lernaufgabe soll der Agent welche Funktionen erfüllen?
Form	Welches Format und welchen Stil soll der Output des Agenten haben?
Teilschritte	Auf welchen Chatverlauf soll der Agent hinwirken?
+ Einschränkungen	Was soll der Agent unterlassen?

**Tab. 1:** Leitfragen zur Entwicklung von didaktischen Agenten nach KRAFT+

Das Lehr-/Lernpotenzial eines didaktischen Agenten hängt davon ab, wie gut die Instruktionen sind, die zu diesen Fragen formuliert werden, und wie intensiv die Agenten getestet werden. Falls es technisch möglich ist, kann zudem z. B. eine bestimmte, zum jeweiligen Agenten besonders gut passende generische KI ausgewählt oder auch die ‚Wissensbasis‘ des Agenten durch die Nutzung geeigneter Daten wie Curricula oder digitale Fachtexte spezifiziert werden. Wenn für didaktische Agenten gilt, was empirische Studien zu pädagogischen Agenten zeigen, kommt es hier auf jedes Detail an (vgl. Schroeder/Davis/Yang 2024).

Im KIMADU-Projekt wird derzeit u. a. der Einsatz des nach dem KRAFT+-Modell konfigurierten didaktischen Agenten „LOU, die Lernpartnerin“ erprobt. Ziel dieses Agenten ist es, Schüler:innen der Sekundarstufe I beim Verstehen und Bearbeiten von Schreibaufgaben zu unterstützen. In Tabelle 2 werden exemplarisch Auszüge aus der Kommunikation mit LOU im Vergleich zur Kommunikation mit der generischen KI Claude Sonnet im gleichen Aufgabenkontext zitiert.

Der letzte Output der KI in beiden Spalten verdeutlicht, dass bei der Kommunikation mit der generischen KI eher eine Ghost-Rolle zu erkennen ist und bei der Kommunikation mit dem didaktischen Agenten eher eine Partner-Rolle. Bezieht man diese Beobachtung auf das oben explizierte übergreifende Kompetenzziel, KI als Lehr-/Lernpartner zu nutzen, lässt der didaktische Agent ein größeres Lehr-/Lernpotenzial erkennen. KIMADU hat zum Ziel, professionelle Kompetenzen zu fördern, die es Lehrkräften ermöglichen, KI in dieser Weise fach- und aufgabenspezifisch im Unterricht einzusetzen. Dabei wird von Erkenntnissen zu professionellen Kompetenzen von Lehrkräften und effektiven Maßnahmen zur Förderung dieser Kompetenzen ausgegangen.

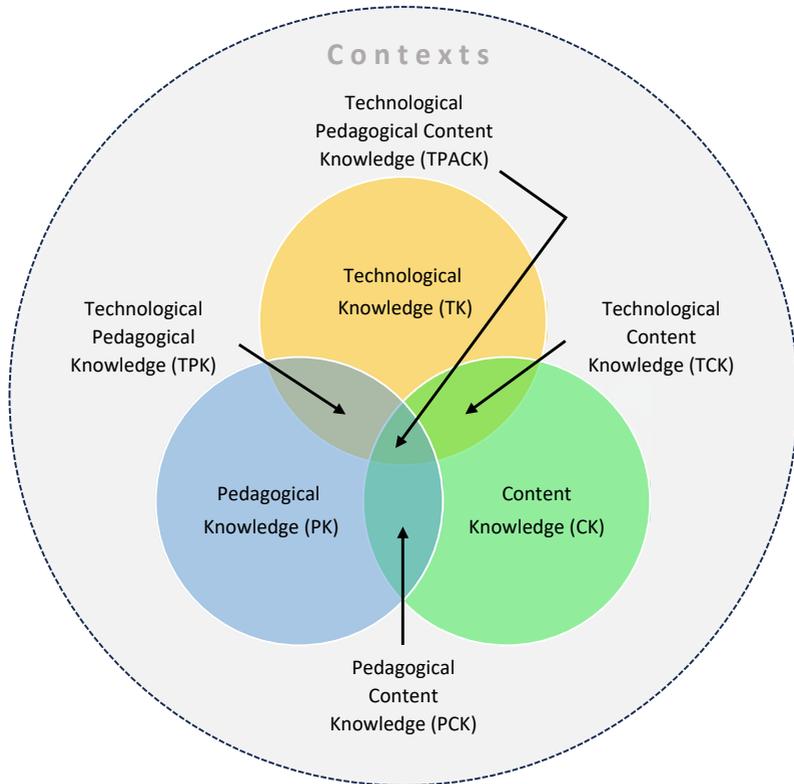
<b>Anwendung</b>	<b>Generische KI – Claude Sonnet 3.7</b>	<b>Didaktische Agentin – LOU, die Lernpartnerin</b>
Mensch	Hallo!	Hallo LOU!
KI	Hallo. Wie kann ich dir heute helfen?	Hallo! Schön, dass du da bist. Welche Schreibaufgabe möchtest du heute mit mir lösen?

Anwendung	Generische KI – Claude Sonnet 3.7	Didaktische Agentin – LOU, die Lernpartnerin
Mensch	„Nimm Stellung zu folgender Aussage: Beauty Filter sollten in sozialen Medien genutzt werden“	„Nimm Stellung zu folgender Aussage: Beauty Filter sollten in sozialen Medien genutzt werden“
KI	„Stellungnahme zu Thema ‚Beauty Filter sollten in sozialen Medien genutzt werden.‘ Gerne helfe ich dir bei der Strukturierung zu diesem aktuellen Thema. Ich erstelle dir zunächst eine Übersicht mit verschiedenen Argumenten für beide Seiten und dann ein Gerüst für deine Stellungnahme.“	Danke für deine Aufgabe. Du sollst also zu Beauty Filtern Stellung nehmen. Findest du Beauty Filter eher gut oder problematisch? Für deine Stellungnahme könntest du mit Redemitteln wie „Meiner Meinung nach ...“ oder „Ich vertrete die Position, dass ...“ arbeiten.

**Tab. 2:** Auszug aus einem Mensch-KI-Chat mit der generischen KI Claude Sonnet 3.7 (links) und der didaktischen Agentin LOU (rechts) (Februar 2025)

## 5 Professionelle Kompetenzen von Lehrkräften nach dem TPACK-Modell

Der Einfluss der Lehrkraft auf die Unterrichtsqualität und die Schülerleistung ist unumstritten (Hattie 2013). Deshalb steht im KIMADU-Projekt das professionelle Unterrichtshandeln von Lehrkräften im Vordergrund des Forschungsinteresses. Professionelle Kompetenzen von Lehrkräften werden oft anhand von Kompetenzdimensionen beschrieben, z. B. kognitiven Dimensionen (Wissen) und affektiv-motivationalen Dimensionen (Überzeugungen, Motivation, Selbstregulation) (vgl. Baumert/Kunter 2011). Was die Beschreibung der kognitiven Dimensionen angeht, hat sich bei technologisch gestützten Unterrichtsszenarien das TPACK-Modell (Technological Pedagogical and Content Knowledge) nach Koehler/Mishra (2009) bewährt, das die Wissensbereiche aus Shulman (1986) zugrunde legt. Das Modell ist nicht auf eine bestimmte Technologie beschränkt, sondern beschreibt allgemein die Verzahnung von technologischem, fachspezifischem und pädagogischem Wissen (s. Abb. 4).



**Abb. 4:** The TPACK framework and its knowledge components – eigene Darstellung nach Koehler/Mishra (2009: 63)

Das technologische Wissen (TK) bezieht sich auf das Verständnis und die Nutzung von Technologien. Das Fachwissen (CK) betrifft das Wissen über die Inhalte des zu unterrichtenden Fachs. Das fachdidaktische Wissen (PCK) bezieht sich auf das Wissen über Aufgaben (Anforderung, Sequenzierung, Arrangement u. a.), über Schülervorstellungen (typische Lernwege oder Fehlkonzepte, Diagnose, Lernunterstützung u. a.) und über geeignete Erklärungs- und Übungsansätze (Veranschaulichung, Übungsformate u. a.). Und das pädagogische Wissen (PK) bezieht sich auf das Klassenmanagement, die Unterrichtsplanung, die Lernmotivation und weitere, das Lernen im Allgemeinen betreffende Aspekte. Das technologische Wissen (TK) findet im Modell inte-

griert Anwendung – als technologisches Fachwissen (TCK), technologisches fachdidaktisches Wissen (TPCK) und technologisches pädagogisches Wissen (TPK). Die gemeinsame Schnittmenge dieser Wissensdimensionen verdeutlicht das Professionalisierungsziel: ein verzahntes technologisch-fachwissenschaftlich-fachdidaktisch-pädagogisches Wissen (TPACK).

Das TPACK-Modell erlaubt es, die Wissensdimensionen (CK, PCK, PK) für den Einsatz bestimmter Technologien (T) zu spezifizieren – auch und gerade für KI: „Changes in technology push us to reconsider how we think about the content to be taught as well as the pedagogical approach that is most appropriate. In this new context, it is critical that we recognize how new GenAI tools are fundamentally different than the technologies that have come before (analog or digital)“ (Mishra/Warr/Islam 2023: 239). Zum technologischen professionellen Wissen von Lehrkräften (vgl. KMK 2016, Rubach/Lazarides 2019) ist nun auch mit Fachwissen, fachdidaktischen und pädagogischem Wissen verknüpftes Wissen über KI zu zählen, das eine zielgerichtete Anwendung und einen reflektierten Umgang dieser Technologie im Unterricht erlaubt.

Hinzu kommen die affektiv-motivationalen Komponenten professioneller Kompetenzen: die Überzeugungen, die Motivation und die Selbstregulation (vgl. Baumert/Kunter 2011). Der Umstand, dass sich gerade diese Kompetenzen i. d. R. über einen vergleichsweise langen Zeitraum entwickeln, verdeutlicht die Dringlichkeit, auf die schnelle Etablierung von KI rasch mit einer umfassenden Professionalisierungsoffensive und einer entsprechenden Begleitforschung zu reagieren. Eine solche Offensive kann wesentlich dazu beitragen, dass Lehrkräfte kognitive und affektiv-motivationale Kompetenzen im Bereich KI erwerben, die es ihnen erlauben, KI kompetent im Unterricht einzusetzen und ihren Schüler:innen entsprechende Lernzuwächse zu ermöglichen.

## 6 Effektive Professionalisierung von Lehrkräften

Effektive Lehrerfortbildungen sind dadurch gekennzeichnet, dass sie sich über einen längeren Zeitraum erstrecken, einen engen fachlichen Fokus haben, eine Verknüpfung mit der Praxis herstellen, Gelegenheiten zur Erprobung der Fortbildungsinhalte einplanen, Möglichkeiten zum Erleben der eigenen Wirksam-

keit eröffnen, Feedback für Lehrpersonen durch Coaching bereitstellen, eine Verankerung in der Schule begünstigen, sich an Merkmalen lernwirksamen Unterrichts orientieren (insbesondere kognitive Aktivierung und adaptives Feedback) und die Zusammenarbeit mit Kolleg:innen intensivieren (vgl. Lipowsky 2021, Lipowsky/Rzejak 2017).

In jüngerer Zeit wurden verschiedene Transferprojekte durchgeführt, in denen solche Professionalisierungsformate durch eine Stärkung der Zusammenarbeit zwischen Forschung, Schuladministration und Schule ermöglicht wurden (vgl. Holtappels 2019). Im Bereich Sprachbildung ist v. a. die bundesweite Initiative BiSS-Transfer zu nennen (vgl. Becker-Mrotzek et al. 2024). Solche Transferprojekte nutzen bereits existierende Professionalisierungsstrukturen. Neue Konzepte werden breit gestreut, z. B. über die Schulämter an eine Vielzahl von Schulen, und die Angebote sind für die Lehrkräfte leicht wahrnehmbar und in die Schul- und Unterrichtsentwicklung eingeplant. Die Konzepte werden in engem Austausch von Expert:innen der Universitäten, der Schuladministration und der Schulen in einem Spiralverlauf aus Erprobung, Reflexion, Überarbeitung entwickelt und evaluiert, und die konkreten Erfordernisse des Unterrichts werden von Anfang an mitgedacht. So erhöht sich nicht nur die Qualität, sondern auch die Akzeptanz der Konzepte.

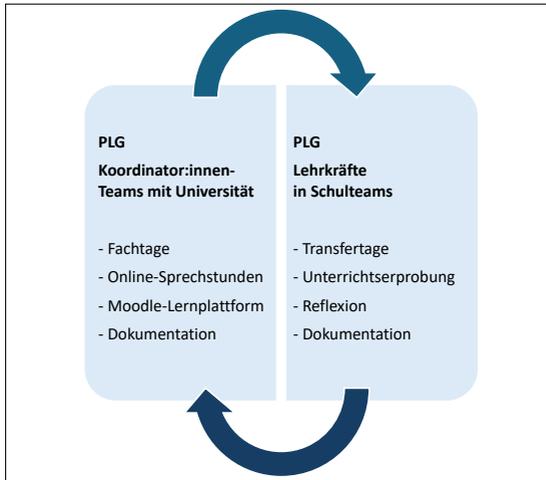
Ein weiteres effektives Format einer nachhaltigen Unterrichts- und Schulentwicklung sind professionelle Lerngemeinschaften (PLG) (vgl. Kansteiner/Welther/Schmid 2023, Wenger/Snyder 2000). Eine PLG zielt auf eine erfolgreiche Unterrichts- und/oder Schulentwicklung durch eine regelmäßige Zusammenarbeit des Kollegiums. Nach Wahlstrom/Louis (2008) zeichnen sich PLGs durch die folgenden Aspekte aus: (1) ein gemeinsames Professionalisierungsinteresse, (2) einen Fokus auf inhaltliches Lernen, (3) den Willen zur Zusammenarbeit mit Kolleg:innen, (4) die Bereitschaft, den eigenen Unterricht zu öffnen und mit Kolleg:innen zu diskutieren, sowie die Bereitschaft (5), einen Beitrag zur regelmäßigen kollektiven Reflexion über Lehr- und Lernprozesse zu leisten. PLG erhöhen die Wahrscheinlichkeit, dass Fachwissen in der Praxis umgesetzt wird. Empirische Ergebnisse aus dem deutschsprachigen Raum zur PLG zeigen positive Auswirkungen von PLG auf (a) das berufliche Wohlbefinden, (b) den fachlichen Kompetenzaufbau, (c) die Intensivierung innovativen Denkens und (d) Veränderungen in der pädagogischen Praxis. Darüber hinaus

wurde weder die zeitliche Beanspruchung noch die Vorgabe, an einer PLG mitzuwirken, von den Lehrkräften als negativ empfunden (vgl. Kansteiner/Stamann/Rist 2020, Kansteiner/Welther/Schmid 2023).

## 7 Anlage des KIMADU-Projekts (Deutsch)

Das KIMADU-Projekt ist ein Transferprojekt, das an den Bedarf von Schulen nach lernförderlichen Konzepten zum Einsatz von KI anschließt. Es erbringt Erkenntnisse zu KI-Lernarrangements und didaktischen Agenten in den Hauptfächern Mathematik und Deutsch. Dafür setzt es auf einen ko-konstruktiven Entwicklungsprozess von Universität, Schuladministration und Schulen. Sein Schwerpunkt liegt auf der Förderung professioneller Kompetenzen von Lehrkräften im Bereich KI mit einem Fokus auf Überzeugungen (Beliefs), die in einem Prä-Post-Design mithilfe von Fragebögen und Leitfaden-Interviews erfasst werden. Es wird mit 25 weiterführenden Schulen in Nordrhein-Westfalen zusammengearbeitet, die sich in einem Bewerbungsverfahren für die Teilnahme qualifiziert haben und dafür u. a. nachweisen mussten, dass sie hohe technische und fachliche Voraussetzungen erfüllen. Es ist anzunehmen, dass sie ein großes Interesse und eine hohe Motivation mitbringen und die Bedingungen für die Umsetzung der Projektziele gut sind.

Das für das Projekt entwickelte Transferkonzept setzt auf einen engen Wissens- und Erfahrungsaustausch zwischen Forscher:innen und Lehrkräften. So werden am Ende ca. 200 Lehrkräfte und 7.000 Schüler:innen erreicht. An jeder Schule fungieren je zwei Lehrer:innen als „Koordinator:innen“ – eine:r für das Fach Mathematik und eine:r für das Fach Deutsch. Sie vermitteln zwischen Forschung und Schule und gehören zwei professionellen Lerngemeinschaften an: fachspezifischen „Koordinator:innen-Teams“, in denen sie mit den anderen Koordinator:innen ihres Fachs in Gruppen standort-, schulform- und jahrgangsstufenübergreifend zusammenarbeiten, und schulinternen „Schulteams“, in denen sie mit ihren Fachkolleg:innen Unterrichtskonzepte entwickeln und erproben. An diese PLGs angeschlossen sind zudem die Medienberater:innen und Digitalisierungsbeauftragten der Schulen. Die Partizipation in den beiden Teams ist an bestimmte Formate gekoppelt (s. Abb. 5).



**Abb. 5:** Transferkonzept des KIMADU-Projekts

Die Koordinator:innen-Teams kommen mit den Forscher:innen der Universität Siegen im Projektverlauf zu zwei fächerübergreifenden und pro Fach zu je vier fachspezifischen, eintägigen „Fachtagen“ zusammen, die von der Qualitäts- und Unterstützungsagentur Nordrhein-Westfalen (QUA-LiS NRW) organisiert und in Präsenz durchgeführt werden. Dort werden u. a. Forschungserkenntnisse diskutiert, Workshops durchgeführt und Erfahrungen, die in den Schulen gemacht wurden, reflektiert. Hinzu kommen regelmäßige Online-Sprechstunden und ein kontinuierlicher Austausch auf einer Moodle-Plattform. Die Resultate der Arbeit in den verschiedenen Formaten werden von den Forscher:innen dokumentiert. Die Schulteams kommen im Rahmen regelmäßiger „Transfertage“ zusammen, erproben die Konzepte im Unterricht und reflektieren und dokumentieren sie.

Darüber hinaus besuchen die Forscher:innen in fächerübergreifenden Teams jede Schule einmal, um mit Schulleitungen, Koordinator:innen, Fachlehrer:innen und Schüler:innen zusammenzukommen, und in Abhängigkeit vom Forschungsprozess weitere Male, um den Unterricht zu beobachten und zu evaluieren. Der Öffentlichkeit werden des Weiteren ausgewählte didaktische Konzepte und Materialien als Open Educational Resources und eine Best-Practice-Handreichung zur Verfügung gestellt.

## 8 Erste Ergebnisse zu Überzeugungen von Lehrkräften zu KI

Im Rahmen von KIMADU wurde ein Fragebogen zu Überzeugungen zu KI im Deutschunterricht entwickelt. Die Items wurden theoriegeleitet festgelegt (vgl. Lindner/Romeike 2019, Schmidt et al. 2009, Velander et al. 2024, Yue et al. 2024). Es wurden folgende Konstrukte erfasst: allgemeine digitale Kompetenzen, Praktikabilität des KI-Einsatzes im Deutschunterricht, persönliche Nutzung von KI, Kommunikation zwischen Schüler:innen und der KI, Erstellen von Aufgaben und Differenzierung mit KI, Lernen und Aufgabenlösen mit KI, Prüfen und Bewerten mit KI sowie Selbsteinschätzung eigener Kompetenzen im Bereich KI-Nutzung.

Dem Fragebogen liegt eine siebenstufige Likert-Skala zugrunde: 0 (stimme überhaupt nicht zu) bis 6 (stimme voll und ganz zu). Die Option, sich enthalten zu können, wurde in Abstimmung mit der Mathematikdidaktik durch eine mittlere Skalenkategorie umgesetzt, die im Fragebogen als „neutral“ bezeichnet ist. Die Stichprobe der Pilotierung umfasste 25 Lehrkräfte ( $N = 25$ ) aus dem ganzen Bundesgebiet, die randomisiert ausgewählt wurden und nicht an KIMADU-Schulen tätig sind. Die Erhebung fand im März 2025 statt und wurde digital durchgeführt.

Erste deskriptive Ergebnisse zeigen, dass die befragten Lehrkräfte eine differenzierte Haltung gegenüber dem Einsatz von generativer KI im Deutschunterricht einnehmen. Die Prompting-Kompetenzen ihrer Schüler:innen schätzen sie als eher gering ein ( $M = 2,16$ ,  $SD = 1,60$ ). Deren Fähigkeit, effektiv mit generativen KI-Lernpartnern zu kommunizieren, bewerten sie eher zurückhaltend ( $M = 2,36$ ,  $SD = 1,70$ ). Zurückhaltend bis neutral bewerten sie zudem das Potenzial von KI zur Unterstützung des Leseverstehens ihrer Schüler:innen ( $M = 2,84$ ,  $SD = 1,49$ ). Besonders auffällig ist ihre Einschätzung des Ablenkungspotenzials von KI im Deutschunterricht ( $M = 4,84$ ,  $SD = 0,99$ ). Sie zeigen sich besorgt über mögliche Beeinträchtigungen der Konzentration und Aufmerksamkeit der Schüler:innen. Positiv sehen sie hingegen das Potenzial von KI, Schüler:innen bei der Textplanung, -formulierung und -überarbeitung zu unterstützen: Hier stimmen sie eher zu ( $M = 4,44$ ,  $SD = 1,29$ ). Eher positiv

ist zudem ihre Haltung zur Nutzung von KI für die eigene Unterrichtsvorbereitung ( $M = 3,92$ ,  $SD = 1,90$ ).

Die Ergebnisse lassen erkennen, dass die befragten Lehrkräfte die schülerseitigen Voraussetzungen für einen effektiven Einsatz von KI eher kritisch einschätzen. Hier geht es letztlich um die für einen kompetenzorientierten Unterricht mit KI zentrale und noch zu wenig diskutierte Frage, was Schüler:innen können müssen, um KI lernförderlich zu nutzen. Daran wiederum schließt sich die Frage an, wie KI passgenau auf dieses Können abgestimmt werden kann, z. B. durch didaktische Agenten. Gleichzeitig identifizieren die Lehrkräfte aber auch Bereiche, in denen sie den Einsatz von KI als sinnvoll einschätzen: für die Förderung der Schreibkompetenzen ihrer Schüler:innen und für ihre eigene Unterrichtsvorbereitung.

## 9 Fazit

Kommunikation mit KI findet in ganz unterschiedlichen Rollenkonstellationen (Ghost/Client, Partner/Explorer, Tutor/Learner) und Chatpraktiken (Chat-to-Generate, Chat-to-Create, Chat-to-Chat) statt, und es ist offen, welche Konstellationen und Praktiken didaktisch besonders aussichtsreich sind. Aktuelle Curricula und fachdidaktische Erwägungen legen eine Nutzung von KI als Partner in Chat-to-Create-Praktiken nahe – als Lehrpartner für Lehrkräfte und Lernpartner für Schüler:innen. Um KI so einzusetzen, benötigen Lehrkräfte relevantes technologisches, fachliches und pädagogisches Wissen und Überzeugungen, die eine lernförderliche und verantwortungsvolle Anwendung dieses Wissens unterstützen. Im KIMADU-Projekt werden solche Kompetenzen in den Fächern Deutsch und Mathematik entwickelt, indem in engem Austausch mit 25 Schulen KI-Lernarrangements erarbeitet, erprobt und evaluiert werden, in deren Mittelpunkt didaktische Agenten stehen, d. h. KIs für spezifische Lehr- und Lernprozesse. In diesem gemeinschaftlichen Vorgehen sehen wir einen sinnvollen Weg der Professionalisierung für den Einsatz von KI – auch für den Umgang mit berechtigten Zweifeln von Lehrkräften, wie sie in Abschnitt 8 referiert worden sind.

## Literatur

- BAUMERT, JÜRGEN/KUNTER, MAREIKE (2011): Das mathematikspezifische Wissen von Lehrkräften, kognitive Aktivierung im Unterricht und Lernfortschritte von Schülerinnen und Schülern. In: Kunter, Mareike/Baumert, Jürgen/Blum, Werner/Klusmann, Uta/Krauss, Stefan/Neubrand, Michael (Hg.): *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann, S. 163–192.
- BECKER-MROTZEK, MICHAEL/HASSELHORN, MARKUS/ROTH, HANSJOACHIM/STANAT, PETRA (2024): Warum BiSS-Transfer? Beweggründe, Ziele und Grundlagen. In: Trägerkonsortium BiSS-Transfer (Hg.): *BiSS-Transfer: Potenziale sichern. Arbeitsergebnisse, Länderperspektiven und Impulse*. Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache, S. 6–8. <https://www.doi.org/10.3278/9783763978014>.
- BEEGE, MAIK (2024): Pädagogische Agenten. In: Rey, Günter Daniel (Hg.): *Lehren und Lernen mit digitalen Medien. Theorien und Design*. Bern: Hogrefe, S. 89–108.
- BODORA, ANNA LENA/DECKER, LENA/FUHLROTT, MAREIKE/NOLDEN, ABYGAIL/STEINHOFF, TORSTEN (2025): Wie schreiben Schüler:innen in der 8. Klasse mit ChatGPT? Einblicke in das Design-Based-Research-Projekt *KI-Schreibarrangements*. In: *Leseräume* 11, S. 1–4. <https://xn--leserume-4za.de/wp-content/uploads/2025/06/Bodora-et-al-2025-LR-JG12-H11.pdf> (letzter Zugriff: 18.7.2025).
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* 5, S. 2–13.
- HATTIE, JOHN (2013): *Lernen sichtbar machen*. Baltmannsweiler: Schneider Hohengehren.
- HECTOR, TIM (2022): Smart Speaker in der Praxis: Methodologische Überlegungen zur medienlinguistischen Erforschung stationärer Sprachassistenzsysteme. In: *Sprache und Literatur* 51 (2), S. 197–229. <https://doi.org/10.30965/25890859-05002021>.
- HIRSCHAUER, STEFAN (2016): Verhalten, Handeln, Interagieren. Zu den mikrosoziologischen Grundlagen der Praxistheorie. In: Schäfer, Hilmar (Hg.): *Praxistheorie. Ein soziologisches Forschungsprogramm*. Bielefeld: transcript, S. 45–67.

- HOLTAPPELS, HEINZ GÜNTER (2019): Transfer in der Schulentwicklung: Ansätze und Gelingensbedingungen aus der Perspektive von Schulentwicklungstheorie und -forschung. In: *DDS – Die Deutsche Schule* 111 (3), S. 274–293. <https://doi.org/10.31244/dds.2019.03.03>.
- KANSTEINER, KATJA/STAMANN, CHRISTOPH/RIST, MARIA (2020): Merkmale professioneller Lerngemeinschaften. In: Buhren, Claus G./Kansteiner, Katja/Stamann, Christoph/Theurl, Peter (Hg.): *Professionelle Lerngemeinschaften als Entwicklungsinstrument im Bildungswesen*. Weinheim: Juventa S. 16–36.
- KANSTEINER, KATJA/WELTHER, SABINE/SCHMID, SUSANNE (2023): *Professionelle Lerngemeinschaften für Schulleitungen und Lehrkräfte. Chancen des Kooperationsformats für Schulentwicklung und Professionalisierung*. Weinheim: Beltz Juventa.
- KARPOUZIS, KOSTAS/PANTAZATOS, DIMITRIS/TAOUKI, JOANNA/MELI, KALLIOPHI (2024): *Tailoring Education with GenAI: A New Horizon in Lesson Planning*. <https://doi.org/10.48550/arXiv.2403.12071>.
- KOEHLER, MATTHEW J./MISHRA, PUNYA (2009): What Is Technological Pedagogical Content Knowledge? In: *Contemporary Issues in Technology and Teacher Education* 9, S. 60–70.
- Kultusministerkonferenz (KMK) (2016): *Bildung in der digitalen Welt. Strategie der Kultusministerkonferenz*. URL: [https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2017/Strategie\\_neu\\_2017\\_datum\\_1.pdf](https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2017/Strategie_neu_2017_datum_1.pdf) (letzter Zugriff: 18.7.2025).
- Kultusministerkonferenz (KMK) (2024): *Handlungsempfehlung für die Bildungverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen*. URL: [https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2024/2024\\_10\\_10-Handlungsempfehlung-KI.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf) (letzter Zugriff: 18.7.2025).
- LINDNER, ANNABEL/ROMEIKE, RALF (2019): Teachers' Perspective on Artificial Intelligence. *Conference: ISSEP 2019: 12th International conference on informatics in schools. Situation, evaluation and perspectives*. Larnaca, Cyprus. URL: <https://cyprusconferences.org/issep2019/wp-content/uploads/2019/10/LocalISSEP-v5.pdf> (letzter Zugriff: 19.5.2025)
- LIPOWSKY, FRANK (2021): *Fortbildungen für Lehrpersonen wirksam gestalten*. Gütersloh: Bertelsmann.

- LIPOWSKY, FRANK/RZEJAK, DANIELA (2017): Fortbildungen für Lehrkräfte wirksam gestalten – erfolgsverprechende Wege und Konzepte aus Sicht der empirischen Bildungsforschung. In: *Bildung und Erziehung* 70 (4), S. 369–492.
- MISHRA, PUNYA/WARR, MELISSA/ISLAM, REZWANA (2023): TPACK in the age of ChatGPT and Generative AI. In: *Journal of Digital Learning in Teacher Education* 39 (4), S. 235–251.
- RUBACH, CHARLOTT/LAZARIDES, REBECCA (2019): Eine Skala zur Selbsteinschätzung digitaler Kompetenzen bei Lehramtsstudierenden. In: *Zeitschrift für Bildungsforschung* 9, S. 345–374.
- SCHORCHT, SEBASTIAN/PETERS, FRANZISKA/KRIEGEL, JULIAN (2024): Communicative AI Agents in Mathematical Task Design: A Qualitative Study of GPT Network Acting as a Multi-professional Team. In: *Digital Experiences in Mathematics Education*, o. S. <https://doi.org/10.1007/s40751-024-00161-w>.
- SCHROEDER, NOAH L./DAVIS, ROBERT O./YANG, EUNBYUL (2024): Designing and Learning with Pedagogical Agents: An Umbrella Review. In: *Journal of Educational Computing Research* 62 (8), 2127–2156. <https://doi.org/10.1177/07356331241288476>.
- SCHULHOFF, SANDER/ILLIE, MICHAEL/BALEPUR, NISHANT/KAHADZE, KONSTANTINE/LIU, AMANDA/SI, CHENGLEI/LI, YINHENG/GUPTA, AAYUSH/HAN, HYOUNG/JUNG/SCHULHOFF, SEVIEN/DULEPET, PRANAV SANDEEP/VIDYADHARA, SAURAV/KI, DAYEON/AGRAWAL, SWETA/PHAM, CHAU/KROIZ, GERSON/LI, FEILEEN/TAO, HUDSON/SRIVASTAVA, ASHAY/DA COSTA, HEVANDER/GUPTA, SALONI/ROGERS, MEGAN L./GONCEARENCO, INNA/SARLI, GIUSEPPE/GALYNKER, IGOR/PESKOFF, DENIS/CARPUAT, MARINE/WHITE, JULES/ANAKDAT, SHYAMAL/HOYLE, ALEXANDER/RESNIK, PHILIP (2025): *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. <https://doi.org/10.48550/arXiv.2406.06608>.
- SCHMIDT, DENIS A./BARAN EVRIM/THOMPSON, ANNA D./MISHRA, PUNYA/KOELLER, MATTHEW J./SHIN, TAE SEOB (2009): Technological Pedagogical Content Knowledge (TPACK). In: *Journal of Research on Technology in Education* 42 (2), 123–149. URL: <https://doi.org/10.1080/15391523.2009.10782544>.
- SEARLE, J. (1969): *Speech acts. An essay in the philosophy of language*. Cambridge: Cambridge University Press.
- SHULMAN, LEE S. (1986): Those who understand: Knowledge growth in teaching. In: *Educational Researcher* 15 (2), S. 4–31.

- STEINHOFF, TORSTEN (2018): Schreibarrangements. Impulse für einen lernförderlichen Schreibunterricht. In: *Der Deutschunterricht* 3, S. 2–10.
- STEINHOFF, TORSTEN (2025): Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel von ChatGPT. In: Albrecht, Christian/Brüggemann, Jörn/Kretschmann, Tabea/Krommer, Axel/Meier, Christel (Hg.): *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven*. Stuttgart: Metzler, S. 85–99.
- STEINHOFF, TORSTEN/LEHNEN, KATRIN (2025): Schreiben mit Künstlicher Intelligenz: Das GPT-Modell (Ghost, Partner, Tutor). In: *Leseräume* 11, S. 1–14. <https://xn--leserume-4za.de/wp-content/uploads/2025/06/Steinhoff-Lehnen-2025-LR-JG12-H11.pdf> (letzter Zugriff: 18.7.2025).
- Ständige Wissenschaftliche Kommission (SWK) (2023): *Large Language Models und ihre Potenziale im Bildungssystem. Impulspapier der Ständigen Wissenschaftlichen Kommission (SWK) der Kultusministerkonferenz*. <http://dx.doi.org/10.25656/01:28303>.
- VELANDER, JOHANNE/TAIYE, MOHAMMED AHMED/OTERO, NUNO/MILRAD, MARCELO (2024): Artificial Intelligence in K-12 Education: eliciting and reflecting on Swedish teachers' understanding of AI and its implications for teaching & learning. In: *Education and Information Technology* 29, S. 4085–4105. <https://doi.org/10.1007/s10639-023-11990-4>.
- WAHLSTROM, KYLA/LOUIS, KAREN SEASHORE (2008): How teachers perceive principal leadership. In: *Educational Administration Quarterly* 44 (4), S. 498–445.
- WENGER, ETIENNE C./SNYDER, WILLIAM M. (2000): Communities of Practice. In: *Harvard Business Review* 7 (2), S. 139–145.
- WHITE, JULES/FU, QUCHEN/HAYS, SAM/SANDBORN, MICHAEL/OLEA, CARLOS/GILBERT, HENRY/ELNASHAR, ASHRAF/SPENCER-SMITH, JESSE/SCHMIDT, DOUGLAS C. (2023): *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. <https://doi.org/10.48550/arXiv.2302.11382>.
- YUE, MIAO/JONG, MORRIS/TSZ KIT NG, DAVY (2024): Understanding K–12 teachers' technological pedagogical content knowledge readiness and attitudes toward artificial intelligence education. In: *Education and Information Technologies* 29, S. 19505–19536. <https://doi.org/10.1007/s10639-024-12621-2>.

FRIEDRICH BACH / SEBASTIAN BERNHARDT /  
SILVIA REUVEKAMP / NINJA SCHMIEDGEN

## **SHIFT happens**

Lernen mit und von textgenerierender KI

### **Abstract**

Mit dem Einsatz von textgenerierender KI und Large Language Models in Bildungskontexten gehen einige Probleme einher: So wurden ChatGPT und Co. durch gewinnorientierte Unternehmen entwickelt, womit die Bildungseinrichtung in eine ökonomische Abhängigkeit gerät und ihre digitale Souveränität aufgibt. Zudem ist die Datenbasis herkömmlicher textgenerierender KI nicht steuerbar. Entsprechend besteht die Gefahr politischer Beeinflussung, der Perpetuierung von Machtaffirmation und der Falschinformation. An der Universität Münster haben wir uns diesen Herausforderungen mit dem interdisziplinären Projekt *Tutor.AI* gestellt. Dabei handelt es sich um eine textgenerierende KI, die auf universitätseigenen Servern läuft, also die digitale und ökonomische Souveränität sichert und eine Einspeisung von Quellen, Präsentationen und Skripten erlaubt, sodass die Datenbasis selbstständig ergänzt werden kann.

In diesem Beitrag geben wir Einblicke in unsere Beta-Testphase von *Tutor.AI* in der Lehrkräftebildung im Fach Deutsch (Literatur- und Mediendidaktik). Wir stellen Seminarkonzepte zur Einbindung von *Tutor.AI* vor, mit denen wir angehende Lehrkräfte auf den Umgang mit textgenerierender KI vorbereiten und bieten erste Einblicke in unsere empirische Evaluation der Konzepte.

**Schlagwörter:** LLM, digitale Souveränität, Datenbasis, Verstehenssupport, kritische Medienkompetenz

Auch im Jahre 2025 sind die gesellschaftlichen und bildungspolitischen Debatten um Large Language Models (LLMs), KI-Chatbots und deren Einsatz in Lehr-/Lernszenarien nicht abgeklungen (vgl. zu diesen Debatten etwa Walter 2024). Im Zeichen der rasanten Weiterentwicklung und der Fortschritte textgenerierender KI stellt sich zunehmend die Frage, wie Schulen und Universitäten mit diesen Entwicklungen umgehen sollten. An der Universität Münster haben wir unter dem Titel *Tutor.AI* ein interdisziplinäres Projekt initiiert, mit dem wir den Herausforderungen der Entwicklung auf dem Feld der textgenerierenden KI begegnen und deren Potenziale in der akademischen Lehre ausloten. Initiiert aus dem *Fachbereich 09: Philologie* an der Universität Münster sollte fachbereichsintern dem häufig vorgetragenen Wunsch der Studierenden Rechnung getragen werden, aktuelle und praxisrelevante Lehrinhalte in die Lehramtsausbildung zu implementieren. Zudem wurde deutlich, dass es gerade in den Geisteswissenschaften darum geht, Teil des akademischen Diskurses rund um den Umgang mit textgenerierenden KI-Anwendungen zu sein und hierfür eigene Erfahrungswerte aufzubauen. Die didaktische Aufbereitung im Unterrichtsfach Deutsch ist in dem Projekt durch die Mitarbeit von Sebastian Bernhardt vertreten, der das Tool als digitale Unterstützung in seinen literatur- und mediendidaktischen Seminaren einsetzt, reflektiert und evaluiert. In diesem Artikel werden wir das Projekt und die Implementierung des Tools in die lehramtsbezogene universitäre Lehre für das Unterrichtsfach Deutsch vorstellen und erste Einblicke in die empirische Evaluation geben.

## **1 Einsatz eines textgenerierenden KI-Chatbots in der akademischen Lehre (Schwerpunkt germanistische Literatur- und Mediendidaktik)**

Im Rahmen des Projekts *Tutor.AI* entwickelt ein interdisziplinäres Team (Germanistik, Informatik, Wertbasiertes Marketing, Maschinelles Lernen und Data Engineering) aus Hochschullehrer:innen, wissenschaftlichen Mitarbeiter:innen und dem Start-up colloc.AI einen KI-gestützten Chatbot, der auf wissenschaftlich kuratierten Inhalten basiert. Das Vorhaben verfolgt das Ziel, den Einsatz textgenerierender KI in der Hochschullehre nicht nur technisch zu

ermöglichen, sondern auch didaktisch zu durchdringen. *Tutor.AI* kombiniert die Potenziale moderner LLMs mit einer kontrollierten, universitätseigenen Infrastruktur, um Studierende in ihrer Reflexionsfähigkeit und Selbstständigkeit zu fördern.

Anders als bei ChatGPT und anderen Modellen besteht bei *Tutor.AI* die Möglichkeit, Foliensätze, wissenschaftliche Artikel oder eigene Skripts unkompliziert einzuspeisen und damit die Wissensbasis des Modells konkret in Hinblick auf die Belange der Veranstaltung oder einer Gruppe von Veranstaltungen zu verändern.<sup>1</sup> Die Wissensbasis des Systems wurde durch die beteiligten Lehrenden gezielt an die jeweiligen Lehrveranstaltungen angepasst, um den Student:innen eine optimale Unterstützung bei der Vor- und Nachbereitung des Lernstoffs zu ermöglichen. Damit sollte die Lehre verbessert und zugleich eine kritische Mediennutzungsreflexion ermöglicht werden.

## 1.1 Textgenerierende KI und die organisatorischen Rahmenbedingungen

Universitäten können und dürfen sich den Entwicklungen auf dem Feld von LLMs und textgenerierender KI nicht versperren (vgl. dazu Lindner/Weßels 2025; Catani 2023; vgl. zu den spezifischen Anforderungen an die universitäre Lehre Exrance 2023). Allerdings wirft der konsequente Einsatz von LLMs ein schwerwiegendes Problem auf: Der Einsatz von herkömmlicher textgenerierender KI wie Chat GPT verstärkt die Abhängigkeiten der Universitäten und Bildungsinstitutionen von kommerziellen Anbieter:innen (vgl. Reuvekamp i. V.). Während Universitäten/Hochschulen in den Jahrzehnten des sogenannten „KI-Winters“ (Hirsch-Kreinsen/Krokowski 2023: 454) maßgeblich zur Entwicklung grundlegender Technologien der KI beitrugen, sind sie nun mit einer Dynamik konfrontiert, in der technologische Fortschritte primär von privatwirtschaftlichen Akteur:innen vorangetrieben werden. Um ein herkömmliches Sprachmodell konsequent nutzen zu können, müssen kostspielige Accounts für die gesamte Universität/Hochschule angemeldet

.....

1 Wir greifen auf ein trainiertes Sprachmodell zurück, haben aber die Möglichkeit, durch das Einspielen die Wissensbasis und deren Schwerpunkte zu modifizieren.

werden. Die Universität/Hochschule macht sich damit von den Unternehmen abhängig und ist auf deren Geschäftsbedingungen angewiesen, was die digitale Souveränität bedroht (vgl. Theurl/Bach 2025). Längst ist deutlich, dass der internationale Wettlauf um die technologische Hegemonie einen stärkeren geopolitischen Charakter bekommt, dass jedoch das Herkunftsland der Anwendungen auch neue Fragen aufwirft (vgl. Jacobides/Brusoni/Candelon 2021: 425). So kommt die Wissensbasis von ChatGPT und Co. für die Nutzer:innen einer Black-Box gleich (vgl. Bajohr 2022: 153), sie ist nicht steuer- und kontrollierbar. Entsprechend stellt sich das Problem, dass der KI-Chatbot gemäß dem ideologischen, moralischen oder politischen Bias der eingespeisten und trainierten Wissensbasis antwortet (vgl. Bajohr 2024: 191). Wie Führer und Gerjets herausarbeiten, besteht entsprechend die Möglichkeit, dass sich rassistische, sexistische oder anderweitig geartete Machtaffirmationen in die Kommunikation einschreiben. Wenn ChatGPT als Begleiter für Lernende genutzt werde, könne dadurch eine Legitimation derartiger Affirmationen herbeigeführt werden (vgl. Führer/Gerjets 2024: 7).

*Tutor.AI* verfolgt daher einen alternativen Ansatz: Das System nutzt fortgeschrittene Methoden der natürlichen Sprachverarbeitung (Natural Language Processing, NLP) und wird auf einer universitären Cloud-Infrastruktur betrieben (vgl. zur Cloud-Infrastruktur an der Universität Münster Vogl/Rudolph/Thoring 2019; sowie Blank-Burian/Hölters/Vogl 2021).<sup>2</sup> Diese Architektur gewährleistet eine skalierbare und datenschutzkonforme Nutzung, die unabhängig von externen kommerziellen Anbieter:innen bleibt. Ein zentrales Merkmal von *Tutor.AI* ist seine Fähigkeit, kontextbezogene Informationen effizient zu verarbeiten und nutzer:innengerechte Antworten zu generieren. Dadurch wird nicht nur die Qualität der Interaktion verbessert, sondern auch eine gezielte inhaltliche Unterstützung für Studierende ermöglicht.

.....

- 2 Die gesamte Architektur ist in die universitäre Cloud-Infrastruktur eingebettet und nutzt Kubernetes für die Bereitstellung und Skalierung der verschiedenen Systemkomponenten. Dies ermöglicht eine effiziente Ressourcennutzung, hohe Ausfallsicherheit und eine strikte Einhaltung datenschutzrechtlicher Vorgaben.

## 1.2 Das Projekt *Tutor.AI*

Das Projekt *Tutor.AI* kombiniert verschiedene Schlüsseltechnologien, um eine wissenschaftlich fundierte und transparente KI-gestützte Lernumgebung bereitzustellen. Die Dozierenden erhalten die Möglichkeit, wissenschaftliche Texte (nach Maßgabe der urheberrechtlichen Möglichkeiten) im PDF-Format hochzuladen, eigene Skripte oder Powerpoint-Präsentationen einzuspeisen und damit die Wissensbasis zu verändern. *Tutor.AI* nutzt für diesen Prozess semantischer Repräsentationen von Textinhalten Methoden der Wort- bzw. Satz-Einbettung (Embeddings) (vgl. zur Methode der Locally Linear Embeddings Roweis/Saul 2000; vgl. zur Erzeugung semantisch aussagekräftiger Satz-Einbettungen Reimers/Gurevych 2019), um eine adäquate Auswahl relevanter Informationen sicherzustellen. Dabei werden die eingespeisten Materialien in numerische Vektorräume überführt, wodurch inhaltliche Zusammenhänge analysiert und gezielt zur Beantwortung von Nutzer:innenanfragen herangezogen werden können. Wenn also für ein Seminar Texte zur Verfügung gestellt werden, dann werden diese Texte in die Erzeugung der Antworten einbezogen. Dadurch lässt sich verhindern, dass die Antworten auf Basis unwissenschaftlicher Wissensbasen erfolgen.<sup>3</sup>

Die eigentliche Antwortgenerierung im Chat-Interface erfolgt durch das zugrundeliegende LLM. *Tutor.AI* nutzt dabei ein Open-Source Mistral-Modell, welches von der Universität Münster bereitgestellt wird (vgl. Radas/Risse/Vogl 2025: 109). Diese eigenständige Modellinfrastruktur gewährleistet eine größtmögliche Unabhängigkeit von externen Anbieter:innen. Die gezielte Ein-

.....

- 3 Zum technischen Kontext: Anders als bei klassischen dialogbasierten Systemen erfolgt die Antwortgenerierung nicht allein auf Basis interner Modellgewichte, sondern durch eine gezielte Einbettung relevanter Quelleninformationen in den Antwortprozess. Dieser Mechanismus basiert auf dem Retrieval-Augmented Generation (RAG)-Ansatz, einer Technik, die die Stärken großer Sprachmodelle mit einer gezielten Informationsabfrage aus externen Datenbanken kombiniert (vgl. dazu Ram et al. 2023). Im RAG-Verfahren wird eine Nutzer:innenanfrage zunächst mit vorhandenen Lehrmaterialien, Forschungstexten oder Datenbanken abgeglichen, um relevante Abschnitte zu identifizieren. Diese Inhalte werden anschließend in die Generierung der Antwort integriert, sodass das Sprachmodell nicht nur auf seine interne Trainingsbasis zurückgreift, sondern auch aktuelle und fachspezifische Informationen berücksichtigen kann.

bindung fachspezifischer Inhalte und die Möglichkeit, die Antwortgenerierung nachvollziehbar zu gestalten, eröffnet neue Perspektiven für den Einsatz textgenerierender KI in der Hochschullehre. Mit der Möglichkeit, die Wissensbasis zu steuern und gezielt durch Einspeisung von Texten, Informationen und Materialien die Antworten zu beeinflussen, geht in ethischer Hinsicht allerdings wiederum eine große Verantwortung auf Seiten der Dozierenden einher: Es muss grundsätzlich in allen Lehr-/Lernszenarien transparent gemacht werden, welche Datenbasis vorliegt und welche Texte eingepflegt wurden. Es darf auch nicht zu einer einseitigen Einspeisung von Positionen kommen.

### **Parameter der Entwicklung von *Tutor.AI***

Folgende Parameter waren leitend für unsere konzeptionellen Überlegungen und wurden in der Entwicklung von *Tutor.AI* umgesetzt:

1. Schaffung einer textgenerierenden KI-Umgebung auf universitätseigenen Servern, um das Problem der wirtschaftlichen Abhängigkeiten zu umgehen und den Datenschutz zu gewährleisten
2. Möglichkeit, die Wissensbasis zu beeinflussen, indem wir eine textgenerierende KI mit Texten sowie Präsentationen informieren und dadurch das Problem der unklaren Wissensbasis und die Gefahr eines ideologischen Bias umgehen
3. Entwicklung von Möglichkeiten, Studierenden eine kritische Mediennutzungskompetenz zu vermitteln und sie in die Lage zu versetzen, diese kritische Mediennutzungskompetenz in Bezug auf Unterrichtshandeln anzuwenden
4. Konkretisierung dieser Überlegungen in Bezug auf die Literatur- und Mediendidaktik

## **2 Praxisbeispiele und Einsatz in der literaturvermittelnden Lehre**

Im Sommersemester 2024 wurde *Tutor.AI* im Rahmen eines Beta-Tests in acht Lehrveranstaltungen eingesetzt und evaluiert. In unserem Beitrag beziehen wir uns auf fünf Seminare, die in der germanistischen Literatur- und Medi-

endidaktik ausgebracht wurden. Es handelte sich dabei um drei Seminare im Aufbaumodul für das Grundschullehramt, ein Seminar im Master Vertiefungsmodul für das Grundschullehramt und ein Seminar im Master-Kompetenzmodul Literatur- und Mediendidaktik für Gymnasien, Gesamtschulen und Berufskollegs, in denen drei unterschiedliche Konzepte zur Einbindung von *Tutor.AI* exploriert wurden.

Gerade in Bezug auf den Umgang mit Literatur und Medien und noch stärker in Bezug auf die entsprechenden Vermittlungskontexte stellen sich besondere Herausforderungen an ein digitales Tutor:innensystem auf Basis einer textgenerierenden KI: Literatur bzw. ästhetische Medien und ihre Rezeption stellen nämlich aufgrund ihrer Fiktionalität einen komplexen Gegenstand dar (vgl. dazu Bernhardt/Hardtke 2022; Bernhardt 2024; Rosebrock 2019). Dadurch, dass Literatur nicht im eigentlichen Sinne verstanden werden kann (vgl. dazu Baum 2019; Heimböckel/Pavlik 2022), ergeben sich auch besonders große Herausforderungen für ein LLM (vgl. Führer/Nix 2023: 339f.; vgl. auch Führer/Gerjets 2024). Ein Fokus muss auf Unbestimmtheitsstellen, Mehrdeutigkeit und Polyvalenz liegen, soll der Gegenstand nicht verfehlt werden. Gerade hier sind die großen LLMs wie ChatGPT oftmals überfordert, weil sie auf Basis der vorliegenden Daten versuchen, eindeutige Wissensbestände zu postulieren und in Bezug auf die literarischen Texte mit Autor:innenkonzepten, Wikipediawissen und als eindeutig gesetzten Informationen operieren, auch gelegentlich halluzinieren, ohne dass die dahinterstehende Wissensbasis transparent ist.

In der akademischen Lehre geht es uns darum, durch das Angebot des digitalen Tutor:innensystems Möglichkeiten zu schaffen, die Studierenden kognitiv zu aktivieren und sie in die selbsttätige Befassung mit den behandelten Themen zu bringen. Die Möglichkeit, mit dem LLM zu interagieren, sollte unserer Zielsetzung nach eine Abkehr von der sonst gelegentlich im Studium zu monierenden Auswendiglern-Mentalität ermöglichen.

Am Beispiel dreier Seminkonzepte zur Einbindung von *Tutor.AI* werden wir unsere hochschuldidaktischen Überlegungen vorstellen und jeweils erste Beobachtungen aus den Seminaren skizzieren. Diese Beobachtungen basieren auf einer teilnehmenden Beobachtung. So wurden in jeder Sitzung Notizen des Verhaltens der Studierenden angefertigt, um erste Eindrücke des Umgangs

und der Effekte des Einsatzes zu erlangen. Diese teilnehmende Beobachtung ist nicht empirisch validiert. Da die Seminare für den Einsatz von *Tutor.AI* komplett neu konzipiert wurden, kann auch kein quantifizierender Vergleich der Lernergebnisse mit einer Gruppe von Studierenden vorgenommen werden, die nicht mit *Tutor.AI* arbeiten konnte. Die Beobachtungen erlauben aber erste Beschreibungen von Effekten und Hypothesen für weitere seminaristische Entwicklungen.

### **Beispiel 1: Aufbaumodul Literatur- und Mediendidaktik im Bachelor Grundschule (Sommersemester 2024, Sebastian Bernhardt)**

In einem Bachelorseminar im Aufbaumodul wurden basale Theorien zum literarischen Lernen behandelt und damit die Spezifika des Feldes der Literatur- und Mediendidaktik abgesteckt. In der ersten Sitzung gab es zunächst Erläuterungen zur Funktionsweise und zu den technischen Rahmenbedingungen von *Tutor.AI* und anderen LLMs. Es wurde dann anhand von Beispielfragen gezeigt, wie *Tutor.AI* arbeitet. Als Vorbereitung zur Sitzung zwei gab es zwei Fachtexte, einen literaturwissenschaftlichen und einen, der didaktische Anschlussüberlegungen darstellt. Dazu wurden Leitfragen bereitgestellt, die helfen sollten, den Text zu strukturieren und aufzubereiten. Vorab wurde sichergestellt, dass die Fragen so gestellt sind, dass *Tutor.AI* in Bezug auf sein Antwortverhalten an seine Grenzen stößt.

Angekündigt war, dass in Sitzung zwei ein Test geschrieben wird, in dem einige der Leitfragen abgeprüft werden. *Tutor.AI* sei als Hilfsmittel erlaubt, es solle aber bedacht werden, welche Grenzen das hat. Das Ergebnis war, dass sich viele Studierende trotz der Warnungen aus Sitzung eins auf *Tutor.AI* verlassen hatten. Da Sprachmodelle auf Basis stochastischer Wahrscheinlichkeiten antworten, aber grundsätzlich auch in ihren Antworten variieren, zudem der vorangehende Chatverlauf – sofern er nicht gelöscht wird – Einfluss auf die Antworten hat, kamen entsprechend auch mehr oder weniger stark voneinander abweichende Antworten zustande (vgl. dazu auch Fürstenberg/Müller 2023: 328–335). Die Korrektur der Tests konnte schnell angefertigt werden, während die Studierenden in eine Gruppenarbeitsphase gingen. Am Ende der Sitzung erfolgte eine Auswertung. Das Ergebnis war, dass teilweise zwei nebeneinandersitzende Studierende denselben Prompt eingegeben, aber aufgrund

ihres Gesprächsverlaufs und aufgrund der von LLMs angestrebten Varianzen unterschiedliche Antworten und damit auch unterschiedliche Punktzahlen erhalten haben. Der Test wurde im Nachhinein nicht benotet und diente nur einer ersten Sensibilisierung für die Perspektiven und Grenzen textgenerierender KI.

Der Schreckmoment nach der Korrektur und Rückmeldung trug dazu bei, dass die Studierenden eine Einsicht in die Grenzen der Nutzung einer LLM erhielten und seither sehr konzise Reflexionen der Antworten von *Tutor.AI* angestellt haben. Im Semesterverlauf gab es vier weitere Tests. Die Studierenden haben hier aber deutlich mehr eigene Reflexionen an den Tag gelegt, teilweise *Tutor.AI* nicht mehr benutzt oder auch geschrieben, was das System behauptet hat, diese Aussage dann aber argumentativ betrachtet und widerlegt, modifiziert oder bestätigt. Die Studierenden haben immer stärker die Einsicht erworben, dass sie im Umgang mit LLMs nicht direkt Antworten erhalten, sondern sich Zusammenhänge erschließen. Neben dieser Medienreflexion erwies sich der Fortgang der entsprechenden Sitzungen als fruchtbar: Im Seminar konnten wir stärker als ich es aus vergleichbaren Seminaren gewohnt bin in die Reflexion und Anwendung bestehender Theorien eintauchen. Entsprechend half das Format *Tutor.AI*, das aktiv behandelt und kritisch hinterfragt wurde, offenbar bei der Vorbereitung der Theorien. Inwiefern allerdings selbsttätiges Lernen über die Auseinandersetzung mit den Leitfragen hinaus erfolgte, lässt sich auf Basis der Beobachtungen nicht feststellen. Der Eindruck der guten Vorbereitung kann auch darauf zurückzuführen sein, dass die Studierenden schlicht und ergreifend neugierig auf die Technologie waren und ihr die vorzubereitenden Leitfragen stellten. Inwiefern tatsächlich aktive und selbsttätige Befassungen mit dem Gegenstand angestoßen wurden, wird im Rahmen der Auswertung unserer Daten genauer zu beleuchten sein.

Am Ende des Seminars stand eine Modulabschlussprüfung in Form einer mündlichen Prüfung. Zu dieser Prüfung sollten Thesen entwickelt werden, die das Gespräch vorstrukturieren. Den Studierenden wurde explizit erlaubt, für die Entwicklung der Thesen und auch für die inhaltliche Vorbereitung *Tutor.AI* zurate zu ziehen. Wir erhofften uns eine kreative und dialogische Interaktion der Studierenden mit *Tutor.AI*, sodass es sich bei dem Tool um einen wirklichen Verstehenssupport handelt. In unserer Auswertung der Prompts

werden wir betrachten, inwiefern sich diese Erwartung im tatsächlichen Nutzer:innenverhalten widerspiegelt.

**Beispiel 2: Aufbaumodul Literatur- und Mediendidaktik; Vertiefungsmodul Literatur- und Mediendidaktik im Master Grundschule (Sommersemester 2024, Sebastian Bernhardt)**

In einem weiteren Seminar im dritten BA-Modul und einem Seminar im Vertiefungsmodul im MA für das Grundschullehramt wurde *Tutor.AI* für die Vermittlung und für die Mediennutzungskompetenz eingesetzt. In den Seminaren wurden Theorien literarischen Lernens und literarischen Verstehens behandelt, kritisch diskutiert und auf unterschiedliche Fallbeispiele angewendet. Die Theorien sowie erläuternde Quellen wurden *Tutor.AI* als Datenbasis zur Verfügung gestellt und den Studierenden transparent gemacht, sodass die Informationsbasis passend zu den Themen des Seminars war. Die Studierenden wurden auch hier im Rahmen der Vorbereitung zu den Sitzungen aufgefordert, die *Tutor.AI* bereitgestellten Artikel mithilfe von Leitfragen vorzubereiten. Dabei sollte aktiv und dezidiert auch *Tutor.AI* genutzt werden, um dann im Rahmen der Sitzung zu vergleichen, welche Antworten es gegeben hatte und an welchen Stellen kritische Hinterfragungen sich als notwendig erwiesen. Die Leitfragen waren bewusst so gestellt, dass *Tutor.AI* nicht immer präzise Antworten geben konnte.

Der Umgang mit dem LLM sollte die Student:innen in Bezug auf ihre eigenständige kritische Reflexionskompetenz fördern. In der ersten Sitzung gab es auch in diesen Seminaren Erläuterungen zur Funktionsweise und zu den technischen Rahmenbedingungen. Als Vorbereitung auf die zweite Sitzung dienten zwei Fachtexte, zu denen Leitfragen bereitgestellt wurden. Die Leitfragen durften gern *Tutor.AI* gestellt werden. In Sitzung zwei erfolgte der kritische Abgleich, inwiefern die Antworten als plausibel erscheinen. Die Student:innen waren von vornherein aufgefordert, mit *Tutor.AI* über den Text zu diskutieren und ihren eigenen Leseindruck mit den Antworten abzugleichen. Hierbei war bereits im Voraus darauf geachtet worden, dass beide Texte in Teilen unterschiedliche Sichtweisen auf Literatur und literarästhetische Erfahrungen mit sich brachten. Das Ergebnis: Die Student:innen bemerkten, dass die Zusammenhänge durch *Tutor.AI* nicht immer zuverlässig erkannt wurden, dass

sich allerdings in Abhängigkeit der jeweiligen Prompts auch unterschiedliche Antworten ergaben.

Die unterschiedlichen Antworten von *Tutor.AI* wurden zunächst präsentiert. Die Studierenden wurden in Kleingruppen dazu aufgefordert, die Antworten anhand eines Close Readings zu überprüfen. Auf Basis dieses Close Readings im Abgleich mit den vorzubereitenden Texten wiesen die Studierenden nach, welche Schwierigkeiten *Tutor.AI* hatte. Besonders erhellend wirkte dabei, dass die Antworten stark differierten. So konnte die entsprechende kritische Distanz in Bezug auf derartige Ergebnisse eingeübt werden. Einige Studierende der MA-Seminargruppe hatten großen Spaß daran, die Grenzen auszutesten und mir per Mail Screenshots von Beispielkonversationen zu schicken, in denen *Tutor.AI* aus ihrer Sicht Falschinformationen vertrat. Auch hier lässt sich als erste teilnehmende Beobachtung festhalten, dass die Studierenden in eine kritische Befassung mit den Antworten von *Tutor.AI* gelangten. So wurde im Rahmen des Seminars mehrfach über die Plausibilität von Antworten diskutiert und entsprechend eine kritische Medienkompetenz eingeübt. Inwiefern allerdings auch eine eigentätige Auseinandersetzung mit den Seminarinhalten über die explizit gestellten Aufgaben hinaus erfolgte, wird Thema der Prompt-Auswertungen sein.

Die Prüfungsleistung bestand aus einer Hausarbeit. Die Studierenden wurden ermuntert, für Ideen und Gliederungspunkte mit *Tutor.AI* zu sprechen, bevor sie zu mir in die Sprechstunde kamen. In der Beratung der Studierenden zeichnete sich ein gemischtes Bild ab: Einige Studierende brachten die Ideen von *Tutor.AI* mit und erklärten mir, dass sie nicht bewerten könnten, wie sinnvoll das sei. Andere schilderten mir, dass es ihnen als erste Ideensammlung geholfen habe. Viele Studierende verzichteten allerdings auch auf die Nutzung.

### **Beispiel 3: Aufbaumodul Bachelor Grundschule und Vertiefungsseminar Literatur- und Mediendidaktik im Master Gymnasium/Gesamtschulen (Sommersemester 2024, Sebastian Bernhardt)**

Im dritten Seminarkonzept diente *Tutor.AI* der reinen Lernbegleitung. Hier gab es im Laufe des Semesters einzelne Sitzungen, in denen die Funktionsweise erläutert wurde und die Studierenden wurden immer wieder ermuntert, sich bei der Seminarvorbereitung in einem kritisch-reflektierenden Sinne mit

*Tutor.AI* auseinanderzusetzen. Am Ende gab es als Prüfungsleistung die Option, eine mündliche Prüfung mit Entwicklung eines Thesepapiers oder eine Hausarbeit zu schreiben.

### **3 Begleitung des Beta-Tests mit Umfragen unter den Studierenden**

Da es sich um eine Pilotierung handelt, begleiteten wir *Tutor.AI* auch wissenschaftlich einerseits durch Umfragen zu Beginn, in der Mitte und am Ende des Semesters und perspektivisch durch eine anonymisierte Auswertung des Prompt- und Gesprächsverhaltens der Studierenden mit *Tutor.AI*. Im Folgenden werden die Ergebnisse der fünf Seminare dargestellt, die im Sommersemester 2024 von Sebastian Bernhardt alle in die Beta-Testung einbezogen wurden und für Lehramtsstudierende des Fachs Deutsch auf verschiedenen Niveaus (Bachelor und Master) stattfanden.

Es wurden im zeitlichen Verlauf der Nutzung der Studierenden sowohl die Erwartungen als auch ihre praktischen Erfahrungen mit dem System erfasst. Sowohl die Anwendung des KI-Chatbots als auch die Teilnahme an den begleitenden Umfragen war freiwillig. Basierend auf drei Befragungen im Semesterverlauf ( $n = 148$ ,  $n = 117$ ,  $n = 82$ ) wurde eine differenzierte Analyse der Nutzungsmuster, wahrgenommenen Vorteile und Herausforderungen des Tools durchgeführt. Vor der Implementierung des Systems wurden Studierende befragt, um ihre Erwartungen an *Tutor.AI* zu erfassen ( $n = 148$ ). Die Mehrheit der Teilnehmenden (83,1 %) hatte bereits Erfahrungen mit textgenerierenden KI-Systemen wie ChatGPT, insbesondere zur Ideenfindung (79,7 %), als Formulierungshilfe (58,5 %) oder zur Klärung von Verständnisfragen (58,5 %). Entsprechend hoch waren die Erwartungen an *Tutor.AI*: 82,4 % der Befragten wünschten sich eine fehlerfreie und verständliche Aufbereitung relevanter Lehrinhalte, während 50,7 % eine allgemeine Verbesserung des Lernerlebnisses erwarteten. Zudem wurde das Tool als potenzielle Unterstützung bei der Prüfungsvorbereitung (54,7 %) und als Möglichkeit zur Reduzierung des eigenen Leseaufwands (53,4 %) gesehen. Allerdings war die Nutzung textgenerierender KI-Systeme in der Hochschullehre bisher wenig verbreitet: 89,2 %

der Befragten hatten in ihren bisherigen Lehrveranstaltungen noch nicht aktiv mit KI-Technologien gearbeitet. Dennoch äußerte die Mehrheit Interesse an *Tutor.AI*, wobei 59,5% planten, das System zu nutzen, während 39,9% eine Nutzung zumindest in Erwägung zogen.

Die ersten praktischen Erfahrungen mit *Tutor.AI* wurden in einer separaten Befragung mit  $n = 117$  Studierenden erfasst. Dabei zeigte sich, dass 82,1% das System ausprobiert hatten, während 17,9% darauf verzichtet hatten. Die häufigsten Nutzungsszenarien umfassten die Bearbeitung von Aufgaben aus den Lehrveranstaltungen (76%), die Vor- und Nachbereitung von Lehrinhalten (60,4%) sowie das Stellen von Verständnisfragen (55,2%). Zudem nutzten 49% das Tool gezielt als Arbeitserleichterung, um den eigenen Leseaufwand zu reduzieren. Die Bewertung des allgemeinen Nutzens von *Tutor.AI* fiel gemischt aus: 13,5% hielten das System für sehr hilfreich, 24% nutzten es regelmäßig mit positiven Effekten, während 60,4% den Nutzen als mittelmäßig einschätzten – sie empfanden das Tool für manche Aufgaben als hilfreich, aber nicht als universell einsetzbar. Lediglich 2,1% berichteten von Verwirrung durch *Tutor.AI* und keine der befragten Personen äußerte eine vollständige Enttäuschung über das System.

Hinsichtlich der Nutzungsfrequenz zeigte sich, dass 11,5% *Tutor.AI* mindestens einmal pro Woche verwendeten, während 62,5% das Tool mindestens einmal pro Monat nutzten. 26% griffen seltener darauf zurück. Die allgemeine Zufriedenheit mit dem System war überwiegend positiv: 8,3% waren sehr zufrieden, 61,5% ziemlich zufrieden, während 28,1% eine neutrale Haltung einnahmen. Nur 2,1% äußerten sich eher unzufrieden und keine der befragten Personen war sehr unzufrieden. Die abschließende Befragung zu den Erfahrungen mit *Tutor.AI* ( $n = 82$ ) bestätigte viele der zuvor festgestellten Muster. 82,9% der Teilnehmenden hatten das System ausprobiert und die häufigsten Anwendungsfälle waren erneut die Vor- und Nachbereitung von Lehrveranstaltungen (66,2%) sowie die Bearbeitung von Aufgaben (64,7%). Ein Großteil der Studierenden wünschte sich eine Weiterentwicklung des Tools: 89,7% gaben an, dass *Tutor.AI* ein sinnvolles Projekt sei, aber noch verbessert werden müsse. Die häufigsten Kritikpunkte in den Freitextfeldern betrafen die Zuverlässigkeit der Antworten, da es gelegentlich zu ungenauen oder inkonsistenten Aussagen kam. Ein weiterer häufig geäußerter Wunsch betraf die

Transparenz der Quellen, auf die sich die jeweils durch das System gegebene Antwort stützt. Dies sei insbesondere bei der Prüfungsvorbereitung und für das wissenschaftliche Arbeiten essenziell.

Aus den Ergebnissen dieser Befragung geht indirekt hervor, dass die Erwartungen an das Tool sehr stark an pragmatischen Zwecken orientiert sind. Die Studierenden erwarten vor allem eine konkrete Erleichterung in Bezug auf die eigenen Lernprozesse. Das ist insofern auch bei der Auswertung der Prompts genau zu betrachten, als sich dadurch die Frage nach der pragmatischen und informatorischen Haltung in Bezug auf die Seminargegenstände abzeichnet. Es wird eine Aufgabe sein, bei der Sichtung der Prompts diese Haltung gezielt zu evaluieren und Reflexionen in Bezug auf die weitere Gestaltung der Lehre anzustellen, um eine Nutzung zu erreichen, die stärker dialog- und supportorientiert ist und weniger als reine Vorbereitung auf konkrete Aufgabenstellungen verstanden wird.

## 4 Exemplarische Auswertung von Prompts

Die Auswertung der Umgangsweisen der Studierenden basiert auf den Gesprächen der Studierenden mit *Tutor.AI*. Im Laufe des Sommersemesters 2024 kamen bei der Auswertung des Materials für die fünf Seminare insgesamt 2.271 Prompts zusammen, die wir in einem ersten Schritt ergebnisoffen sichten. Daraus ergeben sich erste Tendenzen. Die finale Erstellung eines Kategoriensystems zur feingliedrigen Zuordnung und Einordnung des Nutzungsverhaltens steht aktuell noch aus und wird von uns im Sommersemester 2025 fortgeführt. Allerdings fällt schon bei der ersten Sichtung auf, dass sehr viele Prompts 1:1 mit den vorab bereitgestellten Leitfragen aus dem Seminar identisch sind. So zeigt sich in einer Vielzahl von geführten Gesprächen, dass die Studierenden tatsächlich nach und nach die Fragen durchgingen, die vorbereitend ins Learnweb gestellt wurden. In diesen Fällen kamen keine Gespräche zustande. Allenfalls gab es teilweise noch Verständnisfragen oder Bitten, Dinge noch einmal knapper oder ausführlicher zu erläutern, aber grundsätzlich wurden die Fragen gestellt, Antworten abgewartet und danach ohne Überleitung die nächste im Learnweb gestellte Aufgabe platziert (576).

Deutlich weniger vertreten sind Einsatzszenarien, bei denen die Studierenden *Tutor.AI* einsetzen, um Hilfestellungen in Bezug auf die Prüfungsformate zu erlangen. So wurden in den Seminaren, in denen am Ende Thesenpapiere für die mündliche Prüfung erstellt werden sollten, vereinzelt Anfragen dahingehend gestellt, derartige Thesen zu erarbeiten (45). Zudem wurden teilweise Impulse für Hausarbeiten oder Strukturierungsideen erfragt, wobei die Zahl dieser Anfragen sehr begrenzt ist (17). Eigenständige Interpretationsfragen oder Fragen zur didaktischen Anwendbarkeit auf konkrete Textbeispiele sind in den Prompts gar nicht abzulesen.<sup>4</sup> Das deutet darauf hin, dass *Tutor.AI* vornehmlich als Vorbereitung für die Seminarsitzungen oder auch Leistungsabfragen und nicht als dialogorientierter Verstehenssupport eingesetzt wird. Ohnehin war festzustellen, dass die behandelten literarischen Beispiele kaum mit *Tutor.AI* besprochen wurden. Wir hatten angenommen, dass auch über mögliche Deutungen von Literatur gesprochen oder beraten würde oder zumindest verschiedene Interpretationen gegeneinander abgewogen würden. Das war aber nicht der Fall. Gerade in den Seminaren, in denen *Tutor.AI* nur als Lernbegleitung genutzt wurde, war die Nutzungsfrequenz und Intensität signifikant geringer als in den Seminaren mit direkter Einbettung. Überhaupt steigt jeweils vor Tests oder Prüfungen die Nutzungsfrequenz signifikant an.

## 5 Ausblick

Mit unserem Projekt konnten wir sowohl auf Seiten der Student:innen als auch auf Seiten der Dozent:innen eine Sensibilisierung für die Perspektiven und Grenzen sowie die Funktionsweisen von textgenerierender KI herbeiführen. Wichtig sind die weitere Optimierung der Technologie und die Implementierung unterschiedlicher hochschuldidaktischer Impulse, um die Potenziale noch besser auszuschöpfen. Als größte Herausforderung erweist es sich,

.....

- 4 In Bezug auf die weiteren Prompts wird das Kategoriensystem aktuell noch erarbeitet und verfeinert. Die bisherigen Sichtungen deuten aber darauf hin, dass einfache Definitionsfragen vorhanden sind, Nachfragen nach konkreten Konzepten oder Zusammenfassungen behandelte Theorien gestellt wurden. Die feinjustierte Auswertung der Prompts wird im Sommer vorangetrieben und validiert.

*Tutor.AI* noch klarer als dialogischen Support für die inhaltliche Durchdringung und für Möglichkeiten der Textarbeit zu etablieren und weniger als Instrument für das rein pragmatische Erreichen von Anforderungen in konkreten Seminar- und Prüfungssituationen. Gerade im Kontext der Literatur- und Mediendidaktik muss also verhindert werden, dass der Einsatz von textgenerierender KI ästhetische Zugänge verstellt.

Führer und Nix arbeiten für schulische Kontexte pointiert heraus, dass der Einsatz von textgenerierender KI für den Umgang mit Literatur nicht per se positive Auswirkungen auf die Ausprägung einer ästhetischen Rezeptionskompetenz hat (vgl. Führer/Nix 2023: 16f.). Basierend auf der Auswertung eines Unterrichtssettings zeichnet sich die Gefahr ab, dass durch den Einsatz der KI ein vornehmlich informatorischer Lesemodus an den Tag gelegt wird, der einer ästhetischen Lesehaltung entgegensteht. Führer und Gerjets betonen vor diesem Hintergrund, dass es gezielter didaktischer Arrangements bedürfe, um Schüler:innen zu einer individualisierten und ästhetischen Texterschließung zu bringen (vgl. Führer/Gerjets 2024: 6). Für die akademische Lehre ist diese Beobachtung hochgradig relevant: Erstens besteht auch im Kontext der Veranstaltungen die Gefahr, dass die Bereitstellung eines digitalen Tutor:innenmodells einen solchen pragmatischen Verstehensmodus mit sich bringt und zweitens muss auf einer Meta-Ebene mit den Studierenden darüber reflektiert werden, wie sie später als Lehrkräfte ihre Schüler:innen auch zu aktiven und nicht nur informatorischen Umgangsformen mit textgenerierender KI anregen. Um zu genauen Reflexionen der Studierenden, aber auch der Dozierenden in Bezug auf die Seminarkonzeption zu gelangen, ist in einem Folgeschritt zu eruieren, welche Gründe die Studierenden zu ihrem rein pragmatischen Nutzungsverhalten bringen, um darauf aufbauend Settings zu schaffen, die stärker zu einer Supportorientierung einladen.

## Literatur

- BAJOHR, HANNES (2022): *Schreibenlassen. Texte zur Literatur im Digitalen*. Berlin: Matthes & Seitz.
- BAJOHR, HANNES (2024): Whoever Controls Language Models Controls Politics. In: Arns, Inke et al. (Hg.): *Training the Archive*. Köln: Walther König, S. 189–195.
- BAUM, MICHAEL (2019): *Der Widerstand gegen Literatur. Dekonstruktive Lektüren zur Literaturdidaktik*. Bielefeld: transcript.
- BERNHARDT, SEBASTIAN (2024): Einleitung: literarästhetische Unzuverlässigkeit. Diskussionen, Desiderate, didaktische Anschlussfähigkeiten. In: ders. (Hg.): *Unzuverlässiges Erzählen in Literatur und Medien. Didaktische Perspektiven*. Berlin: Frank & Timme, S. 9–45. [https://doi.org/10.57088/978-3-7329-8920-1\\_1](https://doi.org/10.57088/978-3-7329-8920-1_1).
- BERNHARDT, SEBASTIAN/HARDTKE, THOMAS (2022): Einleitung. Interpretation. Literaturdidaktische Perspektiven. In: dies: (Hg.): *Interpretation. Didaktische Perspektiven*. Berlin: Frank & Timme, S. 7–24. [https://doi.org/10.57088/978-3-7329-9143-3\\_1](https://doi.org/10.57088/978-3-7329-9143-3_1).
- BLANK-BURIAN, MARKUS/HÖLTERS, JÜRGEN/VOGL, RAIMUND (2021): JupyterHub on an on-premises cloud – a special focus on GPU Accelerated Machine Learning and 3D Visualization. In: Bolis, Spiros/Desnos, Jean-François/Merakos, Lazaros F./Vogl, Raimund (Hg.): *Proceedings of the European University Information Systems Conference 2021, EUNIS 2021, Virtual Event, Greece, 9–11 June, 2021*. EPiC Series in Computing: EasyChair Vol. 76, S. 69–76. <https://doi.org/10.29007/F8VP>.
- CATANI, STEFANIE (2023): Mit KI schreiben – über KI schreiben. Künstliche Intelligenz als Thema im literaturwissenschaftlichen Studium. In: Fürstenberg, Maurice/Müller, Hans-Georg (Hg.): *Mitteilungen des Deutschen Germanistenverbandes 4/2023: Digitale Umbrüche. Sprache, Literatur und Deutschunterricht in Zeiten von Big Data und KI*. Göttingen: Brill, S. 393–405.
- EXTANCE, ANDY (2023): ChatGPT Has Entered the Classroom: How LLMs Could Transform Education. In: *Nature* 623/7987, S. 474–477. <https://doi.org/10.1038/d41586-023-03507-3>.
- FÜHRER, CAROLIN/GERJETS, PETER (2024): How to understand & write literature with AI? Potentiale und Risiken von KI-Tools für literarisches Lesen und Schreiben. In: *Medien im Deutschunterricht* 1/2024, S. 1–18. <https://doi.org/10.18716/OJS/MIDU/2024.1.3>.

- FÜHRER, CAROLIN/NIX, DANIEL (2023): Literarische Interaktionen mit ChatGPT – Kann der Einsatz von Künstlicher Intelligenz zur Entwicklung literarischer Lesehaltungen beitragen? In: Carl, Mark-Oliver/Jörgens, Moritz/Schulze, Tina (Hg.): *Literarische Texte lesen – Texte literarisch lesen*. Heidelberg: Metzler, S. 335–368.
- FÜRSTENBERG, MARICE/MÜLLER, HANS-GEORG (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: dies. (Hg.): *Mitteilungen des Deutschen Germanistenverbandes 4/2023: Digitale Umbrüche. Sprache, Literatur und Deutschunterricht in Zeiten von Big Data und KI*. Göttingen: Brill, S. 327–345.
- HEIMBÖCKEL, HENDRIK/PAVLIK, JENNIFER (2022): Ästhetisches Verstehen und Nichtverstehen. Positionen und Schnittpunkte des literaturdidaktischen Diskurses. In: dies. (Hg.): *Verstehen und Nicht-Verstehen. Aktuelle Zugänge in Literatur- und Mediendidaktik*. Bielefeld: transcript, S. 9–30.
- HIRSCH-KREINSEN, HARTMUT/KROKOWSKI, THORBEN (2023): Technologieversprechen Künstliche Intelligenz. Vergangene und gegenwärtige Konjunkturen in der Bundesrepublik. In: *Berliner Journal für Soziologie*, Vol. 33, 453–484. <https://doi.org/10.1007/s11609-023-00504-1>.
- JACOBIDES, MICHAEL G./BRUSONI, STEFANO/CANDELON, FRANCOIS (2021): The Evolutionary Dynamics of the Artificial Intelligence Ecosystem. In: *Strategy Science*, Vol. 6, Issue 4, S. 412–35. <https://doi.org/10.1287/stsc.2021.0148>.
- LINDNER, MARLIT ANNALENA/WEBELS, DORIS (2025): Vom Problemfall zur Lösung. Zur Ausgestaltung von Richtlinien zur Nutzung generativer Künstlicher Intelligenz an Hochschulen. In: *Forschung & Lehre 2/2025*, S. 32–35.
- RADAS, JONATHAN/RISSE, BENJAMIN/VOGL, RAIMUND (2025): Building UniGPT: A Customizable On-Premise LLM-Solution for Universities. In: Bolis, Spiros/Desnos Jean-François/Desnos, Laurence/Ferrell, Gill/Merakos, Lazaros/Roumeliotis, Manos/Tsili, Effie/Vogl, Raimund(Hg.): *Proceedings of EUNIS 2024 annual congress in Athens*. Vol. 105, S 108–116. <https://doi.org/10.29007/jv11>.
- RAM, ORI et al. (2023): In-Context Retrieval-Augmented Language Models. In: *Transactions of the Association for Computational Linguistics*, Vol. 11, S. 1316–1331. [https://doi.org/10.1162/tacl\\_a\\_00605](https://doi.org/10.1162/tacl_a_00605).
- REIMERS, NILS/GUREVYCH, IRYNA (2019): Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Confe-*

- rence on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, S. 3980–3990. <https://doi.org/10.18653/v1/D19-1410>.
- REUVEKAMP, SILVIA. (i. V.): Artes und disciplinae oder die Kunst, das Wissenswerte zu gestalten. In: Oestmann, Peter (Hg.): *Bildung zur Freiheit. Universitäre Lehre in der Corona-Zeit*. Verlag n. n.
- ROSEBROCK, CORNELIA (2019): Eine literarische Lesehaltung einnehmen, demonstrieren, entwickeln: Baustein der literaturdidaktischen Professionalisierung. In: *Didaktik Deutsch*, 46/24 (2019), S. 32–46.
- ROWEIS, SAM T./SAUL, LAWRENCE K. (2000): Nonlinear Dimensionality Reduction by Locally Linear Embedding. In: *Science*, Vol. 290, Issue 5500, S. 2323–2326. <https://doi.org/10.1126/science.290.5500.2323>.
- THEURL, THERESIA/BACH, FRIEDRICH (2025): KI-Ökosysteme: Fundamente, Elemente, Konsequenzen. In: Bruhn, Manfred/Hadwich, Karsten (Hg.): *Digital Analytics im Dienstleistungsmanagement*. Wiesbaden: Springer Fachmedien, S. 491–524. [https://doi.org/10.1007/978-3-658-48325-8\\_16](https://doi.org/10.1007/978-3-658-48325-8_16).
- VOGL, RAIMUND/RUDOLPH, DOMINIK/THORING, ANNE (2019): Bringing Structure to Research Data Management Through a Pervasive, Scalable and Sustainable Research Data Infrastructure. In: Bergener, Katrin/Räckers, Michael/Stein, Armin (Hg.): *The Art of Structuring: Bridging the Gap Between Information Systems Research and Practice*. Cham: Springer International Publishing, S. 501–512. [https://doi.org/10.1007/978-3-030-06234-7\\_47](https://doi.org/10.1007/978-3-030-06234-7_47).
- WALTER, YOSHIJA (2024): Embracing the Future of Artificial Intelligence in the Classroom: The Relevance of AI Literacy, Prompt Engineering, and Critical Thinking in Modern Education. In: *International Journal of Educational Technology in Higher Education*, Issue 21, Article Number 15. <https://doi.org/10.1186/s41239-024-00448-3>.



## **Schreibend und lesend Texte schreiben mit dem *Writing-Ko-Aktanten* ChatGPT**

Über das schriftliche Instruieren *von* und den Dialog  
*mit* Large Language Models

### **Abstract**

Der vorliegende Artikel untersucht den dialogischen Schreibprozess zwischen menschlichen Nutzer:innen und KI-gestützten Chatbots. Chatbots wie ChatGPT fungieren dabei als *Writing-Ko-Aktant*, die in den Schreibprozess eingebunden werden und neue Formen des gemeinsamen Schreibens ermöglichen. Nach einer sprach- und kompetenztheoretischen Diskussion zu dialogischen KI-Systemen werden erste Ergebnisse einer empirischen Studie vorgestellt, in der die Nutzung von ChatGPT durch Schüler:innen einer 10. Klasse bei der Überarbeitung von Inhaltsangaben untersucht wird. Die Analyse zeigt zwei unterschiedliche Strategien im Umgang mit der KI: eine explorativ-sukzessive sowie eine fokussiert-antizipative. Bei beiden Strategien wird deutlich, dass die zielgerichtete Nutzung von Chatbots das Vorhandensein schriftsprachlicher und dialogischer Kompetenzen nicht nur für das Verfassen von Prompts, sondern auch für das Lesen, Verstehen und Evaluieren des von dem Bot generierten Textes erfordert.

**Schlagwörter:** ChatGPT, Sprachkompetenz, z. B. Medienkompetenz, z. B. Digitale Schreibforschung, z. B. Internetlinguistik

## **1 Einleitung**

In dem vorliegenden Artikel werden einleitend theoretische Überlegungen zu dem im Medium Schrift ablaufenden Quasi-Dialog – wir sprechen im Fort-

gang nur noch von einem Dialog oder einer Wechselrede<sup>1</sup> – zwischen einer menschlichen Nutzer:in und einem KI-gestützten Chatbot vorgestellt und mit Blick auf die Kompetenzen diskutiert, die für die souveräne Verwendung dieser sprachverarbeitenden Programmanwendung benötigt werden. Im Anschluss an die sprach-, kultur- und kommunikationswissenschaftlichen sowie didaktischen Ausführungen stellen wir erste Ergebnisse einer empirischen Studie vor, in der die Praktiken und Strategien von Schüler:innen einer 10. Klasse untersucht werden, die diese bei der gemeinsamen Textarbeit mit dem Chatbot ChatGPT-4o-mini einsetzen.

## 2 Theoretische Vorüberlegungen zu Chatbots

Stalder bemerkt bereits 2016 und damit sechs Jahre vor Einführung von ChatGPT im November 2022, dass „Tätigkeiten, die noch vor Kurzem unzweifelhaft der menschlichen Intelligenz vorbehalten schienen, beispielsweise das Verfassen von Texten oder die Inhaltsanalyse von Bildern, [...] inzwischen immer häufiger Maschinen“ übernehmen (Stalder 2016: 173). Anders als die von Stalder erwähnten, auf einem Large Language Model (LLM) basierenden Textgenerierungsmodelle, die häufig in Kombination mit ebenfalls algorithmisch basierten Text- und Bildanalyseverfahren bspw. im Bereich der Börsen- und Finanzberichterstattung oder auch der Sportberichterstattung kurze Nachrichtentexte für die Publikation in etablierten Organisationen des Mediensystems generieren (vgl. ebd.: 174), stellen ChatGPT und vergleichbare KI die vorläufige Spitze der KI-Evolution dar (vgl. Albrecht 2024: 16). Die Nutzung von Chatbots wie ChatGPT des US-amerikanischen Entwicklers und Startups OpenAI ist nicht mehr exklusiv an die Funktionen und finanziellen Potenziale gesellschaftlicher Teilsysteme und ihrer Organisationen gebunden. Solche Chatbots

.....

1 Wir schließen uns aus einer phänomenologischen Perspektive somit der im Alltag gebräuchlichen Rede von einem *Dialog mit ChatGPT* an und sehen es als Aufgabe an, die von den menschlichen Akteuren in diesem Dialog eingesetzten Praktiken verstehend zu rekonstruieren.

- a. können prinzipiell von jeder in einer Schriftkultur sozialisierten Einzelperson mit einem internetfähigen Endgerät verwendet werden. Das Voranstehende einschränkend ist jedoch anzumerken, dass die Nutzung der im Frühjahr 2023 öffentlich zugänglich gemachten und bis dato neuesten Version des Chatbots von OpenAI, ChatGPT 4, limitiert sein kann, indem bspw. die Anzahl der Anfragen für einen bestimmten Zeitraum begrenzt wird. Solche Einschränkungen fallen bei den kostenpflichtigen Versionen von Chatbots (ChatGPT Plus, Microsoft Copilot Pro, Jasper Chat u. a.) weg, allerdings hängt deren Nutzung von den pekuniären Kapazitäten oder Kosten-Nutzen-Berechnungen möglicher Interessent:innen ab.
- b. emulieren eine kooperative Gesprächspartner:in, mit der in einem schriftlichen Dialog Anliegen der menschlichen Nutzer:in über eine potentiell unbegrenzte Anzahl von „Gesprächszügen“ dienstfertig bearbeitet werden können, wenngleich es Beobachtungen gibt, dass der Chatbot bei längeren Dialogen dazu neige, „den Fokus zu verlieren und immer stärker vom erwünschten Kommunikationsverhalten abzuweichen“ (Albrecht 2024: 18). Es ist im Übrigen dieses Angebot einer Interaktion in Form einer Konversation bzw. eines Dialogs zwischen Nutzer:in und ChatGPT, mit dem der Entwickler OpenAI sein Produkt im November 2022 auf seiner Homepage vorstellt (vgl. [OpenAI]).
- c. simulieren den Besitz eines die menschliche Vorstellungskraft überschreitenden enzyklopädischen Wissens, das vermittelt über die lexikogrammatischen Konstruktionen und Muster des einzigartigen Gedächtnismediums Schrift<sup>2</sup> (Assmann 2010: 190) in das dem Chatbot zugrunde liegende *Large Language Modell* – und der Sprache einer Gesellschaft überhaupt – eingewebt ist. Dieses

.....

- 2 Die Soziolinguistin Schneider bemerkt in dem ChatGPT zugrunde liegenden LLM das Vorhandensein eines *Written Language Bias*: „Data sets are made up of monolingual written text, mostly scraped from the web. [...] In effect LLMs probabilistically produce strings of words based on patterns observed in the numbers that represent written letter sequences, which, in turn, derive from texts rendered in a standardized phonetic script.“ (2024, 5)

in der Schriftsprache einkodierte Wissen kann mit Hilfe der von menschlichen Nutzer:innen formulierten *Prompts – Instruktionen*, die entweder als syntaktisch wohlgeformte Fragen, Aufgabenstellungen und Bitten oder weniger zielgenau als syntaktisch unverbundene Aneinanderreihung von Schlagwörtern formuliert werden – in Form von Texten und Bildern verfügbar gemacht und prinzipiell als Ausgangspunkt für neue Ideen und Gedanken produktiv gemacht werden (s. hierzu Assmann 2010: 193).

Chatbots wie ChatGPT können als digitales Faktotum verwendet werden, dem die Nutzer:in der KI eine Reihe von schriftlich zu erledigenden, unterschiedlich anspruchsvollen Aufgaben in Form von Prompts zuweisen kann, um ihren Arbeitsaufwand zu reduzieren, Arbeitsprozesse zu beschleunigen und somit ihre limitierten Arbeitskapazitäten effizienter nutzen zu können. Nicht zuletzt können Chatbots helfen, bei der Erstellung von Texten verschiedener Gattungen Qualitätsstandards zu erreichen, die zu realisieren der Nutzer:in ohne die Unterstützung der KI nicht möglich wären (vgl. für das wissenschaftliche Schreiben Reinmann/Watanabe 2024: 39). Auch für die Beantwortung von Wissensfragen oder für die Recherche von Informationen sind Chatbots einsetzbar, darin den Nutzungspotenzialen von einschlägigen Suchmaschinen wie Google, Bing, DuckDuckGo usw. nicht unähnlich.<sup>3</sup> Anders als bei der durch die Eingabe von einem oder mehreren syntaktisch unverbundenen Suchbegriffen relativ niedrighschwelliger Informationsrecherche mit Suchmaschinen erfordert die zielgenaue Informationsrecherche mit Chatbots wie ChatGPT die Formulierung der Suchanfrage in einer tendenziell elaborierten, satzförmigen Gestalt – in dem Eingabefeld von ChatGPT steht die Aufforderung „Stelle eine Frage“ – sowie die Fähigkeit und Bereitschaft der Nutzer:innen, einen auf den *Prompt* respondierenden, möglicherweise längeren Text zu lesen, diesen mit Blick auf den antizipierten Output abzu-

.....  
3 Fürstenberg/Müller weisen kritisch darauf hin, dass den von ChatGPT generierten Produkten inhaltlich nicht getraut werden könne. Aufgrund der „probabilistische[n] Anlage eines Transformernetzwerkes“ (2024: 5) können die von einer KI generierten Texte nie wahr, sondern nur wahrscheinlich sein (vgl. ebd.).

gleichen und ggf. den Dialog unter Bezugnahme auf zuvor Geschriebenes und Gelesenes fortzusetzen. Dienen Suchmaschinen also dem Suchen und Finden von Informationen aus einer Vielzahl der vom Algorithmus der jeweiligen Suchmaschine selegierten Websites, Artikel, Datenbanken und anderen Ressourcen, ist u. E. das Führen einer Wechselrede als die zentrale Affordanz dialogischer KI für die nach Unterstützung bei der Bearbeitung von Aufgaben oder Problemen suchenden Nutzer:in anzusehen. Allerdings erfordert das Führen eines den Zwecken der menschlichen Nutzer:innen dienenden Dialogs mit Chatbots auf Seiten der Nutzer:innen das Vorhandensein textproduktiver und -rezeptiver Kompetenzen, da, um die Potentiale der KI optimal ausschöpfen zu können, nicht nur zielgenaue Prompts zu formulieren sind, sondern auch ein Verstehen und Evaluieren des generierten textuellen Outputs erforderlich ist.

Dieser schriftlich oder – für die Erledigung eher beiläufiger Anfragen – auch mündlich zu realisierende Dialog zwischen menschlicher Nutzer:in und Chatbot sowie die dafür benötigten Kompetenzen werden in dem folgenden Abschnitt genauer beleuchtet.

### **3 Kontingenzerfahrungen mit schriftlichen Prompts und Kontingenzzreduktionen durch Anschlusskommunikation**

Die Verwendung eines Chatbots ist von den Intentionen und Handlungszielen der menschlichen Nutzer:in motiviert. Der in das Dialogfeld geschriebene Prompt instruiert den Bot, einen von der Nutzer:in in Grundzügen antizipierten textuellen Output zu erstellen. In Abhängigkeit von den Aufgaben, die Chatbots mittels Prompts zugewiesen werden, können die Erwartungen der Anwender:in an das von der KI zu generierende Textprodukt stark variieren. Weist die Nutzer:in die KI an, den Text einer bestimmten Textsorte zu erstellen, mit deren Merkmalen die Nutzer:in vertraut ist, so antizipiert sie bei der Formulierung des Prompts den Text, den die KI erstellen soll, und ist in der Lage, das von der KI vorgeschlagene Resultat mit ihrem Textsortenwissen abzugleichen. In vielen Fällen wird das zu erstellende Textprodukt

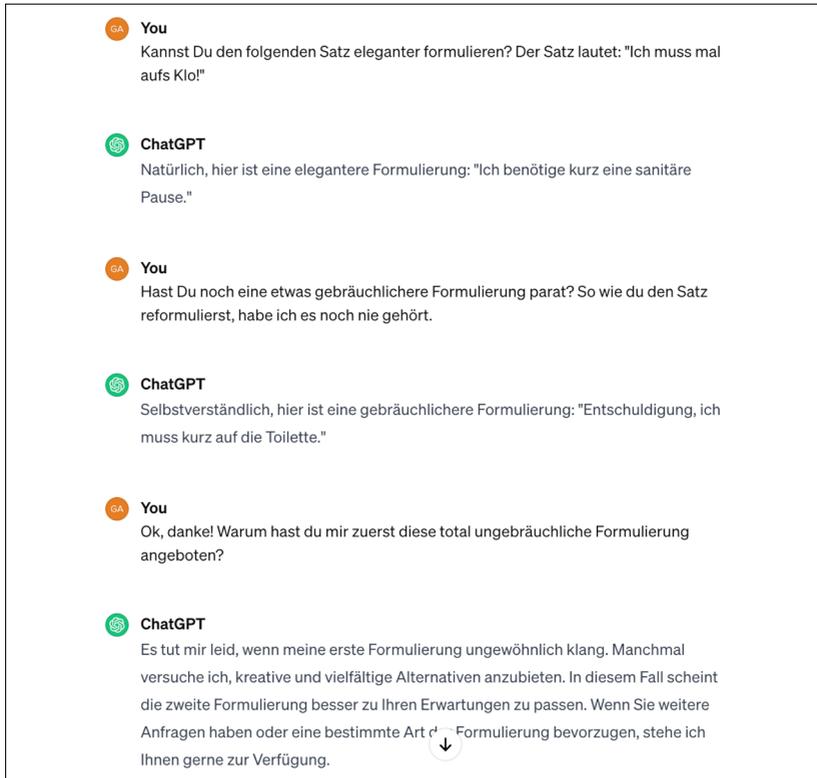
von der Nutzer:in allerdings nur auf Grundlage eher impliziter Erwartungen antizipiert, was die Formulierung eines zielgenauen, der Realisierung der Handlungsziele dienenden Prompts erschwert. Analog zu den Ausführungen des Hermeneutikers Hans-Georg Gadamer zur *Frage* ist es für einen souveränen Umgang mit Chatbots erforderlich, dass Prompts eine „eindeutige Sinnrichtung [auf]weisen“ (Gadamer 1993:195), damit eine „sinnvolle, sinngemäße Antwort“ (Gadamer 1990: 368) generiert werden kann. Fehle der Frage eine Sinnrichtung, so sei sie nach Gadamer eine schlechte Frage und schwer oder gar nicht zu beantworten (vgl. ebd.). Ähnliches kann für das Verhältnis von Prompts zu dem von ihnen generierten Output behauptet werden: Zwischen dem von der Nutzer:in gewünschten „Nutzwert“ des textuellen Outputs eines Chatbots und dem diesen evozierenden Prompt besteht ein Bedingungsverhältnis.

Vor dem Hintergrund dieser Ausführungen wird nachvollziehbar, warum sich parallel zu der viralen Verbreitung des Chatbots ChatGPT nach seiner Einführung Ende 2022 ein rapide wachsender Markt entwickelt hat, in dem Dienstleistungen zur optimalen Nutzung von Chatbots angeboten und nachgefragt werden (vgl. Becker 2024). Insbesondere die „Verfeinerung von Prompts“ (ebd.: 89) mit Hilfe sog. *Prompt-Engineers* – die nach heutigem Stand neuen *Literati* des postdigitalen Zeitalters – soll bspw. Unternehmen oder auch Bildungseinrichtungen helfen, die Potenziale der KI für ihre jeweiligen Zwecke auszuschöpfen. Die Modellierung von Prompts zur zielgenauen und effizienten Nutzung von ChatGPT ist demnach voraussetzungsreich und erfordert das Vorhandensein einer „Prompting-Kompetenz“ (Reinmann/Watanabe 2024: 39), die gerade auch von den Sprach- und Kulturwissenschaftler:innen sowie der Text- und Gesprächsdidaktik zu modellieren wäre. Zu dieser textproduktiven Kompetenz gehören wesensnotwendig bestimmte sprachlich-textuelle Fähigkeiten, die man im weitesten Sinne als bildungssprachlich charakterisieren kann; konstitutiver Bestandteil dieser Kompetenz ist zudem ein differenziertes Verständnis von der Funktionsweise, den Potenzialen und Limitierungen von Chatbots als Interface zwischen menschlicher Nutzer:in und einem LLM.

Kann die prinzipielle Kontingenz des Outputs von Chatbots mithilfe reflektiert und differenziert formulierter Prompts also reduziert werden, so bleiben dialogische KI-Systeme doch eine „algorithmische Black-Box“ (Ifenthaler 2023: 80). Wird im Dialog mit ChatGPT eine sprachliche Äußerung von der KI als Output angeboten, „ist es schwer nachzuvollziehen, warum genau dieser Text generiert wurde und welche internen Gewichtungen und Muster dazu geführt haben.“ (Arnold 2024: 69). Dass nur schwer nachzuvollziehen ist, welche algorithmischen Prozesse dem Output eines Chatbots zugrunde liegen, heißt allerdings nicht, dass die Rekonstruktion seiner Generierung zumindest approximativ nicht möglich ist, ja, für die Ausbildung einer Prompting-Kompetenz sogar hilfreich sein kann. Wenn ChatGPT als „Ko-Aktant“ (vgl. Steinhoff 2024) beim Prozess des Schreibens von Texten im Deutschunterricht fungieren soll, müssen Schüler:innen mit Hilfe entsprechender didaktischer Settings folglich lernen,

1. ein Bewusstsein für mögliche Kontingenzerfahrungen zu entwickeln,
2. welche Stereotype und Vorurteile in das LLM „einprogrammiert“ sind, das dem jeweiligen Chatbot zugrunde liegt,
3. welche sprachlichen Strategien und Praktiken bei der Erstellung von Prompts geeignet sind, solche Kontingenzen zu reduzieren.

Das im Folgenden erläuterte Beispiel veranschaulicht, wie eine den Dialog zwischen einem menschlichen Nutzer und ChatGPT initiiierende Aufgabenformulierung zu einem kontingenten Output führen kann. Der Austausch wurde 2023 als vorwissenschaftliches Experiment zwischen einem der Verfasser zusammen mit seinem 13-jährigen Sohn und dem Chatbot ChatGPT in der Version 3.5 geführt:



**Abb. 1:** Sprachwissenschaftliches Quasi-Experiment mit ChatGPT

Der Nutzer wird nicht nur mit einem erwartungswidrigen und idiosynkratischen Formulierungsvorschlag konfrontiert, sondern auch die erbetene Erklärung für den Vorschlag dieser Formulierung überrascht: „Manchmal versuche ich, kreative und vielfältige Alternativen anzubieten.“ Für den Nutzer ist diese Erklärung nicht nur deshalb beunruhigend, weil der Bot, ohne dies explizit zu machen, kreative und vielfältige Alternativen anbietet, sondern auch und gerade deshalb, weil er diese „manchmal“ anbietet.

Das Beispiel macht u. E. deutlich, dass der von dem Prompt evozierte textuelle Output – in dem vorliegenden Fall eine alternative Formulierung für die im Stil einer informellen, alltagssprachlichen Mündlichkeit formulierte Ankündigung der Verrichtung einer Notdurft – mit Blick auf die in dem Prompt in

einem bestimmten Maße antizipierte Antwort der KI evaluiert werden muss.<sup>4</sup> Fällt, wie in dem vorliegenden Fall, die Evaluation negativ aus, werden die Antwortmöglichkeiten der KI mit einem weiteren Prompt bspw. durch die Einführung von Antwortkriterien eingeschränkt. Die von dem Chatbot als verlaufsoffen und zeitlich unbegrenzt angebotene Konsultation erlaubt dem Nutzer somit eine schrittweise Annäherung an eine Antwort, bis dieser den Zweck der Konsultation mit der KI erreicht sieht.

Wenn weiter oben von *Prompting-Kompetenz* gesprochen wurde, dann machen die voranstehenden Ausführungen deutlich, dass diese textproduktive Kompetenz von einer schriftlichen Interaktionskompetenz – Steinhoff spricht von einer „Koaktivitätskompetenz“ (2024: 96) – begleitet sein muss, sollen die Potenziale textgenerativer Chatbots möglichst optimal ausgeschöpft werden. Vor dem Hintergrund dieser Überlegungen kann ein Blick in die Gesprächslinguistik und -didaktik instruktiv sein. In seiner Modellierung von Gesprächskompetenz unterscheidet Becker-Mrotzek u. a. die Teilkompetenz „Fähigkeit

.....

4 Weiter oben wurde bereits erwähnt, dass der souveräne Umgang mit Chatbots eine genauere Kenntnis des dem jeweiligen Bot zugrunde liegenden LLM erfordert. Die Evaluation des von einem Prompt generierten Outputs hat demnach nicht nur mit Fokus auf den Prompt und seine Eignung zu erfolgen, den benötigten textuellen Output zu generieren, sondern in die Reflexion einzubeziehen ist auch das „hinter“ dem Chatbot liegende LLM. So sind nämlich in das LLM von ChatGPT, wie eine aktuelle Studie von Gandalf Niklas und Aylin Caliskan (2024) mit Hilfe statistischer Methoden nachweist, kulturelle Stereotypen zu unterschiedlichen sozialen Kategorien (Gender & Sexuality, sozialer Status, Moral u. a. m.) einkodiert, die sich im Dialog zwischen Mensch und Chatbot wiederum reproduzieren – im schlimmsten Fall unbemerkt, die Stereotype affirmierend. Vor dem Hintergrund dieser Ausführungen könnte eine Erklärung für die von ChatGPT in unserem kleinen Versuchsdialog vorgeschlagene – tatsächlich, wie im Prompt gewünscht, elegante – Formulierung „Ich benötige kurz eine sanitäre Pause“ lauten, dass ein in allen Kulturen seit jeher schambehaftetes Thema wie die Ankündigung der Verrichtung der Notdurft nur sprachlich verhüllt zur Sprache kommen sollte. Als Beleg für diese Deutung kann eine Untersuchung des Ethnologen Hans Peter Duerr dienen, der mit Hilfe von zahlreichen Quellen nachweist, wie bereits im Mittelalter Euphemismen für diese intime Verrichtung auf dem „Ab-Ort“ in den europäischen Sprachen verbreitet waren: Man sprach von dem Klosett bspw. als dem „haymlich gmäch“, dem „necessarium“ oder „prifet“ (1992, 211); die sprachliche Tabuisierung ging in Frankreich sogar so weit, dass man von dem Klosett von einem „so ekelhaften Ort“ sprach, „daß man ihn nicht beim Namen nennen kann“ (ebd.). Auch heute noch werden Euphemismen zur Bezeichnung der Verrichtung der Notdurft verbreitet genutzt: So teilen bspw. Angestellte im Einzelhandel ihren Kolleg:innen mit, dass sie „auf 17 gehen“. ChatGPT schreibt mit der Formulierung „sanitäre Pause“ die Geschichte der schicklichen bzw. schamhaften Umschreibung des Klosetts und seiner Nutzung erkennbar fort!

zum Einsatz von Unterstützungsverfahren“ (2020: 80). Der Einsatz solcher Unterstützungsverfahren im Gespräch wird erforderlich, wenn Gesprächsteilnehmer:in<sub>x</sub> anhand der Beiträge von Gesprächsteilnehmer:in<sub>y</sub> nicht nur feststellt, sondern auch richtigstellen möchte, dass Gesprächsteilnehmer:in<sub>y</sub> einen vorangegangenen Gesprächsbeitrag nicht in dem intendierten Sinne versteht. Das heißt, dass Gesprächsteilnehmer:in<sub>x</sub> zunächst das Verstehen eines Gesprächsbeitrags<sub>n</sub> von Gesprächsteilnehmer:in<sub>y</sub> als Verstehensleistung eines vorangegangenen Beitrags<sub>n-1</sub> von Gesprächsteilnehmer:in<sub>x</sub> leisten muss, um erforderlichenfalls im nächsten Schritt das von Gesprächsteilnehmer:in<sub>x</sub> intendierte Verstehen ihres Gesprächsbeitrags<sub>n-1</sub> durch Gesprächsteilnehmer:in<sub>y</sub> mit Hilfe eines das intendierte Verstehen unterstützenden Gesprächsbeitrags<sub>n+1</sub> herzustellen. Erwartungsgemäß lässt sich dieser dialogische Aushandlungsprozess in struktureller Hinsicht auch im Schriftdialog mit ChatGPT (Abb. 1) beobachten: Der auf die Initiierung des Dialogs durch den menschlichen Nutzer erfolgende erste Output („sanitäre Pause“) des Chatbots veranlasst den Nutzer im Fortgang, einen *Reparatur-Prompt* zu formulieren (nachzuprompten), in dem der von dem ersten Prompt evozierte Formulierungsvorschlag nicht nur als erwartungswidrig evaluiert wird, sondern in dem auch konkretere Erwartungen an die angeforderte Alternativformulierung eingeführt werden („Usualität“).

Prompts, die als Antwort der Nutzer:innen von Chatbots auf den von vorangegangenen Prompts evozierten sprachlich-textuellen Output verfasst werden, können i. S. der Lesesozialisationsforschung demnach als schriftliche Anschlusskommunikationen verstanden werden, in denen sich das Verstehen des textuellen Outputs durch die Nutzer:in manifestiert. Kompetenztheoretisch gewendet heißt das, dass die Nutzer:innen von Chatbots die Fähigkeit benötigen, den durch einen von ihnen formulierten Prompt

- a. evozierten textuellen Output verstehend lesen,
- b. diesen mit den antizipierten Erwartungen abgleichen und
- c. auf Grundlage dieses Abgleichs ggf. einen inhaltlich, logisch konsekutiven Prompt formulieren

zu können.

Diese Analyse eines kurzen Dialogs mit ChatGPT und der Blick in einschlägige Forschungsliteratur zeigt Dreierlei:

1. Der von einem Prompt evozierte textuelle Output des Chatbots kann mit Blick auf die Handlungsziele der menschlichen Nutzer:in und den von ihr formulierten Prompt kontingent und inadäquat ausfallen.
2. Die von ChatGPT auf Grundlage eines Prompts generierten Texte oder Sätze erfordern von der kompetenten Nutzer:in der KI, dass sie den Output mit Blick auf das die Formulierung des Prompts leitende Erkenntnisinteresse evaluiert und mit ihren Erwartungen an eine Antwort abgleichen kann.
3. Auf Grundlage der Evaluation des sprachlich-textuellen Outputs des Chatbots kann die Kontingenz durch Anschlusskommunikation in Form von „Nachprompten“ sukzessive reduziert werden.

## 4 Textschreiben mit ChatGPT im Deutschunterricht

In dem vorliegenden Kapitel werden zunächst Einsatzmöglichkeiten von Chatbots skizziert, bevor wir auf das Schreiben von Schülertexten unter Zuhilfenahme von ChatGPT zu sprechen kommen. Dafür gehen wir von Tekins (2024) Systematik aus, die für den Kontext des Unterrichts von Deutsch als Fremdsprache vier Rollen differenziert, die die Nutzer:in ChatGPT mithilfe von Prompts zuweisen kann: *Generator*, *Evaluator*, *Transformator* und *Kommunikator* (vgl. ebd.: 138). Geht es bei den drei zuerst genannten Rollen *Generator*, *Evaluator* und *Transformator* im Kern um die Arbeit mit Schrifttexten, also um deren Erstellung, Überarbeitung und Veränderung, um Beurteilung, Bewertung und Korrektur, verhält es sich mit der Verwendung von ChatGPT als Kommunikator im DaF-Unterricht etwas anders. Zunächst stellt Tekin das Offenkundige klar, dass ChatGPT nämlich ständig mit uns ‚kommuniziert‘ (vgl. 152), um dann die spezifischen Potenziale von ChatGPT in der Rolle als *Kommunikator* für das Lernen von Deutsch als Fremdsprache zu umreißen: So sei der Einsatz von ChatGPT im DaF-Unterricht eine Lerngelegenheit, ein schriftliches Gespräch in der Zielsprache bspw. zu Themen der Zielkul-

tur zu führen (ebd.: 152) oder grammatische und semantische Fragen zu der Zielsprache im Diskurs mit der KI zu bearbeiten (ebd.: 156). In den beiden Szenarien liegt das Lernziel der DaF-Lerner:innen in dem Aufbau von Diskurskompetenzen sowie in der diskursiven Vermittlung von Wissen in der Zielsprache und -kultur, *indem* sie einen Dialog mit dem Chatbot führen. Die Verwendung des Chatbots als Generator, Evaluator oder Transformator und die zu diesen Zwecken geführten Dialoge zwischen Nutzer:in und KI zielen dagegen auf die Erstellung von Texten, auf die Korrektur, Bewertung oder Übersetzung von Texten u. v. a. m. ab. Wir sprechen bei diesen Verwendungsweisen hier und im Folgenden deshalb von *textbezogenen Dialogen*.

Die dargestellten Typen möglicher Verwendungen von ChatGPT im DaF-Unterricht können weiterführend als Heuristik für die Didaktik des Textschreibens im Deutschunterricht dienen, wobei zum einen die Potenziale und Chancen von ChatGPT für die Seite der Lehrenden, zum anderen die Chancen und Herausforderungen für die Lernenden und deren Lernprozesse herausgearbeitet werden können. In dem vorliegenden Artikel gilt unser Interesse der Seite der Lerner:innen und des Lernens mit Chatbots, und wir richten unseren Fokus auf die der KI zugewiesene Rolle des *Transformators*, in der nach Tekin Texte zusammengefasst, vereinfacht oder übersetzt, der Schreibstil von Texten oder einfach die Formatierung eines Textes verändert werden (vgl. 2024: 138). Ausdrücklich für den DaF-Unterricht entwickelt, erwähnt Tekin bei der Beschreibung der Verwendungsoption von Chatbots als Transformator etwas überraschend nicht die *Textüberarbeitung*, die nicht nur in den bekannten Schreibprozessmodellen von zentraler Bedeutung ist, sondern die auch aus den Bildungsplänen für das Fach Deutsch nicht wegzudenken ist. An dieser Stelle sind die Ausführungen Steinhoffs weiterführend, der von der sog. Partizipanden-Rolle des Chatbots als „Writing Partner“ *von* Schüler:innen oder als „Writing Tutor“ *für* Schüler:innen spricht, bei der die menschliche Nutzer:in die KI „bewusst und gezielt in eine Schreibpraktik einbezieht“ (Steinhoff 2024: 95) – im erstgenannten Fall als gleichberechtigter Akteur, im zweiten Fall als „Quasi-Lehrkraft“ (ebd.: 94). Steinhoff zufolge kann die KI in beiden Partizipanden-Rollen die Schüler:innen bei der Überarbeitung von Texten unterstützen (ebd.: 94, 95). Dabei hat die Unterscheidung der beiden Rollen u. E. zwar heuristischen Nutzen, bei näherer Betrachtung beschreiben die beiden Partizipanden-Rollen jedoch eher die zwei

Enden einer bipolaren und kontinuierlichen Skala, auf der die Eigenständigkeit und Textsouveränität der menschlichen Schreiber:in an dem Pol „Writing Partner“ maximal und an dem entgegengesetzten Pol „Writing Tutor“ minimal ausgeprägt ist. Von uns erhobene Daten zeigen zudem, dass sich die Nutzer:in im Verlauf eines Dialogs mit einem textgenerativen Chatbot auf dem skalaren Kontinuum zwischen den beiden Polen seinen jeweiligen Anforderungen und Handlungszielen entsprechend variabel bewegt. Wir sprechen im Folgenden deshalb von dem Chatbot als *Writing-Ko-Aktant* (im Folgenden abgekürzt mit WKA), wenn dieser von einer menschlichen Nutzer:in für einen gemeinsamen Schreibprozess genutzt wird.

Der Chatbot bzw. die ihm zugrundeliegende LLM-Technologie bilden zusammen mit der menschlichen Nutzer:in demnach einen neuartigen Typ „gemeinsamer Schreibeinheit in Autorenteam“ (Lehnen 2014: 425). Wenn, wie Lehnen es für schriftliche Verfahren der kooperativen Textherstellung allgemein festhält, die Beteiligten solcher Teams „einen komplexen Prozess koordinieren und Aufgaben aufeinander abstimmen“ (ebd.) müssen, dann stellt sich nicht nur die theoretisch und empirisch zu beantwortende Frage, wie die Prozesse und Aufgaben beim Schreiben eines Textes gemeinsam mit einer KI als WKA durch den menschlichen Ko-Aktanten koordiniert und sprachlich realisiert werden, sondern auch, welche spezifischen Chancen und Herausforderungen diese Form der Partnerarbeit für die Schüler:innen bereithält.

## 5 Studie

In dem vorliegenden Kapitel werden erste Resultate der Teilstudie einer größeren Datenerhebung vorgestellt, in der die Praktiken und die diesen zugrunde liegenden Kompetenzen von Schüler:innen einer 10. Klasse des gymnasialen Zweigs einer Gesamtschule bei der aufgabengesteuerten Überarbeitung von Inhaltsangaben als tradierter Aufsatzgattung des Deutschunterrichts mit Hilfe des WKA ChatGPT-4o-mini untersucht werden.<sup>5</sup>

.....  
5 Der kostenfreie Zugang zu ChatGPT-4-mini wurde den Wissenschaftler:innen dankenswerterweise von dem Unternehmen *fobizz* ermöglicht.

## 5.1 Studiendesign

Das vorrangige Erkenntnisinteresse der Studie gilt den sprachlichen Praktiken sowie den Strategien, die von Nutzer:innen in dem schriftlichen Dialog KI-gestützten Chatbots eingesetzt werden, um die gemeinsame Bearbeitung der Aufgabe zu bewerkstelligen. Ziel der Studie ist die Identifikation rekurrenter sprachlicher und textueller Handlungsmuster sowie die Typisierung strategischer Verfahren der menschlichen Nutzer:innen bei der gemeinsamen Textarbeit mit LLMs via Chatbots.

Wir haben uns in der Studie für eine Aufgabenstellung zur Textüberarbeitung entschieden, mit der die Komplexität des Textschreibprozesses adressiert werden kann. Das Überarbeiten ist nämlich nicht als Teilprozess eines übergeordneten Textschreibprozesses zu verstehen, sondern stellt einen eigenständigen Schreibprozess mit entsprechenden Teilprozessen dar:

It involves planning a solution to the problem (in written form or not), translating that solution into language, and transcribing that language into new text to replace the old text. In this view, revision, like planning, is seen not as a separate writing process parallel to the other writing processes identified [...] but rather as a special application of the writing model. (Hayes 2012, 276)

Die Aufgabe zur Überarbeitung eines Ganztextes erlaubt den Schüler:innen, selbst zu entscheiden, ob sie eine möglichst arbeitssparsame Strategie wählen, indem sie die Bearbeitung der Aufgabe und die Verantwortung für das Resultat primär oder vollständig an ihren WKA ChatGPT delegieren, oder ob sie die Aufgabe als Möglichkeit wahrnehmen, gemeinsam mit der KI den gesamten Schreibprozess (s. das Zitat von Hayes) einer Überarbeitung zu durchlaufen, um souverän und gewissermaßen als „Herr des Verfahrens“ eine überarbeitete und verbesserte Textversion zu erstellen.

Die „didaktische Gattung“ (Dix 2017, 57 ff.) der Inhaltsangabe für eine Studie mit Schüler:innen einer 10. Klassenstufe wurde herangezogen, weil diese im Aufsatzunterricht des Faches Deutsch der Sekundarstufe I fast durchgehend unterrichtet wird. Die Aufgabe, die Inhaltsangabe einer anderen Schüler:in

mit dem WKA ChatGPT zu überarbeiten, kann als *interessanter* „type of task“ (Boscolo/Hidi 2007: 6) bestimmt werden: „A student [...] finds writing interesting if the instructional situation allows him/her to discover and practice the attractive, unusual, and challenging aspects of the activity, which may not, and usually do not emerge from traditional writing tasks.“ (ebd.) Die mit der Aufgabenstellung u. a. intendierte Herstellung situativen Interesses soll eine hohe Motivation der an der Studie teilnehmenden Schüler:innen (im Folgenden *Proband:innen*) gewährleisten.

Kurz einige Worte zur Datenerhebung: Die *Proband:innen* hatten die Aufgabe, die Inhaltsangabe einer Schülerin aus der 9. Klasse einer Gesamtschule zu der Kurzgeschichte „Kahlschnitt“ von Kirsten Boie (1993) zusammen mit einem Chatbot zu überarbeiten. Um sicherzustellen, dass die an der Studie teilnehmenden Proband:innen mit der Kurzgeschichte ausreichend vertraut waren, haben diese vor Bearbeitung der Überarbeitungsaufgabe [Auftrag<sub>2</sub>] selbst eine Inhaltsangabe der Kurzgeschichte an einem Tablet mit Hilfe eines handelsüblichen Textverarbeitungsprogramms verfasst [Auftrag<sub>1</sub>]. Für die Bearbeitung von Arbeitsauftrag<sub>1</sub> erhielten die Proband:innen die Aufgabenstellung, die Merkmale einer Inhaltsangabe in Listenform sowie die Kurzgeschichte von Kirsten Boie auf einem ausgedruckten Arbeitszettel ausgehändigt. Für die Überarbeitung der Inhaltsangabe [Auftrag<sub>2</sub>] durften die Proband:innen auf das Arbeitsblatt aus Aufgabe<sub>1</sub> zurückgreifen. Arbeitsauftrag<sub>2</sub> wurde den Proband:innen in Form eines digitalen Dokuments eines handelsüblichen Textverarbeitungsprogramms auf dem Tablet zur Verfügung gestellt; in das Dokument sollte auch das Resultat der Überarbeitung hineinkopiert werden.

Für die Bearbeitung des oben umrissenen Erkenntnisinteresses wurden mit Hilfe eines Fragebogens personenbezogene Daten, Interviewdaten sowie verschiedene Schreibprozess- und Schreibproduktdata erhoben, auf die im Fortgang auszugsweise zurückgegriffen wird. Die genannten Interviewdaten wurden im Anschluss an die Erhebung der Schreibdaten als Gruppeninterview mit 21 Proband:innen erhoben. In dem offen geführten Interview erhielten die Proband:innen die Möglichkeit, ihre Erfahrungen und Eindrücke von der Studie im Speziellen und zum Thema KI, Chatbots und Digitalisierung im Allgemeinen zu thematisieren. Aus den Aussagen der Proband:innen kristallisierten sich Hinweise auf zwei gegensätzliche Vorgehensweisen bei der

Bearbeitung der Aufgabe heraus. Auf Grundlage dieser im Gruppeninterview thematisierten gegensätzlichen Vorgehensweisen wurde im Anschluss an eine erste Sichtung und Sortierung der Daten ein Sample mit zwei Datensätzen aus einem Untersuchungskorpus mit aktuell 37 Datensätzen gezogen. Von diesen zwei Datensätzen wurden wiederum die textbezogenen Dialoge, die in Form von Chatprotokollen vorliegen, einer genaueren Analyse unterzogen (s. Anlage I und II).<sup>6</sup>

## 5.2 Zwei Typen textbezogener Dialoge

Die Analyse der beiden in Form von ChatGPT-Protokollen vorliegenden textbezogenen Dialoge zeigt zwei mit Blick auf

- die Verwendung sprachlicher Mittel,
- die Gestaltung des kollaborativen Arbeitsprozesses mit dem Chatbot ChatGPT
- das In-Beziehung-Treten der menschlichen Nutzer:in mit dem Chatbot

kriteriell gut zu unterscheidende Ansätze. Wir unterscheiden im Folgenden zwischen

- a. *explorativ-sukzessiven* und
- b. *fokussiert-antizipativen*

textbezogenen Dialogen.

Ad a: Der explorativ-sukzessive textbezogene Dialog (s. Anhang I) zeichnet sich durch einen Prozess mit zahlreichen Überarbeitungsschleifen aus. Proband:in<sub>09</sub> formuliert im Stil informeller Mündlichkeit einen Initialprompt [Prompt<sub>1</sub>], bei dem es sich um die unvollständige Ankündigung eines Arbeitsauftrags handelt. Trotz der Unvollständigkeit des Prompts generiert der

.....

6 Die anhängenden Chatprotokolle wurden für die Veröffentlichung orthographisch und grammatikalisch nicht korrigiert.

Chatbot in seiner Rolle als WKA eine kooperative Antwort, in der er nicht nur antizipiert, dass eine Textzusammenfassung gewünscht ist, sondern in der er Proband:in<sub>09</sub> auch auffordert, den Text, der zusammengefasst werden soll, der KI mitzuteilen [Output<sub>1</sub>]. Der WKA gibt mit seiner Rückmeldung gewissermaßen den Folgeprompt von Proband:in<sub>09</sub> vor und strukturiert auf diese Weise den Dialog und den Arbeitsprozess. Diese formuliert daran anschließend einen vervollständigten Initialprompt [Prompt<sub>2</sub>], in dem sie von ihrem ursprünglichen Plan abweicht, dem Chatbot als ersten Schritt die Inhaltsangabe zu „geben“ [Prompt<sub>1</sub>]. Stattdessen „gibt“ sie diesem jetzt zuerst die Kurzgeschichte. Diese fügt sie direkt an ihre einleitende Formulierung mittels Copy & Paste in das Dialogfeld ein. Direkt im Anschluss daran gibt sie ihrem WKA Kontextinformationen, expliziert das intertextuelle Verhältnis der beiden Texte und formuliert eine Anweisung, was mit der Inhaltsangabe zu tun ist. Mittels Copy & Paste kopiert sie sodann die zur Überarbeitung vorgesehene Inhaltsangabe in das Dialogfeld, bevor sie den Chatbot erneut deiktisch orientiert und auch den Arbeitsauftrag in Frageform leicht abgewandelt erneut formuliert. Prompt<sub>2</sub> ist demnach eine intern komplex strukturierte sprachliche Instruktionshandlung, in der nicht nur eine Anweisung formuliert wird, sondern in der auch basale Informationen über die zwei in den Prompt in einer sinnvollen Reihenfolge hineinkopierten Texte sowie die intertextuelle Relation zwischen ihnen expliziert wird. Im Anschluss an Prompt<sub>2</sub> fordert Proband:in<sub>09</sub> gleich dreimal Nachbesserungswünsche [Prompt<sub>3-5</sub>] zu den vorgeschlagenen Überarbeitungsversionen [Output<sub>2-4</sub>] der KI ein. Zusammengefasst zeichnet sich der explorativ-sukzessive textbezogene Dialog zwischen Nutzer:in und Chatbot durch folgende Merkmale aus:

- Proband:in<sub>09</sub> tritt mit dem Chatbot mithilfe des Einsatzes informeller Mündlichkeit, anthropomorphisierender Anreden sowie mithilfe von Höflichkeitsformen in Beziehung, deren Verwendung prototypisch für zwischenmenschliche Dialoge ist. Belegstellen:
  - Expressive Nutzung von Interpunktionszeichen (s. Prompt<sub>2</sub>: „kannst du sie bitte korrigieren????“)
  - Vertraulichkeit suggerierende Anrede des Chatbots mit der zweiten Pers. Sing. („Du“)

- Informelle Grußformeln in Prompt<sub>1</sub> und Prompt<sub>2</sub> („hey“)
- Handlungsaufforderungen, die als Bitte formuliert werden (Prompt<sub>3</sub> und Prompt<sub>4</sub>)
- Proband:in<sub>09</sub> verwendet grammatische Strukturen, mit denen der Chatbot als menschlicher Gegenüber konzeptualisiert wird. So wird in Prompt<sub>1</sub> das Verb *geben* verwendet, dessen semantische Rollen Agens und Rezipient das semantische Merkmal *lebendig* tragen („ich gebe dir ...“)
- Proband:in<sub>09</sub> gibt zu dem Output der KI Feedback, liefert Informationen nach und fordert Anpassungen ein. Die Bearbeitung der Aufgabenstellung erfolgt somit schrittweise, schleifenartig und unter Ausnutzung der keinen zeitlichen Limitierungen oder sozialen Normen des zwischenmenschlichen Umgangs unterliegenden dialogischen Struktur, die ChatGPT als Interface bei der Arbeit mit dem ihm zugrunde liegenden LLM anbietet.

Ad b: Der *fokussiert-antizipativ textbezogene Dialog* zeichnet sich dadurch aus, dass in ihm die Aufgabenbearbeitung nicht in einen mehrzügigen, dialogischen Aushandlungsprozess verlagert wird, sondern die Formulierung des ersten, auf Vollständigkeit abzielenden instruktiven Textprompts zielt auf unmittelbare Aufgabenerfüllung ab. In dem aus dem Datensatz gezogenen textbezogenen Dialog von Proband:in<sub>17</sub> (s. Anhang II) ist der Initialprompt [Prompt<sub>1</sub>] unvollständig, aber erneut zeigt sich der Chatbot als geduldiger und kooperativer WKA, wenn er die Proband:in<sub>17</sub> auffordert, die zu verbessernde Inhaltsangabe mitzuteilen. Die Proband:in formuliert im Anschluss daran den folgenden instruktiven Prompt [Prompt<sub>2</sub>]:

**PU** Proband17 UniKassel

Verbesser Folgende Inhaltsangabe anhand folgender Kriterien: der Einleitungssatz besteht aus Titel, Autor, Textart, Erscheinungsjahr, Thema; Der Hauptteil ist sachlich im Präsens in der 3. person konzentriere dich auf die wesentlichen informationen der stil ist sachlich und nicht wertend. Hier die Inhaltsangabe die nzu verbessern Kurzgeschichte „Kahlschnitt“ von Kristen Boie, die 1933 verfasst und erschienenen ist.

**Abb. 2:** Instruktiver Prompt von Proband:in<sub>17</sub>

Der Prompt wird erkennbar unter Zuhilfenahme der den Teilnehmer:innen der Studie noch aus der Bearbeitung von Arbeitsauftrag<sub>1</sub> vorliegenden Liste mit Merkmalen von Inhaltsangaben realisiert. Ähnlich wie Proband:in<sub>09</sub> formuliert auch Proband:in<sub>17</sub> in seinem eigentlichen Initialprompt nicht nur einen Arbeitsauftrag sowie die für die Bearbeitung dieses Auftrags relevanten Kriterien, sondern auch sie nimmt Orientierungshandlungen die beiden in das Dialogfeld kopierten Texte betreffend vor: „hier die Inhaltsangabe die nzu verbesser“ und „hier der originaltext als referenz“. Das intertextuelle Verhältnis der beiden Texte sowie die Funktion des Originaltextes für die Aufgabenbearbeitung („als referenz“) werden nicht explizit gemacht, sondern sind aus der verwendeten Lexik zu schlussfolgern. Die KI bekommt somit eine annähernd maximal spezifizierte Anweisung erteilt. Lediglich ein weiterer Auftrag wird mit Prompt<sub>3</sub> nachgereicht, der allerdings nicht die Gattungsmerkmale einer „Inhaltsangabe“ betrifft, sondern mit dem der WKA ChatGPT angewiesen wird, das Sprachniveau der in Output2 vorgeschlagenen Inhaltsangabe an das einer 9. Klasse anzupassen (Arbeitsauftrag<sub>2</sub> der Studie zu entnehmen).

Der *fokussiert-antizipativ textbezogene Dialog* lässt sich wie folgt charakterisieren:

- ChatGPT als WKA wird sprachlich als digitales Werkzeug konzeptualisiert, dessen Output mit natürlichsprachlichen Mitteln „programmiert“ werden kann. Elemente phatischer Kommunikation (z. B. Begrüßungen) und andere Anthropomorphismen fehlen; die zwei Arbeitsanweisungen werden als subjektlose Imperativsätze (Verberststellung), zudem ohne Verwendung von Höflichkeitsformeln formuliert. Die beiden deiktischen Orientierungshandlungen in Prompt<sub>2</sub> werden mittels verbloser Konstruktionen realisiert, ohne dass die zu orientierende Entität – also der Chatbot – explizit adressiert wird.
- Proband:in<sub>17</sub> formuliert zielgerichtete Instruktionen (Prompt<sub>1</sub> und Prompt<sub>2</sub>: „verbessere“ sowie Prompt<sub>3</sub>: „verändere“), aus denen die Erwartungen an den textuellen Output des Chatbots unmissverständlich hervorgehen. Der den Dialog initiiierende Textprompt [Prompt<sub>2</sub>] wird von Proband:in<sub>17</sub> tendenziell im Sinne der Textschreibdidaktik

als Lösung eines kommunikativen Problems interpretiert, dessen erfolgreiche Bearbeitung durch die exhaustive Aufzählung der Kriterien für Inhaltsangaben antizipierbar wird.

Nur der Vollständigkeit halber: Die in beiden Dialogen feststellbare schwindelerregende Nachlässigkeit in den Bereichen Grammatik, Interpunktion und Orthografie resultiert mutmaßlich nicht aus mangelnden Kompetenzen: Beide Proband:innen werden von der Lehrkraft zu den guten Schüler:innen gezählt und haben gute Zeugnisnoten im Fach Deutsch im gymnasialen Zweig einer Gesamtschule vorzuweisen. Die beiden Schüler:innen erlauben sich die formalen Lässlichkeiten mutmaßlich aufgrund ihrer Erfahrungen mit der KI als kooperatives, unterstützendes und fehlertolerantes Gegenüber sowie unter Einberechnung der der KI unterstellten Fähigkeit zum Verarbeiten auch formal fehlerhafter Eingaben.

## 6 Schlussfolgerungen

Die vorgestellten Analysen textbezogener Dialoge zwischen menschlichen Nutzer:innen und dem Chatbot ChatGPT in der Rolle als WKA deuten an, dass die den Zielen der Nutzer:innen dienende Arbeit mit einem LLM das Vorhandensein textproduktiver Kompetenzen für das Schreiben von Prompts erfordert. Zu dieser Prompting-Kompetenz gehören einerseits Formulierungskompetenzen, andererseits das Vorhandensein einer für die Erstellung schriftlicher Instruktionen benötigten Fähigkeit, zu antizipieren, welche Informationen in welcher Granularität und welche Handlungsanweisungen die KI benötigt, und wie diese Informationen und Anweisungen textuell zu strukturieren sind, damit der von der Nutzer:in erwartete textuelle Output generiert wird. Das Vorhandensein solch einer „Kontextualisierungskompetenz“ ist für das Schreiben von Instruktionen generell als besonders relevant anzusehen (vgl. Bachmann 2014), da erst ihr Vorhandensein Schreiber:innen befähigt, den kommunikativen Zweck einer Instruktion „raum- und zeitunabhängig – also ohne weitere Kontextinformationen“ (ebd., 273) – verständlich zu machen.

Die dargestellten Herausforderungen beim Abfassen eines instruierenden Prompts werden in der gemeinsamen Arbeit der Nutzer:in mit der KI als WKA bedingt durch die dialogische Struktur der Chatbots allerdings relativiert; dafür rücken andere sprachlich-kommunikative Kompetenzen in den Fokus. So hat insbesondere die Analyse des explorativ-sukzessiven textbezogenen Dialogs gezeigt, dass die Rückmeldung durch die KI das Lesen des textuellen Outputs sowie den Abgleich des dabei Verstandenen mit den Erwartungen der Nutzer:innen bedarf. Ggf. werden diese Erwartungen durch die Rückmeldung der KI der menschlichen Nutzer:in erst richtig bewusst. Folglich sind die Formulierungen konsekutiver Prompts als „Verstehensselektionen“ (i. S. v. Luhmann 1984) vorangegangener, von der KI generierter textueller Outputs zu analysieren, mit denen die Kontingenz weiterer Outputs reduziert werden soll bzw. mit denen die an den textuellen Output des Chatbots gerichteten Erwartungen möglichst umfänglich erfüllt werden sollen. Ähnlich wie im Fall von Dialogen zwischen Menschen ist zudem zu fragen, ob und in welcher Weise die von ChatGPT generierten Responses auf Prompts dazu führen, die Erwartungen und Handlungsziele der Nutzer:innen zu modifizieren bzw. zu verschieben. Hinweise auf solche Einflüsse der KI auf ihre Nutzer:innen finden sich ebenfalls in dem textbezogenen Dialog von Proband:in<sub>09</sub> mit dem Chatbot, wenn sie ihr Vorgehen nach der ersten Rückmeldung [Prompt<sub>2</sub>] durch ChatGPT augenscheinlich modifiziert (s. o.).

Die ersten und vorläufigen Analysen zweier textbezogener Dialoge legen zudem nahe, der Frage weiter nachzugehen, ob und in welcher Weise die oben skizzierten unterschiedlichen Formen des *In-Beziehung-Tretens* (anthropomorphisierend vs. technisch-direktiv) des Menschen mit der KI zusammenhängen mit bestimmten Verwendungsweisen, Nutzungspräferenzen und Erwartungshaltungen an diese Technik.

## Literatur

- ALBRECHT, STEFFEN (2024): ChatGPT als doppelte Herausforderung für die Wissenschaft. In: Schreiber, Gerhard/Ohly, Lukas (Hg.): *KI: Text. Diskurse über KI-Textgeneratoren*. Berlin; New York: de Gruyter, S. 13–27.
- ARNOLD, THOMAS O. (2024): Herausforderungen in der Forschung: Mangelnde Reproduzierbarkeit und Erklärbarkeit. In: Schreiber, Gerhard/Ohly, Lukas (Hg.): *KI-Text. Diskurse über KI-Textgeneratoren*. Berlin; New York: de Gruyter, S. 67–80.
- ASSMANN, ALEIDA (2010): *Erinnerungsräume. Formen und Wandlungen des kulturellen Gedächtnisses*. 5., durchgesehene Aufl. München: Beck.
- BACHMANN, THOMAS (2014): Schriftliches Instruieren. In: Feilke, Helmuth/Pohl, Thorsten (Hg.): *Schriftlicher Sprachgebrauch. Texte verfassen*. Baltmannsweiler: Schneider Hohengehren (Bd. 4 Deutschunterricht in Theorie und Praxis. hg. von Winfried Ulrich), S. 270–286.
- BECKER, JENIFER (2024): Können Chatbots Romane schreiben? Der Einfluss von KI auf kreatives Schreiben und Erzählen. Schreiber, Gerhard/Ohly, Lukas (Hg.): *KI-Text. Diskurse über KI-Textgeneratoren*. Berlin; New York: de Gruyter, S. 83–100.
- BECKER-MROTZEK, MICHAEL (2020): Mündliche Kommunikationskompetenz. In: Ders. (Hg.): *Mündliche Kommunikation und Gesprächsdidaktik*. Baltmannsweiler: Schneider Hohengehren (Bd. 2 Deutschunterricht in Theorie und Praxis. hg. von Winfried Ulrich), S. 66–83.
- BOSCOLO, PIETRO/HIDI, SUZANNE (2007): The Multiple Meaning of Motivation to Write. In: Hidi, Suzanne/Boscolo, Peitro (Hg.): *Writing and Motivation*. Amsterdam u. a.: Elsevier, S. 1–14.
- DIX, ANNIKA (2017): *Berichte und Berichten als didaktische Gattungen. Eine Textform zwischen Erwerb und schulischer Norm*. Baltmannsweiler: Schneider Hohengehren.
- DUERR, HANS PETER (1992): *Nacktheit und Scham. Der Mythos vom Zivilisationsprozess*. 4. Aufl. Frankfurt/Main: Suhrkamp.
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* Nr. 5/2024, S. 2–13.
- GADAMER, HANS-GEORG (1990): *Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik*. Bd. 1. Tübingen: Mohr Siebeck.

- GADAMER, HANS-GEORG (1993): *Wahrheit und Methode. Ergänzungen und Register*. Bd. 2. Tübingen: Mohr Siebeck.
- GANDALF, NICOLAS/CALISKAN, AYLIN (2024): A Taxonomy of Stereotype Content in Large Language Models. URL: <https://arxiv.org/abs/2408.00162> (letzter Zugriff: 08.3.2025)
- HAYES, JOHN R. (2012): Modeling and Remodeling Writing. In: *Written Communication* 29, S. 369–388.
- IFENTHALER, DIRK (2023): Ethische Perspektiven auf Künstliche Intelligenz im Kontext der Hochschule. In: Schmohl, Tobias/Watanabe, Alice/Schelling, Kathrin (Hg.): *Künstliche Intelligenz in der Hochschulbildung. Chancen und Grenzen des KI-gestützten Lernens und Lehrens*. Bielefeld: transcript, S. 71–86.
- LEHNEN, KATRIN (2014): Gemeinsames Schreiben. In: Feilke, Helmuth/Pohl, Thorsten (Hg.): *Schriftlicher Sprachgebrauch. Texte verfassen*. Baltmannsweiler: Schneider Hohengehren (Bd. 4 Deutschunterricht in Theorie und Praxis. hg. von Winfried Ulrich), S. 414–431.
- LUHMANN, NIKLAS (1984): *Soziale Systeme. Grundriß einer allgemeinen Theorie*. Frankfurt/Main: Suhrkamp.
- [Open AI]: <https://openai.com/index/chatgpt/> (letzter Zugriff: 11.3.2025).
- REINMANN, GABI/WATANABE, ALICE (2024): KI in der universitären Lehre. Vom Spannungs- zum Gestaltungsfeld. In: Schreiber, Gerhard/Ohly, Lukas (Hg.): *KI-Text. Diskurse über KI-Textgeneratoren*. Berlin; New York: de Gruyter, S. 29–46.
- SCHNEIDER, BRITTA (2024): A Sociolinguist's Look at the „Language“ in Large Language Models. In: *CriticalAI* (2024) 2 (1). Url: <https://doi.org/10.1215/2834703X-11205168> (letzter Zugriff: 13.3.2025).
- STALDER, FELIX (2016/<sup>s</sup>2021): *Kultur der Digitalität*. Frankfurt/Main: Suhrkamp.
- STEINHOFF, TORSTEN (2024): Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel des Schreibens mit ChatGPT in der 8. Klasse. In: Albrecht, C./Brüggemann, J./Kretschmann, T./Meier, C. (Hg.): *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven*. Stuttgart: Metzler, S. 85–99.
- TEKIN, ÖZLEM (2023): ChatGPT im Unterricht Deutsch als Fremdsprache. In: *Alman Dilive Kültürü Araştırmaları Dergisi* (ALKAD) 5 (2), S. 135–163.

## Anhang I: Beispiel für explorativ-sukzessiv textbezogene Dialoge

Dialogpartner & Zählung	Dialogzüge	In das Dialogfeld kopierte Texte bzw. von GPT generierte Texte
Proband:in <sub>09</sub> [Prompt <sub>1</sub> ]	hey ich gebe dir jetzt eine inhaltsangabe zu dem text:	
GPT als WKA [Output <sub>1</sub> ]	Klar, ich bin bereit! Bitte teile mir die Inhaltsangabe oder den Text mit, den du zusammenfassen möchtest.	
Proband:in <sub>09</sub> [Prompt <sub>2</sub> ]	hey, ich gebe dir jetzt einen text:	[Kurzgeschichte: Kahlschnitt von KB]
	Zu diesem text hat ein schüler eine inhaltsangabe geschrieben, dieser soll jetzt bitte auf rechtschreibung und auf die kriterien einer inhaltsangabe korrigiert werden. hier die inhaltsangabe, die korrigiert werden soll:	
		[Inhaltsangabe <sub>ÜB</sub> ]
	Dies ist die anhaltsangabe, kannst due sie bitte korrigieren????	
GPT als WKA [Output <sub>2</sub> ]	Hier ist die korrigierte Version der Inhaltsangabe:	
		[Überarbeitete Version I von Inhaltsangabe <sub>ÜB</sub> ]
	Ich habe die Rechtschreibung und Grammatik korrigiert sowie einige Formulierungen für mehr Klarheit angepasst.	

<b>Dialogpartner &amp; Zählung</b>	<b>Dialogzüge</b>	<b>In das Dialogfeld kopierte Texte bzw. von GPT generierte Texte</b>
Proband:in <sub>09</sub> [Prompt <sub>3</sub> ]		[Foto mit Kriterien einer Inhaltsangabe von Arbeitsblatt]
	Bitte nochmal auf diese Kriterien kontrollieren	
GPT als WKA [Output <sub>3</sub> ]	Hier ist die überarbeitete Inhaltsangabe, die den angegebenen Kriterien entspricht:	
		[Überarbeitete Version II von Inhaltsangabe <sub>ÜB</sub> ]
	Diese Version berücksichtigt die Kriterien für eine Inhaltsangabe: Sie ist in der dritten Person verfasst, konzentriert sich auf die wesentlichen Informationen, verwendet das Präsens und vermeidet unwichtige Details.	
Proband:in <sub>09</sub> [Prompt <sub>4</sub> ]	Bitte mit einem richtigen einleitungssatz	
GPT als WKA [Output <sub>4</sub> ]	Hier ist die überarbeitete Inhaltsangabe mit einem passenden Einleitungssatz:	
		[Überarbeitete Version III von Inhaltsangabe <sub>ÜB</sub> ]
	Dieser Einleitungssatz bietet einen klaren Überblick über das Thema der Geschichte.	
Proband:in <sub>09</sub> [Prompt <sub>5</sub> ]	Ohne die eigenen meinung am ende	
GPT als WKA [Output <sub>5</sub> ]	Hier ist die überarbeitete Inhaltsangabe ohne die eigene Meinung am Ende:	
		[Überarbeitete Version IV von Inhaltsangabe <sub>ÜB</sub> ]
	Diese Version enthält keine persönliche Meinung am Ende und bleibt sachlich.	

## Anhang II: Beispiel für fokussiert-antizipativ textbezogene Dialoge

Dialogpartner & Zählung	Dialogzüge	In das Dialogfeld kopierte Texte bzw. von GPT generierte Texte
Proband:in <sub>17</sub> [Prompt <sub>1</sub> ]	Verbessere folgende Inhaltsangabe anhand folgender Kriterien: einleitungssatz besteht aus titel autor tzextart erscheinungsjahr thema	
GPT als WKA [Output <sub>1</sub> ]	Gerne! Bitte teile mir die Inhaltsangabe mit, die du verbessern möchtest, und ich werde sie entsprechend deinen Kriterien anpassen.	
Proband:in <sub>17</sub> [Prompt <sub>2</sub> ]	Verbesser Folgende Inhaltsangabe anhand folgender Kriterien: der Einleitungssatz besteht aus Titel, Autor, Textart, Erscheinungsjahr, Thema; Der Hauptteil ist sachlich im Präsens in der 3. person konzentriere dich auf die wesentlichen informationen der stil ist sachlich und nicht wertend. Hier die Inhaltsangabe die nzu verbessern Kurzgeschichte „Kahlschnitt“ von Kristen Boie, die 1933 verfasst und erschienenen ist.	
		[Kurzgeschichte: Kahlschnitt von KB]
	hier der originaltext als referenz	
		[Inhaltsangabe <sub>ÜB</sub> ]

<b>Dialogpartner &amp; Zählung</b>	<b>Dialogzüge</b>	<b>In das Dialogfeld kopierte Texte bzw. von GPT generierte Texte</b>
GPT als WKA [Output <sub>2</sub> ]	Hier ist die verbesserte Inhaltsangabe gemäß deinen Kriterien:	
		[Überarbeitete Version I von Inhaltsangabe <sub>ÜB</sub> ]
	Diese Version entspricht den genannten Kriterien und präsentiert die Informationen sachlich und neutral.	
Proband:in <sub>17</sub> [Prompt <sub>3</sub> ]	verändere das sprachlevel auf das einer 9 klässlerin	
GPT als WKA [Output <sub>3</sub> ]	Hier ist die überarbeitete Inhaltsangabe auf dem Sprachniveau einer 9. Klässlerin:	
		[Überarbeitete Version II von Inhaltsangabe <sub>ÜB</sub> ]
	Diese Version ist einfacher und klarer formuliert, um dem Sprachniveau einer 9. Klässlerin gerecht zu werden.	



# Large Language Models – Chancen und Grenzen großer Sprachmodelle für die schulische Nutzung in sprachlich heterogenen Lerngruppen

## Abstract

Die Nutzung von *Large Language Models* (LLM) wie ChatGPT im schulischen Kontext birgt große Potenziale, insbesondere für eine personalisierte Lernbegleitung oder zeitnahes Feedback. Allerdings ist die erfolgreiche und lernwirksame Nutzung von Tools, die auf künstlicher Intelligenz (KI) basieren, nicht voraussetzungsfrei. Vielmehr ist diese stark von personenbezogenen Merkmalen der Schüler:innen – insbesondere von ihren sprachlichen Kompetenzen – abhängig. Dies schlägt sich auch in ersten empirischen Befunden nieder, die zeigen, dass eine schulische LLM-Nutzung nicht unbedingt kompensatorisch wirkt, sondern zum *Digital Divide* beitragen könnte. Der vorliegende Beitrag illustriert die Effekte sprachlich unterschiedlicher Ausgangsvoraussetzungen bei der LLM-Nutzung anhand von Chatverläufen von Schüler:innen und formuliert auf Basis dieser Beispiele Empfehlungen für die Gestaltung von Eingabeaufforderungen (Prompts). Ziel ist hierbei, einer KI-basierten Verstärkung von Bildungsungleichheiten entgegenzuwirken und eine gewinnbringende sowie lernförderliche LLM-Nutzung für alle Schüler:innen zu ermöglichen, indem diese auch im Deutschunterricht angebahnt wird.

**Schlagwörter:** ChatGPT, LLM, generative KI, Digital Divide, Prompt-Engineering

## 1 Einleitung

Die gegenwärtige digitale Transformation durchdringt alle gesellschaftlichen Bereiche und wirkt sich dementsprechend auch auf das schulische Lehren und

Lernen aus. Eine besonders prominente digitale Innovation stellt die rasante und dynamische Entwicklung generativer Künstlicher Intelligenz (KI) dar, welche im Bildungskontext sowohl mit großen Hoffnungen als auch mit großen Befürchtungen verknüpft ist (Böhme/Mesenhöller 2024; KMK 2024; SWK 2024). Zeitlich fällt diese digitale Transformation mit weiteren gesellschaftlichen Herausforderungen wie einer zunehmend heterogenen Schüler:innen-schaft (Stanat et al. 2022, 2023a) zusammen, sodass Unterrichtsgestaltung sowie Schulstruktur und -organisation in Teilen neu gedacht werden müssen.

Die zunehmende Heterogenität der Lernenden spiegelt sich in der Vielfalt ihrer Lernvoraussetzungen, Lernbedürfnisse und Lernerfolge wider (Stanat et al. 2022, 2023a). Ein wichtiger Heterogenitätsaspekt ist hierbei die sprachliche Vielfalt der Lernenden, die eine sprachlich differenzierte und sprachensible Gestaltung von Lerngelegenheiten in allen Fächern – so auch im Deutschunterricht – erfordert. Lehrkräfte stehen somit vor der Aufgabe, auf ihre vielfältige Schüler:innenschaft einzugehen und individuell hilfreiche, differenzierte Lernangebote zu unterbreiten (Lütje-Klose 2023). Die Gestaltung individualisierter Lerngelegenheiten und insbesondere individuelle Rückmeldungen zu Lernverläufen und Lernfortschritten sind zwar lernwirksam (Hattie/Timperley 2007), aber gleichzeitig auch zeitaufwändig und ressourcenintensiv (Wambsgans et al. 2020).

In dieser Situation besteht die Hoffnung, dass KI-basierte Anwendungen im Bildungskontext innovative Zugänge ermöglichen. Zudem verändert die alltägliche Verfügbarkeit von KI-basierten Anwendungen das schulische Lehren und Lernen in grundlegender Weise. KI-basierte Tools eröffnen neue Möglichkeiten, Lernprozesse individuell zu gestalten – bringen aber auch zusätzliche Herausforderungen bspw. in Hinblick auf die angestrebte Stärkung der Bildungsgerechtigkeit mit sich (Dumont/Ready 2023).

Vor diesem Hintergrund beleuchtet der vorliegende Beitrag die Möglichkeiten und Grenzen großer Sprachmodelle (*Large Language Models*, LLM; vgl. Simbeck i. d. B.) für die Gestaltung individueller Lerngelegenheiten und personalisierte Rückmeldung in sprachlich heterogenen Lerngruppen. Als zentrale theoretische und konzeptionelle Grundlage wird zunächst kurz der Heterogenitätsbegriffs im schulischen Kontext erläutert. Im Anschluss wird skizziert, was unter generativer KI zu verstehen ist, wobei sich die Ausfüh-

rungen auf LLM fokussieren. Ferner wird der Forschungsstand zu einem (sprachlich) kompetenten Umgang mit LLM sowie zur LLM-Nutzung in sprachlich heterogenen Lerngruppen und dem sogenannten *Digital Divide* vorgestellt. Im nachfolgenden Abschnitt illustriert der Beitrag beispielhaft, wie sich unterschiedliche sprachliche Voraussetzungen und Unterschiede in der Medien- bzw. KI-Kompetenz von Schüler:innen in der Interaktion mit und damit in der lernförderlichen Nutzung von LLM niederschlagen können. Diese Unterschiede werden vor dem Hintergrund einer LLM-Nutzung in sprachlich heterogenen Lerngruppen diskutiert und münden in Hinweisen für das Prompt-Engineering. Der Beitrag schließt mit einem Fazit zu einem möglichen Beitrag des Deutschunterrichts für eine gelingende LLM-Nutzung in sprachlich heterogenen Lerngruppen.

## 2 Theoretische Grundlagen

### 2.1 Sprachliche Heterogenität im schulischen Kontext

Der Begriff *Heterogenität* beschreibt die Vielfalt der Schüler:innen – etwa in Bezug auf ihre Lernvoraussetzungen, Lernbedürfnisse oder den Erfolg im Kompetenzerwerb sowie ihre sozialen, kulturellen oder auch sprachlichen Hintergründe (Lütje-Klose 2023). Ein weites Inklusionsverständnis schließt alle Lernenden und die unterschiedlichsten Heterogenitätsdimensionen mit ein (Lindmeier/Lütje-Klose 2019). Ein besonderer Fokus des inklusiven Lernens sollte dabei auf der Unterstützung vulnerabler Gruppen – etwa Lernende mit Beeinträchtigungen, Armutserfahrungen oder Fluchtbiografien – liegen (Lütje-Klose 2023). Auf Basis eines solchen weiten Inklusionsverständnisses erfolgt die Gestaltung schulischer Lerngelegenheiten im Idealfall adaptiv – diese werden also flexibel an die individuellen Bedarfe der Schüler:innen angepasst (Grosche/Lüke 2020).

Aktuelle Studien belegen, dass die Heterogenität der Schüler:innen im deutschen Bildungssystem in den letzten Jahren zugenommen hat (Stanat et al. 2022, 2023a). So stieg bspw. der Anteil von Schüler:innen der 9. Jahrgangsstufe mit familiärem Zuwanderungshintergrund zwischen 2009 und

2022 um gut 11 Prozentpunkte auf knapp 38% (Henschel et al. 2023: 303), während gleichzeitig auch soziale und zuwanderungsbezogene Disparitäten in den sprachlichen Kompetenzen der Schüler:innen zunahm (Stanat et al., 2023b: 456f.). Die für das fachliche Lernen zentrale Lesekompetenz von Schüler:innen der 9. Jahrgangsstufe verringerte sich bspw. von 2009 bis 2022 im Mittel um insgesamt 31 Punkte auf der „Pisa-Metrik“ und lag im Jahr 2022 signifikant unter den Werten der Jahre 2015 und 2009 (Boemmel/Schneider 2023: 160 f.). Die Kompetenz im Bereich Zuhören reduzierte sich im selben Zeitraum sogar um 52 Punkte (ebd.: 160f.). Diese ungünstigen Trends sind nicht nur für die Kompetenzmittelwerte, sondern auch für die Streuung der erreichten Kompetenzen zu verzeichnen. In der Gesamtpopulation der Neuntklässler:innen in Deutschland hat die Streuung der Kompetenzwerte in den untersuchten Bereichen zwischen 2015 und 2022 signifikant zugenommen (ebd.: 171). Eine zunehmende Streuung bedeutet, dass die Heterogenität der sprachlichen Kompetenzen der Schüler:innen zunimmt.

Die sprachliche Heterogenität im Schulkontext umfasst neben unterschiedlichen sprachbezogenen Lernerfolgen aber auch zahlreiche andere Facetten. So ist das schulische Lernen in Deutschland heutzutage z. B. auch durch die Mehrsprachigkeit der Schüler:innen geprägt. Viele Schüler:innen mit Zuwanderungshintergrund wachsen mehrsprachig auf und nutzen neben dem Deutschen oder auch statt des Deutschen zu Hause eine andere Familiensprache (Stanat et al. 2022, 2023a). Besonders durch Fluchtmigration ist der Anteil von Lernenden mit anderen Familiensprachen in den letzten Jahren weiter gestiegen (Robert Bosch Stiftung 2022).

Unabhängig von Mehrsprachigkeit haben die Bildungsnähe und der sozioökonomische Hintergrund der Familien bereits früh in der kindlichen Entwicklung Einfluss auf die sprachlichen Kompetenzen (Hart/Risley 1995).

Neben unterschiedlichen sprachlichen Lernvoraussetzungen weisen einige Schüler:innen – unabhängig von Mehrsprachigkeit und Bildungsnähe der Familien – besondere Schwierigkeiten im Lesen und Rechtschreiben oder einen sonderpädagogischen Förderbedarf im Bereich Sprache und Kommunikation auf.

Zusammenfassend ist die sprachliche Heterogenität von Schüler:innen durch viele Facetten geprägt. Damit alle Schüler:innen ungeachtet ihrer

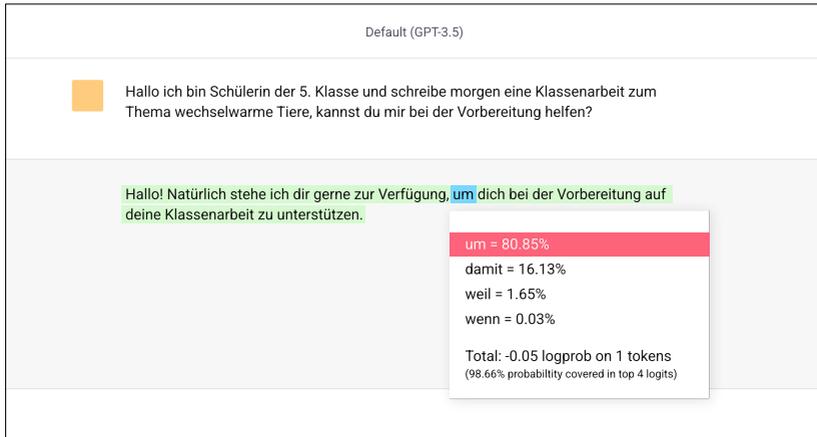
sprachlichen Lernvoraussetzungen erfolgreich am Unterricht teilhaben und auch sprachlich anspruchsvolle – oft bildungssprachlich geprägte – Lernangebote verstehen und nutzen können, sind eine sprachensible Unterrichtsgestaltung, sprachfördernde Maßnahmen und sprachlich differenzierte Lernmaterialien notwendig. An dieser Stelle könnten KI-basierte Anwendungen und insbesondere große Sprachmodelle ihr Potenzial entfalten.

## 2.2 Große Sprachmodelle als schulisch bedeutsame Variante generativer KI

KI bezeichnet technische Systeme, die dazu in der Lage sind, Aufgaben zu übernehmen, die normalerweise menschliches Denken erfordern – etwa das Erkennen von Mustern, das Produzieren natürlich wirkender Sprache oder das Treffen einfacher Entscheidungen (Görz et al. 2021). Innerhalb dieses Feldes bezeichnet der Begriff *generative KI* solche Systeme, die nicht nur analysieren, sondern auf Basis von Trainingsdaten eigenständig neue Inhalte – bspw. Texte, Bilder oder Musik – erzeugen können (Feuerriegel et al. 2024). Eine für den schulischen Kontext besonders relevante Variante generativer KI sind LLM, also große Sprachmodelle. Sie sind auf die Verarbeitung und Erzeugung von Sprache spezialisiert, wurden mit umfangreichen Textdaten trainiert und können in Chats dialogisch kommunizieren, Fragen beantworten und selbständig vollständige Texte erzeugen (Albrecht 2023).

Die Ausgabe eines LLM ist eine sehr wahrscheinliche Abfolge von Texteinheiten wie Buchstaben, Wörtern oder Satzzeichen. Als Reaktion auf die von einem Menschen erteilte Eingabeaufforderung werden Texte generiert, die einen dialogischen Austausch mit dem menschlichen Interaktionspartner ermöglichen. Hierbei wird durch die Auswahl der wahrscheinlichen nächsten Texteinheit – sogenannter *Token* – ein zusammenhängender Text generiert (Hildebrandt 2023). Dabei erzeugen identische Eingabeaufforderungen (leicht) variierende Output-Antworten, die inhaltlich nicht immer korrekt sein müssen. Abbildung 1 verdeutlicht die Auswahl verschieden wahrscheinlicher Token. Im dargestellten Beispiel folgt auf die Eingabeaufforderung im Output

des LLM (hier: ChatGPT-3.5) nach dem Wort „Verfügung“, gefolgt von einem Komma, mit einer 81-prozentigen Wahrscheinlichkeit das Wort „um“.<sup>1</sup>



**Abb. 1:** Darstellung der Wahrscheinlichkeit nachfolgender Token in einer Ausgabe von ChatGPT-3.5

### 2.3 LLM in sprachlich heterogenen Lerngruppen

LLM bergen großes Potenzial für das Lernen in sprachlich heterogenen Gruppen. So können LLM Inhalte neben dem Deutschen auch in den verschiedenen Familiensprachen der Schüler:innen bereitstellen oder Texte entlasten und auf diese Weise an unterschiedliche Lesekompetenzniveaus der Schüler:innen anpassen (Martínez et al. 2024). Dies kann sowohl den sprachsensiblen Fachunterricht durch eine Berücksichtigung der Mehrsprachigkeit der Schüler:innen bereichern als auch einen insgesamt inklusiveren Zugang zu gemeinsamen Lerngegenständen gewährleisten. Wang und Kolleg:innen (2024) betonen in

.....

1 Die Funktion, Token-Wahrscheinlichkeiten für die Ausgabe anzeigen zu lassen, war für frühere GPT-Versionen (bis GPT-3.5) im AI Playground ([platform.openai.com/playground](https://platform.openai.com/playground)) verfügbar und wurde von Entwickler:innen sowie Wissenschaftler:innen genutzt, um Einblicke in das Entscheidungsverhalten des Modells zu erhalten. In der aktuellen Web-Oberfläche von ChatGPT sowie im heutigen AI Playground (ab GPT-4) gibt es keine direkte Möglichkeit mehr, sich Token-Wahrscheinlichkeiten im User Interface anzeigen zu lassen.

diesem Kontext die Bedeutung mehrsprachiger LLM, um Bildungsressourcen in verschiedenen Familiensprachen verfügbar zu machen und dabei auch unterschiedliche kulturelle Kontexte zu berücksichtigen. Für den deutschsprachigen Raum skizziert Heiden (2024) Chancen von LLM im mehrsprachigen Unterricht und die Möglichkeit, unterrichtliche Anforderungen individuell an unterschiedliche sprachliche Kompetenzniveaus der Schüler:innen anzupassen, indem das LLM als digitaler Tutor fungiert (Fürstenberg/Müller 2024). Gleichzeitig wird in der Forschungsgemeinschaft kritisch diskutiert, wie zuverlässig und unverfälscht LLM in weniger prominenten Sprachen oder Dialekten agieren, die in den Trainingsdaten (gegenwärtig noch) unterrepräsentiert sind (Wang et al. 2024).

Bereits gut empirisch untersucht ist der Einsatz von LLM als Unterstützungstool beim freien Schreiben und für personalisiertes Feedback im Schreibprozess (Jelson et al. 2025; Meyer et al. 2024; vgl. auch Fürstenberg i. d. B.). So können LLM Rechtschreib- und Grammatikfehler korrigieren oder stilistische Verbesserungsvorschläge machen. Beispielsweise für das Fach Englisch in der Sekundarstufe II zeigten Meyer und Kolleg:innen (2024), dass LLM-generiertes Feedback die Überarbeitungskompetenz, die Aufgabenmotivation sowie positive Emotionen in Zusammenhang mit der Schreibaufgabe gegenüber einer Bedingung ohne Feedback deutlich steigern konnte.

In Anknüpfung an die erfolgreiche Begleitung im Schreibprozess besteht im Bildungsbereich die Hoffnung, dass LLM auch generell als individuelle Lernbegleitung genutzt werden können (Kasnecki et. al 2023) – ähnlich wie Intelligente Tutorielle Systeme (ITS), die mit einem ständig verfügbaren, persönlichen Tutor vergleichbar sind. Solche LLM-basierten ITS-Systeme könnten Erklärungen liefern, Fragen beantworten und den Lernweg an die Lernvoraussetzung und Lernbedürfnisse der einzelnen Schüler:in anpassen. Murgia und Kolleg:innen (2023) demonstrierten entsprechend in einer Fallstudie, dass ChatGPT in der Einzelförderung tatsächlich auf den individuellen Sprachstand der Schüler:innen eingehen kann und z. B. Erklärungen in der Familiensprache gibt und so als persönlicher Sprachcoach fungiert. Kasnecki und Kolleg:innen (2023) diskutieren die Perspektive einer LLM-Nutzung als persönlicher Tutor und sehen sowohl Chancen (etwa unmittelbares Feedback,

adaptive Hilfestellungen) als auch Herausforderungen (wie mögliche Fehlinformationen).

Auch Sharma und Kolleg:innen (2025) heben in einem systematischen Review, das 55 Studien einbezieht, die zwischen 2020 und 2024 veröffentlicht wurden, das erhebliche Potenzial von LLM hervor. So waren LLM im Vergleich mit traditionellen Bildungsangeboten besser geeignet, das Engagement der Schüler:innen zu steigern, ihre emotionale und soziale Entwicklung zu fördern und den Lernfortschritt in Echtzeit zu verfolgen und dadurch das Lernen zu unterstützen.

## 2.4 ELSI der schulischen KI-Nutzung und Digital Divide

Technologische Innovationen bergen viele Potenziale, berühren jedoch mitunter auch gesellschaftliche und kulturelle Aspekte, die im Vorhinein nicht immer ausreichend antizipiert wurden. So rückt auch die Nutzung von KI in Schule und Unterricht neben technologischen und anwendungsbezogenen Fragen mögliche gesellschaftlich-kulturelle Auswirkungen dieser Nutzung in den Fokus der Aufmerksamkeit. Entsprechende ethische, rechtliche und soziale Themen werden unter dem Begriff ELSI (*Ethical, Legal and Social Implications*; Boden et al. 2021) verhandelt und setzen sich mit den gesellschaftlichen Konsequenzen der Interaktion von Mensch und Maschine auseinander.

Für eine lernförderliche Interaktion mit LLM ist bspw. ein hohes Maß an produktiver und rezeptiver (Schrift-)Sprachkompetenz erforderlich, um die Generierung von Outputs über Eingabeaufforderungen (*Prompts*) anstoßen, die generierten Inhalte verstehen, kritisch reflektieren und sinnvoll in den eigenen Lernprozess integrieren zu können. Aus der Schulpraxis gibt es erste Hinweise, dass Schüler:innen mit ausgeprägten sprachlichen Kompetenzen die von LLM erstellten Texte besser verstehen und bewerten können (SWK 2024).

Dieses Phänomen kann als Variante des Matthäus-Effekts („*Wer hat, dem wird gegeben.*“) interpretiert werden. In ähnlicher Weise wurde ein für das Fach Deutsch relevanter Matthäus-Effekt bereits von Steinmetz (2020) für den Literaturunterricht beschrieben. Magirius und Kolleg:innen (2024) thematisieren diesen Effekt auch im Kontext des LLM-Einsatzes im Literaturunterricht. Ein vielfach diskutiertes Thema im Kontext KI-bezogener ELSI ist daher die Frage,

ob eine schulische KI-Nutzung, speziell die von LLM, zu mehr Bildungsgerechtigkeit führen kann oder ob durch diese Technologien bestehende Unterschiede im Sinne eines Matthäus-Effekts weiter verstärkt werden.

Im Zusammenhang mit digitalen Technologien wird für diesbezügliche Disparitäten auch der Begriff *Digital Divide* verwendet. Forschungsarbeiten zum Digital Divide untersuchen Ungleichheiten im Zugang zu und der Nutzung von digitalen Technologien verschiedener gesellschaftlicher Gruppen. Dabei umfasst das Konzept des Digital Divide drei Ebenen: Die erste Ebene bezieht sich auf Unterschiede im Zugang zu digitalen Technologien. Die zweite Ebene bezieht sich auf Kompetenzunterschiede und Unterschiede im Ausmaß, mit dem verschiedene gesellschaftliche Gruppen vorhandene digitale Technologien nutzen (Van Deursen/Van Dijk 2019). Hargittai und Kolleg:innen (2019) betonen diesbezüglich, dass ein bloßer physischer Zugang zu Technologien wirkungslos bleibt, wenn nicht auch die nötigen Kompetenzen und unterstützende Strukturen vorhanden sind, um die Technologien effektiv nutzen zu können. Die dritte Ebene des Digital Divide beschreibt ungleiche Ergebnisse, die Angehörige verschiedener gesellschaftlicher Gruppen durch die Nutzung digitaler Technologien erzielen können, und versucht zu erklären, wie es dazu kommt, dass selbst bei gleichem Zugang und vergleichbaren digitalen Kompetenzen privilegierte Nutzer:innen mehr Gewinn aus digitalen Technologien ziehen können als weniger privilegierte Nutzer:innen (Van Dijk 2020: 119).

In diesem Kontext untersuchten Hannan und Eynon (2025) die ITS-Nutzung von über 65.000 australischen Schüler:innen der Jahrgangsstufen 7 bis 10. Dabei zeigte sich, dass Schüler:innen aus wohlhabenderen und städtisch gelegenen Schulen die ITS-Plattform früher und über längere Zeiträume nutzten und insgesamt bessere Lernergebnisse als ihre ländlichen, weniger privilegierten Peers erzielten. Die Autor:innen konnten ferner zeigen, dass die ITS-Nutzung den Zusammenhang zwischen dem sozioökonomischen Status und dem Lernerfolg der Schüler:innen vermittelt: Die ITS-Nutzung verstärkt sowohl die positiven Effekte eines günstigen Schulumfelds als auch die negativen Effekte eines benachteiligten, ländlichen Schulkontexts. Die Ergebnisse deuten somit darauf hin, dass der Einsatz von adaptiven ITS-Plattformen bestehende Leistungsunterschiede zwischen privilegierten und benachteiligten Schulen nicht verringert. Vielmehr lässt sich ein Matthäus-Effekt in dem Sinne

beobachten, dass Schüler:innen an privilegierten Schulen die Systeme effektiver nutzen, was wiederum zu einer weiteren Vergrößerung der Leistungsunterschiede beiträgt.

Auch hinsichtlich der Nutzung von LLM konnte ein Einfluss des sozioökonomischen Hintergrunds der Lernenden nachgewiesen werden. Yu und Kolleg:innen (2024) analysierten für den US-amerikanischen Raum über eine Million akademische Texte und konnten zeigen, dass die Einführung von LLM vor allem bei Studierenden mit höherem sozioökonomischem Status zu einer Verbesserung der Schreibqualität führte. Dieser Befund deutet somit ebenfalls darauf hin, dass bestehende Bildungsungleichheiten durch den schulischen KI-Einsatz – und auch konkret durch LLM – verstärkt werden könnten. Empirische Untersuchungen, die sich spezifisch auf den Schulkontext beziehen, liegen allerdings gegenwärtig noch nicht vor.

Eine mögliche lernförderliche Nutzung von LLM kann somit von verschiedenen Faktoren abhängen, unter anderem von Merkmalen der Schüler:innen, wie bspw. ihrem sozioökonomischen Status. Dieser wiederum kovariiert auch mit ihren sprachlichen Kompetenzen in dem Sinne, dass Schüler:innen aus bildungsnahen Familien im Mittel über besser ausgeprägte sprachliche Kompetenzen verfügen. Die etwaige Abhängigkeit einer lernförderlichen LLM-Nutzung von schüler:innenseitigen Voraussetzungen – wie den sprachlichen Kompetenzen, aber auch dem Vorwissen und der Reflexionsfähigkeit – soll im folgenden Abschnitt anhand einiger konkreter LLM-Nutzungen durch Schüler:innen beispielhaft demonstriert werden.

### **3 Beispielhafte Erprobung von ChatGPT für den Schulkontext**

Zum Zeitpunkt der nachfolgend berichteten Erprobung im Frühjahr 2023 war die Version GPT-3.5 (OpenAI 2022) kostenlos verfügbar. Bereits im März 2023 wurde von OpenAI jedoch das – zunächst kostenpflichtige – deutlich leistungsfähigere Modell ChatGPT-4 eingeführt (OpenAI 2023a), welches auf eine aktuellere Datenbasis und damit breiteres Weltwissen zugreifen konnte. Außerdem berichtete OpenAI über eine im Vergleich zu GPT-3.5 spürbare

Reduktion halluzinierter Inhalte (OpenAI 2023b). Insgesamt lässt sich – auch für andere, hier nicht näher dargelegte LLM – beobachten, dass sich seit 2022 nicht nur die Leistungsfähigkeit der Sprachmodelle deutlich gesteigert hat, sondern auch deren Anpassbarkeit an spezifische Anwendungsfelder optimiert wurde. Für einen umfassenderen Überblick zu großen Sprachmodellen vgl. auch Simbeck (i. d. B.).

Zielstellung der nachfolgend berichteten Erprobung war es, zu untersuchen, inwieweit Schüler:innen ohne Anleitung einer Lehrkraft ein LLM für das Lernen zu Hause nutzen können und welchen möglichen Mehrwert sie in dieser Form einer tutorähnlichen Lernbegleitung sehen. Ferner war für uns von Interesse, wie stark schüler:innenseitige Voraussetzungen – konkret hinsichtlich des KI-bezogenen Vorwissens und unterschiedlich ausgeprägter sprachlicher Kompetenzen – die Nutzung und den Nutzen des LLM-Einsatzes beeinflussen.

Für die Erprobung haben wir die zum damaligen Zeitpunkt kostenfreie ChatGPT-Version verwendet, um die Situation der Schüler:innen zu Hause möglichst realistisch nachbilden zu können. Eine Nutzung zu Hause wurde als Kontext gewählt, da der schulische Einsatz von LLM im Frühjahr 2023 noch nicht administrativ reguliert war, entsprechende Zugänge für Schulen nicht vorhanden waren und die weit überwiegende Mehrheit der Lehrkräfte auch noch keinen Kontakt zu großen Sprachmodellen hatte, um eine solche Nutzung kompetent begleiten zu können. Eine schulische Nutzung vor Ort im Unterricht war zum damaligen Zeitpunkt somit kein realistisches Szenario.

In dem von uns gewählten Erprobungskontext bereiteten sich die Schüler:innen mithilfe von ChatGPT-3.5 zu Hause auf eine Klassenarbeit zum Thema *wechselwarme Tiere* im naturwissenschaftlichen Unterricht (NaWi-Unterricht) vor. Dieses Thema wird üblicherweise in den Jahrgangsstufen 5 oder 6 thematisiert. Um einen möglichen Zusammenhang zwischen den mit der besuchten Jahrgangsstufe variierenden sprachlichen Kompetenzen und dem medienbezogenen Vorwissen der Schüler:innen einerseits und der Komplexität der von den Schüler:innen genutzten Prompts andererseits untersuchen zu können, wurden aber auch Schüler:innen höherer Jahrgangsstufen gebeten, an der Erprobung teilzunehmen.

Nachfolgend werden zwei Dialoge von Schüler:innen mit ChatGPT-3.5 beispielhaft in Auszügen wiedergegeben. Die Schülerin „Mia“<sup>2</sup> (10 Jahre) besuchte zum Zeitpunkt der ChatGPT-Erprobung die 6. Jahrgangsstufe einer Grundschule<sup>3</sup>. Der Schüler „Ben“ (12 Jahre) besuchte die 8. Jahrgangsstufe eines Gymnasiums. Beide Schüler:innen zeigten in den Naturwissenschaften sehr gute Lernergebnisse und hatten bereits einmal eine Klassenarbeit zum Thema wechselwarme Tiere geschrieben. Sie verfügten aber – wie intendiert – aufgrund unterschiedlicher schulischer Lernerfahrungen im Bereich Medienbildung über unterschiedlich ausgeprägte Vorerfahrungen mit digitalen Medien, wobei Ben deutlich mehr Erfahrung mitbrachte. Beide hatten ein Grundverständnis davon, wie ChatGPT funktioniert, hatten das Tool aber noch nie genutzt.

Beide Schüler:innen wurden wie folgt instruiert:

- Stell dir vor, du schreibst eine Klassenarbeit zum Thema *wechselwarme Tiere* und willst ChatGPT nutzen, um dich auf diese Klassenarbeit vorzubereiten.
- Die Idee ist, dass ChatGPT dir Fragen zum Thema *wechselwarme Tiere* stellt, die du dann beantwortest. ChatGPT soll dir dann Feedback zu deinen Antworten geben.
- Was würdest du eintippen?

Die Instruktion wurde den Schüler:innen von einer Versuchsleiterin vorgelesen und alle von den Schüler:innen mündlich gegebenen Antworten wurden für die Schüler:innen durch die Versuchsleiterin eingetippt. Nachfolgend ist in Tabelle 1 Mias Prompt sowie ein Ausschnitt des darauffolgenden Outputs von ChatGPT zu sehen.

.....

- 2 Um die Anonymität der Schüler:innen gewährleisten zu können, wurden die Namen geändert.
- 3 In den Ländern Berlin und Brandenburg beginnt die weiterführende Schule erst mit der 7. Jahrgangsstufe.

Eingabe/Ausgabe	Chatverlauf
Prompt Mia	<i>Stelle mir Fragen, die mich auf eine Klassenarbeit zum Thema wechselwarme Tiere vorbereiten und gebe mir zu meinen Antworten Feedback.</i>
Ausgabe ChatGPT (Ausschnitt)	<i>Natürlich, ich kann dir einige Fragen stellen, um dich auf eine Klassenarbeit zum Thema wechselwarme Tiere vorzubereiten. Bitte gib mir deine Antworten, und ich werde dir anschließend Feedback geben.</i> <ol style="list-style-type: none"> <li>1. <i>Was sind wechselwarme Tiere, und wie unterscheiden sie sich von gleichwarmen Tieren?</i></li> <li>2. <i>Welche Vorteile bietet die wechselwarme Regulation für Tiere in Bezug auf Energieeffizienz und Überlebensfähigkeit?</i></li> <li>3. <i>Erkläre den Begriff „Torpor“ und wie er bei wechselwarmen Tieren genutzt wird.</i></li> <li>4. ...</li> </ol>

**Tab. 1:** Prompteingeabe von Mia (6. Jahrgangsstufe) sowie Ausschnitt des Outputs von ChatGPT

Mia forderte ChatGPT direkt auf, ihr Fragen zum Thema wechselwarme Tiere sowie eine Rückmeldung zu den eigenen Antworten auszugeben. Auf diese Eingabe von Mia generierte ChatGPT zehn Fragen. Der dargestellte Ausschnitt des Outputs macht deutlich, dass diese das sprachliche Kompetenzniveau von Lernenden der 5./6. Jahrgangsstufe übersteigen. Begriffe wie „Energieeffizienz“ und „Torpor“ entsprechen nicht dem fachsprachlichen Niveau der Schüler:innen dieser Jahrgangsstufen. Im Gespräch mit der Testleiterin wurde deutlich, dass sich die 10-jährige Schülerin sowohl von der Menge als auch der Formulierung der Fragen überfordert fühlte. Sie versuchte dennoch, die erste Frage („Was sind wechselwarme Tiere, und wie unterscheiden sie sich von gleichwarmen Tieren?“) zu beantworten (siehe Tabelle 2).

Eingabe/Ausgabe	Chatverlauf
2. Prompt Mia	<i>Sie passen ihre Körpertemperatur an die Temperatur der Umgebung an.</i>
2. Ausgabe ChatGPT	<p><i>Vielen Dank für deine Antwort. Du hast eine grundlegende Erklärung für wechselwarme Tiere gegeben. Lass uns jedoch deine Antwort genauer betrachten und einige Ergänzungen vornehmen:</i></p> <p><i>Wechselwarme Tiere passen ihre Körpertemperatur an die Temperatur der Umgebung an. Dies bedeutet, dass ihre Körpertemperatur nicht konstant ist, sondern sich entsprechend der Umgebungstemperatur ändert. Diese Anpassung ermöglicht es ihnen, Energie zu sparen, da sie keine Wärme erzeugen müssen, wenn die Umgebungstemperatur ausreichend warm ist. Auf der anderen Seite sind sie bei niedrigeren Temperaturen weniger aktiv, da ihre Körperfunktionen langsamer ablaufen. Wechselwarme Tiere nutzen verschiedene Mechanismen, um ihre Temperatur zu regulieren, darunter Sonnenbaden, Gruppenverhalten, Aktivität während wärmerer Tageszeiten und Torpor.</i></p> <p><i>Wenn du weitere Fragen hast oder mehr Details benötigst, stehe ich zur Verfügung.</i></p>
Mündlicher Kommentar Mia	<i>Das ist doch viel zu lang, was er gesagt hat. Das kann ich doch nicht in eine Klassenarbeit schreiben. Ich denke, da steht dann nämlich so was wie: „Beschreibe kurz ...“.</i>

**Tab. 2:** Antwort von Mia und nachfolgender Chatverlauf mit ChatGPT

Im weiteren Verlauf des mündlichen Dialogs wurde deutlich, dass Mia sowohl die Fragen zu schwer als auch die Rückmeldung von ChatGPT zu komplex und lang fand. Nach Beantwortung der zweiten Frage hatte sie keine Lust mehr. Sie empfand die Kommunikation mit ChatGPT als nicht hilfreich.

Anders gestaltete sich der mit ChatGPT geführte Dialog des 12-jährigen Schülers Ben. Wie Tabelle 3 deutlich macht, lieferte der ältere Schüler ChatGPT in seinem ersten Prompt selbstständig Informationen zum Schwierigkeitsgrad der Übungsfragen, indem er in seinem Prompt die passende Jahrgangsstufe angab („Ich bin Fünftklässler“). Bereits vor der Nutzung von ChatGPT zeigte sich ferner, dass auch die Reflexion über die Interaktion mit ChatGPT und der

Umgang mit dessen Output bei Ben ausgeprägter und differenzierter war. Er merkte z. B. mündlich an, dass er die von ChatGPT generierten Inhalte dahingehend überprüfen würde, ob sie den richtigen Schwierigkeitsgrad haben und ob sie tatsächlich der Wahrheit entsprechen. Er schlug dazu vor, den Output von ChatGPT mit den ersten drei Ergebnissen einer Google-Abfrage abzugleichen, um sicherzustellen, dass ChatGPT keine falschen Antworten liefert. Insgesamt ging Ben nicht davon aus, dass ChatGPT sofort die für ihn passenden Übungsfragen generieren kann, sondern nahm an, dass er erst mehrere verschiedene Prompts testen müsse, bis er ein für ihn passendes Ergebnis erhält.

Eingabe/Ausgabe	Chatverlauf
Prompt Ben	<p><i>Hallo ChatGPT. Ich bin Fünftklässler und schreibe eine Klassenarbeit zum Thema wechselwarme Tiere und muss mich noch darauf vorbereiten. Könntest Du mir einige Übungsaufgaben für die Klassenarbeit stellen, die ich dann beantworte und die Du dann kontrollierst?</i></p>
Ausgabe ChatGPT	<p><i>Hallo! Natürlich kann ich dir einige Übungsaufgaben zum Thema wechselwarme Tiere stellen. Nachdem du die Fragen beantwortet hast, werde ich deine Antworten überprüfen und dir Feedback geben. Los geht's:</i></p> <p><b>Frage 1:</b> Was sind wechselwarme Tiere? Erkläre diesen Begriff.</p> <p><b>Frage 2:</b> Nenne mindestens drei Beispiele für wechselwarme Tiere und erkläre, wie sie ihre Körpertemperatur regulieren.</p> <p>...</p> <p><b>Frage 5:</b> Erkläre den Begriff „Hibernation“ und gib ein Beispiel für ein Tier, das diesen Überlebensmechanismus nutzt. Antworte auf diese Fragen, und ich werde gerne überprüfen, wie gut du dich vorbereitet hast. Viel Erfolg bei deiner Klassenarbeit!</p>

**Tab. 3:** Prompteingabe von Ben (8. Jahrgangsstufe) sowie Ausschnitt des Outputs von ChatGPT

Die von ChatGPT generierten Fragen (siehe Tabelle 3) entsprachen bereits weitestgehend einem sprachlichen Kompetenzniveau, das für Lernende einer 5. oder 6. Jahrgangsstufe geeignet ist. Auch enthielt der Output eine andere Tonalität und war stärker motivierend formuliert („Antworte auf diese Fragen, und ich werde gerne überprüfen, wie gut du dich vorbereitet hast. Viel Erfolg

bei deiner Klassenarbeit!“). Dennoch wurden trotz Bens Prompt („Ich bin Fünftklässler und schreibe eine Klassenarbeit ...“) Begriffe und Konzepte aufgegriffen, von denen anzunehmen ist, dass sie von Lernenden der intendierten Jahrgangsstufe nicht verstanden werden, wie bspw. der Begriff „Hibernation“.

Der weitere Dialog zwischen Ben und ChatGPT ist in Tabelle 4 wiedergegeben.

<b>Eingabe/Ausgabe</b>	<b>Chatverlauf</b>
2. Prompt Ben	<i><b>Frage 1:</b> Wechselwarme Tiere sind Tiere, deren Körpertemperatur von der Umgebungstemperatur abhängig ist.</i>
2. Ausgabe ChatGPT	<i>Richtig! Wechselwarme Tiere sind Tiere, deren Körpertemperatur von der Umgebungstemperatur abhängig ist. Jetzt kannst du mit den anderen Fragen fortfahren.</i>
Mündlicher Kommentar von Ben	<i>Hätte ich nicht erwartet, dass das direkt so gut klappt. Und: Antwort überprüfen im Internet.</i>
...	
3. Prompt Ben	<i><b>Frage 2</b> (Nenne mindestens drei Beispiele für wechselwarme Tiere und erkläre, wie sie ihre Körpertemperatur regulieren): Eidechsen, Schlangen, Amphibien.</i>
3. Ausgabe ChatGPT	<i>Das sind gute Beispiele! Jetzt könntest du noch kurz erläutern, wie diese Tiere ihre Körpertemperatur regulieren.</i>
3. Prompt Ben	<i>Sie erwärmen sich zum Beispiel, indem sie sich in die Sonne setzen.</i>
3. Ausgabe ChatGPT	<i>Das ist korrekt! Diese Tiere können ihre Körpertemperatur erhöhen, indem sie sich in die Sonne setzen oder an warmen Orten aufhalten, um die Wärme aufzunehmen. Weiter so! Möchtest du die nächsten Fragen beantworten?</i>
Mündlicher Kommentar von Ben	<i>Ey, das läuft gut!</i>

**Tab. 4:** Antwort von Ben und nachfolgender Chatverlauf mit ChatGPT (Auszug)

Im Dialog mit ChatGPT wird deutlich, dass das LLM adaptiv auf die Eingabe hin Rückmeldung gibt. Bei der Beantwortung von Frage 2 unterstützt es bspw. dahingehend, dass es Ben vorschlägt, noch detaillierter auf die Frage zu antworten, woraufhin der Schüler ein Beispiel angibt. Ben schätzte dieses Feed-

back als nützlich ein und war erstaunt darüber, wie gut der Dialog mit dem KI-Tool funktioniert. Zu Beginn hatte er nicht erwartet, direkt eine hilfreiche Antwort zu erhalten.

Die Unterschiede im Output von ChatGPT auf die Prompts von Mia und Ben machen deutlich, wie entscheidend die Spezifität des Prompts für die Qualität von Output und Feedback ist. Allein die Angabe von Ben „Ich bin Fünftklässler“ bewirkt einen entscheidenden Unterschied im Output, der dazu führt, dass Ben die Nutzung von ChatGPT als hilfreich bewertet, während Mia aufgrund der Komplexität der Fragen und der Nutzung von für sie nicht nachvollziehbarer Bildungssprache bereits nach zwei Fragen abbricht.

## 4 Prompt Engineering

Die in Abschnitt 3 beispielhaft wiedergegebenen Dialoge zwischen Schüler:innen und ChatGPT in der Version 3.5 haben aufgezeigt, wie eine tutorähnliche LLM-Nutzung für individualisiertes Feedback denkbar wäre. Der Verlauf der dialogischen Interaktion ist in Abbildung 2 noch einmal schematisch illustriert.



**Abb. 2:** Möglicher Verlauf einer dialogischen Interaktion bei der Bearbeitung von Übungsaufgaben

Nachfolgend soll nun noch einmal detaillierter darauf eingegangen werden, welche Rolle die Prompts in der dialogischen Interaktion von Schüler:innen – oder auch Lehrkräften – mit LLM spielen und wie Prompts so angepasst wer-

den können, dass sie den bestmöglichen Output generieren. Auch hierbei führt die dynamische Entwicklung von KI insgesamt und spezifisch die der großen Sprachmodelle dazu, dass entsprechende Empfehlungen eine kurze Halbwertszeit haben und schnell überholt bzw. nur in bestimmten Kontexten nützlich sind.

Um herkömmliche Prompts effizienter zu gestalten, können verschiedene Parameter angepasst werden, die einen Einfluss auf den Output haben. Tabelle 5 gibt einen Überblick zu Möglichkeiten, herkömmliche Prompts anzupassen.

Parameter	Beispielhafte Angaben im Prompt
Schwierigkeitsgrad und inhaltliche Spezifik der Aufgabe	<ul style="list-style-type: none"> <li>• <i>Jahrgangsstufe</i></li> <li>• <i>Unterthemen</i></li> </ul>
Aufgabentyp	<ul style="list-style-type: none"> <li>• <i>Multiple Choice</i></li> <li>• <i>Lückentext</i></li> <li>• <i>Offene Fragen</i></li> <li>• ...</li> </ul>
Sprachniveau	<ul style="list-style-type: none"> <li>• <i>sprachliche Komplexität (leichte/einfache Sprache)</i></li> <li>• <i>Aus- und Eingabe sowohl in anderer Familiensprache als auch auf Deutsch</i></li> </ul>
Tonalität	<ul style="list-style-type: none"> <li>• <i>freundlich</i></li> <li>• <i>motivierend</i></li> <li>• <i>„Tu so, als wärst du meine Lehrerin ...“</i></li> </ul>
Art des Feedbacks	<ul style="list-style-type: none"> <li>• <i>lernförderlich</i></li> <li>• <i>empathisch</i></li> <li>• <i>aktivierend</i></li> <li>• <i>fachlich korrekt</i></li> </ul>

**Tab. 5:** Beispielhafte Informationen, die bestimmte Parameter des Dialogs mit ChatGPT spezifizieren (adaptiert nach Flick 2023; Jacobsen/Weber 2023; Prilop et al. 2020; Wu/Schunn 2021)

In der nachfolgenden Tabelle 6 wurde für den in Abschnitt 3 dargestellten Anwendungskontext beispielhaft ein Prompt konstruiert, der möglichst viele Parameter eines lernförderlichen Dialogs von vornherein festlegt. Entsprechende Prompts könnten Lehrkräfte im Vorfeld einer geplanten Nutzung von LLM für

alle Schüler:innen bereitstellen. Weitere Anregungen und Hilfestellungen für die Formulierung solcher Prompts finden Lehrkräfte bspw. bei Flick (2025).

Parameter	Konstruierte Prompteingabe
Inhaltliche Spezifik (Kontext)	Ich bin <b>Schülerin der 5. Klasse</b> und schreibe morgen eine Klassenarbeit zum Thema wechselwarme Tiere im NaWi-Unterricht. Auf diese Klassenarbeit möchte ich mich vorbereiten und brauche dafür deine Unterstützung.
Schwierigkeitsgrad (Jahrgangsstufe)	Bitte gib mir Fragen zum Thema wechselwarme Tiere aus. Die <b>Schwierigkeit der Fragen muss so sein, dass ich sie als Schülerin der 5. Klasse verstehe und beantworten kann.</b>
Inhaltliche Spezifik der Aufgabe (Unterthemen)	Ich beantworte dann die Fragen. <b>Folgende Unterthemen</b> haben wir im Unterricht behandelt: Was sind wechselwarme Tiere, Körpermerkmale von wechselwarmen Tieren, Bedeutung der Umgebungstemperatur, Jahreszeiten und Winterruhe, Fortpflanzung bei wechselwarmen Tieren, Anpassung und Lebensräume von wechselwarmen Tieren.
Aufgabentyp (offene Fragen, Multiple Choice)	Gib mir dazu insgesamt <b>5 offene Fragen und 5 Single-Choice-Fragen</b> aus. Die offenen Fragen und die Single-Choice-Fragen sollten nicht zu ähnlich sein. Gib keine Antworten aus, sondern gib mir nur Rückmeldung zu meinen Antworten.
Tonalität (Lehrerin, Adjektive)	Die Rückmeldung soll wie folgt gestaltet sein: <b>Tu so, als wärst du meine NaWi-Lehrerin</b> und gestalte die Rückmeldung genau, <b>empathisch, aktivierend</b> (d. h., dass du Fragen stellen und Hinweise zur Verbesserung geben sollst) <b>und korrekt</b> (d. h., dass du zu verbessernde Punkte benennen und auf die Richtigkeit deiner Rückmeldung achten sollst).
Art des Feedbacks	Deine Rückmeldung soll <b>mein Lernen fördern, mich motivieren und mir helfen, meine Fehler gut zu verstehen.</b>

**Tab. 6:** Konstruierter Prompt für einen lernförderlichen Dialog mit ChatGPT

Der erste Teil des Prompts in Tabelle 6 gibt Informationen zur Jahrgangsstufe sowie Informationen darüber, was die grundlegende Aufgabe des LLM sein wird. Der zweite Abschnitt konkretisiert den Auftrag, indem Angaben zur Schwierigkeitsstufe der Aufgaben gemacht, die spezifischen Unterthemen benannt und die Aufgabentypen beschrieben werden. Der dritte Abschnitt des Prompts macht Angaben zur intendierten Tonalität und beschreibt die Art und Weise, wie das Feedback gestaltet sein soll, damit es lernförderlich ist.

In Tabelle 7 findet sich der Output, den ChatGPT-3.5 auf den Prompt in Tabelle 6 hin generiert hat.

Eingabe/Ausgabe	Chatverlauf
Ausgabe von ChatGPT	<p><i>Natürlich, ich helfe dir gerne bei der Vorbereitung! Hier sind die Fragen, die du benötigst:</i></p> <p><i>Offene Fragen:</i></p> <ol style="list-style-type: none"> <li>1. Was sind wechselwarme Tiere und wie unterscheiden sie sich von gleichwarmen Tieren?</li> <li>2. Erläutere drei Körpermerkmale von wechselwarmen Tieren und warum sie für ihre Lebensweise wichtig sind.</li> <li>3. ...</li> </ol> <p><i>Single-Choice-Fragen:</i></p> <ol style="list-style-type: none"> <li>4. Welches Merkmal ist typisch für wechselwarme Tiere?             <ol style="list-style-type: none"> <li><i>a. Sie können ihre Körpertemperatur konstant halten.</i></li> <li><i>b. Ihre Körpertemperatur ändert sich je nach Umgebung.</i></li> <li><i>c. Sie haben kein Blut.</i></li> </ol> </li> <li>5. Welche Bedeutung hat die Umgebungstemperatur für wechselwarme Tiere?             <ol style="list-style-type: none"> <li><i>a. Sie hat keinen Einfluss auf ihr Wohlbefinden.</i></li> <li><i>b. Sie beeinflusst ihre Aktivität und Stoffwechselrate.</i></li> <li><i>c. Sie bestimmt ihre Farbe.</i></li> </ol> </li> <li>6. ...</li> </ol> <p><i>Nun bist du an der Reihe, die Fragen zu beantworten. Falls du Fragen oder Unsicherheiten hast, stehe ich gerne zur Verfügung. Viel Erfolg bei deiner Klassenarbeit!</i></p>

**Tab. 7:** Ausgabe von ChatGPT als Antwort auf den Prompt aus Tabelle 6 (Auszug)

Im Output von ChatGPT zeigt sich die große Stärke des LLM, unmittelbar individualisiert auf die Eingabe zu reagieren und eine motivierende und lobende Rückmeldung auf die Schüler:innenantwort zu erzeugen („Sehr gut!“, „Das ist eine ausgezeichnete Ergänzung!“, „Weiter so, du machst das wirklich gut!“). Auch nutzt ChatGPT wie gefordert Fragen („Hast du noch weitere Punkte, die du hinzufügen möchtest?“) und bietet Hilfe an („Wenn du weitere Fragen beantworten möchtest oder wenn du Hilfe zu den anderen Fragen benötigst, lass es mich einfach wissen. Ich stehe dir gerne zur Verfügung.“).

Im weiteren (hier nicht dargestellten) Verlauf zeigt sich, dass ChatGPT prinzipiell in der Lage ist, eine sowohl sprachlich angemessene als auch inhaltlich elaborierte Rückmeldung zu geben, Fehler zu erkennen, dahinterliegende Fehlvorstellungen zu thematisieren und Verbesserungsvorschläge zu unterbreiten. Es wird aber auch deutlich, dass ChatGPT-3.5 zum Teil fachlich falsche Rückmeldung gibt, bspw. da es inkorrekte Körpermerkmale von wechselwarmen Tieren aufgreift und nicht korrigiert (z. B. „Fell“). In diesem Zusammenhang wird auch eine inkorrekte Antwort der Lernenden verstärkt („Sehr gut! Du hast die Körpermerkmale gut beschrieben.“). Ferner ist ChatGPT nicht immer in der Lage, in seiner Rückmeldung zwischen verschiedenen Teilen einer zu kommentierenden Antwort zu differenzieren. Wird nur auf einen Teil der Frage geantwortet, auf andere Teile jedoch nicht, berücksichtigt ChatGPT dies in der Rückmeldung nicht immer und gibt mitunter das Feedback, dass die gesamte Frage korrekt beantwortet wurde. Rechtschreib- und Grammatikfehler werden in der deutschen Sprache, auch wenn dies noch zusätzlich in den Prompt aufgenommen wird, durch ChatGPT (in der Version 3.5) nicht verlässlich korrigiert. Außerdem gelingt es ChatGPT nicht immer, ein angemessenes Sprachniveau zu wählen. So werden teilweise Begriffe und Konzepte in Fragen und Rückmeldungen verwendet (z. B. „Stoffwechselrate“), die von Schüler:innen der 5. oder 6. Jahrgangsstufe gegebenenfalls noch nicht verstanden werden.

## 5 Fazit und Implikationen für den Deutschunterricht

LLM bieten innovative Möglichkeiten für die Unterstützung des Lernprozesses in allen Fächern. Die effektive und lernförderliche Nutzung großer Sprachmodelle hängt jedoch von verschiedenen Faktoren ab – unter anderem von den sprachlichen Fähigkeiten der Lernenden. Um Bildungsungleichheiten nicht zu verstärken, ist es entscheidend, diesen Faktor bei der Implementierung von LLM im schulischen Kontext zu berücksichtigen und entsprechende Unterstützungsmaßnahmen zu etablieren.

Hierbei kommt dem Deutschunterricht eine verantwortungsvolle Rolle für die Vorbereitung einer lernförderlichen und chancengerechten LLM-Nutzung zu. Als zentrales Sprachfach ist der Deutschunterricht in besonderer Weise dafür prädestiniert, alle Schüler:innen – unter Berücksichtigung ihrer sprachlichen Ausgangsvoraussetzungen – zu einem reflektierten, kritischen und kompetenten Umgang mit KI-basierten Anwendungen wie LLM zu befähigen. Hierfür können im Deutschunterricht gezielt sprachliche Kompetenzen gefördert werden, die für eine lernwirksame LLM-Nutzung entscheidend sind, so etwa das Verstehen und Erstellen präziser Prompts, das kritische Bewerten generierter Outputs sowie das Formulieren von Rückfragen oder Korrekturen. Damit kann der Deutschunterricht dazu beitragen, einer möglichen Verstärkung bestehender Bildungsungleichheiten entgegenzuwirken.

Neben der Professionalisierung der Lehrkräfte in Hinblick auf KI-bezogene Kompetenzen ist in diesem Kontext auch die Haltung der Lehrkräfte entscheidend. Ziel sollte es sein, den Deutschunterricht als Lernort zu gestalten, an dem sprachliche Bildung und Medienbildung zusammengedacht werden, um Inklusion und Teilhabe in der digitalen Transformation aktiv zu gestalten. Eine grundlegende Bereitschaft von Lehrkräften, sich konstruktiv mit dem Einsatz von LLM auseinanderzusetzen, ist an dieser Stelle entscheidend, um diese Teilhabe zu ermöglichen.

## Literatur

- ALBRECHT, STEFFEN (2023): *ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen* (TAB-Hintergrundpapier Nr. 26). <https://doi.org/10.5445/IR/1000158070>.
- BODEN, ALEXANDER/LIEGL, MICHAEL/BÜSCHER, MONIKA (2021): Ethische, rechtliche und soziale Implikationen (ELSI). In: Reuter, Christian (Hg.): *Sicherheitskri-tische Mensch-Computer-Interaktion. Interaktive Technologien und Soziale Medien im Krisen- und Sicherheitsmanagement*. Wiesbaden: Springer Vieweg, S. 185–205. [https://doi.org/10.1007/978-3-658-32795-8\\_9](https://doi.org/10.1007/978-3-658-32795-8_9).
- BOEMMEL, QUIRIN/SCHNEIDER, REBECCA (2023): Mittelwerte und Streuungen der erreichten Kompetenzen im Fach Deutsch. In: Stanat, Petra/Schipolowski, Stefan/Schneider, Rebecca/Weirich, Sebastian/Henschel, Sofie/Sachse, Karoline A. (Hg.) (2023): *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*. Münster: Waxmann, S. 149–172. <https://doi.org/10.31244/9783830997771>.
- BÖHME, KATRIN/MESENHÖLLER, JANNE (2024): Meine Kollegin, die KI – Wie die Nutzung von Künstlicher Intelligenz das schulische Lehren und Lernen verändert. In: Schork, Sabrina (Hg.): *Vertrauen in Künstliche Intelligenz. Eine multi-perspektivische Betrachtung*. Wiesbaden: Springer Vieweg, S. 79–99. [https://doi.org/10.1007/978-3-658-43816-6\\_5](https://doi.org/10.1007/978-3-658-43816-6_5).
- DUMONT, HANNAH/READY, DOUGLAS D. (2023): On the promise of personalized learning for educational equity. In: *NPJ science of learning*, 8/1, Article 26. <https://doi.org/10.1038/s41539-023-00174-x>.
- FEUERRIEGEL, STEFAN/HARTMANN, JOCHEN/JANIESCH, CHRISTIAN/ZSCHECH, PATRICK (2024): Generative AI [Generative KI]. In: *Business & Information Systems Engineering*, 66/1, p. 111–126. <https://doi.org/10.1007/s12599-023-00834-7>.
- FLICK, MANUEL (2023): ChatGPT im Unterricht: Prüfungsvorbereitung mit KI. In: *manuelflick.de*. URL: <https://www.manuelflick.de/blog/pruefungsvorbereitung-mit-chatgpt> (letzter Zugriff: 25.05.2025).
- FLICK, MANUEL (2025): Der ChatGPT-Guide für Lehrkräfte. In: *manuelflick.notion*. URL: <https://manuelflick.notion.site/Der-ChatGPT-Guide-f-r-Lehrkr-fte-f214379898ce405089ac05555f06ba04> (letzter Zugriff: 25.05.2025).

- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* 5/2024, S. 2–13.
- GÖRZ, GÜNTHER/SCHMID, UTE/BRAUN, TANYA (2021): *Handbuch der Künstlichen Intelligenz*. 6. Aufl. Berlin: De Gruyter. <https://doi.org/10.1515/9783110659948>.
- GROSCHKE, MICHAEL/LÜKE, TIMO (2020): Vier Vorschläge zur Verortung quantitativer Forschungsergebnisse über schulische Inklusion im internationalen Inklusionsdiskurs. In: Gresch, Cornelia/Kuhl, Poldi/Grosche, Michael/Sälzer, Christine/Stanat, Petra (Hg.): *Schüler\*innen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen*. Wiesbaden: Springer VS, S. 29–54. [https://doi.org/10.1007/978-3-658-27608-9\\_2](https://doi.org/10.1007/978-3-658-27608-9_2).
- HANNAN, BRODY/EYNON, REBECCA (2025): Widening the Digital Divide: The mediating role of Intelligent Tutoring Systems in the relationship between rurality, socioeducational advantage, and mathematics learning outcomes. In: *Computers & Education*, 233/105312. <https://doi.org/10.1016/j.compedu.2025.105312>.
- HARGITTAI, ESZTER/PIPER, ANNE MARIE/MORRIS, MEREDITH RINGEL (2019): From internet access to internet skills: Digital inequality among older adults. In: *Universal Access in the Information Society*, 18, S. 881–890. <https://doi.org/10.1007/s10209-018-0617-5>.
- HART, BETTY/RISLEY, TODD R. (1995): *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes Publishing.
- HATTIE, JOHN/TIMPERLEY, HELEN (2007): The Power of Feedback. In: *Review of Educational Research*, 77/1, S. 81–112. <https://doi.org/10.3102/003465430298487>.
- HEIDEN, ANNA (2024): Mehrsprachigkeit und Künstliche Intelligenz: Chancen und Herausforderungen in der Lehrer\*innenbildung. In: *k:ON – Kölner Online Journal für Lehrer\*innenbildung*, 8, S. 312–327. <https://doi.org/10.18716/ojs/kON/2024.19>.
- HENSCHKE, SOPHIE/HEPPT, BIRGIT/WEIRICH, SEBASTIAN (2023): Zuwanderungsbezogene Disparitäten. In: Stanat, Petra/Schipolowski, Stefan/ Schneider, Rebecca/Weirich, Sebastian/Henschel, Sofie/Sachse, Karoline A. (Hg.) (2023): *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*. Münster: Waxmann, S. 299–344. <https://doi.org/10.31244/9783830997771>.
- HILDEBRANDT, ANDREAS (2023): Wie funktioniert ChatGPT? KI-basierte Schreibwerkzeuge aus technischer Sicht. In: *Institute for Computer Science, Johannes*

Gutenberg University Mainz. URL: <https://lehre.uni-mainz.de/files/2023/03/Textgenerierende-KI-in-der-Hochschulbildung-27.02.2023.pdf> (letzter Zugriff: 25.05.2025).

JACOBSEN, LUCAS JASPER/WEBER, KIRA ELENA (2023): *Optimierung von Feedback in der Hochschulbildung: Eine explorative Studie zu Prompt Engineering und der Qualität von KI-Feedback*. Vortrag auf der AEPF 2023, 13.–15. September 2023, Potsdam, Deutschland.

JELSON, ANDREW/MANESH, DANIEL/JANG, ALICE/DUNLAP, DANIEL/LEE, SANG WON (2025): *An Empirical Study to Understand How Students Use ChatGPT for Writing Essays*. <https://doi.org/10.48550/arXiv.2501.10551>.

KASNECI, ENKELEJDA/SEBLER, KATHRIN/KÜCHEMANN, STEFAN/BANNERT, MARIA/DEMENTIEVA, DARYNA/FISCHER, FRANK/GASSER, URS/GROH, GEORG/GÜNNE-MANN, STEPHAN/HÜLLERMEIER, EYKE/KRUSCHE, STEPHAN/KUTYNIOK, GITTA/MICHAELI, TILMAN/NERDEL, CLAUDIA/PFEFFER, JÜRGEN/POQUET, OLEKSANDRA/SAILER, MICHAEL/SCHMIDT, ALBRECHT/SEIDEL, TINA/... KASNECI, GJERGJI (2023): *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*. <https://doi.org/10.35542/osf.io/5er8f>.

KMK: Sekretariat der Ständigen Konferenz der Kultusminister der Länder der Bundesrepublik Deutschland (2024): *Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit künstlicher Intelligenz in schulischen Bildungsprozessen: Themenspezifische Handlungsempfehlung (Beschluss der Bildungsministerkonferenz vom 10. Oktober 2024)*. URL: [https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2024/2024\\_10\\_10-Handlungsempfehlung-KI.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf) (letzter Zugriff: 25.05.2025).

LINDMEIER, CHRISTIAN/LÜTJE-KLOSE, BIRGIT (2019): Inklusion. In: Haring, Marius/Rohlf, Carsten/Gläser-Zikuda, Michaela (Hg.): *Handbuch Schulpädagogik*. Münster: Waxmann, S. 586–596. <https://doi.org/10.36198/9783838587967>.

LÜTJE-KLOSE, BIRGIT (2023): Schulische Inklusion und sonderpädagogische Professionalität. Chancen und Herausforderungen der Digitalisierung. In: Ferencik-Lehmkuhl, Daria/Huynh, Ilham/Laubmeister, Clara/Lee, Curie/Melzer, Conny/Schwank, Inge/Weck, Hannah/Ziemen, Kerstin (Hg.): *Inklusion digital! Chancen und Herausforderungen inklusiver Bildung im Kontext von Digitalisierung*. Bad Heilbrunn: Julius Klinkhardt, S. 17–32. <https://doi.org/10.35468/5990-02>.

- MAGIRIUS, MARCO/HESSE, FLORIAN/HELM, GERRIT/SCHERR, DANIEL (2024): KI im Literaturunterricht. Chancen und Herausforderungen zwei Jahre nach der Veröffentlichung von ChatGPT. In: *Der Deutschunterricht* 5/2024, S. 14–23.
- MARTÍNEZ, PALOMA/RAMOS, ALBERTO/MORENO, LOURDES (2024): Exploring Large Language Models to generate Easy to Read content. In: *Frontiers in Computer Science*, 6, Article 1394705. <https://doi.org/10.3389/fcomp.2024.1394705>.
- MEYER, JENNIFER/JANSEN, THORBEN/SCHILLER, RONJA/LIEBENOW, LUCAS W./STEINBACH, MARLENE/HORBACH, ANDREA/FLECKENSTEIN, JOHANNA (2024): Using LLM to Bring Evidence-Based Feedback into the Classroom: AI-Generated Feedback Increases Secondary Students' Text Revision, Motivation, and Positive Emotions. In: *Computers and Education: Artificial Intelligence* 6 (2024), 100199. <https://doi.org/10.1016/j.caeai.2023.100199>.
- MURGIA, EMILIANA/SOLEDAD PERA, MARIA/LANDONI, MONICA/HUIBERS, THEO (2023): Children on ChatGPT Readability in an Educational Context: Myth or Opportunity? In: Association for Computing Machinery (ed.): *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23 Adjunct)*, S. 311–316. <https://doi.org/10.1145/3563359.3596996>.
- OpenAI (2022). GPT-3.5 [Large language model]. In: *openai.com*. URL: <https://platform.openai.com/> (letzter Zugriff: 25.05.2025).
- OpenAI (2023a): GPT-4 Technical Report. In: *openai.com*. URL: <https://openai.com/research/gpt-4> (letzter Zugriff: 25.05.2025).
- OpenAI (2023b): GPT-4 System Card. In: *openai.com*. URL: <https://cdn.openai.com/papers/gpt-4-system-card.pdf> (letzter Zugriff: 25.05.2025).
- PRILOP, CHRISTOPHER NEIL/WEBER, KIRA ELENA/KLEINKNECHT, MARC (2020): Effects of digital video-based feedback environments on pre-service teachers' feedback competence. In: *Computers in Human Behavior*, 102, S. 120–131. <https://doi.org/10.1016/j.chb.2019.08.011>.
- Robert Bosch Stiftung (2022): *Das Deutsche Schulbarometer Spezial: Geflüchtete ukrainische Schüler:innen an deutschen Schulen*. URL: [https://www.bosch-stiftung.de/sites/default/files/publications/pdf/2022-05/Deutsches\\_Schulbarometer\\_Spezial\\_Ukraine.pdf](https://www.bosch-stiftung.de/sites/default/files/publications/pdf/2022-05/Deutsches_Schulbarometer_Spezial_Ukraine.pdf) (letzter Zugriff: 25.05.2025).
- SHARMA, SAHIL/MITTAL, PUNEET/KUMAR, MUKESH/BHARDWAJ, VIVEK (2025): The role of large language models in personalized learning. A systematic review

- of educational impact. In: *Discov Sustain*, 6, Article 243. <https://doi.org/10.1007/s43621-025-01094-z>.
- STANAT, PETRA/SCHIPOLOWSKI, STEFAN/SCHNEIDER, REBECCA/SACHSE, KAROLINE A./WEIRICH, SEBASTIAN/HENSCHEL, SOFIE (Hg.) (2022): *IQB-Bildungstrend 2021. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im dritten Ländervergleich*. Münster: Waxmann. <https://doi.org/10.25656/01:26077>.
- STANAT, PETRA/SCHIPOLOWSKI, STEFAN/SCHNEIDER, REBECCA/WEIRICH, SEBASTIAN/HENSCHEL, SOFIE/SACHSE KAROLINE A. (Hg.) (2023a): *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*. Münster: Waxmann. <https://doi.org/10.31244/9783830997771>.
- STANAT, PETRA/SCHIPOLOWSKI, STEFAN/SCHNEIDER, REBECCA/WEIRICH, SEBASTIAN/HENSCHEL, SOPHIE/SACHSE, KAROLINE A. (2023b): Zusammenfassung und Einordnung der Befunde. In: Stanat, Petra/Schipolowski, Stefan/ Schneider, Rebecca/Weirich, Sebastian/Henschel, Sofie/Sachse, Karoline A. (Hg.) (2023): *IQB-Bildungstrend 2022. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im dritten Ländervergleich*. Münster: Waxmann, S. 435–467. <https://doi.org/10.31244/9783830997771>.
- STEINMETZ, MICHAEL (2020): Zum Matthäus-Effekt bei Support im Literaturunterricht. In: Dawidowski, Christian/Hoffmann, Anna Rebecca/Stolle, Angelika Ruth/Witte, Jennifer (Hg.): *Schulische Literaturvermittlungsprozesse im Fokus empirischer Forschung*. Frankfurt am Main: Peter Lang, S. 253–273.
- SWK: Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (Hg.) (2024): *Large Language Models und ihre Potenziale im Bildungssystem: Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz*. Bonn: SWK. <http://dx.doi.org/10.25656/01:28303>.
- VAN DEURSEN, ALEXANDER/VAN DIJK, JAN (2019): The first-level digital divide shifts from inequalities in physical access to inequalities in material access. In: *New Media & Society*, 21/2, S. 354–375. <https://doi.org/10.1177/1461444818797082>.
- VAN DIJK, JAN (2020): *The Digital Divide*. Cambridge, UK: Polity Press.
- WAMBSGANS, THIEMO/NIKLAUS, CHRISTINA/CETTO, MATTHIAS/SÖLLNER, MATTHIAS/HANDSCHUH, SIEGFRIED/LEIMEISTER, JAN MARCO (2020): AL: An Adaptive Learning Support System for Argumentation Skills. In: Bernhaupt, Regina/Mueller, Florian ,Floyd/Verweij, David/Andres, Josh/McGrenere, Joanna/Cock-

- burn, Andy/Avellino, Ignacio/Goguey, Alix/Bjøn, Pernille/Zhao, Shengdong/Samson, Briane Paul/Kocielnik, Rafal (Hg.): *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, S. 1–14.
- WANG, SHEN/XU, TIANLONG/LI, HANG/ZHANG, CHAOLI/LIANG, JOLEEN/TANG, JILIANG/YU, PHILIP S./WEN, QINGSONG (2024): *Large Language Models for Education. A Survey and Outlook*. <https://arxiv.org/html/2403.18105v1>.
- WU, YONG/SCHUNN, CHRISTIAN D. (2021): From plans to actions: A process model for why feedback features influence feedback implementation. In: *Instructional Science*, 49/3, S. 365–394. <https://doi.org/10.1007/s11251-021-09546-5>.
- YU, RENZHE/XU, ZHEN/CH-WANG, SKY/ARUM, RICHARD (2024): *Whose ChatGPT? Unveiling Real-World Educational Inequalities Introduced by Large Language Models*. <https://doi.org/10.48550/arXiv.2410.22282>.

## Zur Qualität KI-generierten Feedbacks

Ein explorativer Vergleich menschlicher  
und künstlicher Intelligenzen

### Abstract

Ziel des vorliegenden Beitrags ist, die Qualität der Rückmeldungen eines didaktisch systemgeprompteten, generativen Sprachmodells zu Schüler:innentexten zu untersuchen. Hierzu schrieben in einer explorativen Studie 19 Schüler:innen eine erste Textversion, erhielten KI-generiertes Feedback und überarbeiteten daraufhin noch einmal ihre Texte, die wiederum ein maschinelles Feedback bekamen. Die kriteriengeleiteten Rückmeldungen des KI-basierten Systems wurden im Anschluss mit menschlichen Expertenurteilen qualitativ und quantitativ verglichen. Darüber hinaus wurden die Überzeugungen der Schüler:innen untersucht. Die Arbeit liefert vier zentrale Ergebnisse: Erstens verbesserten die Schüler:innen die Qualität ihrer Texte durch die Überarbeitung – ob diese Verbesserung *dank* oder *trotz* des KI-generierten Feedbacks erwirkt werden konnte, muss ob der fehlenden Kontrollgruppe unbeantwortet bleiben. Zweitens weist der Vergleich der analytischen Urteile zwischen einem menschlichen Experten und dem KI-System nur auf schwache Übereinstimmungen ( $ICC = 0,279$ ,  $p = 0,001$ ) hin. Drittens deuten die qualitativen Analysen an, dass lernförderliches Feedback zu Schüler:innentexten durch das KI-System zwar möglich ist, es treten aber auch eindeutige Probleme in Bezug auf die Konsistenz und die inhaltliche Richtigkeit des Feedbacks zutage. Viertens haben diese Fehler zur Folge, dass einige Schüler:innen dem Feedback kein Vertrauen schenken, was eine der zentralen Herausforderungen von KI-generiertem Feedback ist und bleiben wird.

**Schlagwörter:** Künstliche Intelligenz, Sprachmodelle, Feedback, Argumentieren, Feedback

## 1 Hinführung

Im Bildungssektor löste der ChatGPT-Hype – analog zu anderen Entwicklungen der Digitalisierung – eine vorschnelle Absage an etablierte Teile des Systems aus: Hausaufgaben (Emmerich 2023) und Prüfungsformate wie Seminararbeiten (Batzlen 2023) werde es so nicht mehr geben. In Lehrerzimmern rund um den Globus fragt(-e) man sich, wie damit umzugehen sei, dass plötzlich alle Zugriff auf eine Technologie haben, die bessere Texte als Schüler:innen (Herbold et al. 2023) produziert, die nicht einmal mehr von erfahrenen Lehrkräften als maschinell erstellte Texte entdeckt werden können (Fleckenstein et al. 2024). Nach den anfänglichen Sorgen wurden auch Potenziale aufgedeckt, wie diese stochastischen Papageien (Bender et al. 2021) für das Bildungssystem genutzt werden könnten.

Für das Unterrichtsfach Deutsch, das im Zentrum dieses Beitrags steht, aber auch andere korrekturintensive Fächer wurde schon oft genug betont, dass Aufsatzkorrekturen Lehrkräfte enorm beanspruchen (Mußmann et al. 2017, Mußmann et al. 2020). Dies wird durch eine einfache Rechnung offenbar: Geht man von einer durchschnittlichen Klassengröße von 25 Schüler:innen pro Klasse und vier großen schriftlichen Leistungsnachweisen inklusive eines jeweils vorbereitenden Übungsaufsatzes aus, summiert sich die Anzahl der zu korrigierenden Text allein für das Fach Deutsch und bei ‚nur‘ zwei Deutschklassen bereits auf 400 Texte pro Schuljahr. Deutschlehrkräfte müssen folglich mehr als einen Text pro Tag im Jahr korrigieren. Es kommt erschwerend hinzu, dass dieser hohe Aufwand nicht unbedingt didaktisch zielführend ist, weil Kriterien lernförderlichen Feedbacks nicht eingehalten werden (Müller/Utesch/Busse 2023) und das Feedback teils mit starker zeitlicher Verzögerung erfolgt.

An eben dieser Herausforderung setzt die vorliegende Untersuchung an: Sprachmodelle produzieren im Handumdrehen sprachlich einwandfreies Textfeedback und könnten damit einerseits Lehrkräfte entlasten und andererseits Schüler:innen in kürzester Zeit lernförderliches Feedback geben. Der inhaltlichen Qualität dieses Feedbacks geht der vorliegende Text nach.

## 2 Forschungseinblick

Der Gedanke, manuelle Beurteilung von Texten mithilfe von Computern zu automatisieren, ist nicht neu (vgl. auch den Überblick in Wendt (2023)): Page (1966) beschäftigte sich schon vor fast 60 Jahren mit automatisierter Bewertung von Schüler:innentexten. Die meisten Entwicklungen in diesem Bereich, die insbesondere im angloamerikanischen Raum stattfanden, sind unter *Automatic Short Answer Grading* oder *Automated Essay Scoring* zu finden. Dabei stützten sich dafür eingesetzte Systeme bis 2022 bei der Bewertung von (kurzen) Texten vor allem auf sprachliche Oberflächenphänomene (Ramesh/Sanampudi 2022). Relativ neu ist der Einsatz von Sprachmodellen, die durch Verwendung hochdimensionaler Vektorräume (Simbeck i. d. B.) einen qualitativen Sprung versprechen.

So setzen beispielsweise Sawatzki et al. (2022) ein auf Wikipedia und Open Legal Data trainiertes und mit 233 Fragen und Antworten aus der Betriebswirtschaftslehre (Moodle-Prüfungen) spezialisiertes BERT-Modell ein. Das Modell sollte Punktbewertungen (0–6/8/10) zu Kurzantworten ( $M_{\text{Wörter}}=87,6$ ) vorhersagen. Die Modellgenauigkeit wurde statt mit dem für ordinale Daten geeigneten Spearman- mit dem Pearson-Korrelationskoeffizient geschätzt und liegt bei nicht gesehen Testsets bei  $r = 0,78$ , was die Autoren so interpretieren, dass BERT-Modelle effektiv zur automatischen Bewertung von Kurzantworten genutzt werden können, ohne jedoch einen p-Wert anzugeben.

Padò et al. (2023) prüfen die Inter-Rater-Reliabilität von speziell trainierten SBERT-Modellen und menschlichen Ratings beim Short Answer Grading (richtige vs. falsche Antwort). Das beste Modell erreichte eine Genauigkeit von  $M=71,4\%$  [ $MIN=64,7$ ,  $MAX=86,3$ ].

Zahlreiche Studien haben den Einsatz von nicht speziell trainierten Modellen für automatisierte Rückmeldung zu englischen Texten untersucht (u. a. Mizumoto/Eguchi (2023), Naismith et al. (2023)). Beispielsweise verglichen Chiang und Lee (2023) die Bewertungen von drei Lehrkräften mit denen von GPT-3. In ihrer Studie wurden 400 englische Textfragmente ( $M_{\text{Wörter}}=150$ ) analysiert. Die Beurteilung der Texte durch Mensch und Modell erfolgte anhand von vier Kategorien (*Grammaticality*, *Cohesion*, *Likability* und *Relevance*) auf einer fünfstufigen Likert-Skala. Die Ergebnisse zeigten lediglich für die Kate-

gorie *Relevance* (Aufgabenpassung des Textes) einen moderaten Zusammenhang zwischen den menschlichen und den GPT-3-Bewertungen, mit einem Kendall's Tau von  $\tau = 0,38$ .

Eine darauf aufbauende Untersuchung von Stahl et al. (2024) vergleicht verschiedene Prompting-Strategien, um zu überprüfen, wie gut Mistral Rückmeldungen zu englischen Texten geben kann. Dabei wurden die generierten Rückmeldungen mit Bewertungen von 12 Laien sowie mit den Rückmeldungen von LLaMA-2 und Mistral selbst verglichen. Die Autoren begründeten ihre Entscheidung, das Urteil über die Qualität des Textfeedbacks eines Sprachmodell durch ein anderes Sprachmodell bewerten zu lassen, mit dem vermeintlichen Ergebnis von Chiang und Lee (2023): „*Using LLMs to assess the quality of generated texts has been shown to be consistent with human expert annotations for some free-text generation tasks.*“ (Stahl et al. 2024: 7)

Die Studie von Seßler et al. (2025) untersucht die Fähigkeit großer Sprachmodelle, Schüler:innenaufsätze anhand von zehn Kriterien zu beurteilen. Dabei wurden fünf Modelle (GPT-3.5, GPT-4, o1, LLaMA 3-70B und Mixtral 8x7B) mit den Bewertungen von 37 Lehrkräften verglichen. Die Modelle und die Lehrkräfte bewerteten 20 Schüler:innenaufsätze auf einer sechs-stufigen Likert-Skala. Die Modelle wurden mit einem Zero-Shot-Prompt instruiert und es wurden zu jedem Text jeweils zehn Bewertungsdurchläufe durchgeführt, um die interne Konsistenz der Modellbewertungen zu testen. Die Untersuchung zeigt, dass die Urteile der geschlossenen Modelle (GPT-3.5, GPT-4, o1), insbesondere aber o1, eine hohe Korrelation mit denen der Lehrkräfte aufweisen. Besonders hoch ist die Übereinstimmung bei oberflächennahen Kriterien wie *Ausdruck*, *Satzstruktur* und *Rechtschreibung*. Das Modell o1 erreichte eine Übereinstimmung mit menschlichen Bewertungen von  $r = 0,74$  und eine interne Konsistenz von  $ICC = 0,80$ . Offen zugängliche Modelle (LLaMA 3-70B, Mixtral 8x7B) zeigten hingegen eine geringe Reliabilität und keine bis geringe Korrelation mit den Lehrerbewertungen. Zudem neigten alle Sprachmodelle dazu, Schüler:innen-aufsätze im Vergleich zu Lehrkräften systematisch milder zu bewerten.

Die besprochenen Studien, die stellvertretend für einen Teil der aktuellen Forschung zur automatisierten Beurteilung von Texten stehen, nutzen verschiedene quantitative Bewertungsmaße, mit denen sie ein überaus komplexes Phänomen (Textqualität) zu quantifizieren suchen, untersuchen fast nie

deutsche, kaum Texte von Lernenden und verwenden nur selten authentische Expertenurteile (Lehrkräfte, Sprachdidaktiker:innen oder Linguist:innen), um die Qualität der automatisierten Beurteilungen zu prüfen. Trotz vergleichsweise einfacher Bewertungsaufgaben (richtig oder falsch, Punktebewertung zu definierten Kategorien) liefern die Modelle dennoch oft keine zufriedenstellenden Ergebnisse, was den Eindruck vermitteln kann, dass der Weg zu lernförderlichem Feedback in Textform, das didaktischer Zielpunkt all dieser Bemühungen ist, noch überaus weit ist, wie auch Seßler et al (2025: 471) betonen.

Die Qualität der maschinellen Beurteilungen wird in allen vorgestellten Studien daran festgemacht, wie hoch sie mit einer menschlichen Beurteilung übereinstimmen. Dadurch entstehen mindestens zwei Probleme: Erstens sind sich Menschen bei der Beurteilung von Textqualität nicht unbedingt einig, auch wenn sie vergleichbar qualifizierte Expert:innen (z. B. Lehrkräfte) sind (Birkel/Birkel 2002, Schröter et al. 2023) – ihre Urteile über Textqualität sind also nicht unbedingt ein zuverlässiger „Goldstandard“. Zweitens wird in den Studien die qualitative Analyse der einzelnen Texte und Rückmeldungen durch Expert:innen ausgespart. Dadurch gerät das aus didaktischer Sicht relevanteste Qualitätskriterium aus dem Blick: der potenzielle Beitrag des Feedbacks zum Lernerfolg der Lernenden. Es ist daher dringend geboten, den quantitativen Vergleichen zwischen Mensch und Maschine auch qualitative Analysen von Textfeedback aus fachdidaktischer Perspektive an die Seite zu stellen und auch die Perspektive der Lernenden zu integrieren.

Obwohl verlässliche Qualitätsanalysen maschineller Rückmeldungen zu Schüler:innentexten also noch ausstehen, finden schon vielfach Systeme Einsatz – bisher allerdings meist<sup>1</sup> ohne fachdidaktische Begleitung oder Überprüfung –, die ein eben solches Feedback zur Verfügung stellen (u. a. *PEER*, *FelloFish*, *jobizz*, *cornelsen-AI*). Der Beitrag geht daher der Frage nach der Qualität solcher Rückmeldungen nach. Dazu werden im folgenden Kapitel die für diese Untersuchung notwendigen forschungsmethodischen Entscheidungen dargestellt, bevor zentrale Ergebnisse und deren Auswertung vorgestellt und diskutiert werden.

.....

1 Der Autor gehört seit dem Frühjahr 2024 dem wissenschaftlichen Beirat der FelloFish GmbH an.

## 3 Methodik

### 3.1 Stichprobe

Da Sprachmodelle maschinell lesbare Texte<sup>2</sup> benötigen und junge Schüler:innen teils noch größere Probleme mit der Schreibflüssigkeit<sup>3</sup> bei digitalen Texten haben, wurde die Studie mit einer neunten Klasse ( $N=19$ ;  $n_w=7$ ,  $n_m=12$ ) durchgeführt. Die teilnehmende Klasse gehörte einem bayerischen Gymnasium einer Kleinstadt im ländlichen Raum an. Die Stichprobe ist selbstredend weder randomisiert noch repräsentativ für die Grundgesamtheit, weshalb die vorzustellenden Ergebnisse zwar nicht verallgemeinerbar sind, wohl aber relevante methodische und inhaltliche Ergebnisse für zukünftige Studien liefert.

### 3.2 Erhebungsinstrumente

#### 3.2.1 Textsorte

Zur Untersuchung der Qualität von KI-generierten Rückmeldungen wurde die Textsorte *Materialgestütztes Argumentieren* (Feilke/Topfink 2017) gewählt. Das (schriftliche) Argumentieren und die Vermittlung entsprechender Kompetenzen führt die Kultusministerkonferenz gleich zu Beginn ihrer Bildungsstandards für das Fach Deutsch für die Allgemeine Hochschulreife (KMK 2014: 13) unter allgemeinen Zielen des Faches auf und verleiht dieser Kompetenz auch in der Formulierung eine hohe Relevanz: „Besonderes Gewicht erhält die Entwicklung der Argumentations- und Reflexionsfähigkeit“. Diese nimmt spätestens ab der siebten Jahrgangsstufe eine zentrale Rolle im Schreibprogramm des Gymnasiums ein, die es bis zum Abitur nicht einbüßt.<sup>4</sup> KI-Systeme bieten insbesondere in diesem Bereich didaktische Anknüpfungspunkte weit über das

.....

- 2 Zwar machen es OCR-Techniken auch möglich, die Handschrift von Schüler:innen erkennen zu lassen, allerdings produzieren sie nach wie vor keine verlässlichen Ergebnisse.
- 3 Schreibflüssigkeit gilt als Prädiktor für Textqualität (Sturm/Schneider 2021: 40 f.).
- 4 Hier zeichnet sich in der Grundschule ein Wandel ab: Waren argumentierende Kompetenzen in den Bildungsstandards für die Primarstufe in der Fassung von 2004 (KMK 2005) noch vor allem auf den Kompetenzbereich *Sprechen und Zuhören* fokussiert, weist die neue Fassung von 2022 auch schriftliche Argumentationskompetenzen aus (KMK 2022: 15).

Feedback hinaus, welche die Unnatürlichkeit des Argumentierens per Aufsatz auflösen können (Fürstenberg/Matz 2025).

### 3.2.2 KI-System

Doch auch über die Auflösung der zerdehnten Kommunikationssituation hinaus können sich KI-Systeme als lernförderlich erweisen: Insbesondere die Unmittelbarkeit des Feedbacks, aber auch Rückmeldungen zu eher formal orientierten Kompetenzen wie dem Einsatz von Informationen oder Zitierten Materialien, aber auch die Einhaltung der sprachlichen Normen sollten durch KI-Systeme schnell und effektiv überprüft werden können (Neff 2023). Gleichzeitig gelten Sprachmodelle im Vergleich zu Lehrkräften als objektiver, was Vor- und Nachteil zugleich ist: In Bezug auf die Fairness der Beurteilung wäre das zu begrüßen. Andererseits haben KI-Systeme (noch) keinen Zugriff auf den „Kontext des Feedbacks“ (Philipp 2023: 13), wissen also nicht, wie der bisherige Kompetenzerwerb der jeweiligen Person abgelaufen ist, und können daher das Produkt nicht vor dem Hintergrund der Lernbiografie beurteilen. Zudem muss kritisch auf den Einsatz von Sprachmodellen als Feedbackmaschinen geblickt werden, denn „[e] liegt auf der Hand, dass [...] Feedback hohes metakognitives Wissen zum Schreiben (darunter auch: schreibdidaktisches Wissen) impliziert, um sinnvolle Handlungsoptionen vorzuschlagen.“ (ebd.: 12) Mit Blick auf das Training von großen Sprachmodellen (Simbeck i. d. B.; Müller/Fürstenberg 2023) ist es sehr wahrscheinlich, dass dabei „deutschdidaktisch relevante Inhalte allenfalls einen Bruchteil ausgemacht haben“ (Fürstenberg/Müller 2024: 2f.). Allgemeine Sprachmodelle wie ChatGPT können also nicht auf solches Wissen zugreifen. Daher wurde ein didaktisch systemgepromptes KI-System für die Studie genutzt. Von den bei der Erhebungsplanung zur Verfügung stehenden Werkzeugen war *fiete.ai* das ausgereifteste, da es nicht nur Bewertungskriterien, die Aufgabenstellung, Jahrgangsstufe und Material in Bezug auf das Feedback miteinbezieht, sondern zusätzlich mithilfe von System-Prompts für das Feedbackgeben spezialisiert wurde (Haverkamp/Hecht/Schindler 2024). Das KI-System gibt den Schüler:innen neben qualitativem Feedback auch eine quantitative Einschätzung der Leistung in Form von Fortschrittsbalken (s. Abb. 1):

Du hast einen guten Anfang gemacht, aber es gibt noch einige Bereiche, in denen du dich verbessern kannst. Konzentriere dich besonders auf die Verwendung von Materialien zur Unterstützung deiner Argumente und die Struktur deines Textes. Du hast das Potenzial, einen ausgezeichneten Beitrag zu schreiben, also gib nicht auf!



1. *Die Einleitung steigt mit einer sprachlichen Auffälligkeit (Frage, These, Zitat) ein, skizziert die Thematik und stellt kurz die Kontroverse dar. Sie führt dann knapp zur Themafrage hin, mit der in den Hauptteil übergeleitet wird.*  
Du hast eine gute Einleitung geschrieben, die das Thema einführt und die Kontroverse darstellt. Du könntest jedoch die Themenfrage klarer formulieren, um besser in den Hauptteil überzuleiten.

2. *Der Hauptteil ist in Gegenposition, Eigenposition und Synthese gegliedert, wobei die Argumente der Gegenseite vom stärksten zum schwächsten und bei der Eigenposition vom schwächsten zum stärksten Argument sortiert sind.*  
Dein Hauptteil ist gut strukturiert, aber du könntest die Argumente besser sortieren. Versuche, die Argumente der Gegenseite vom stärksten zum schwächsten und bei der Eigenposition vom schwächsten zum stärksten Argument zu sortieren.

3. *Die Argumente sind vollständig (enthalten jeweils Behauptung, Begründung und ein konkretes Beispiel), sind inhaltlich richtig, passen zur Erörterungsfrage, nutzen eine sachliche und unterstützende Sprache und sind inhaltlich kohärent.*  
Deine Argumente sind vollständig und inhaltlich richtig, aber du könntest mehr Beispiele und Materialien zur Unterstützung verwenden. Versuche, mehr Materialien zu zitieren und konkrete Beispiele zu geben.

Abb. 1: KI-Feedback zu einem Schüler:innentext

Der Ausschnitt zeigt einen Teil der KI-generierten Rückmeldungen zu einem der abgegebenen Texte. Es wird zunächst ein allgemeiner Eindruck wiedergegeben, der zur Überarbeitung des Textes motivieren soll (Guter Anfang, aber Verbesserungspotential) und stellt zwei Kriterien (Materialnutzung und Aufbau) in den Mittelpunkt der Kritik. Im Anschluss wird jedes Kriterium erst wiederholt (kursiv), dann wird qualitatives und im Anschluss quantitatives Feedback gegeben. Durch die farbigen Fortschrittsbalken können sich die Schüler:innen schnell orientieren, wo noch am meisten Handlungsbedarf besteht. Das qualitative Feedback nennt stets zuerst Positives im Sinne eines *Feed backs* (Philipp 2023: 11f.) und formuliert im Anschluss Kritik in Form eines *Feed forwards* (ebd.) – beides basiert auf den zugrundeliegenden Beurteilungskriterien.

### 3.3 Beurteilungskriterien

U.a. Philipp (2015: 34) zeigt, dass der Aufbau argumentativer Textkompetenz von einer kriteriengeleiteten Überarbeitung des eigenen Textes profitieren kann. Solche Kriterien erarbeitete die Klasse in einer Unterrichtssequenz, die hier nur skizziert wird: Die Teile des Zieltextes wurden dazu von den Schüler:innen produziert, im Anschluss mithilfe einer Mischung<sup>5</sup> aus Schreibkonferenz (Spitta 1992) und Textlupe (Böttcher/Wagner 1993) miteinander sowie mit KI-generierten Lösungen verglichen. Daraus wurden folgende Kriterien abgeleitet:

ID	Kriterium (aufgeteilt in Unterkriterien)
1	a) Die Einleitung steigt mit einer sprachlichen Auffälligkeit (Frage, These, Zitat) ein, b) skizziert die Thematik und c) stellt kurz die Kontroverse dar. d) Sie führt dann knapp zur Themafrage hin, e) mit der in den Hauptteil übergeleitet wird.
2	a) Der Hauptteil ist in Gegenposition, Eigenposition und Synthese gegliedert, b) wobei die Argumente der Gegenseite vom stärksten zum schwächsten und c) bei der Eigenposition vom schwächsten zum stärksten Argument sortiert sind.
3	a) Die Argumente sind vollständig b) (enthalten jeweils Behauptung, c) Begründung und d) ein konkretes Beispiel), e) sind inhaltlich richtig, f) passen zur Erörterungsfrage, g) nutzen eine sachliche und unterstützende Sprache und h) sind inhaltlich kohärent.
4	a) Die Synthese wägt vor dem Schluss nochmal die stärksten Argumente gegeneinander ab und b) begründet die eigene Position abschließend.
5	a) Der Schluss schließt an die Einleitung an und b) enthält einen Appell oder einen Ausblick.
6	a) Es werden alle notwendigen Kommas gesetzt und b) die großzuschreibenden Wörter werden großgeschrieben.
7	a) Aus dem Material wird mit Anführungszeichen und b) unter Angabe der Quelle zitiert.

**Tab. 1:** Bewertungskriterien materialgestütztes schriftliches Argumentieren

- .....
- 5 Die engere Führung durch die Textlupe kann für Schüler:innen hilfreich sein (Reichardt/Kruse/Lipowsky 2014: 81).

Die ersten fünf Kriterien sind weitgehend textchronologisch angeordnet. Kriterien 6 und 7 zielen auf eher formalsprachliche Aspekte ab. Die Einzelkriterien werden hier nicht detailliert besprochen. Sie wurden von der Lehrkraft im Unterricht vermittelt und enthalten sowohl gelungene wie auch durchaus strittige Inhalte: So kann mithilfe der überleitenden Themafrage lokale Textkohärenz (Averintseva-Klisch 2018: 15) hergestellt werden und die Qualitätskriterien der Argumente sind an Schwarze (2016) angelehnt. Das in der Schule beliebte Sanduhr-Prinzip hingegen darf mit Dax (2023: 254 f.) als umstritten<sup>6</sup> gelten und die Kriterien beziehen sich weder auf prozessuale (z. B. metakognitive) noch auf sonstige personale Fähigkeiten. Darüber hinaus sind es zwar auf erster Ebene nur sieben Kriterien, die jedoch bei näherer Betrachtung in 24 Einzelkriterien zerfallen. Es sollte allerdings nicht unberücksichtigt bleiben, dass viele der Anforderungen den Schüler:innen aus früheren Jahrgangsstufen bereits bekannt gewesen sein sollten.

### 3.4 Untersuchungsablauf

Die Schüler:innen schrieben den Übungsaufsatz auf Basis einer ansatzweise profilierten Schreibaufgabe (Bachmann/Becker-Mrotzek 2010):

*Die Schülerzeitung am Gymnasium XY ist noch auf der Suche nach einem abwägenden Beitrag zur Frage, ob Influencer ein Traumberuf ist. Schreibe diesen Beitrag, indem du die gegebenen Materialien (M1–7) und eigenes Wissen nutzt. Wähle eine geeignete Überschrift und beziehe gegen Ende deines Textes Position.*

**Abb. 2:** Schreibaufgabe

Die Aufsätze wurden auf den schuleigenen iPads mit dem Textverarbeitungsprogramm *Microsoft Word* verfasst. Direkt nach der Abgabe analysierte das KI-System die Texte und gab in weniger als einer Minute ein entsprechendes Feedback (s. Abb. 1). Die Klasse hatte am Ende der ersten Stunde und in der

.....

6 Eine zentrale Herausforderung bildet die Gewichtung der Argumente nach Güte, denn hierbei sind Wissen und Kompetenzen gefragt, die der Deutschunterricht eher sekundär ausbildet. Ganz konkret mussten die Schüler:innen entscheiden (s. Abb. 2), ob bspw. die hohe Verdienstmöglichkeit von Influencern oder die große Freiheit in der Ausübung des Berufes ein wichtigeres Argument darstellt.

Folgestunde Zeit, um dieses Feedback zu analysieren, die Aufsätze zu überarbeiten und erneut einzureichen. Auch nach dieser zweiten Abgabe gab das KI-System Feedback auf die überarbeitete Version. Beide Textversionen wurden anonymisiert an den Untersuchungsleiter weitergeleitet, der im Anschluss alle Abgaben der Klasse (N=347) randomisiert anordnete und die Qualität jedes Aufsatzes sechs Wochen später mit einer Punktzahl (0–15) holistisch beurteilte. In derselben Reihenfolge benotete auch ein weiterer schreibdidaktischer Experte die Arbeiten. Einen Monat später beurteilte nur der Untersuchungsleiter auf derselben Skala (1–11) wie das KI-System im Unterricht für alle 34 Texte anhand von 7 Kriterien (s. Tab. 1) analytisch die Qualität jedes Aufsatzes. Daraus ergibt sich folgender Überblick:

Rater	Beurteilungsform	Likert-Skala
KI-System	analytisch (7 Kriterien)	1–11
Untersuchungsleiter	analytisch (7 Kriterien) und holistisch	1–11 und 0–15
Schreibdidaktik-Experte	holistisch	0–15

**Tab. 2:** Überblick über die Rater

Somit wird ein Vergleich der analytischen Urteile von KI-System und Untersuchungsleiter sowie der holistischen Urteile von Untersuchungsleiter und Schreibdidaktik-Experte möglich.

Die folgenden Ergebnisse wurden mit *R* berechnet. Die Ergebnisse des Fragebogens wurden direkt in *Google-Forms* ausgewertet und in *Excel* teilweise aufbereitet.

- .....
- 7 Zwei Schüler:innen waren in der Überarbeitungsstunde nicht anwesend und holten die Überarbeitung zu Hause nach. Diese wurden für die Untersuchung ausgeschlossen.

## 4 Ergebnisse und Diskussion

### 4.1 Qualitätsentwicklung der Texte

Ein kurzfristiges Ziel von Feedback zu Texten ist, dass die Schüler:innen diese überhaupt überarbeiten: Jansen et al. (2025) stellen in Übereinstimmung mit dem Forschungsstand fest, dass lediglich 48 % der Schüler:innen ihre Texte nach einem KI-generierten Feedback verändern. Sollten die Schüler:innen überarbeiten, sind grundsätzlich drei Folgen dieser Tätigkeit denkbar: Die Texte können verschlechtert werden, auf einem ähnlichen Niveau bleiben oder verbessert werden.

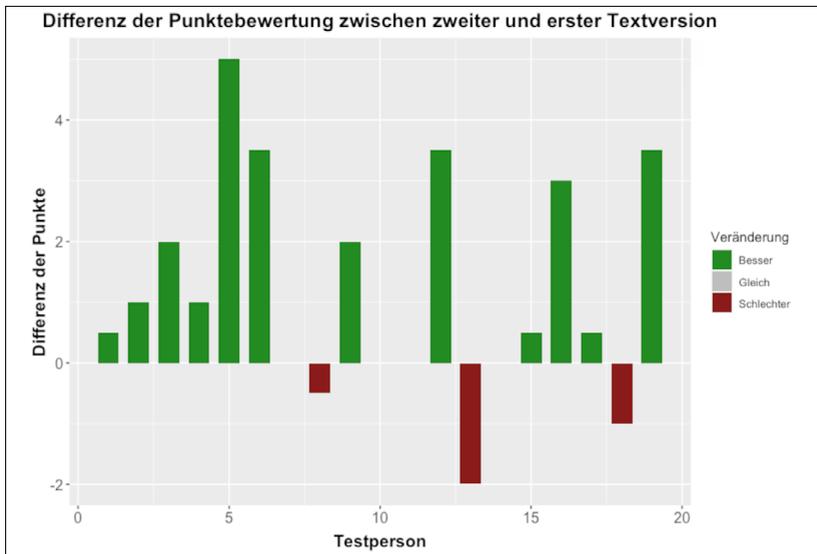


Abb. 3: Säulendiagramm zum Vergleich der beiden Textversionen

Um ein möglichst genaues Bild zeichnen zu können, wurde für Abb. 3, die eben diese Entwicklung für 17 Schüler:innen in einem Säulendiagramm zeigt, ein Mittelwert aus den Punkten der beiden menschlichen Rater<sup>8</sup> für 34 Texte

.....

8 Die holistischen Gesamturteile der beiden menschlichen Rater liegen bei einem ICC von 0,754 ( $p = 0,001$ ), was auf eine hohe Übereinstimmung hindeutet.

gebildet. Im Anschluss wurde der mittlere Punktwert der zweiten Abgabe vom mittleren Punktwert der ersten Abgabe subtrahiert, wodurch die Veränderung von der ersten zur zweiten Abgabe sichtbar wird.

Von den 17 Schüler:innen verschlechterten drei ihren Text, zwei Texte blieben auf demselben Niveau und zwölf Texte wurden bei der zweiten Abgabe besser. Zwei Ausprägungen der kategorialen Variable *Testperson* auf der x-Achse bleiben leer, weil diese keine zweite Version einreichten. Damit verbesserten 70 % der Proband:innen durch die Überarbeitung ihren Text. Fünf Testpersonen waren sogar in der Lage, mit der überarbeiteten Textversion drei bis fünf Punkte mehr zu erreichen, was im Mittel einer Leistungssteigerung von einer ganzen Notenstufe entspricht. Selbstredend kann die Verbesserung nicht direkt auf die Qualität des Feedbacks durch das KI-System rückgeführt werden. Dazu wären eine Kontrollgruppe und mehr Testpersonen nötig. Nichtsdestotrotz kann festgehalten, dass das KI-generierte, kriteriengeleitete Feedback bei den meisten Schüler:innen nicht zu einer Leistungsverschlechterung geführt hat. Womit die Verbesserung der meisten Texte zusammenhängt, muss hingegen ungeklärt bleiben. Die explorative Untersuchung zeigt, dass es durchaus lohnenswert sein könnte, das Experiment mit mehr Testpersonen zu replizieren. Es bleibt das ermutigende Ergebnis, dass die meisten Schüler:innen kompetent genug waren, ihre Texte entweder mithilfe oder vielleicht sogar entgegen dem Feedback durch das KI-System zu verbessern.

## 4.2 Quantitatives Feedback

Um die Qualität des KI-generierten Feedbacks zu bestimmen, wird in vielen Fällen auf den Vergleich zwischen menschlichem und maschinellem Feedback zurückgegriffen. Wenngleich diese Qualitätsprüfung problematisch ist, stellt es für diese frühe Phase der Analyse von KI-Feedback dennoch eine pragmatische Zwischenlösung dar.

Dazu wurden die sieben Bewertungskriterien (s. Tabelle 1) vom Untersuchungsleiter und dem KI-System jeweils auf einer Skala (1–11 Punkte) bewertet. Rein deskriptiv lässt sich für jede Kategorie zählen, wie häufig die beiden Rater (KI und Untersuchungsleiter) in Bezug auf die Punkte wie weit auseinanderliegen:

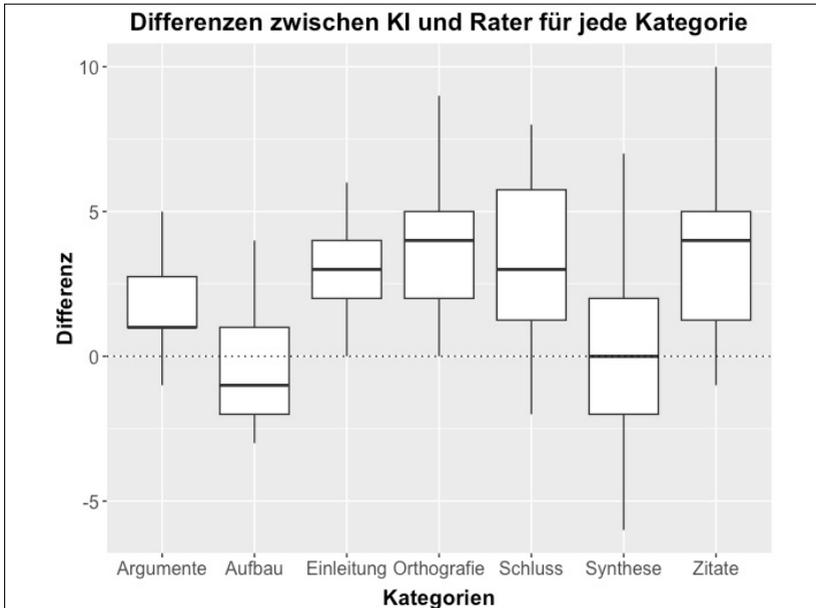
Rater-differenz	0	+/-1	+/-2	+/-3	+/-4	+/-5	+/-6	+/-7	+/-8	+/-9	+/-10
Anzahl	19	53	52	35	30	22	12	11	2	1	1
Prozent	8	22,3	21,8	14,7	12,6	9,2	5	4,6	0,8	0,4	0,4

**Tab. 3:** Raterdifferenz über alle Beurteilungskategorien

Maschine und Mensch stimmen bei 30,3 % ( $n = 72$ ) der Beurteilungen über alle Kategorien hinweg entweder perfekt (Differenz = 0) oder annähernd perfekt (Differenz = 1/-1) überein. Bei 159, also etwa in etwa zwei Drittel aller 238 Ratings, stimmen die beiden Rater mit einer Differenz von 3 oder weniger Punkte überein. Das wirkt wie ein sehr guter Wert. Diese intuitive Einschätzung wird aber durch die Angabe eines statistisch verlässlicheren Wertes infrage gestellt: Da es sich um zwei feste Rater und sieben Kategorien handelt, die mit in ihren Abständen sinnvoll interpretierbaren Punkten bewertet wurden, wurde die Interrater-Reliabilität mit der Intraklassen-Korrelation (ICC3) berechnet (Hedderich/Sachs 2020: 434 ff.). Die ICC liegt hier bei 0,279 ( $p = 0,0001$ ), was eine geringe bis mäßige Übereinstimmung zwischen dem menschlichen und maschinellen Rater bei der kategorialen Bewertung der Texte andeutet.

Betrachtet man die Übereinstimmung pro Kategorie, zeigt sich die überaus komplexe Kategorie 3 (Argumentgüte) mit 53 % als diejenige, mit dem höchsten Zusammenhang zwischen Mensch und Maschine. Es ist also möglich, dass innerhalb der globalen Einschätzung der Übereinstimmung von Mensch und Maschine einzelne, besonders schlecht oder gut übereinstimmende Kategorien enthalten sind, die durch das Gesamtergebnis verschleiert werden.

Weiter oben wurde bereits darauf hingewiesen, dass Sprachmodelle ob ihrer Funktionsweise gerade die Kategorien, welche sich auf die sprachliche Oberfläche (z. B. Orthografie oder Zitate) beziehen, besonders sicher bewerten sollten. Daher beschreiben die Boxplots in Abb. 4 für jede der sieben Kategorien, wie hoch die Differenz zwischen menschlichem und maschinelltem Rater ist. Dazu wurde für alle 34 Texte in jeder der sieben Kategorien jeweils die Differenz zwischen maschinelltem und menschlichem Rater berechnet. Der Plot fasst die wichtigsten deskriptiven Daten der Abweichungen in Punkten (y-Achse) für jede Kategorie (x-Achse) zusammen.



**Abb. 4:** Boxplots zur Abweichung von menschlichem und maschinellem Rater nach Bewertungskategorie

Die maschinelle Bewertung ist bei fünf der sieben Kategorien höher als die menschliche, weshalb die Boxen über der gestrichelten Nulllinie liegen. Die KI hat also die Texte der Schüler:innen grundsätzlich eher milder beurteilt als der Untersuchungsleiter, was zu den Ergebnissen von Seßler et al. (2025: 468) passt. Lediglich bei der Kategorie *Aufbau* war der menschliche Rater strenger und bei der Kategorie *Synthese* liegt der Median exakt auf der Null-Linie – hier stimmen also die menschlichen und maschinellen Beurteilungen im Mittel überein. Das ist insofern weniger erstaunlich, als die Kategorie *Synthese* mit der Abwägung der stärksten Argumente gegeneinander und dem Begründen der eigenen Position durchaus sprachliche Handlungen (abwägen, begründen) fordert, die konkrete Oberflächenmerkmale wie adversative Konjunkionaladverbien (*allerdings, einerseits ... andererseits* etc.) oder kausale Subjunktionen (*weil, da* etc.) mit sich bringen, die von Sprachmodellen besonders gut erkannt werden. Auch die stärksten Argumente sind bei den meisten Texten sprachlich (z. B. *am wichtigsten aber ist*) und/oder formal (durch Absätze)

gekennzeichnet. Jedoch ist die Box der Kategorie *Synthese* an sich relativ groß und auch die Whisker-Linien (vertikale Linien, enthalten den Großteil der Daten außerhalb der Box) zeigen eine breite Streuung der Daten an. Wie bereits die ICC kategorienübergreifend gezeigt haben, kann von Einigkeit also nicht die Rede sein. Eine Reliabilitätsanalyse der einzelnen Ergebnisse kam zu keinen signifikanten Ergebnissen, was vor allem auf die geringe Anzahl an Daten pro Kategorie zurückzuführen sein dürfte.

Das mit Abstand komplexeste Beurteilungskriterium ist die Qualität der *Argumente*, da es acht – vor allem inhaltlich geprägte – Unterkriterien enthält. Es wurde vermutet, dass das KI-System mit der Beurteilung dieser komplexen Kategorie Probleme haben sollte. Die prozentualen Daten zur Rater-Übereinstimmung ließen schon vermuten, dass diese Kategorie aber die höchste Übereinstimmung aufweist, was die Boxplots entsprechend spiegeln, da die Box am kleinsten ist und die Whisker-Linien am kürzesten sind. Da die Box nur nahe der Null-Linie liegt, haben Mensch und KI hier zwar tendenziell ähnlich, die KI aber mit durchschnittlich mehr Punkten beurteilt. Ähnliches gilt noch für die zweitkomplexeste Kategorie (*Einleitung*), bei der sich allerdings die Box noch weiter von der Null-Linie entfernt, die KI also noch einmal im Schnitt mit einer höheren Punktzahl bewertet hat, wenn auch die Tendenz ähnlich ist wie beim menschlichen Rater. Die Kategorien *Schluss*, *Orthografie* und der Einsatz von *Zitaten* sind allesamt überaus wenig übereinstimmend, was insbesondere mit Blick auf die beiden sprachformalen Kategorien, *Orthografie* und *Zitate*, überrascht.

Insgesamt lässt sich keine zufriedenstellende Übereinstimmung zwischen der kategorialen Bewertung durch den Untersuchungsleiter und dem Sprachmodell festhalten. Zwar liegen einzelne und teilweise auch inhaltlich komplexe Kriterien näher an einer Übereinstimmung. Die ICC-Analyse legt aber nicht nahe, dass das eingesetzte KI-System eine der menschlichen Bewertung ähnliche Beurteilung vornehmen würde. Die Punkte des Sprachmodells lagen im Mittel fast durchgehend höher als die Bepunktung durch den Untersuchungsleiter. Eine durchaus denkbare Begründung für diesen Umstand ist, dass potenzielle Attribuierungen im System-Prompt wie *Du bist ein freundlicher Lehrer, der Texte beurteilen soll*. dazu führen, dass *freundlich* in den Vektorräumen des Sprachmodells auch mit besserer Benotung assoziiert ist. Ebenso plausibel ist,

dass Sprachmodelle grundsätzlich einen Höflichkeits-Bias aufweisen, der sich hier und bei Seßler et al. 2025 in (zu) hohen Bewertungen von Schüler:innen-texten zeigt.

### 4.3 Qualitatives Feedback

Um den quantitativen Abweichungen zwischen menschlichen und maschinellen Ratings qualitativ auf den Grund zu gehen, werden kursorisch einige der Textteile (orthografisch unverändert) mitsamt dem Feedback dargestellt, um Herausforderungen KI-generierten Feedbacks zu entdecken.<sup>9</sup> Es ist bei diesem Vorgehen wichtig, sich vor Augen zu führen, dass diese Beispiele bewusst ausgewählt wurden, um Herausforderungen herauszuarbeiten, die sich in den Daten zeigen. Es könnten ebenso viele Beispiele gezeigt werden, die potentiell hilfreiches Feedback aufweisen.

In der Gesamtschau über alle qualitativen Textrückmeldungen zeigt sich das KI-System als genau der Phrasenspeicher (Müller/Fürstenberg 2023: 336) bzw. Papagei, als der große Sprachmodelle gelten, und zwar offenbar insbesondere dann, wenn sie durch den Systemprompt einen sehr spezifischen Auftrag erhalten. So werden stets dieselben syntaktischen Strukturen (Matrixsatz, häufig mit einer Aufzählung, gefolgt von einem Satzgefüge mit einer finalen Infinitivgruppe; s. Abb. 5) mit den entsprechenden Wörtern aus der jeweiligen Beurteilungskategorie gefüllt, was insgesamt eher an Produkte eines deterministischen Systems erinnert und wohl auf eine sehr niedrige *temperature* (Fürstenberg/Müller 2024: 85) zurückzuführen ist. In der Einzelbetrachtung zeigen sich dann aber durchaus Unterschiede zwischen den jeweiligen Feedbacktexten, deren inhaltliche Qualität in der Folge beleuchtet wird.

Überraschende Einigkeit zeigte in der quantitativen Untersuchung die komplexe Beurteilungskategorie *Argumente*. Bei der qualitativen Beurteilung moniert das KI-System in jedem zweiten Feedbacktext, die Sprache solle sachlicher und/oder unterstützender sein. Auch die inhaltliche Kohärenz

.....  
 9 Qualitative Analysen zu den Kategorien *Einleitung* und *Synthese* liegen in Fürstenberg (2025) bereit.

der Argumente wird relativ häufig ( $n = 12$ ) kritisiert. Der menschliche Rater hingegen kritisiert am häufigsten fehlende oder zu wenig konkrete Beispiele in den Argumenten und damit die Vollständigkeit der Argumente und weniger sprachliche Faktoren. Es zeigen sich qualitativ also deutliche Unterschiede, wenngleich die quantitative Übereinstimmung in dieser Kategorie am höchsten ist. Hier deutet sich eine Inkonsistenz zwischen quantitativer Beurteilung und qualitativer Beurteilung per textuellem Feedback an, die sich noch an weiteren Stellen zeigt: Bei fünf Texten lässt das qualitative Feedback des KI-Systems keine Verbesserungsoptionen erkennen, beurteilt die Texte also als perfekt. Nicht nur für die Schüler:innen ist es verwirrend, dass das System aber nur in zwei dieser fünf Fälle analog zum qualitativen Feedback auch quantitativ die Höchstpunktzahl ausgibt. Hinzu kommt, dass in beiden Fällen, in denen das KI-System auch elf Punkte erteilt, das menschliche Feedback die Texte deutlich kritischer (sieben bzw. acht Punkte) einschätzt: So fehlen den sowohl qualitativ als auch quantitativ als perfekt beurteilten Argumenten teilweise (konkrete) Beispiele und auch sprachlich sind die Argumente stellenweise zu unsachlich formuliert. Insofern böten beide Texte noch Verbesserungsoptionen, die durch das maschinelle Feedback aber unentdeckt bleiben.

Auch die Kategorie *Schluss* wurde vom KI-System sprachlich überaus homogen beurteilt. So formuliert die KI für jeden der 34 Texte, dass der Schluss (gut) an die Einleitung anschließe, einen Appell/Ausblick enthalte, jedoch vice versa noch Ausblick/Appell formuliert werden könnte, um den Text zu verbessern. Mit Blick auf das Beurteilungskriterium (s. Tab. 1) war jedoch nur ein Ausblick *oder* ein Appell gefordert, weshalb jedes einzelne Feedback inhaltlich mindestens als unpräzise zu bewerten ist. Nur in drei Fällen weicht das Feedback (jeweils bei der zweiten Textversion) ganz leicht von der benannten Struktur ab, indem es auf die Überarbeitung konkret Bezug nimmt (*Dein Schluss schließt immer noch gut an die Einleitung an und enthält einen Ausblick.*).

Der Schluss eines Schüler:innentextes (Herv. MF) rief eine besonders hohe quantitative Unstimmigkeit hervor und die entsprechende Rückmeldung deckt eine Herausforderung maschinell erstellten Feedbacks auf:

<p>Obwohl der Beruf <u>Vorteile und Nachteile hat</u>, ist es die <u>Entscheidung jedes Einzelnen</u>, ob er oder sie sich für den Beruf entscheidet. Allerdings sollte man sich den <u>möglichen Konsequenzen</u>, die die körperliche und mentale Gesundheit beeinträchtigen können, <u>bewusst sein</u>.</p> <p>Dein Schluss schließt gut an die Einleitung an und enthält einen Ausblick. Du könntest jedoch einen stärkeren Appell einbauen, um deine Position zu unterstreichen.</p>
--

**Abb. 5:** Feedbackbeispiel I

Der Schluss wird vom Sprachmodell mit der Höchstzahl bewertet (11 Punkte), während der menschliche Rater lediglich vier Punkte vergibt. Das qualitative KI-Feedback unterstreicht zu Recht und positiv die Verbindung zur Einleitung, moniert aber (trotz Höchstpunktzahl) die Stärke des Appells in Form eines Überarbeitungsvorschlags. Das maschinelle Feedback weist hier erneut mangelnde Konsistenz in der Bewertung auf – in der Diskrepanz zwischen Maximalpunktzahl als Bestnote und der qualitativen Rückmeldung – und lässt inhaltliche Mängel außer Acht, die im Vergleich zum menschlichen Feedback deutlich werden, welches den Schluss ob der vielen Allgemeinplätze als unzureichend ausweist. Da die Bewertungskriterien aber nicht explizit ausweisen (können), was alles nicht gemacht werden sollte, und womöglich auch weil Sprachmodelle selbst qua Entwicklung (Müller/Fürstenberg 2023) einen Hang zu sehr allgemeinen Aussagen aufweisen, erkennt das Sprachmodell ein zentrales Problem dieses Schlusses nicht. Darüber hinaus wird hier auch ein bekanntes Problem kriteriengeleiteter Beurteilung von Texten offenbar, denen das Modell ausgeliefert ist: Über die Kriterien hinaus können Texte sehr individuelle Mängel aufweisen, die aber durch den Kriterienkatalog nicht abgedeckt sind. Hier kommt also eher die kriteriengeleitete Beurteilung als das KI-System an seine Grenzen.

Auch die deskriptiven Daten (s. Abbildung 3) zur Übereinstimmung zwischen Mensch und Maschine innerhalb der Kategorie *Zitate* wiesen auf eine hohe Unstimmigkeit hin, die insofern weniger erwartbar war, als der Einsatz von Zitaten auf der sprachlichen Oberfläche recht eindeutig markiert wird. Das folgende Beispiel aus einem Schüler:innentext weist den zentralen Grund für die Diskrepanz innerhalb dieser Kategorie aus:

Daraufhin ist er dann zwei oder drei Wochen „in“, aber danach interessiert sich keiner mehr für ihn und er hat kein gutes Einkommen mehr, da er nicht so viele Klicks bekommt. Du hast aus dem Material zitiert, aber die Quellenangabe fehlt. Achte darauf, die Quelle unter Angabe der Quelle zu zitieren.
--

**Abb. 6:** Feedbackbeispiel II

Neben der zitierten Textstelle wird nur in der Einleitung mit Verweis auf die Quelle zitiert. Das KI-System vergibt sieben Punkte, während der menschliche Rater den Einsatz von Zitaten über den gesamten Text hinweg mit lediglich zwei Punkten beurteilt. Die vermeintliche Eindeutigkeit der Markierung von Zitaten auf der sprachlichen Oberfläche scheint bei dem Modell zu einer fehlergenerierenden Übergeneralisierung geführt zu haben. So zeigt dieses Beispiel, dass teilweise modalisierende An- und Abführungszeichen als Zitate „fehlinterpretiert“ wurden, was stark an die Urteile rein oberflächenorientierter Systeme erinnert, deren Beschränktheit große Sprachmodelle durch ihre hochdimensionalen Vektorräume und damit einen stärkeren bedeutungsorientierten Zugang zu überwinden versprechen.

Auch in dieser Kategorie zeigen sich wieder Konsistenzprobleme, da zweimal sechs Punkte mit der qualitativen Rückmeldung gepaart werden, dass überhaupt nicht aus dem Material zitiert wurde.

#### 4.4 Fragebogen

Abschließend wurden die Schüler:innen ( $N = 17$ ) zum Feedback durch die KI mithilfe eines Fragebogens befragt. Es werden an dieser Stelle lediglich wenige Auszüge präsentiert.

Elf Schüler:innen sind der Meinung, das Überarbeiten grundsätzlich habe ihren Text verbessert. Bis auf eine Testperson hatten damit auch alle recht. Das zeigt neben der guten Selbsteinschätzung dieser Schüler:innen, dass die Überarbeitung an sich schon wirksam sein kann, so sie denn angegangen wird. KI-Systeme könnten hier durch die Unmittelbarkeit des Feedbacks die Motivation steigern, diese Überarbeitung auch tatsächlich durchzuführen. Die qualitativen Rückmeldungen zum maschinellen Feedback fokussieren vor allem die teilweise fehlende Genauigkeit bzw. Richtigkeit des Feedbacks, aber auch den mangelnden persönlichen Kontakt, den auch Rüdian et al. (2025) in ihren

Daten (Meinungen von Lernenden zu automatisch generiertem Feedback) finden. Auch die mangelnde Möglichkeit der Nachfrage sowie die fehlende Korrektur im Text werden kritisiert. Folgendes Feedback fasst diesen Kritikpunkt zusammen:

Zum Beispiel bekommst du auf Fiete die Nachricht, dass du viele Rechtschreibfehler hast, während das Feedback von Lehrkräften dir in dem Text zeigt, wo du die Fehler gemacht hast. Das finde ich besser weil ich sehr lange nach meinen Fehlern gesucht habe und teilweise nicht gefunden habe.

**Abb. 7:** Schüler:innenfeedback I

In diesem Feedback zeigt sich eine wichtige Herausforderung von KI-Systemen. Denn für die Anbindung an den Text bzw. die konkreten Textstellen wäre im Grunde doch wieder eine Art von Korrektur im Text notwendig, die wiederum ein eher deterministisches Vorgehen benötigen würde. Hier sind die Entwickler:innen dieser Systeme gefordert, Lösungen zu erarbeiten, die es den Schüler:innen möglich machen, das Feedback konkret an bestimmte Textstellen rückzubinden (s. auch das *Lokalisationsproblem* in Fürstenberg 2025).

Durchaus erfreulich ist der Umstand, dass sich Schüler:innen der bereits diskutierten Fehler des KI-Systems bewusst waren und diese auch im Fragebogen monierten:

Generell war es schon ziemlich gut, aber die KI hat mir nicht sonderlich gefallen, da sie manches als falsch oder fehlend gekennzeichnet hat, was aber da war. Das macht einen dann unsicher ob das andere Feedback dann auch richtig ist.

**Abb. 8:** Schüler:innenfeedback II

Der Umstand, dass auch didaktisch systemgepromptete KI-Systeme eindeutige Fehler produzieren und auch quantitativ kaum mit menschlicher Bewertung übereinstimmen, zeigt überdeutlich, dass bei der Entwicklung dieser Modelle Expert:innen aus Gebieten wie der Didaktik, Psychologie, Linguistik und Informatik gemeinsam an Lösungen arbeiten müssten. Denn was die Schüler:in hier noch recht milde mit „nicht sonderlich gefallen“ umschreibt, hat wohl eine der größten Herausforderungen maschinell erstellten Feedbacks in Zukunft zur Folge: das Vertrauen in die KI-Systeme und deren Ausgaben. Entwickeln

Lehrkräfte und Schüler:innen kein Zutrauen in die Modelle, werden sie erst gar nicht eingesetzt oder ihr Feedback wird nicht ernst genommen. Gerade jetzt, in einer Zeit, in der Modelle schon breit rezipiert werden, aber womöglich noch nicht ausreichend gut entwickelt sind, ist eine überaus sensible Phase für dieses Vertrauensverhältnis. Bereits in diesem frühen Stadium das Vertrauen der betroffenen Personen (Schüler:innen und Lehrkräfte) in KI-generiertes Feedback zu verlieren, wäre fatal.

Das letzte Fragebogenitem erfragte die Vorteile KI-gestützten Feedbacks im Vergleich zu Feedback durch die Lehrkraft und vice versa. Tabelle 4 fasst die Ergebnisse zusammen:

<b>Vorteile Lehrerfeedback</b>	<b>N</b>	<b>Vorteile KI-Feedback</b>	<b>N</b>
Korrektheit	9	Unmittelbarkeit	8
Genauigkeit	9	Nichts	4
Rückfragemöglichkeit	2	Visualisierung	4
persönlich	2	Neutralität & Wiederholbarkeit	1

**Tab. 4:** Zusammenfassung Schüler:innenfeedback zu Vorteilen von Feedback durch Lehrkräfte/KI

Die Schüler:innen heben vor allem einen zentralen Vorteil des Feedbacks durch KI-Systeme hervor: die Unmittelbarkeit des Feedbacks. Dies wird auch auf lange Sicht der größte Vorteil automatisch generierten Feedbacks bleiben. Des Weiteren wird die Visualisierung bzw. Übersichtlichkeit der KI-generierten Rückmeldungen betont, womit auf die Balkenanzeigen (s. Abbildung 1) abgehoben wird, welche es den Schüler:innen theoretisch erleichtert, auf die Schnelle besonders gut oder schlecht erfüllte Kriterien zu erfassen. Vor dem Hintergrund der quantitativen Ergebnisse, die zeigen, dass die Balkenanzeige und das qualitative Feedback nicht immer konsistent sind, bleibt dieser zweite Vorteil noch fragwürdig zurück. Beim Vorteil des Feedbacks durch Lehrkräfte wird einerseits die Genauigkeit betont, womit die Schüler:innen die Rückbindung an ihren konkreten Text durch die Korrektur meinten. Andererseits wird die Korrektheit des Feedbacks durch Lehrkräfte von den Schüler:innen besonders betont. Dies ist in dieser Untersuchung durchaus auch als Reaktion auf die bereits thematisierten Fehler zurückzuführen, welche die KI-Systeme

zeigten, und verstärkt noch einmal die Relevanz von inhaltlich richtigem Feedback mit Blick auf die Vertrauenswürdigkeit des Feedbackgebenden, die von zentraler Bedeutung für die Feedbacknehmenden ist.

## 5 Fazit

Die meisten Schüler:innen konnten durch die Überarbeitung ihre Texte verbessern, was allerdings aus Gründen des Forschungsdesigns nicht eindeutig auf das Feedback der KI zurückzuführen ist. Der Vergleich von menschlicher und maschineller Bewertung deutet an, dass es zwischen dem Rater und der eingesetzten KI keine allzu hohen Übereinstimmungen bei kriteriengeleiteter Bewertung gibt. Allerdings muss auch dieses Ergebnis in weiteren Studien an einer breiteren Proband:innengruppe geprüft werden. Ein Problem des maschinellen Feedbacks liegt in der mangelnden Konsistenz, dass also qualitatives Feedback in Form von Text und die quantitative Bewertung durch das Sprachmodell nicht übereinstimmen. Hohe Unstimmigkeitswerte zwischen Mensch und Maschine sind meist auf fehlerhaftes Feedback durch die KI rückführbar, die jedoch neben den hier zitierten Stellen durchaus auch passendes Feedback produzierte, was die deskriptiven Ergebnisse zu den quantitativen Beurteilungen auch durchaus eindrücklich zeigen. In der Befragung loben die Schüler:innen die Unmittelbarkeit des maschinell erstellten Feedbacks. Ein großer Nachteil sind die inhaltlich falschen Feedbacks, welche die Schüler:innen zwar stellenweise als solche erkennen, aber dennoch bleibt es eine aktuell nicht überschreitbare Grenze des Einsatzes Künstlicher Intelligenz für das Feedback zu Schüler:innentexten. Denn der Umstand des inhaltlich falschen Feedbacks durch beispielsweise Halluzinationen bzw. Bullshit (Müller/Fürstenberg 2023; Hicks/Humphries/Slater 2024) ist ein architekturinhärentes Problem (Müller/Fürstenberg 2023: 335–341), das sich nur durch überaus aufwändiges Training oder Feintuning (Simbeck i. d. B.) einschränken, aber vorerst nicht gänzlich beheben lässt.

Die inhaltliche Qualität KI-generierten Feedbacks benötigt dringend weitere qualitative und fachdidaktisch fundierte Forschung, um den aufgezeigten

Herausforderungen zu begegnen und Lehrkräften weitere Möglichkeiten sowie Schüler:innen schnelles und lernförderliches Textfeedback zu ermöglichen.

## Literatur

- AVERINTSEVA-KLISCH, MARIA (2018): *Textkohärenz*. 2., aktualisierte Aufl. Heidelberg: Winter.
- BACHMANN, THOMAS/BECKER-MROTZEK, MICHAEL (2010): Schreibaufgaben situieren und profilieren. In: Pohl, Thorsten/Steinhoff, Torsten (Hg.): *Textformen als Lernformen*. Duisburg: Gilles & Francke, S, 191–209.
- BATZLEN, CHRISTIAN (13.1.2023): *ChatGPT: Das Ende der Hausarbeit?* URL: <https://www.swf.de/swr2/programm/chatgpt-in-der-uni-schreiben-sich-wissenschaftliche-arbeiten-bald-von-alleine-100.html> (letzter Zugriff: 7.2.2024).
- BENDER, EMILY M/GEBRU, TIMNIT/MCMILLAN-MAJOR, ANGELINA/SHMITCHELL, SHMARGARET (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada*, S. 610–623. <https://doi.org/10.1145/3442188.3445922>.
- BIRKEL, PETER/BIRKEL, CLAUDIA (2002): Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. In: *Psychologie in Erziehung und Unterricht* 49 (3), S. 219–224.
- BÖTTCHER, INGRID/WAGNER, MONIKA (1993): Kreative Texte bearbeiten. In: *Praxis Deutsch*, 20 (199), S. 24–27.
- CHIANG, CHENG-HAN/HUNG-YI LEE (2023): Can large language models be an alternative to human evaluations? In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, S. 15607–15631.
- DAX, SIMON (2023): Werteorientiertes Argumentieren im Deutschunterricht. In: *Mitteilungen des Deutschen Germanistenverbandes* (3), S. 254–269.
- EMMERICH, NADINE (25.1.2023): *ChatGPT in der Bildung. „Hausaufgaben sind tot“*. URL: <https://www.gew.de/aktuelles/detailseite/hausaufgaben-sind-tot> (letzter Zugriff: 15.2.2024).
- FEILKE, HELMUTH/TOPHINKE, DORIS (2017): Materialgestütztes Argumentieren. In: *Praxis Deutsch*. 44 (262), S. 4–13.

- FLECKENSTEIN, JOHANNA/MEYER, JENNIFER/JANSEN, THORBEN/KELLER, STEFAN/KÖLLER, OLAF/MÖLLER, JENS (2024): Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. In: *Computers and Education: Artificial Intelligence* (6), S. 100209. <https://doi.org/10.1016/j.caeai.2024.100209>.
- FÜRSTENBERG, MAURICE (2025): Zur Qualität von KI-generiertem Feedback – Ergebnisse einer explorativen Untersuchung. In: *Leseräume* (11).
- FÜRSTENBERG, MAURICE/MATZ, DANIELA (2025): Künstliche Intelligenz als Diskussionspartner? Materialgestütztes Argumentieren mit und reflektieren über KI. In: *Praxis Deutsch* (311), S. 23–30.
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* 76 (5), S. 2–13.
- HAVERKAMP, HENDRIK/HECHT, MALTE/SCHINDLER, KIRSTEN (2024): Lernförderliches Feedback KI-basiert vermitteln. In: *Der Deutschunterricht* 76 (5), S. 60–71.
- HEDDERICH, JÜRGEN/SACHS, LOTHAR (2020): *Angewandte Statistik Methodensammlung mit R*. 17., überarbeitete und ergänzte Aufl.. Berlin: Springer.
- HERBOLD, STEFFEN/HAUTLI-JANISZ, ANNETTE/HEUER, UTE/KIKTEVA, ZLATA/TRAUTSCH, ALEXANDER (2023): A large-scale comparison of human-written versus ChatGPT-generated essays. In: *Scientific Reports* 13 (1), S. 18617. <https://doi.org/10.1038/s41598-023-45644-9>.
- HICKS, MICHAEL/HUMPHRIES, JAMES/SLATER, JOE (2024): ChatGPT is bullshit. In: *Ethics and Information Technology* (26), S. 38. <https://doi.org/10.1007/s10676-024-09785-3>.
- JANSEN, THORBEN/HORBACH, ANDREA/MEYER, JENNIFER (2025): Feedback from Generative AI: Correlates of Student Engagement in Text Revision from 655 Classes from Primary and Secondary School. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, S. 831–836. <https://doi.org/10.1145/3706468.3706494>.
- MIZUMOTO, ATSUSHI/EGUCHI, MASAKI (2023): Exploring the potential of using an ai language model for automated essay scoring. In: *Research Methods in Applied Linguistics* 2 (2), S. 100050.
- MUBMANN, FRANK/HARDWIG, THOMAS/RIETHMÜLLER, MARTIN (2017): *Arbeitszeit und Arbeitsbelastung von Lehrkräften in Niedersachsen: Ergebnisbericht der*

- Arbeitsbelastungsstudie 2016*. Georg-August-Universität Göttingen, Kooperationsstelle Hochschulen und Gewerkschaften.
- MUBMANN, FRANK/HARDWIG, THOMAS/RIETHMÜLLER, MARTIN/KLÖTZER, STEFAN/PETERS, STEFAN (2020): *Arbeitszeit und Arbeitsbelastung von Lehrkräften an Frankfurter Schulen 2020: Ergebnisbericht*. Georg-August-Universität Göttingen, Kooperationsstelle Hochschulen und Gewerkschaften. <https://doi.org/10.3249/ugoe-publ-7>.
- MÜLLER, HANS-GEORG/FÜRSTENBERG, MAURICE (2023): Der Sprachgebrauchsmat. Sieben Thesen, die aus den technischen Grundlagen von GPT folgen. In: *Mitteilungen des Deutschen Germanistenverbandes* (4), S. 327–345.
- MÜLLER, NORA/UTESCH, TILL/BUSSE, VERA (2023): Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs- und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrationsbedingter Mehrsprachigkeit. In: *Unterrichtswissenschaft* (51), S. 169–198. <https://doi.org/10.1007/s42010-023-00173-2>.
- NAISMITH, BEN/MULCAIRE, PHOEBE/BURSTEIN, JILL (2023): Automated evaluation of written discourse coherence using gpt-4. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, S. 394–403.
- NEFF, TINA (2023): Einsatz digitaler Korrekturhilfen im Rechtschreibunterricht. Erste Erkenntnisse einer Pilotstudie in der Primarstufe und Sekundarstufe I. In: *Medien im Deutschunterricht* (1), S. 1–17.
- PADÓ, ULRIKE/ERYILMAZ, YUNUS/KIRSCHNER, LARISSA (2023): Short-Answer Grading for German: Addressing the Challenges. In: *International Journal of Artificial Intelligence in Education* 34 (4), S. 1321–1352. <https://doi.org/10.1007/s40593-023-00383-w>.
- PAGE, ELLIS (1966): The Imminence of... Grading Essays by Computer. In: *The Phi Delta Kappan* 47 (5), S. 238–243. <http://www.jstor.org/stable/20371545>.
- PHILIPP, MAIK (2015): *Schreibkompetenz. Komponenten, Sozialisation und Förderung*. Tübingen: A. Francke.
- PHILIPP, MAIK (2023): Formatives Feedback aus der Sicht des selbstregulierten Lernens. Grundlagen und Grundsätze förderlicher Rückmeldungen. In: *ide* 47 (2), S. 8–17.
- RAMESH, DADI/SANAMPUDI, SURESH (2022): An automated essay scoring systems: a systematic literature review. In: *Artificial Intelligence Review* 55 (3), S. 2495–2527.

- REICHARDT, ANKE/KRUSE, NORBERT/LIPOWSKY, FRANK (2014): Textüberarbeitung mit Schreibkonferenz oder Textlupe. Zum Einfluss der Schreibumgebung auf die Qualität von Schülertexten. In: *Didaktik Deutsch*. 19 (36), S. 65–85.
- RÜDIAN, SYLVIO/PODELO, JULIA/KUŽÍLEK, JAKUB/PINKWART, NIELS (2025): Feedback on Feedback: Student's Perceptions for Feedback from Teachers and Few-Shot LLMs. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, S. 82–92. <https://doi.org/10.1145/3706468.3706479>.
- SAWATZKI, JÖRG/SCHLIPPE, TIM/BENNER-WICKNER, MARIAN (2022): Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. Cheng, Eric/Koul, Rekha/Wang, Tianchong/Yu, Xinguo (Hg.): *Artificial Intelligence in Education: Emerging Technologies, Models and Applications. Proceedings of 2021 2nd International Conference on Artificial Intelligence in Education Technology* (Lecture Notes on Data Engineering and Communications Technologies: 104). Singapore: Springer, S. 65–75.
- SCHRÖTER, PAULINE/SÖLDNER, HANNELORE/HOFFMANN, LARS/RIEMENSCHNEIDER, ANJA/JOST, JÖRG/WIESER, DOROTHEE (2022): Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen. In: Schröter, Pauline/Groß, Alexander/Schmid-Kühn, Svenja/Stanat, Petra/Hoffmann, Lars (Hg.): *Das unvergleichliche Abitur: Entwicklungen-Herausforderungen-Empirische Analysen*. Bielefeld: wbv, S. 213–250.
- SCHWARZE, CORDULA (2016): Was ist ein gutes Argument? – Zu Analyse, Reflexion und Beurteilung mündlichen Argumentierens. In: Hinger, Barbara (Hg.): *Zweite „Tagung der Fachdidaktik“ 2015. Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht* (Innsbrucker Beiträge zur Fachdidaktik: 2). Innsbruck: innsbruck university press, S. 161–190. URL: [https://www.uibk.ac.at/iup/buch\\_pdfs/zweite-fachdidaktik/10.152033122-51-2.pdf](https://www.uibk.ac.at/iup/buch_pdfs/zweite-fachdidaktik/10.152033122-51-2.pdf) (letzter Zugriff: 5.2.2024).
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hg.) (KMK) (2005): *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4). Beschluss vom 15.10.2004*. URL: [https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen\\_beschluesse/2004/2004\\_10\\_15-Bildungsstandards-Deutsch-Primar.pdf](https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Deutsch-Primar.pdf) (letzter Zugriff: 5.2.2024).
- Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hg.) (KMK) (2014): *Bildungsstandards im Fach*

*Deutsch für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012)*. URL: [https://www.kmk.org/fileadmin/veroeffentlichungen\\_beschluesse/2012/2012\\_10\\_18-Bildungsstandards-Deutsch-Abi.pdf](https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf) (letzter Zugriff: 5.2.2024).

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hg.) (KMK) (2022): *Bildungsstandards für das Fach Deutsch Primarbereich (Beschluss der Kultusministerkonferenz vom 15.10.2004, i. d. F. vom 23.06.2022)*. URL: [https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen\\_beschluesse/2022/2022\\_06\\_23-Bista-Primarbereich-Deutsch.pdf](https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2022/2022_06_23-Bista-Primarbereich-Deutsch.pdf) (letzter Zugriff: 5.2.2024).

SEBLER, KATHRIN/FÜRSTENBERG, MAURICE/BÜHLER, BABETTE/KASNECI, ENKELEJDA (2025): Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, S. 462–472. <https://doi.org/10.1145/3706468.3706527>.

SPITTA, GUDRUN (1992): *Schreibkonferenzen in Klasse 3 und 4. Ein Weg vom spontanen Schreiben zum bewussten Verfassen von Texten*. Frankfurt, Main: Cornelsen Scriptor.

STAHL, MAJA/BIERMANN, LEON/NEHRING, ANDREAS/WACHSMUTH, HENNIG (2024): Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In: *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. S. 283–298. URL: <https://aclanthology.org/2024.bea-1.23.pdf> (letzter Zugriff: 13.02.2025).

STURM, AFRA/SCHNEIDER, HANSJAKOB (2021): Flüssiges Formulieren in der Textproduktion (Klasse 4/5). In: *Didaktik Deutsch* 26 (51), S. 28–49.

WENDT, CHARLOTTE (2023): Schreiben lernen mit intelligenter Hilfe. Wie computergestütztes Feedback Schreiblernprozesse verändern kann. In: *ide* (2), S. 38–47.

## **Schreiben mit KI als didaktische Herausforderung**

Empirische Einblicke in Prompting-Praktiken von Lernenden

### **Abstract**

Das Schreiben mit textgenerativer KI stellt die Schreibdidaktik vor neue Herausforderungen. Eine zentrale Rolle nimmt dabei das Prompting ein – die Formulierung von Eingaben zur Steuerung der KI. Basierend auf empirischen Daten untersucht dieser Beitrag, wie Schüler:innen und Studierende beim materialgestützten Schreiben mit KI Prompts formulieren, welche Herausforderungen dabei auftreten und welche didaktischen Forderungen sich daraus ergeben. Die Studie kombiniert eine quantitative sowie qualitative Inhaltsanalyse der erfassten Prompts und zeigt, dass Lernende vor allem Befehle zur Texterstellung und oberflächlichen Überarbeitung von Texten nutzen, während Funktionen bspw. für tiefere Umstrukturierungen oder kritische Bewertungen kaum genutzt wurden. Teilweise lassen sich Kommunikationsmuster aus Messenger-Diensten vermuten, was auf fehlende Konzepte zur gezielten Steuerung der KI hinweist. Der Beitrag versteht Prompting als metatextuelle Handlung, deren Ausübung spezifische Fähigkeiten voraussetzt, und diskutiert, wie diese durch didaktische Impulse gestärkt werden können.

**Schlagwörter:** Prompting, Schreibdidaktik, KI-gestütztes Schreiben, Schreibpraktiken, Materialgestütztes Schreiben

# 1 Einleitung

„Je besser der Prompt wahrscheinlich ist, desto besser wird auch der Text, der von der KI rauskommt“ (Sch\_GS11\_ndx)<sup>1</sup>. So beschreibt ein Schüler der 11. Klasse, wie zentral aus seiner Sicht die Formulierung von Prompts für das Schreiben mit textgenerativer KI ist. Prompting bezeichnet den Vorgang, bei dem eine Person der KI eine Eingabe gibt, um die KI zu steuern und Ausgaben zu erzeugen. Larsen et al. (2022) definieren Prompts als „Anforderungen an das gewählte Modell in natürlicher Sprache“, womit sie eine zentrale Schnittstelle zwischen Nutzenden und KI darstellen. Doch das Prompting ist der Aussage des Schülers nach nicht nur eine technische Notwendigkeit, sondern auch eine anspruchsvolle Tätigkeit. Der Schüler ergänzt, „dass die KI nie wirklich so sein wird, wie man selber das schreiben würde, dass man das einfach umändern muss und nicht einfach den ganzen KI-Text einfügen sollte“ (Sch\_GS11\_ndx). Diese Aussagen verdeutlichen das komplexe Spannungsfeld zwischen automatisierter Textproduktion und menschlicher Teilhabe am Schreibprozess. Ähnlich formuliert auch eine Mitschülerin, „dass [es] auf jeden Fall schlau ist, wirklich sehr, sehr detailliert, sich halt mehr Zeit zu nehmen für diesen Prompt [...], einfach weil dann auch die Antwort, die man kriegt, genauer ist und mehr, wie man sich das vorstellt“ (Sch\_GS11\_txm). Die Aussagen der Lernenden zeigen, dass sie dem Prompten beim Schreiben mit KI eine besondere Bedeutung beimessen. Das erfolgreiche Hervorbringen von Output für den späteren Text hängt für sie offensichtlich eng mit der Qualität der Textarbeit am Prompt zusammen.

Steinhoff (2023: 13) beschreibt Prompting als „neues, interessantes und didaktisch relevantes Gegenstandsfeld“<sup>2</sup>. Gerade im Kontext des schulischen

.....

1 Die verwendeten Codes dienen der Anonymisierung der Teilnehmenden und ermöglichen eine klare Zuordnung zu den Erhebungsgruppen. Das Schema lautet: [Teilnehmergruppe]\_[Zugehörigkeit]\_[Individuelle ID]. Beispielsweise steht „Sch\_GS11\_ndx“ für eine:n Schüler:in der 11. Klasse einer Gesamtschule, während „St\_SoSe24\_chu“ eine:n Studierende:n des Sommersemesters 2024 bezeichnet.

2 Dabei ist jedoch darauf hinzuweisen, dass um die langfristige didaktische Relevanz von Prompt-Kompetenzen eine kritische Diskussion geführt wird, wie beispielsweise von Fürstenberg/Müller (2024) vertreten und ausführlicher in Kapitel 2.3 dargestellt.

Schreibens führt die Einführung von textgenerativer KI in Schreibprozesse zu einer tiefgreifenden Veränderung traditioneller Aufgaben in der Textproduktion und fordert eine Neubewertung dessen, was im Bildungs- und Hochschulkontext unter „Schreiben“ verstanden wird: Wie beeinflusst die neuartige Zusammenarbeit zwischen Mensch und Maschine unser Verständnis von Schreiben? Kann die Textproduktion mit textgenerativer KI noch als „Schreiben“ bezeichnet werden? Und welche Konsequenzen hat das für den Schreibunterricht und die Vermittlung von Schreibkompetenz in der Schule?

Der vorliegende Beitrag geht der Frage nach, wie Schüler:innen textgenerative KI beim materialgestützten Schreiben nutzen und welche Herausforderungen sich dabei zeigen. Dafür werden zunächst theoretische Perspektiven auf den Wandel des Schreibbegriffs und die Rolle von Prompting kritisch diskutiert (Kap. 2). Anschließend werden empirische Erkenntnisse aus dem Forschungsvorhaben „Postdigitale Schreibpraktiken: Eine empirische Studie zum Einsatz textgenerativer KI bei materialgestützten Schreibaufgaben“ (Ansari, i. V.) präsentiert. Hier werden konkrete Muster und Herausforderungen in den Schreibpraktiken der Lernenden untersucht (Kap. 3), bevor abschließend didaktische Implikationen für das materialgestützte Schreiben mit KI abgeleitet werden (Kap. 4).

## **2 Schreibpraktik mit textgenerativer KI**

### **2.1 Wandel der Schreibtheorie**

In der traditionellen Schreibforschung wurde Schreiben lange als ein kognitiver Problemlöseprozess betrachtet (vgl. Bachmann/Becker-Mrotzek 2017: 25). Dabei stand die Idee im Vordergrund, dass ein Mensch durch Denken und Formulieren einen Text erstellt. Hayes und Flower (1980) etablierten mit ihrem Schreibprozess-Modell dieses kognitionspsychologische Verständnis, in dem sie den Schreibprozess in drei Teilprozesse unterteilen: Planen, Formulieren und Überarbeiten. Dieses Modell wurde u. a. für seine Vereinfachung kritisiert, das den komplexen und nichtlinearen Charakter des Schreibens nur

unzureichend abbilde (vgl. Perrin 2016: 432). Ihm folgten eine Reihe weiterer Modellierungsansätze (z. B. Hayes 1996, 2012).

Bachmann und Becker-Mrotzek (2017) definieren Schreiben als eine „spezifische sprachliche Handlung, die im Kern auf die selbstständige Produktion von kommunikativ angemessenen und inhaltlich bedeutungsvollen Texten abzielt“ (ebd.: 25). Diese Definition betont die kommunikative Funktion des Schreibens und fokussiert die zielgerichtete Tätigkeit des Menschen. Perrin (2016: 432) geht über diese Sichtweise hinaus und beschreibt Schreiben als „alle Tätigkeiten innerhalb eines Textproduktionsprozesses“. Er versteht Textproduktion als „das Gesamte der Tätigkeiten zur Erstellung eines (multisemiotischen) Textes“ (ebd.), die auf schriftsprachliche Kommunikationsangebote abzielen (vgl. auch Jakobs/Perrin 2014: 7).

In diesem Verständnis ist Schreiben mehr als ein rein kognitiver Problemlöseprozess. Es lässt sich als Textproduktionspraktik begreifen, die mentale, materielle, mediale, semiotische und soziale Tätigkeiten miteinander verbindet: Schreibpraktiken „umfassen, einfach gesagt, den denkenden Kopf, den tippenden Körper am vernetzten Computer am Arbeitsplatz, die sich verändernden Zeichen am Bildschirm – sowie die durch den Zeichengebrauch verbundenen Gemeinschaften“ (Perrin 2016: 433).

Perrin sieht Schreiben als routinisierte und gleichzeitig dynamische Tätigkeit, die sich durch reflektierte Anpassung an sich verändernde Umwelten entwickelt. In einem ähnlichen Ansatz beschreibt Steinhoff (2022) Schreibpraktiken anhand von sechs Facetten: technisch-medial, physisch, kognitiv, sozial, semiotisch und textuell-diskursiv. Er plädiert dafür, die Aufmerksamkeit beim Schreiben auf alle Entitäten zu lenken, die beim Schreiben involviert sind. Dies gilt insbesondere für die beim Schreiben integrierte Technologie.

Der Wandel des Schreibbegriffs umfasst somit nicht nur die Erweiterung um die Perspektive als multimodale und dynamische Praktik, sondern auch die Einbeziehung soziotechnischer Aspekte, die in früheren Modellen wie dem von Hayes und Flower (1980) noch keinen Eingang fanden. Auch andere Ansätze, wie das Schreibkompetenzmodell von Becker-Mrotzek/Schindler (2007), reduzieren das Schreibmedium auf eine speichernde Funktion und koppeln es vor allem an motorische Fähigkeiten (vgl. ebd.: 12). Diese instrumentelle

Perspektive wird von Steinhoff (2023) kritisiert, der Schreibtechnologien als aktive „Partizipanden“ (ebd.: 3) im Schreibprozess versteht.

Zusammenfassend lässt sich festhalten, dass sich das Verständnis des Schreibbegriffs von einem eher linearen, kognitiven Problemlöseprozess hin zu einem Verständnis als komplexe, dynamische, kommunikativ eingebettete und soziotechnische Praktik entwickelt hat.

## 2.2 Prompting als eigenständige Schreibpraktik

Mit der zunehmenden Digitalisierung des Schreibens rücken also die Technologien in den Fokus, die den Schreibprozess nicht nur unterstützen, sondern aktiv mitgestalten. Textgenerative KI markiert dabei eine neue Dimension: Sie kann nicht nur Rechtschreibkorrekturen und Textformatierungen liefern, sondern auch eigenständig Texte generieren. Die zentrale Frage ist daher nicht mehr nur, wie Menschen schreiben, sondern auch, wie sie mit Maschinen schreiben. Während frühere Schreibtechnologien vor allem als Werkzeuge verstanden wurden, die den Schreibprozess erleichtern oder beschleunigen, verändert KI-gestütztes Schreiben die Rollenverteilung grundlegend. Der Mensch agiert nun in einem dynamischen Interaktionsverhältnis mit der KI. Steinhoff (2023: 10) beschreibt dies als Koaktivität – eine wechselseitige Einflussnahme, bei der sowohl Mensch als auch Maschine in unterschiedlichem Maß am Schreibprozess beteiligt sind. Zudem verweist Steinhoff auf den Begriff der „Gebrauchssuggestion“ aus Hirschauers (2016: 52) Praxeologie, der besagt, dass Dinge gewisse Handlungen von Menschen „präfigurieren und disponieren“ (ebd.). In Bezug auf textgenerative KI wie ChatGPT deutet Steinhoff dies folgendermaßen: „ChatGPT etwa suggeriert dem Menschen keinen kompletten Schreibprozess (Planen, Formulieren, Überarbeiten), sondern ‚Prä-Post-Koaktivitäten‘, also Schreibhandlungen vor oder nach der Content-Generierung“ (Steinhoff, 2023: 12). Wie genau Prä- und Post-Koaktivitäten des Menschen in diesem Schreibprozess aussehen können, ist noch weitestgehend offen. Eine zentrale Koaktivität des Menschen im KI-gestützten Schreiben ist die Steuerung der KI durch Prompting, also die Eingabe eines Textes, der die Maschine einen Text produzieren lässt.

Im Gegensatz zu anderen Technologien – wie z. B. dem Taschenrechner – ist das Prompting im Schreibprozess durch eine besondere Opazität geprägt: Die Eingaben werden in einem proprietären, also geschlossenen System verarbeitet, das Nutzende nicht einsehen und kaum nachvollziehen können, wie der generierte Text entsteht (vgl. Bajohr 2023: 272). Zudem generiert die KI bei jeder Anfrage unter Variation ähnlich wahrscheinlicher Wörter einen neuen Text, sodass auch bei gleichem Prompting immer ein anderer Output entsteht (vgl. Müller/Fürstenberg 2023). Dies bedeutet, dass Schreibende sich teilweise auf experimentelles Vorgehen und rekursives Verfeinern ihrer Prompts verlassen müssen, um gewünschte Ergebnisse zu erzielen. Obwohl systeminterne Faktoren wie bspw. auch Trainingsdaten und Sampling-Strategien die Generierung prägen, stellt das Prompting eine notwendige Bedingung für die Output-Generierung dar und beeinflusst, welche Inhalte generiert, wie sie strukturiert und in welcher sprachlichen Qualität sie ausgegeben werden.

### 2.3 Prompting in der schulischen Praxis

Für den schulischen Kontext stellt sich die Frage, welche Rolle das gezielte Formulieren von Prompts für Lernende tatsächlich spielt. So stellen Fürstenberg/Müller (2024) die langfristige Relevanz von Prompting-Kompetenzen infrage. Sie argumentieren, dass sich durch die rasante technische Entwicklung die Notwendigkeit, Prompts bewusst und strategisch zu formulieren, zunehmend verringern könnte. Außerdem modifizieren im Hintergrund sogenannte System-Prompts die Nutzereingaben, ohne dass dies für die Schreibenden ersichtlich ist: „Es erscheint daher durchaus denkbar, dass künftige GPTs so gut auf menschliche Eingaben vorbereitet werden, dass Promptkompetenz in Zukunft nicht die Relevanz haben wird, die ihr in der aktuellen Debatte zugesprochen wird“ (ebd.: 8).

Unabhängig von der langfristigen Perspektive zeigen erste empirische Untersuchungen, dass das gezielte Formulieren von Prompts für Lernende aktuell eine anspruchsvolle Fähigkeit darstellt. Trotz unterschiedlicher Schwerpunkte in der Lese- und Schreibdidaktik kommen sowohl Buck et al. (2023) als auch Führer/Nix (2024) in ihren explorativen Erhebungen zu einem ähnlichen Ergebnis: Schüler:innen nutzten die textgenerative KI in häufig isolierten An-

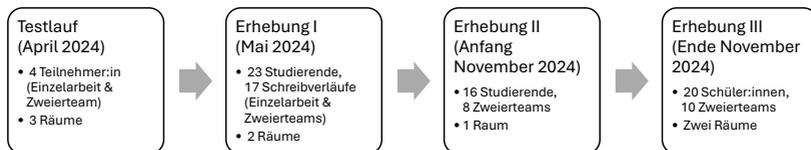
fragen teilweise unreflektiert als eine Art „Suchmaschine“ (Buck et al. 2023: 67; Führer/Nix 2024: 362), ohne das interaktive Potenzial beim Schreiben mit textgenerativer KI auszuschöpfen. Dieses begrenzte Nutzungsverhalten steht im Kontrast zu dem, was Steinhoff/Lehnen (2025: 9) als *Prompt Chaining* beschreiben – „die Fähigkeit, Prompts stufenartig entlang des Outputs immer weiter einzugrenzen und nachzuschärfen, um zu dem gewünschten Ergebnis zu kommen“ (ebd.), wodurch Nutzende eine höhere Kontrolle über den generierten Text gewinnen können. Die Ergebnisse unterstreichen somit die Notwendigkeit didaktischer Impulse, um Schüler:innen gezielt darin zu unterstützen, die KI als interaktive Schreibpartnerin zu nutzen (vgl. Führer/Nix 2024: 363). Diese Herausforderung resultiert nicht nur aus allgemeinen Anforderungen an präzise Anweisungen – wie sie bspw. in Anleitungen gefordert sind –, sondern auch aus der Notwendigkeit, die Eingaben an die spezifische Funktionsweise textgenerativer KI anzupassen. Vor diesem Hintergrund kann das Prompting als eine Art didaktisch relevanter Metatext verstanden werden: ein Text zur Erstellung von Texten.

Das sogenannte *Prompt Scaffolding*, also die Unterstützung von Lernenden bei der Formulierung zielführender Prompts, ist ein Konzept, das im deutschsprachigen Raum bislang kaum untersucht wurde. Jang et al. (2024) führten dazu eine quasi-experimentelle Studie mit 179 koreanischen Oberstufenschüler:innen durch, um den Einfluss von Prompt Scaffolding auf die Qualität argumentativer Texte zu untersuchen. Die Ergebnisse legen nahe, dass eine gezielte Anleitung dazu beitragen kann, dass Lernende kohärentere und argumentativ stärkere Texte schreiben. Zudem deutet die Untersuchung darauf hin, dass eine systematische Schulung im Erstellen zielführender Prompts das Verständnis für argumentative Strukturen vertiefen und die Reflexion über eigene Texte unterstützen kann. Diese Perspektive eröffnet interessante didaktische Ansätze, da sich das Verfassen von Prompts sowohl als eigenständige Schreibkompetenz als auch als Reflexionsinstrument über Textsorten oder Textmusterwissen nutzen lassen könnte.

### 3 Empirische Studie zum Prompting

Die dargestellte empirische Forschung legt nahe, dass Prompting für Lernende eine anspruchsvolle Fähigkeit darstellt, die gezielt didaktisch begleitet werden sollte. Vor diesem Hintergrund setzt die vorliegende Untersuchung an, um systematisch zu analysieren, welche Muster und Herausforderungen sich in den von Lernenden formulierten Prompts identifizieren lassen und welche Implikationen sich daraus für die Schreibdidaktik ergeben. Das Forschungsvorhaben, aus dem im Folgenden Ergebnisse präsentiert werden, untersucht die Schreibpraktiken von Schüler:innen im Kontext des materialgestützten Schreibens mit textgenerativer KI. Um methodische und inhaltliche Erkenntnisse zu gewinnen, wurde das Forschungsdesign zunächst mit Studierenden erprobt. Diese ersten Erhebungen dienten dazu, die Machbarkeit des Settings zu testen und Anpassungen vorzunehmen. Das zirkuläre Forschungsdesign ermöglichte es dabei, das Untersuchungssetting iterativ zu optimieren und praxisnahe Bedingungen für die Hauptuntersuchung mit den Schüler:innen zu schaffen.

Die Untersuchung bestand aus mehreren Erhebungsphasen, die schrittweise weiterentwickelt wurden. Alle Teilnehmenden hatten jeweils 80 bis 85 Minuten Zeit für die Bearbeitung der Schreibaufgaben. Die meisten Teilnehmenden nutzten ihre eigenen iPads, während in wenigen Fällen auch Laptops verwendet wurden. Die Nutzung von textgenerativer KI war den Teilnehmenden nach einer 20-minütigen Lesezeit freigestellt. Ziel war es, herauszufinden, wie und wofür die Teilnehmenden die textgenerative KI einsetzen. Während in der ersten Erhebung ChatGPT und ein Texteditor der Plattform Ilias genutzt wurden, kam in der zweiten und dritten Erhebung die datenschutzkonforme Plattform Fobizz zum Einsatz. Diese bietet zwar ebenfalls Zugang zu ChatGPT, anonymisiert jedoch die Anfragen und stellt zugleich einen integrierten Texteditor bereit.



**Abb. 1:** Zeitliche Abfolge und Organisation der Erhebungen

### 3.1 Forschungsdesign

Im April 2024 wurde ein erster Testlauf mit vier Teilnehmenden durchgeführt, die primär dem Testen technischer Abläufe und des Schreibarrangements dienete und deren Ergebnisse hier nicht weiter vertieft werden. Jedoch zeigte sich bei dieser Umsetzung, dass Unsicherheiten im Umgang mit der KI-Schreibumgebung sowohl während des Schreibprozesses als auch in den anschließenden Reflexionsgesprächen eine zentrale Herausforderung darstellten. Um diesen Schwierigkeiten zu begegnen, wurde für die nachfolgenden Erhebungen eine 90-minütige Input-Sitzung zum Schreiben mit KI eingeführt, die immer weiter angepasst wurde. Konkret wurden zwei zentrale Ziele verfolgt: Erstens wurden die Funktionsweise von KI und ihre eingeschränkte Zuverlässigkeit thematisiert, um ein grundlegendes Verständnis für die technischen Mechanismen und deren Auswirkungen auf die Textproduktion zu schaffen (vgl. Buck et al. 2023: 73). Zweitens erhielten die Teilnehmenden die Möglichkeit, sich in einer Vorübung mit der Schreibumgebung vertraut zu machen, bevor sie ihre eigentliche Schreibaufgabe begannen. Dies sollte insbesondere die technischen Hürden abbauen.



**Abb. 2:** Schreibsituation während Erhebung III: Schüler:innen arbeiten in Zweierteams mit Tablets

Auf dieser Grundlage wurde im Mai 2024 eine erste explorative Erhebung mit 23 Studierenden durchgeführt. Die Teilnehmenden sollten – teilweise in Einzelarbeit, teilweise in Zweiertteams – eine materialgestützte Schreibaufgabe zum Thema „Warnhinweise vor potenziell verstörender Literatur“ bearbeiten. Kritisiert wurden an diesem Arrangement die fehlende Zielgruppenorientierung und eine unrealistische Schreibsituation.

Basierend auf diesen Erkenntnissen wurde das Schreibarrangement überarbeitet und in einer zweiten Erhebung mit 16 Studierenden Anfang November 2024 erneut getestet. Die überarbeitete Aufgabe bestand darin, einen argumentativen Kommentar für die Rubrik „Lehrkräfte diskutieren“ einer regionalen Zeitung zu verfassen. Um den kollaborativen Schreibprozess gezielter zu untersuchen, wurde diese Erhebung ausschließlich in Zweiertteams durchgeführt. Das Thema lautete: „Sollten soziale Medien für Jugendliche stärker reguliert werden?“ Die Materialien umfassten journalistische Beiträge, die unterschiedliche Perspektiven auf das Thema lieferten. Diese Aufgabe wurde gewählt, um ein anspruchsvolles Schreibarrangement zu schaffen, das an die Anforderungen des materialgestützten Schreibens im Abitur angelehnt ist. Zur strukturierten Bearbeitung wurden im Anschluss eine Schreibkonferenz sowie Gruppendiskussionen durchgeführt, in der die Studierenden ihre Argumentationen reflektierten. Diese Gespräche fanden ohne externe Moderation statt. Die dritte Erhebung fand mit 20 Schüler:innen einer 11. Klasse Ende November 2024 statt. Der Aufbau der Erhebung war in der Grundstruktur identisch mit der zweiten Erhebung bei den Studierenden – dieses Mal sollte aber für die Rubrik „Jugend diskutiert“ geschrieben werden.

Die Datenerhebung umfasste zahlreiche Daten, darunter Screen- und Audioaufnahmen der Schreibprozesse, erstellte Texte, Fragebögen sowie Gruppendiskussionen. Im Folgenden wird die Analyse der Prompts in den Fokus gestellt, die im Januar 2025 von mir<sup>3</sup> kodiert wurden.

.....

3 Durch wiederholte Vergleichscodierungen in der Auswertungsphase, theoriegestützte Rückbindungen und die Orientierung an einem konsistenten Kategoriensystem wurde ein hohes Maß an Interpretationsicherheit angestrebt. Dennoch ist nicht auszuschließen, dass subjektive Wahrnehmungen die Analyse beeinflusst haben (vgl. Reichertz 2016, S. 78–81).

## 3.2 Analytisches Vorgehen

Die Analyse der Prompts folgte einem deduktiv-induktiven Ansatz der qualitativen Inhaltsanalyse. Dabei wurde ein vorhandenes theoretisches Konzept von Tekin (2023) zur Systematisierung der Rollen von KI genutzt und durch eine induktive Analyse der empirischen Daten angepasst und ergänzt. Ziel war es, die in den Prompts der Lernenden zum Ausdruck kommenden Funktionen systematisch zu erfassen und zu klassifizieren.

Tekin (2023: 140–157) unterscheidet vier zentrale Rollen, die textgenerative KI im Schreibprozess einnehmen kann:

- Generator: Die KI wird genutzt, um neue Texte zu erzeugen.
- Transformator: Bestehende Texte werden durch die KI modifiziert, z. B. durch Zusammenfassungen, sprachliche Anpassungen, Übersetzungen oder Stilvariationen.
- Evaluator: Die KI übernimmt eine analytische Funktion, indem sie Texte hinsichtlich z. B. sprachlicher Korrektheit oder Verständlichkeit bewertet und Rückmeldungen gibt.
- Kommunikator: Die KI wird als interaktiver Partner genutzt, der Fragen beantwortet, Informationen bereitstellt oder dialogische Schreibprozesse unterstützt. Dazu zählt laut Tekin (2023) das Recherchieren von Informationen: „Was sind die fünf wichtigsten Feiertage in Deutschland?“ (ebd.: 152) und das Schreiben von E-Mails ähnlich einem Sekretariat (vgl. ebd.: 156).

Während Tekin die Rollen der KI aus der Perspektive von Lehrenden im Fremdsprachenunterricht beschreibt, wurde das System hier angepasst, um die Funktionen der Prompts zu codieren. Die vier Hauptkategorien wurden beibehalten, jedoch konzeptionell auf die Analyse der Funktion der Prompts ausgerichtet: *Generieren*, *Transformieren*, *Evaluieren* und *Kommunizieren*.

Der Code *Kommunizieren* wurde dabei anders als bei Tekin definiert. Tekin fasst unter dieser Kategorie auch bspw. informationsgenerierende Prompts oder solche, die das Umformen von Textsorten betreffen (Tekin 2023: 152–157). Dies führt jedoch zu einer Überschneidung mit den Codes *Generieren*

und *Transformieren*, da bspw. Informationsabfragen oft neue Inhalte erzeugen oder Sekretariatsaufgaben bestehende Texte umformulieren. In dieser Analyse wurde *Kommunizieren* daher enger gefasst und ausschließlich auf explizit dialogische Interaktionen zwischen Mensch und KI beschränkt. Dazu gehören z. B. Begrüßungen, metakommunikative Strategien wie Rollenzuweisungen („Agiere wie ein ...“) oder Feedback zu KI-generierten Texten („Das war sehr hilfreich“). Informationsabfragen oder Textumstrukturierungen wurden hingegen den entsprechenden anderen Kategorien zugeordnet.

Da in den Daten dennoch Prompts beobachtet wurden, die sich nicht eindeutig einer der Hauptkategorien zuordnen ließen, wurde zusätzlich die Kategorie *nicht zuordenbar* eingeführt. Diese umfasst Prompts, deren Funktion nicht klar erkennbar waren. Mit fortschreitender Auswertung wurde deutlich, dass ein Teil der zunächst schwer zuzuordnenden Prompts einem wiederkehrenden Muster folgte, das schließlich unter dem Code *Splitting* systematisch erfasst wurde. Der identifizierte Code *Splitting* beschreibt ein spezifisches Prompting-Verhalten, bei dem Lernende größere Inhalte in kleinere Abschnitte zerlegten und diese schrittweise an die KI übermittelten – oft ohne eine explizite Aufgabenstellung in jedem einzelnen Prompt zu formulieren. Die Praktik des *Splittings* wurde bereits von Beißwenger (2015: 287–289) im Kontext von Chatkommunikation beschrieben. Er definiert es als die Praxis, größere Einheiten auf mehrere kleinere Nachrichten aufzuteilen, wodurch der Sender ein temporäres exklusives Senderecht erlangt. Während Beißwenger (2015) die Chatpraktik des *Splitting* primär in Messenger-Diensten beobachtete, lassen sich deutliche Parallelen zum Prompting-Verhalten in der vorliegenden Untersuchung erkennen. Diese Kommunikationsstruktur erschwerte zunächst die Zuordnung einer klaren Funktion, da sie erst in Verbindung mit den vorhergehenden oder nachfolgenden Prompts erkennbar wurde. Mit der Einführung der Kategorie wurde der jeweils erste Prompt mit der Funktion codiert und dem darauffolgenden Prompt die Funktion des *Splittings* zugewiesen.

Während der Kodierung der Prompts wurde deutlich, dass vorgegebene Materialien und formulierte Textbestandteile häufig in die Prompts integriert wurden, dabei jedoch keiner der bestehenden Funktionskategorien eindeutig zugeordnet werden konnten. Um diesen Aspekt unabhängig von der funktionalen Einordnung der Prompts zu erfassen, wurde eine zusätzliche Dimension

in der Analyse eingeführt: der Code *Material und Textprodukt*. Diese Kategorie teilt sich in die Subcodes *Material* (unveränderte Übernahmen vorgegebener Materialien) und *Textprodukt* (Entsprechungen des zu erstellenden Textendprodukt ‚Kommentar‘). Die Einführung dieser zusätzlichen Codes ermöglicht es, die Einbindung von Materialien und bereits formulierten Textbausteinen in die Prompting-Praktiken systematischer zu beschreiben, ohne sie fälschlicherweise einer funktionalen Kategorie zuzuordnen. So lässt sich differenzierter analysieren, in welchem Umfang Lernende z. B. auf vorhandene Texte zurückgreifen und wie sie diese in ihren Interaktionen mit der KI nutzen.

### 3.3 Ergebnisse der Promptanalyse

Die Analyse der Prompts zeigt eine große Bandbreite an Nutzungsverhalten der Teilnehmenden. Tabelle 1 gibt einen Überblick über die Anzahl der Chatverläufe, Prompts und Kodierungen in den verschiedenen Erhebungen.

Erhebung	Erhebung I (N=23 Studierende)	Erhebung II (N=16 Studierende)	Erhebung III (N=20 Schüler:innen)
<b>Anzahl der Chatverläufe</b>	17	8	10
<b>Anzahl der Prompts insgesamt</b>	143	56	63
<b>Anzahl der Kodierungen insgesamt</b>	152	57	72
<b>Durchschnittliche Anzahl der Prompts pro Verlauf</b>	8,4	7	6,4
<b>Niedrigste und höchste Anzahl an Prompts pro Verlauf</b>	2–25	3–17	0–25

**Tab. 1:** Anzahl der Chatverläufe, Prompts und Kodierungen in den Erhebungen

Die durchschnittliche Anzahl an Prompts pro Chatverlauf ist in allen drei Erhebungen ähnlich (zwischen 6,4 und 8,4). Gleichzeitig zeigt sich in allen Erhebungen eine große Spannweite innerhalb der Gruppen: Während einige Teams nur wenige oder gar keine Prompts nutzten, griffen andere bis zu 25-mal auf die KI zurück. In Erhebung III verzichtete eine Gruppe vollständig auf den Einsatz der KI. In der abschließenden Gruppendiskussion begründeten sie dies durch „Paranoia“ (Sch\_GS\_11\_cm9) vor dem Abhören durch KI und fehlender Kontrolle beim Schreiben. Die großen Unterschiede innerhalb der Gruppen deuten darauf hin, dass die Entscheidung zur KI-Nutzung stark von individuellen Faktoren geprägt ist.

Da ein Prompt inhaltlich mehrere Zwecke erfüllen kann, wurden insgesamt 19 Prompts doppelt codiert. Das erklärt, warum in jeder Erhebung die Zahl der Kodierungen höher ist als die Zahl der Prompts. Ein Beispiel hierfür ist der folgende Prompt:<sup>4</sup>

Verfasse einen Kommentar für einen Zeitungsartikel für die Rubrik „Jugend diskutiert“ zu den Themen : „Sollte die Nutzung sozialer Medien bei Jugendlichen reguliert werden?“ und „welche vor- und Nachteile haben strengere Regulierung sozialer Medien für Jugendliche?“, der sich an die breite Leserschaft, einschließlich Jugendlicher und deren Eltern richtet. Argumentiere gegen die Regulierung, schreibe 400 bis 600 Wörter und beziehe dich auf die folgenden Argumente im Ping Pong Prinzip: „Treffen mit Freunden und Nachbarn oder gemeinsame Zeit mit dem Partner sind dagegen seltener geworden“ (S. 4, Z. 8–9) „Spiele fördern oft direkte Kommunikation“ (S. 2, Z. 10) [...] (Sch\_GS11\_kma & Sch\_GS11\_xmf)

Der Prompt fordert einerseits dazu auf, einen neuen Text zu formulieren. Gleichzeitig folgen stichpunktartig Argumente und Ausschnitte aus dem Material, die zu dem neuen Text transformiert werden sollen. Daher wurde dem

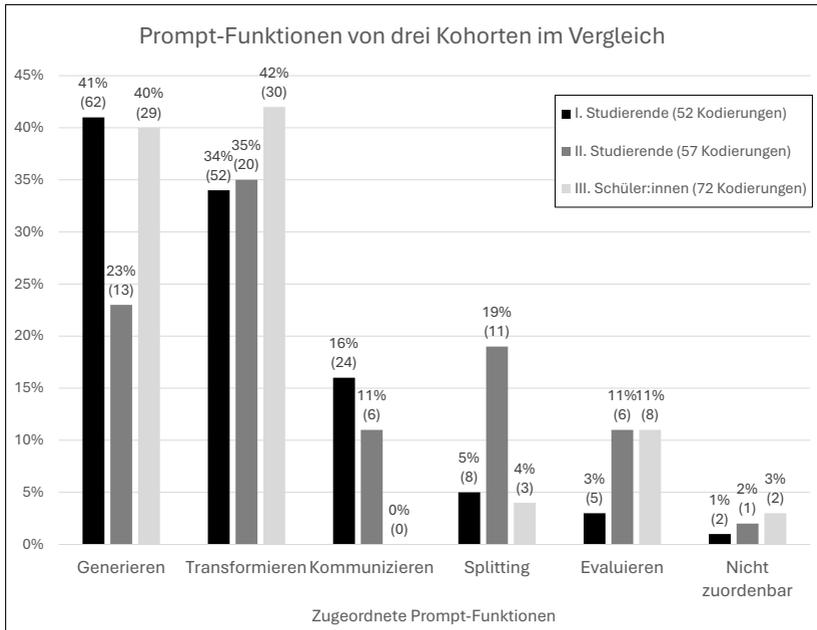
.....

4 Alle von den Lernenden formulierten Prompts werden hier unverändert wiedergegeben, einschließlich möglicher orthografischer oder grammatischer Fehler.

Prompt sowohl die Funktion *Generieren* als auch die Funktion *Transformieren* zugeordnet.

Abbildung 3 verdeutlicht die Verteilung der Prompt-Funktionen in den drei Erhebungen. Zur besseren Vergleichbarkeit wurden alle Kodierungen in Prozent umgerechnet.

### Prompt-Funktionen von drei Kohorten im Vergleich



**Abb. 3:** Quantitative Analyse der Prompt-Funktionen in KI-gestützten Schreibprozessen in drei Erhebungen (absolute Zahlen in Klammern).

Zusätzlich wurde ausgewertet, ob die Prompts der Teilnehmenden bereitgestelltes *Material* oder *Textprodukte* (Kommentar) beinhalteten. Die Analyse zeigt, dass Teilnehmende wesentlich häufiger Inhalte aus bereitgestellten Materialien in ihre Prompts einbezogen als Textendprodukte (bei den Studierenden insgesamt: 57 *Material*-Kodierungen vs. 22 *Textprodukt*-Kodierungen; bei den Schüler:innen: 26 vs. 5). Dies legt nahe, dass die Prompts primär

auf die Verarbeitung und Integration der Materialien ausgerichtet waren und weniger auf eine Überarbeitung oder Weiterentwicklung ihrer Textentwürfe bzw. -produkten. Trotz der unterschiedlichen Rahmenbedingungen und unterschiedlichen Voraussetzungen der Schreibenden lassen sich in der funktionalen Auswertung der Prompts strukturelle Gemeinsamkeiten erkennen, die auf Muster im Umgang mit textgenerativer KI hindeuten. In den nachfolgenden Abschnitten (3.3.1–3.3.5) werden die jeweiligen Prompt-Funktionen differenziert betrachtet, bevor in Kapitel 3.4 zentrale Herausforderungen dieser Praktiken diskutiert werden.

### 3.3.1 Generieren

Die Kategorie *Generieren* ist zusammen mit *Transformieren* die am häufigsten kodierte Funktion. Prompts sind mit diesem Code versehen, wenn sie die KI zur Erstellung neuer Texte oder Textteile auffordern. Diese Aufforderungen reichten von der Generierung einzelner Sätze oder Überschriften bis hin zu vollständigen Texten. Besonders auffällig war die enge Verknüpfung von *Generieren* mit dem Subcode *Material*. Bei Schüler:innen wurde diese Verknüpfung in 24 von 29 *Generieren*-Codierungen beobachtet, bei Studierenden in 40 von 75 Fällen. Die Schreibpraktik des Generierens scheint daher bei materialgestützten Schreibaufgaben eng mit Copy-and-Paste-Verfahren mit den vorgegebenen Materialien verbunden. Es gab jedoch auch eigenständig formulierte und detaillierte Prompts. Ein Beispiel hierfür lautet:

Ich schicke dir nun meinen Artikel. Da ich für Warnhinweise bin, behandelt der erste Abschnitt die Gefahren. Der zweite Abschnitt dann die Chancen. Bitte formuliere eine ansprechende Schlagzeile für den Artikel sowie eine kurze Einleitung. Achte dabei, dass du keine Doppelungen von unten aufgreifst. Es sollte spannend sein, die Relevanz des Themas darstellen und auch direkt die Position zeigen, dass Warnhinweise gut sind (jedoch sehr neutral und journalistisch dargestellt).  
(St\_SoSe24\_sam)

Dieses Beispiel zeigt, dass die Lernende ihr Anliegen zwar ausführlicher beschreibt, ihre Anweisung jedoch nicht vollständig zielgerichtet formuliert. So

bleibt unklar, warum sie zu Beginn ihres Prompts eine kurze Zusammenfassung des folgenden Textproduktes gibt – möglicherweise, um die Einleitung entsprechend gestalten zu lassen, allerdings macht sie diese Erwartung nicht explizit. Zudem fordert sie gleichzeitig eine klare Positionierung für Warnhinweise und eine neutrale, journalistische Darstellung – zwei Anforderungen, die ohne weitere Präzisierung schwer vereinbar sind.

### 3.3.2 Transformieren

Die Kategorie *Transformieren* war die insgesamt häufigste Funktion und umfasst Prompts, die darauf abzielen, bestehende Inhalte zu modifizieren. Diese Art der Interaktion zeigt, dass die Lernenden die KI nicht nur zur Generierung neuer Inhalte nutzen, sondern auch zur Verbesserung oder Anpassung vorhandener Anfragen, Inhalte oder Texte. Dabei lassen sich zwei Hauptformen der Transformation unterscheiden: einerseits die Überarbeitung von bereitgestelltem Material und andererseits die Überarbeitung eigener Textprodukte.

Viele Prompts zielten darauf ab, die vorgegebenen Materialien zu strukturieren, zu verdichten oder neu zu ordnen. Typische Beispiele sind:

- „Fasse den vorliegenden Text stichpunktartig zusammen“ (St\_SoSe24\_cae & St\_SoSe24\_les).
- „Sortiere die im Text stehenden Argumente und Kernaussagen mit Hinsicht auf die oben beschriebene Diskussion in Pro und Kontra“ (St\_SoSe24\_mom & St\_SoSe24\_nit).

Hierbei wurde deutlich, dass die *Transformieren*-Prompts der Studierenden (in 17 von 72 *Transformieren*-Kodierungen) und der Schüler:innen (in 20 von 30 Kodierungen) häufig vorgegebenes Material enthielten, das sie gezielt modifizieren ließen. In diesen Fällen handelt es sich meist um strukturierte und zielgerichtete Aufgaben, wie etwa das Extrahieren von Kernargumenten oder das Umformen von Texten in Stichpunkte.

Weniger häufig wurde die KI zur Modifikation selbst verfasster Texte genutzt. Während Studierende insgesamt in 22 von 72 Fällen ihre eigenen Texte transformieren ließen, trat dieses Verhalten bei den Schüler:innen in nur 2 von 30 Kodierungen auf. Auffällig war zudem, dass viele dieser Prompts unspezi-

fisch formuliert wurden, was darauf hindeutet, dass die Lernenden zwar eine Überarbeitung wünschen, aber nicht genau wissen, woran es liegt oder welche konkreten Anpassungen notwendig sind:

- „nochmal etwas leserfreundlicher“ (St\_WiSe24\_5ln & St\_WiSe24\_bwk).
- „Kann man es noch etwas rhetorischer darstellen? Sonst soll der Text genau gleich bleiben“ (St\_SoSe24\_sam).

Diese Eingaben deuten darauf hin, dass die Teilnehmenden zwar Veränderungen vornehmen lassen möchten, aber ihnen die klare Vorstellung vom Problem oder den Möglichkeiten der Modifikation fehlt. Besonders selten wurden Prompts beobachtet, die gezielt auf makrostrukturelle Veränderungen des Textes abzielten, also auf grundlegende Anpassungen der Argumentationsweise oder des Textaufbaus. Solche Fälle fanden sich ausschließlich bei den Studierenden:

- „Ich möchte das Ping-Pong-Prinzip bei der Argumentation verwenden“ (St\_SoSe24\_chu).
- „Ändere das Fazit, indem du dich für eine angemessene Regulierung aussprichst“ (St\_WiSe25\_omz & St\_WiSe25\_zg2).

Bei den Schüler:innen wurden keine vergleichbaren Eingriffe in die übergeordnete Textstruktur beobachtet.

### 3.3.3 Kommunizieren

Die Funktion *Kommunizieren* umfasst Prompts, die eine dialogische Interaktion mit der KI anstreben. Diese können entweder strategisch – zur gezielten Steuerung der KI – oder nicht strategisch – z. B. als spielerische Interaktion – genutzt werden. Auffällig ist, dass in der Schüler:innengruppe kein einziger solcher Prompt vorkam, während bei den Studierenden 16 % der Prompts in der ersten und 11 % in der zweiten Erhebung kommunikative Elemente enthielten.

Ein Teil der kommunikativen Prompts diente der strategischen Metakommunikation, indem der KI gezielt Rollen zugewiesen wurden, etwa durch Anwei-

sungen wie „Du bist ein erfahrener Journalist und machst bei einem Wettbewerb der ZEIT mit.“ (St\_SoSe24\_sam). Andere Prompts enthielten Rückmeldungen zur Qualität der generierten Texte („Der Text ist sehr gut.“ oder „Okay, nein.“ (St\_SoSe24\_sil)), was eine explizite Bewertung des Outputs der KI zeigt.

Der andere Teil der kommunikativen Prompts deutet auf einen anthropomorphen Umgang mit der KI hin. So fanden sich Begrüßungen wie „Hallo“ (St\_SoSe24\_sam), Dankesbekundungen („Danke“ (7x), „Gut“ (5x)) oder die Antwort auf eine KI-Entschuldigung: „Wir vergeben dir.“ (St\_SoSe24\_mom & St\_SoSe24\_nit).

### 3.3.4 Evaluieren

Die Kategorie *Evaluieren* war insgesamt selten vertreten. Während Studierende lediglich 10 Prompts dieser Art verwendeten, lag die Anzahl bei den Schüler:innen mit 8 Prompts auf einem ähnlich niedrigen Niveau. Ein Großteil dieser Prompts bezog sich auf die Überprüfung der Wortanzahl, was zeigt, dass Lernende die KI – wenn überhaupt – primär zur quantitativen Textbewertung einsetzten. Abweichend davon war jedoch das Vorgehen zweier Schüler:innen, die eine andere Strategie wählten: Sie kopierten den in der vorherigen Sitzung besprochenen Bewertungskatalog schrittweise in die KI und forderten sie auf, den Text daran zu messen („Bewerte den Text anhand dieser Kriterien“ (Sch\_GS11\_9dv & Sch\_GS11\_v48)). Zudem erfragten sie eine numerische Bewertung auf einer Notenskala („Gebe uns eine Note von 1–15“ (ebd.)). An dieser Stelle setzen die Lernenden die KI als prüfende Instanz ein (vgl. Katanneck/Suñer 2024; Fürstenberg i. d. B.).

### 3.3.5 Nicht zuordenbare Prompts und die Identifikation des Musters „Splitting“

Nachdem die Kategorie *Splitting* als eigenständige Kategorie kodiert wurde, verblieben tatsächlich nur noch 1–3 % der Prompts als *nicht zuordenbare* Eingaben. Diese wenigen Prompts könnten auf technische Fehler oder unvollständige Formulierungen zurückzuführen sein, wie etwa: „Lass ds m“ (St\_SoSe24\_amn & St\_SoSe24\_art).

Der identifizierte Code *Splitting* beschreibt ein spezifisches Prompting-Verhalten, bei dem Lernende größere Inhalte in kleinere Abschnitte zerleg-

ten und diese schrittweise an die KI übermittelten – oft ohne eine explizite Aufgabenstellung in jedem einzelnen Prompt zu formulieren. Das Vorgehen wurde in insgesamt 22 Fällen beobachtet und betraf vor allem Prompts, die ausschließlich kopierte Materialien enthielten, jedoch keine explizite Aufgabenstellung formulierten. Einige Proband:innen thematisierten dieses Vorgehen in den Prompts explizit z. B. durch „[d]iese drei Materialien würde ich dir nach und nach reinschicken“ (St\_SoSe24\_sam) oder „[i]ch gebe dir einen Text, ok?“ (St\_WiSe25\_pic & St\_WiSe25\_yfn). Nach diesen Ankündigungen bspw. folgten weitere Prompts mit Abschnitten aus dem vorgegebenen Material – jedoch ohne zusätzliche Instruktionen.

### **3.4 Diskussion: Herausforderungen beim Prompting**

Das entwickelte Kategoriensystem betrachtet Prompts als eigenständige Metatexte. So können Herausforderungen beim KI-gestützten Schreiben gezielt sichtbar gemacht und – unabhängig davon, welchen Output die KI liefert – didaktisch relevante Einblicke in die Formulierungs- und Strukturierungskompetenzen der Lernenden gewonnen werden.

Die Analyse zeigt, dass sich die Nutzung der textgenerativen KI durch die Lernenden erheblich unterscheidet, was sich besonders in der stark variierenden Anzahl der verwendeten Prompts widerspiegelt. Während ein Schreibteam bspw. keine Prompts formulierten, griffen andere intensiv auf die KI zurück. Dieser Befund legt nahe, dass es innerhalb der Lerngruppe zu unterschiedlichen Nutzungspraktiken kommt, deren Ursachen nicht aus der vorliegenden Untersuchung heraus bestimmt werden können. In der Diskussion möglicher Implikationen lässt sich jedoch auf Magirius/Scherf (2023: 414) verweisen, die argumentieren, dass ohne gezielte Unterstützung leistungsstarke Lernende vom Schreiben mit KI besonders profitieren könnten, während andere größere Schwierigkeiten im Umgang mit Chatbots haben könnten. Eine ungleiche Nutzungspraxis könnte langfristig bestehende Differenzen verstärken – ein Aspekt, der in weiteren Studien gezielt untersucht werden sollte.

Ein zentrales Ergebnis ist zudem die vielfach beobachtete Nutzung der KI zur Generierung neuer Inhalte, häufig unter Rückgriff auf Copy-and-Paste-Techniken mit bereitgestelltem Material. Diese Beobachtung deckt sich mit der

von Steinhoff/Lehnen (2025: 6f.) beschriebenen „Chat-to-Generate“-Praktik, bei der die KI als „Ghostwriter“ (ebd.: 6) fungieren kann. Zwar kann auf Basis der hier ausgewerteten Daten nicht überprüft werden, inwiefern generierte Inhalte tatsächlich unverändert in die Textprodukte übernommen werden. Allerdings lässt sich anhand der Prompting-Praktiken rekonstruieren, dass bestimmte epistemische Prozesse bereits in der ‚Prä-Koaktivität‘ (vgl. Steinhoff 2023: 12) an die KI delegiert werden: Auch wenn am so generierten Text Überarbeitungen stattfinden, bleibt fraglich, in welchem Maß Lernende den Schreibprozess strategisch steuern. Hier liegt die Vermutung nahe, dass die beobachtete Delegation des Schreibens an die Maschine mit einer Verschiebung kognitiver und epistemischer Prozesse einhergeht.

Die Analyse zeigt zudem, dass eigenständig formulierte Prompts häufig unspezifisch oder widersprüchlich sind. Transformative Prompts wurden zwar häufig genutzt, beschränkten sich aber überwiegend auf oberflächliche sprachliche Anpassungen oder das Umwandeln von Texten in Stichpunkte. Prompts, die auf eine inhaltliche Umstrukturierung oder tiefgreifende Makrorevision abzielten, blieben hingegen die Ausnahme. Die Frage, ob solche strukturellen Eingriffe notwendig waren, kann auf Grundlage der hier ausgewerteten Daten nicht gezeigt werden. Aus der Betrachtung der Prompts selbst lassen sich aber Potenziale ableiten: So zeigen die vorhandenen Umstrukturierungs-Prompts, dass einige Lernende durchaus das Bedürfnis hatten, Textstrukturen zu verändern – ihnen jedoch möglicherweise das methodische Repertoire fehlte, um dies der KI gegenüber klar zu formulieren. Daraus ergibt sich als mögliche didaktische Anschlussfrage, ob und wie gezielte Strategien zur Prompt-Formulierung im Unterricht thematisiert werden sollten, um die Handlungsfähigkeit der Lernenden im Umgang mit KI zu stärken.

Ein weiteres ungenutztes Potenzial zeigt sich in der Verwendung evaluativer Prompts. Steinhoff/Lehnen (2025) weisen darauf hin, dass KI als „Writing Tutor“ (ebd.: 11) gezielt für inhaltliche Rückmeldungen oder Hilfestellungen eingesetzt werden kann. Zwar stellen Fürstenberg/Müller (2024) kritisch infrage, ob GPTs didaktisch spezialisiert genug sein können, um einen solchen Mehrwert zu bieten. Dennoch bot gerade die hier beschriebene Unterrichtssituation eine günstige Gelegenheit, evaluative Prompts einzusetzen: Da den Lernenden der Bewertungskatalog bereits bekannt war und sie wussten, dass

ihre Texte in der nächsten Sitzung auf Basis dieser Kriterien beurteilt würden, hätten evaluative Prompts genutzt werden können, um vorab eigenständig Rückmeldungen zur Textqualität einzuholen. Diese Möglichkeit wurde jedoch lediglich von einem Schreibteam genutzt. Stattdessen beschränkten sich die meisten Lernenden auf quantitative Abfragen (z. B. Wortanzahl), was problematisch war, da textgenerative KI in der genutzten Version keine exakte Wortzählung gewährleisten konnte und auf solche Anfragen ausschließlich inkorrekte Ergebnisse lieferte. Diese Praxis verweist sowohl auf ein unzureichendes Verständnis der Funktionsweise der genutzten textgenerativen KI als auch auf fehlende Konzepte für die eigenständige Textevaluation.

Die vollständige Abwesenheit kommunikativer Prompts bei den Schüler:innen könnte aus didaktischer Perspektive positiv gewertet werden: So verzichteten sie auf Begrüßungen oder Dankesformeln, die in der Messenger-Kommunikation üblich sind, da sie sich möglicherweise funktional auf die Texterstellung konzentrierten. Bei den Studierenden hingegen könnte teilweise auch eine unreflektierte Übernahme bekannter Chatpraktiken angenommen werden, etwa durch anthropomorphisierende Sprache oder durch die Prompt-Praktik des *Splitting*. Dieses sequenzielle Vorgehen unterscheidet sich elementar vom sogenannten *Prompt Chaining*, also der systematischen Verfeinerung und Eingrenzung von Prompts entlang des Outputs (vgl. Steinhoff/Lehnen 2025: 9), da die Lernenden lediglich das Material auf mehrere Prompts verteilten und keine Anpassungen vornahmen. Vor diesem Hintergrund könnte es vorteilhaft sein, dass die Schüler:innen sich auf eine klar funktionale Interaktion beschränkten.

Gleichzeitig lässt sich die Zurückhaltung der Schüler:innen gegenüber kommunikativen Elementen auch negativ deuten. Das Fehlen interaktionsorientierter Prompts verweist möglicherweise auf eine einseitig textorientierte Schreibhaltung, die sich primär auf die rein funktionale Nutzung der KI konzentriert (vgl. Storrer 2018)<sup>5</sup> und die der dialogischen Oberfläche der KI-Plattform wenig Rechnung trägt. Im Gegensatz dazu könnte bei den Studien-

.....

5 Storrer (2018) unterscheidet zwischen einer *textorientierten* und einer *interaktionsorientierten* Schreibweise. Während textorientiertes Schreiben monologisch, explizit, planend und revidierend erfolgt, ist interaktionsorientiertes Schreiben dialogisch, implizit und spontan.

den ein Bewusstsein für die strategische Gestaltung ihrer Eingaben vermutet werden, etwa durch höfliche Formulierungen, Rückmeldungen und gezielte Rollenzuweisungen. So zeigen Yin et al. (2024), dass unhöfliche Eingaben oft zu schlechteren KI-Antworten führen. Allerdings wurde ein solches Wissen im Vorfeld nicht explizit vermittelt und in den anschließenden Gruppendiskussionen nicht thematisiert. Inwiefern eine kommunikative Promptgestaltung zu einem produktiven KI-Einsatz beiträgt, bleibt vorerst offen. Unmittelbar daran schließt sich jedoch die didaktische Frage an, wie Lernende zu einer reflektierten, differenzierten und strategischen Steuerung ihrer KI-gestützten Schreibprozesse befähigt werden können.

#### **4 Didaktische Implikationen: Vom experimentellen zum strategischen Prompting**

Die Ergebnisse legt nahe, dass Lernende Prompting häufig als isolierte Technik zur Textgenerierung nutzen, anstatt es als integralen Bestandteil eines reflektierten Schreibprozesses zu verstehen. Zwar kann anhand der hier präsentierten Daten nicht gezeigt werden, dass eine größere Vielfalt an Prompts automatisch zu besseren Texten führt, jedoch wird deutlich, dass Lernende wiederholt transformieren und evaluieren möchten – ihnen dafür jedoch vermutlich geeignete Konzepte fehlen. Erste Elemente eines Prompt-Scaffoldings wurden im Rahmen der Untersuchung bereits erprobt, doch eine vertiefende und zeitlich umfangreichere Unterstützung erscheint didaktisch sinnvoll, um die Handlungsspielräume der Lernenden beim Prompten systematisch zu erweitern. Konkret könnten Unterrichtseinheiten folgende Elemente enthalten:

- **Strukturiertes Experimentieren mit Prompting:** Lernende sollten verschiedene Prompt-Typen systematisch erproben – nicht nur zur Textgenerierung, sondern auch strategische und evaluative Prompts. Auch wenn es bspw. spezialisierte Plattformen wie FelloFish (ehemals fiete.ai) für die Textbewertung gibt, ist es sinnvoll, solche Evaluationsprozesse auch mit offenen KI-Chatsystemen wie ChatGPT und Co zu üben. Dies kann dazu beitragen, dass sich Lernende von ei-

ner möglichen Gebrauchssuggestion, mit Copy-and-Paste Texte zu generieren, lösen können und reflektiert entscheiden, wie sie die KI für ihren Schreibprozess einsetzen wollen. Eine kritische Reflexion über die begrenzte Verlässlichkeit bestimmter Prompts – etwa bei der Wortanzahl-Abfrage – kann ebenfalls dazu beitragen, dass Lernende ihre Rolle als Schreibende bewusster wahrnehmen und die Steuerung der KI gezielter übernehmen.

- Fokus auf Textumformungen: Beim KI-gestützten Schreiben kann es auch darum gehen, die inhaltliche Struktur von Texten zu überdenken und überarbeiten. Damit dies gelingt, kann der Unterricht gezielt zentrale Merkmale wichtiger Textsorten thematisieren – beim materialgestützten Schreiben etwa Thesenformulierung, Argumentationsaufbau, Umgang mit Gegenpositionen oder Kohärenz (vgl. Feilke et al. 2018; Philipp 2020). Diese Aspekte lassen sich gemeinsam mit den Schüler:innen in konkreten Prompt-Formulierungen aufgreifen und erproben. Eine solche Auseinandersetzung mit dem Aufbau unterschiedlicher Textsorten kann das Textverständnis vertiefen und die Fähigkeit zur eigenständigen Textgestaltung fördern (vgl. Jang et al. 2024).
- Reflexion über Schreibkonferenzen: Im Peer-Feedback könnten nicht nur die fertigen Textprodukte, sondern auch die zugrundeliegenden Chatverläufe analysiert werden. Die Diskussion verschiedener Promptvarianten und Steuerungsansätze kann Lernenden helfen, sich aus passiven Nutzer:innenrollen zu lösen und gezielt Strategien für eine effektive KI-Nutzung zu entwickeln. Eine solche Reflexion kann nicht nur die kritische Auseinandersetzung mit KI-gestütztem Schreiben, sondern auch die metakommunikativen Fähigkeiten der Lernenden stärken, die es für das Formulieren von Prompts benötigt.

Auch wenn keine direkten Rückschlüsse auf die Qualität der resultierenden Texte gezogen werden können, zeigen die Daten, dass die Lernende unterschiedliche Prompt-Formulierung erproben, diese aber nur begrenzt differenziert und gezielt einsetzen. Vor diesem Hintergrund erscheint es didaktisch anschlussfähig, Prompting als metatextuelle Handlung zu verstehen und diese

im Schreibunterricht gezielt zu thematisieren – etwa durch die Erprobung unterschiedlicher Prompttypen, die Reflexion über Textstruktur und durch kollektive Analyse von Prompting-Prozessen. Die vorliegenden Befunde liefern somit Impulse für die Weiterentwicklung didaktischer Konzepte, die den Schreibprozess im Zusammenspiel mit textgenerativer KI kritisch und reflektiert begleiten.

## Literatur

- ANSARI, ANNA (i. V.): *Postdigitale Schreibpraktiken. Eine empirische Studie zum Einsatz textgenerativer KI im Deutschunterricht*. Dissertationsvorhaben an der Justus-Liebig-Universität Gießen.
- BACHMANN, THOMAS/BECKER-MROTZEK, MICHAEL (2017): Schreibkompetenz und Textproduktion modellieren. In: Becker-Mrotzek, Michael/Grabowski, Joachim/Steinhoff, Torsten (Hg.): *Forschungshandbuch empirische Schreibdidaktik*. Münster: Waxman, S. 25–53.
- BAJOHR, HANNES (2023): Autorschaft und Künstliche Intelligenz. In: Catani, Stephanie (Hg.): *KI und die Künste*, S. 265–280. Berlin: De Gruyter.
- BECKER-MROTZEK, MICHAEL/SCHINDLER, KIRSTEN (2007): Schreibkompetenz modellieren. In: dies. (Hg.): *Texte schreiben*. Duisburg: Gilles & Francke, S. 7–26. URL: [http://koebes.phil-fak.uni-koeln.de/sites/koebes/user\\_upload/koebes\\_05\\_2007.pdf](http://koebes.phil-fak.uni-koeln.de/sites/koebes/user_upload/koebes_05_2007.pdf) (letzter Zugriff: 03.03.2025).
- BEISWENGER, MICHAEL (2015): Praktiken in der internetbasierten Kommunikation. In: Deppermann, Arnulf/Feilke, Helmuth/Linke, Angelika (Hg.): *Sprachliche und kommunikative Praktiken*. Berlin/New York, S. 279–308. DOI: 10.1515/9783110451542-012.
- BUCK, ISABELLA/LIMBURG, ANIKA/MUNDORF, MARGRET (2023): Faszination, Skepsis und Enttäuschung. Eine explorative Studie zur epistemischen Nutzung von ChatGPT unter Schüler:innen. In: *Der Deutschunterricht* 5, S. 64–74.
- FEILKE, HELMUTH/LEHNEN, KATRIN/REZAT, SARA/STEINMETZ, MICHAEL (Hg.) (2018): *Materialgestütztes Schreiben – Erfahrungen aus der Praxis und Perspektiven der Forschung*. Stuttgart: Fillibach bei Klett.

- FÜHRER, CAROLIN/NIX, DANIEL (2024): Literarische Interaktionen mit ChatGPT – Kann der Einsatz von Künstlicher Intelligenz zur Entwicklung literarischer Lesehaltungen beitragen? In: *Literarische Texte lesen – Texte literarisch lesen: Festschrift für Cornelia Rosebrock*. Berlin, Heidelberg: J.B. Metzler, S. 335–368. DOI: 10.1007/978-3-662-67816-9\_19.
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): Wie funktioniert ChatGPT und was sind die Folgen für den Deutschunterricht? In: *Der Deutschunterricht* 5, S. 2–13.
- HAYES, JOHN R. (1996): A new framework for understanding cognition and affect in writing. In: Levy, Michael C./Ransdell, Sarah R. (Hg.): *The science of writing: Theories, methods, individual differences, and applications*, S. 1–28.
- HAYES, JOHN R. (2012): Modeling and Remodeling Writing. In: *Written Communication* 29 (3), S. 369–388. DOI: 10.1177/0741088312451260.
- HAYES, JOHN/FLOWER, LINDA (1980): Identifying the organization of writing processes. In: Gregg, Lee W./Steinberg, Erwin R. (Hg.): *Cognitive processes in writing*. Hillsdale, NJ: Erlbaum, S. 3–30.
- HIRSCHAUER, STEFAN (2016): Verhalten, Handeln, Interagieren. In: Schäfer, Hilmar (Hg.): *Ein soziologisches Forschungsprogramm*. Bielefeld: transcript, S. 45–67. DOI: 10.1515/9783839424049-003.
- JAKOBS, EVA-MARIA/PERRIN, DANIEL (2014): Introduction and research roadmap. Writing and text production. In: dies. (Hg.): *Handbook of writing and text production*. Berlin, New York: De Gruyter, S. 1–24.
- JANG, JINHO/EUN, SEUNMIN/LEE, HYUNWOONG/CHOI, JEEWON/CHO, YOUNG HOAN (2024): The Effects of Prompt Scaffolding on Learning to Write Arguments with ChatGPT. In: *Proceedings of the International Society of the Learning Sciences (ISLS)*, S. 1502–1505. DOI: 10.22318/icls2024.831011.
- KATANNECK, STANISLAV/SUÑER, FERRAN (2024): Selbstreguliertes Lernen mit Feedback von ChatGPT: Die Rolle individueller Einstellungen und Wahrnehmungen beim selbstregulierten Lernen mit KI-basiertem Feedback. In: *Kontexte: Internationales Journal zur Professionalisierung in Deutsch als Fremdsprache* 2 (2), S. 14–32. DOI: 10.24403/jp.1394575.
- LARSEN, MORITZ/LUCHT, MELANIE/WEBELS, DORIS (2022): Chain of Thought Prompting – Ein Weg in die Gedankenwelt der KI Sprachmodelle. In: *KI Campus*.

- URL: <https://ki-campus.org/blog/chain-of-thought-prompting> (letzter Zugriff: 03.03.2025).
- MAGIRIUS, MARCO/SCHERF, DANIEL (2023): Studierende interpretieren Gedichte mit ChatGPT– Chancen und Herausforderungen von KI-Tools im Lehramtsstudium Deutsch. In: *Mitteilungen des Deutschen Germanistenverbandes* 70 (4), S. 406–415.
- MÜLLER, HANS-GEORG/FÜRSTENBERG, MAURICE (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Deutschen Germanistenverbandes* 70 (4), S. 327–345.
- PERRIN, DANIEL (2016): Vom vielschichtigen Planen. Textproduktions-Praxis empirisch erforscht. In: Deppermann, Arnulf/Feilke, Helmuth/Linke, Angelika (Hg.): *Sprachliche und kommunikative Praktiken*. Berlin/Boston: De Gruyter, S. 431–456.
- PHILIPP, MAIK (2020): *Grundlagen der effektiven Schreibdidaktik und der systematischen schulischen Schreibförderung*. 8., erw. Aufl. Baltmannsweiler: Schneider Hohengehren.
- REICHERTZ, JO (2016): *Qualitative und interpretative Sozialforschung. Eine Einladung*. Wiesbaden: Springer Fachmedien. DOI: 10.1007/978-3-658-13462-4.
- STEINHOFF, TORSTEN (2022): Die digitale Transformation des Schreibens. In: Knopp, Matthias et al. (Hg.): *Sprachliche Bildung in der digitalisierten Gesellschaft. Was wir in Zukunft wissen und können müssen*. Münster/New York: Waxmann, S. 143–159.
- STEINHOFF, TORSTEN (2023): Der Computer schreibt (mit): digitales Schreiben mit Word, Whatsapp, ChatGPT & Co. als Koaktivität von Mensch und Maschine. In: *MiDU – Medien im Deutschunterricht* 5 (1), 1–16. URL: <https://journals.ub.uni-koeln.de/index.php/midu/article/view/1912> (letzter Zugriff: 03.03.2025).
- STEINHOFF, TORSTEN/LEHNEN, KATRIN (2025): Schreiben mit Künstlicher Intelligenz: Das GPTModell (Ghost, Partner, Tutor). Erscheint in: *leseräume.de*. Preprint: DOI: 10.13140/RG.2.2.23519.37286.
- STORRER, ANGELIKA (2018): Interaktionsorientiertes Schreiben im Internet. In: Deppermann, Arnulf/Reineke, Sibylle (Hg.): *Sprache im kommunikativen, interaktiven und kulturellen Kontext*. Berlin: De Gruyter, S. 219–244. DOI: 10.1515/9783110538601-010.
- TEKIN, ÖZLEM (2023): ChatGPT im Unterricht Deutsch als Fremdsprache. In: *Alman Dili ve Kültürü Araştırmaları Dergisi* 5 (2), S. 135–163. DOI: 10.55143/alkad.1390420.

YIN, ZIQI/WANG, HAO/HORIO, KAITO/KAWAHARA, DAISUKE/SEKINE, SATOSHI (2024): Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. In: *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, S. 9–35. DOI: 10.18653/v1/2024.sicon-1.2.

# ChatGPT-4o im Grammatikunterricht

## Möglichkeiten und Grenzen des Tools

### Abstract

In der vorliegenden Studie wurde die Performance des KI-Tools ChatGPT-4o im Bereich der deutschen Grammatik untersucht. Die Leistungen des Tools wurden durch verschiedene Aufgabenformate und Fragestellungen getestet. Die dadurch gewonnenen Einblicke in seine Möglichkeiten und Grenzen geben Anlass zur Diskussion, ob und ggf. wie ChatGPT-4o als Hilfsmittel im schulischen Grammatikunterricht eingesetzt werden kann. Die Ergebnisse zeigen, dass ChatGPT-4o bei Ergänzungs-, Kategorisierungs- und Korrekturaufgaben gut abschneidet, jedoch Probleme mit der metasprachlichen Ebene aufweist. Das Tool eignet sich also insbesondere zur Förderung sprachpraktischer Fähigkeiten und zur Bearbeitung von Zuordnungsaufgaben. Was die Bereiche des Grammatikunterrichts anbelangt, die einen tieferen Einblick in die sprachlichen Zusammenhänge erfordern, ist aufgrund der festgestellten Mängel Vorsicht geboten.

**Schlagwörter:** Grammatikdidaktik, Grammatikunterricht, Grammatikwissen, künstliche Intelligenz, sprachheterogene Lerngruppen

## 1 Einleitung

Seit der Freigabe von ChatGPT-3.5 im November 2022 ist die Anwendung der künstlichen Intelligenz im schulischen Kontext zum Gegenstand der bildungspolitischen Diskussion (KMK 2024; SWK 2024) und der sprachdidaktischen Forschung geworden (Leifheit et al. 2024). Im deutschsprachigen Raum wird der Einsatz von KI-gestützten Tools in der Schule insbesondere

re im Hinblick auf die Schreibdidaktik untersucht (Lehnen/Steinhoff 2024; Steinhoff 2024a, 2024b, 2025). Dabei liegt der Fokus auf der Frage, ob und ggf. wie Modelle wie ChatGPT zur Förderung von Schreibkompetenzen verwendet werden können. In diesem Zusammenhang wird auch über den Einsatz von KI-Anwendungen als Werkzeug zum Erstellen von individualisierten und lernförderlichen Feedbacks diskutiert (Haverkamp et al. 2024; Dai et al. 2023; Kasneci et al. 2023). Feedbacktools wie FelloFish (<https://www.fellofish.com/>) sollen die Lehrkräfte entlasten und die Schüler:innen dabei unterstützen, ihre Schreibkompetenzen zu entfalten (Haverkamp et al., 2024, vgl. auch Fürstenberg i. d. B.). Die Arbeit mit generativen KI-Modellen in anderen Kompetenzbereichen des Deutschunterrichts wurde bislang kaum erforscht (für einen Überblick s. Fürstenberg/Müller 2024). Besonders wenig Beachtung wird dem Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ geschenkt. Gleichwohl handelt es sich dabei um einen Bereich, der sowohl den Schüler:innen als auch den Lehrkräften große Schwierigkeiten bereitet. Wie empirische Studien der letzten 30 Jahre zeigen, verfügen viele Schüler:innen, Schulabsolvent:innen und angehende Deutschlehrer:innen über geringe bis mangelhafte Grammatikkenntnisse (Ivo/Neuland 1991; Risel 1999; Kammler/Switalla 2001; Frentz/Lehmann 2003; Schmitz 2003; Schulze 2011; Binanzer et al. 2015; Dämmer 2019; Müller/Geyer 2020). Als mögliche Ursachen für diese Tendenz werden die Unterrichtsgestaltung (Dürscheid 2007; Granzow-Emden 2013) sowie negative Einstellungen zum Grammatikunterricht seitens der Schüler:innen und der Lehrkräfte (Bremerich-Vos 1981, 1999; Fix 2000) angesehen. Dies wirft die Frage nach neuen Vermittlungsmethoden und Vermittlungstools auf, die den Grammatikunterricht attraktiver bzw. zugänglicher und somit wirksamer machen könnten. Im Hinblick auf den Zweit- und Fremdsprachunterricht zeigt der aktuelle Forschungsstand, dass KI-basierte Tools wie Chatbots nicht nur zur Verbesserung von sprachlichen Kompetenzen der Schüler:innen beitragen, sondern auch zur Entfaltung der Sprachbewusstheit in ihrer Zweitsprache, im Besonderen des sog. Fehlerbewusstseins (Fryer/Carpenter 2006; Dodigovic 2007). Des Weiteren kann die Arbeit mit solchen Tools die Motivation der Schüler:innen steigern und ihnen helfen, ihre Hemmungen im Gebrauch ihrer Zweitsprache abzubauen (Coniam 2008; Goda et al. 2014; Fryer et al. 2017).

Vor diesem Hintergrund wird in dem vorliegenden Beitrag diskutiert, ob und ggf. wie das KI-Sprachmodell ChatGPT-4o als Hilfsmittel im schulischen Grammatikunterricht eingesetzt werden kann. Dafür wird zunächst getestet, welchen Output das Tool zu linguistischen Fragen und grammatischen Aufgaben generiert. Dabei wird zwischen unterschiedlichen Wissensformen und den damit einhergehenden Sprachbetrachtungsaktivitäten differenziert. Dies soll zeigen, in welchen Teilbereichen des Grammatikunterrichts sich die Arbeit mit ChatGPT-4o als besonders lernförderlich erweisen kann.

Bevor auf die Ergebnisse des durchgeführten Experiments eingegangen wird, werden in Kapitel 2 Formen sprachlichen Wissens und deren Bedeutung für den schulischen Grammatikunterricht besprochen. Dabei liegt der Fokus auf dem Spannungsverhältnis zwischen den in den Bildungsstandards definierten Zielen und den Lernbedürfnissen der Schüler:innen in sprachheterogenen Gruppen. In Kapitel 3 wird erläutert, wie die Performance von ChatGPT-4o operationalisiert wurde. Die Ergebnisse werden in Kapitel 4 dargestellt. Der Beitrag schließt in Kapitel 5 mit didaktischen Implikationen für die Arbeit in sprachheterogenen Gruppen.

## 2 Sprachwissen und Grammatikunterricht

Sowohl im Kontext des Erst- und Zweitspracherwerbs als auch in der sprachdidaktischen Forschung wird oftmals zwischen zwei verschiedenen Formen sprachlichen Wissens differenziert (vgl. u. a. Andresen/Funke 2006; Paradis 2009; Bredel 2013; Funke 2014, 2018; Lütke 2014; Müller/Unterholzner 2023). Als *implizit* wird eine primärsprachliche Wissensform bezeichnet, die im Laufe des Spracherwerbs ungesteuert erworben wird und die Sprecher:innen befähigt, sprachliche Äußerungen in einer Sprache zu rezipieren und zu produzieren. Da diese Wissensform dem Bewusstsein weitgehend unzugänglich bleibt, kann sie nicht verbalisiert werden. Im Gegensatz dazu handelt es sich beim *expliziten* Sprachwissen um ein bewusstes Wissen über den Aufbau eines Sprachsystems. Die beiden Wissensformen manifestieren sich bei unterschiedlichen Sprachbetrachtungsaktivitäten. Bredel (2013) unterscheidet in diesem Zusammenhang zwischen operativen und deklarativen Handlungen. Unter

operativen Handlungen werden Sprachbetrachtungsaktivitäten verstanden, bei denen implizite Sprachverarbeitungsprozesse aktiviert und spezifischen Manipulationen unterzogen werden. Dazu gehören u. a. Fremdkorrekturen oder Grammatikalitätsurteile. Deklarative Handlungen werden *offline* ausgeführt d. h. ohne dass primärsprachliche Sprachverarbeitungsprozesse aktiviert werden. Bei solchen Sprachbetrachtungsaktivitäten steht also ein rein metasprachliches Wissen im Fokus. Dabei handelt es sich bspw. um Benennen, Beschreiben und Kategorisieren sprachlicher Strukturen (Bredel 2013: 106).

Die Ausdifferenzierung sprachlichen Wissens ist besonders relevant für den frühen Grammatikunterricht in sprachheterogenen Lerngruppen, da zu Beginn der Primarstufe nicht alle Kinder über gleiche sprachpraktische Fähigkeiten verfügen (Grimm/Schulz 2014). Während Schüler:innen mit Deutsch als Muttersprache beginnen können, ihre Wissensbestände zu systematisieren, arbeiten Schüler:innen mit Deutsch als Zweitsprache oft noch an ihren basalen Sprachkenntnissen. Dies wirkt sich auf ihre Lernbedürfnisse und somit auf die Lehr- und Lernziele des Grammatikunterrichts aus. Ein Grammatikunterricht, der beiden Gruppen gerecht werden will, müsste dementsprechend nicht nur auf den Aufbau expliziten Sprachwissens, sondern auch auf den Ausbau impliziten Sprachwissens abzielen (Binanzer/Langlotz 2018; Binanzer/Wecker 2020).

Die Ziele des Grammatikunterrichts in der Grundschule werden in den Bildungsstandards für die Primarstufe (KMK, 2022) unter dem Kompetenzbereich „Sprache und Sprachgebrauch untersuchen“ dargestellt. Im Rahmen dieses Kompetenzbereichs sollen die Schüler:innen sprachliche Verständigung, sprachliche Vielfalt und sprachliche Strukturen untersuchen. Zentral dabei sind die Beschreibung und Benennung von sprachlichen Strukturen und Prozessen. Die Untersuchung linguistischer Phänomene soll den Schüler:innen helfen, ihre Sprachbewusstheit und metasprachlichen Kompetenzen zu entwickeln. Die so gewonnenen Einblicke in Aufbau und Funktion von Sprache sollen des Weiteren zur Entwicklung von orthografischen und grammatischen Kompetenzen in der Schriftsprache beitragen (KMK, 2022). Das zentrale Ziel des schulischen Grammatikunterrichts besteht demzufolge darin, das implizite Sprachwissen der Schüler:innen zu explizieren und zu systematisieren. Somit werden im Grammatikunterricht bestimmte sprachpraktische Kompetenzen vorausgesetzt. Besonders deutlich zeigt sich dies bei Grammatikaufgaben in

gängigen Lehrwerken. So müssen die Schüler:innen bspw. oft Konjugations- und Deklinationsmuster bereits kennen, um Aufgaben richtig zu lösen. Durch den starken Fokus auf den Erwerb expliziten Wissens bleiben die Lernbedürfnisse der Schüler:innen mit Deutsch als Zweitsprache unberücksichtigt. Dabei ist es fraglich, ob und ggf. inwiefern explizit vermitteltes Sprachwissen die Weiterentwicklung des impliziten Sprachwissens voranbringen kann. Sowohl in der Erstsprachdidaktik (Andresen/Funke 2006; Funke 2014; 2018; Feilke/Tophinke 2016) als auch in der Zweitsprachdidaktik (Diehl et al. 2000; Lütke 2014) wird intensiv diskutiert, welcher Grammatikunterricht tatsächlich zum Ausbau impliziten Wissens beitragen könnte. Die Frage nach der zielführendsten Form von Grammatikunterricht stellt sich auch in Bezug auf die Vermittlung expliziten Sprachwissens. Studien zum grammatischen Wissen von Schüler:innen und angehenden Lehrkräften (s. Kapitel 1) deuten nämlich an, dass die in den Bildungsstandards formulierten Ziele und Ansprüche nur zum Teil erfüllt werden. Diese Ergebnisse können allerdings durch die Unterschiede in methodischen Zugängen bedingt sein. Wird Grammatikwissen auf rein terminologisches Wissen reduziert, so schneiden die Proband:innen relativ schlecht ab. Geht es bei den Grammatiktests nicht um eine reine Wortarten- und Satzgliedbestimmung, sondern vielmehr um sprachreflexive Aufgaben, so sind die Leistungen deutlich besser (Funke 2005, 2018). Demzufolge ist es relevant, sowohl für derlei Studien als auch für den Grammatikunterricht selbst zu präzisieren, was unter Grammatikwissen und Grammatikkenntnissen verstanden wird.

### 3 Methodik

In der vorliegenden Studie wurden die Leistungen des KI-gestützten Modells ChatGPT-4o im Bereich der deutschen Grammatik untersucht. Bei dem getesteten Tool handelt es sich um eine im Mai 2024 freigegebene Version, die kostenlos nur im limitierten Umfang verwendet werden kann. Für das durchgeführte Experiment wurde ein neues Konto erstellt, sodass ChatGPT nicht auf das bereits Gelernte zurückgreifen konnte. Sein Output basiert somit ausschließlich auf seinen eigenen Daten und nicht auf den Interaktionen mit dem User.

In Bezug auf die Performance von ChatGPT-4o wurde zwischen Aufgaben und Fragestellungen differenziert, die implizites oder explizites Sprachwissen erfordern (s. Kapitel 2). Wie Experimente mit KI-basierten Tools zeigen, hängen die Qualität und Korrektheit des generierten Outputs maßgeblich von der Fähigkeit ab, gute Prompts zu schreiben (Acar, 2024; Chen et al., 2024; Schulhoff et al., 2025, vgl. auch Simbeck i. d. B.). So erhöhen präzise und kontextreiche Prompts die Wahrscheinlichkeit, dass KI-Modelle gezielte Antworten liefern. Auch Beispiele und Schritt-für-Schritt Anweisungen können sich positiv auf die Qualität des Outputs auswirken. Die sog. Prompting-Kompetenzen können allerdings nicht bei allen Usern vorausgesetzt werden. SWK (2024) und KMK (2024) zählen Prompting-Kompetenzen zu Schlüsselkompetenzen, die erst im schulischen Rahmen entwickelt werden sollten (zur Kritik s. Fürstenberg/Müller 2024: 8). Aus diesem Grund wurden die Prompts in der vorliegenden Studie aus der Perspektive einer Person formuliert, die wenig Erfahrung im Umgang mit der KI hat. Dabei handelt es sich um einfache und kurze Formulierungen ohne zusätzliche Beispiele und Kontextualisierung. Es wurden auch keine Theorien und Kategorisierungsmuster vorgegeben, auf denen ChatGPT seinen Output basieren könnte. Diese Vorgehensweise sollte zeigen, mit welchen Ergebnissen Schüler:innen konfrontiert werden, wenn sie das Tool ohne Vorbereitung und Begleitung seitens der Lehrkräfte verwenden. So können potenzielle Probleme und Schwächen des Tools identifiziert und anschließend Lösungsvorschläge bzw. didaktische Implikationen für dessen Einsatz im Grammatikunterricht formuliert werden.

Im Folgenden wird detailliert erläutert, wie die Performance von ChatGPT-4o operationalisiert wurde.

### 3.1 Implizites Sprachwissen

Leistungen im Bereich des impliziten Sprachwissens wurden durch C-Tests überprüft, die mit dem Didaktisierungstool LingoFox (<https://lingofox.dw.com/index.php?url=c-test>, 16.12.2024) erstellt wurden. Klassische C-Tests bestehen aus vier kurzen Texten, in denen bei jedem zweiten Wort die Hälfte getilgt ist. Die Aufgabe der Testpersonen besteht darin, die fehlenden Wortteile sinngemäß zu ergänzen. Mittlerweile gelten C-Tests als ein globales Tool

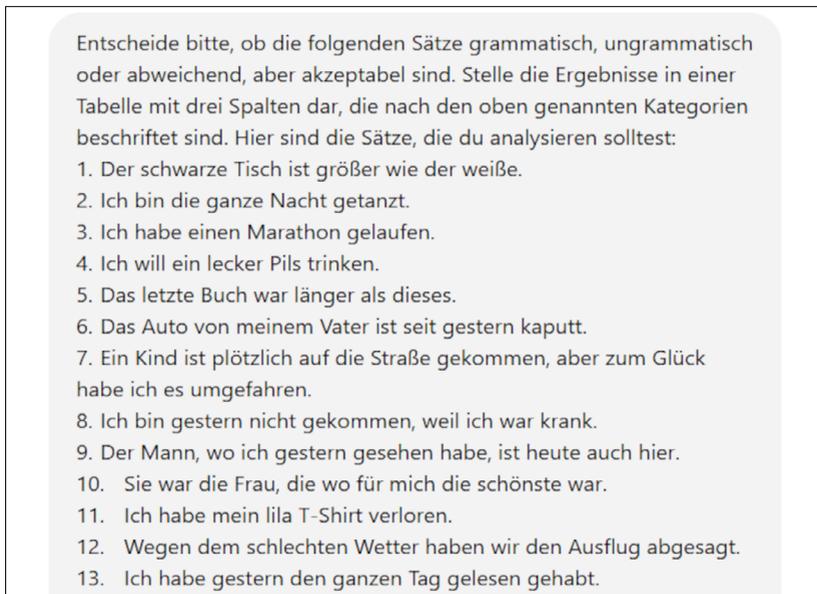
zur Einschätzung der allgemeinen Sprachkompetenz in der Erst-, Zweit- und Fremdsprache. Der Vorgänger des C-Tests, der sog. Cloze-Test (Taylor, 1953), wird auch als Evaluationstool für große Sprachmodelle (LLMs) verwendet (Lewis et al. 2019; Fabio et al. 2019). Die für die vorliegende Studie entwickelten C-Tests basieren auf demselben Tilgungsprinzip (Abb. 1) und umfassen vier verschiedene Textsorten. Dabei handelt es sich um einen Ausschnitt aus dem Roman *Die Klavierspielerin* (Jelinek, 2011: 40), einen Ausschnitt aus dem Essay *Haben oder Sein* (Fromm, 2019: 57), einen Ausschnitt aus der Reportage *Die Erde ist ein gewalttätiges Paradies* (Kapuściński, 2010: 296) und um ein Fragment des Zeitungsartikels *Teuer erkaufte Ruhe* (Brucklacher, 2015) zum Thema Kinder und Smartphones. Diese vielfältige Auswahl an Texten sollte zeigen, ob ChatGPT-4o bei allen Lückentexten gleich gut abscheidet oder seine Performance durch die Textsorte beeinflusst wird. Die Leistung des Tools wurde anhand der zu ergänzenden Lücken gemessen. Da generative Sprachmodelle darauf trainiert werden, für eine gegebene Datenreihe den nächstwahrscheinlichen Wert zu erzeugen (Fürstenberg/Müller 2024), wird bei C-Tests eine gute Performance erwartet.

Ergänze bitte die Lücken in dem folgenden Text

Meine alten Beka \_\_\_\_\_ 1) verlassen mich a \_\_\_\_\_ 2) zwei Arten: d \_\_\_\_\_ 3) einen verschwinden a \_\_\_\_\_ 4) den Friedhof f \_\_\_\_\_ 5) immer, die and \_\_\_\_\_ 6) bleiben am Le \_\_\_\_\_ 7), doch ich verl \_\_\_\_\_ 8) den geistigen Kon \_\_\_\_\_ 9) zu ihnen. Man \_\_\_\_\_ 10) erwecken den Eind \_\_\_\_\_ 11), sie seien ni \_\_\_\_\_ 12) älter geworden, hät \_\_\_\_\_ 13) keine Erfahrungen gesa \_\_\_\_\_ 14). Sie sprechen no \_\_\_\_\_ 15) dieselbe Sprache w \_\_\_\_\_ 16) vor Jahren u \_\_\_\_\_ 17) denken in dens \_\_\_\_\_ 18) starren Bahnen. I \_\_\_\_\_ 19) weiß nicht, w \_\_\_\_\_ 20) ich ihnen antw \_\_\_\_\_ 21), wie ich z \_\_\_\_\_ 22) ihnen sprechen so \_\_\_\_\_ 23). Das Bewusstsein die \_\_\_\_\_ 24) Menschen erinnert mi \_\_\_\_\_ 25) a das Obj \_\_\_\_\_ 26) eines Fotoapparates, i \_\_\_\_\_ 27) dem nur ein \_\_\_\_\_ 28) die Blende geöf \_\_\_\_\_ 29) wurde. Der Fi \_\_\_\_\_ 30) notiert ein gewi \_\_\_\_\_ 31) Bild der We \_\_\_\_\_ 32), dann schließt si \_\_\_\_\_ 33) die Blende, u \_\_\_\_\_ 34) so bleibt e \_\_\_\_\_ 35).

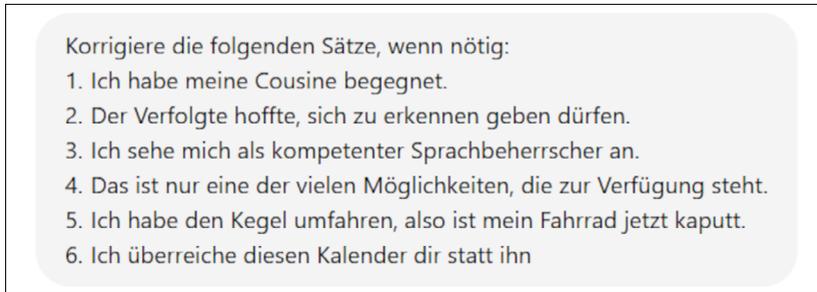
**Abb. 1:** C-Test auf der Grundlage von *Die Erde ist ein gewalttätiges Paradies* (Kapuściński, 2010: 296)

Des Weiteren wurden Fragen zur Grammatikalität und Akzeptabilität sprachlicher Äußerungen sowie Korrekturaufgaben konzipiert. Die Grammatikalitätsurteile wurden mithilfe von 31 Sätzen erhoben. Die Sätze wurden ChatGPT-4o als Liste präsentiert und mussten anschließend auf der folgenden Skala eingeordnet werden: *grammatisch korrekt*, *grammatisch inkorrekt* und *nicht standardsprachlich, aber unter Umständen akzeptabel*. Die letzte Kategorie umfasst eher dialektal gefärbte (*Der schwarze Tisch ist größer wie der weiße*) und umgangssprachliche Konstruktionen (*Das Auto von meinem Vater ist seit gestern kaputt*) als grammatische Zweifelsfälle. Abbildung 2 stellt die Formulierung des Prompts und einen Ausschnitt aus dem zu analysierenden Material dar. Die Grammatikalitätsurteile von ChatGPT-4o wurden in der Folge sprachlichen Normen aus Referenzgrammatiken (Dudenredaktion 2009; Dudenredaktion 2011) sowie aus dem Dudenkalender zu grammatischen Zweifelsfällen (Balcik/Hesse 2023) gegenübergestellt. Des Weiteren wurde analysiert, inwiefern sich aus den generierten Antworten Muster bzw. Regelmäßigkeiten ableiten lassen.



**Abb. 2:** Prompt zu Grammatikalitätsurteilen

Für die Korrekturaufgaben wurden sechs Sätze mit grammatischen Fehlern (Abb. 3) und zwei authentische Schülertexte aus dem Scriptoria Korpus der Forschungsdatenbank FD-LEX (Becker-Mrotzek/Grabowski 2018) verwendet. Bei den Texten handelt es sich um einen kurzen Bericht über einen Autounfall. Die getesteten Schüler:innen waren in der fünften Klasse einer integrierten Gesamtschule und haben Deutsch als Zweitsprache erworben<sup>1</sup>. Die Aufgabe von ChatGPT-4o bestand darin, das sprachliche Material in Bezug auf die Grammatik und Rechtschreibung zu korrigieren. Seine Verbesserungs- und Korrekturvorschläge wurden anschließend auf ihre sprachliche Angemessenheit hin geprüft.



**Abb. 3:** Prompt zur Korrektur isolierter Sätze

### 3.2 Explizites Wissen

Da es sich beim expliziten Sprachwissen um deklaratives Wissen über grammatische Strukturen und linguistische Phänomene handelt, wurden die Leistungen in diesem Bereich durch Aufgaben zu Wortarten, Wortgruppen und Satzgliedern überprüft. So musste ChatGPT-4o in den vorgegebenen Sätzen und Texten Wortarten und Satzglieder bestimmen. Bei einer weiteren Aufgabe musste das Tool Präpositionalphrasen identifizieren und anschließend ihre syntaktischen Funktionen bestimmen (Abb. 4). Darüber hinaus wurde eine Reihe von offenen Fragen formuliert, die auf die Begründung für bestimmte

.....

- 1 Die Daten zur Kontaktdauer zum Deutschen liegen nicht vor. Der Spracherwerbstyp kann also nicht genauer bestimmt werden.

Konstruktionen und Korrekturvorschläge von ChatGPT-4o abzielen. Dies sollte zeigen, ob die metalinguistischen Erklärungen des Tools sprachwissenschaftlich fundiert sind.

Bestimme, welche Satzgliedfunktion die präpositionalen Phrasen in den folgenden Sätzen haben.

Wir freuen uns über ihren Lottogewinn.

An der Ecke wartet er auf seinen Freund.

Unser Interesse an der Lösung dieser Fragen ist sehr groß.

Er ist auf einem Ohr fast taub.

**Abb. 4:** Prompt zu Präpositionalphrasen

Die vorgestellten Aufgaben wurden um ausgewählte Grammatikübungen aus Lehrwerken für die Grundschule ergänzt. Die Aufgaben aus dem Primarbereich eignen sich besonders gut zum Testen expliziten Sprachwissens, da ihr Fokus auf der Wortarten- und Satzgliedbestimmung liegt. Gleichzeitig ist ihr Schwierigkeitsgrad dem Grundschulniveau angepasst und somit niederschwelliger als bei den anderen Aufgaben, was in Bezug auf die Performance von ChatGPT einen Vergleich ermöglicht.

## 4 Ergebnisse

In diesem Kapitel wird gezeigt, wie ChatGPT-4o mit den im vorausgehenden Kapitel dargestellten Aufgabenformaten umgegangen ist. Die Ergebnisse werden in Bezug auf die in Kapitel 2 genannten Wissensformen besprochen.

### 4.1 Implizites Sprachwissen

Aus den Ergebnissen der durchgeführten C-Tests geht hervor, dass das Tool bei Sachtexten deutlich besser abschnidet, als bei literarischen Texten ( $\chi^2(3) = 20.43, p < .001$ ). Während der Anteil korrekter Ergänzungen bei der Reportage und dem Essay bei über 97 % liegt, beträgt er bei dem Roman 68 %. Auch bei

dem Zeitungsartikel war der Anteil von zutreffend gefüllten Lücken mit 93 % hoch, dennoch etwas niedriger als bei den zwei anderen Sachtexten. Tabelle 1 fasst alle Ergebnisse zusammen und verschafft einen Überblick über das Verhältnis zwischen den zu ergänzenden Lücken und korrekten Antworten.

Textsorte	Anteil von korrekten Antworten	korrekt ergänzte Lücken/ alle Lücken
Roman	68 %	22/32
Essay	97 %	37/38
Reportage	97 %	34/35
Zeitungsartikel	93 %	30/32

**Tab. 1:** Ergebnisse der durchgeführten C-Tests

Diese Ergebnisse sind insofern überraschend, als generative Sprachmodelle wie ChatGPT-4o darauf trainiert werden, die Wahrscheinlichkeit des nächsten Tokens zu berechnen. In Anbetracht dessen könnte vermutet werden, dass ChatGPT-4o bei allen Textsorten entweder gleich gut oder gleich schlecht abschneidet. Kleine Abweichungen von dem Ausgangstext, wie diese bei Sachtexten beobachtet werden, werden auch in C-Tests erwachsener Muttersprachler festgestellt (Mashkovskaya, 2013). Im Fall des Romans liegt die Angemessenheit des Outputs allerdings unter 70 % und kann deshalb nicht auf individuelle Präferenzen zurückgeführt werden. Es ist nicht ausgeschlossen, dass die Sprache der verwendeten Quellentexte dabei eine entscheidende Rolle spielt. Ob sich vergleichbare Diskrepanzen auch im Englischen zeigen würden, d. h. in der Sprache der meisten Trainingsdaten, bleibt an der Stelle offen. Durch die Präsenz von stilistischen Mitteln wie Metaphern, Ellipsen oder Ironie weichen literarische Texte überdies von prototypischen Formulierungen ab, was sich auf die Angemessenheit des generierten Outputs auswirken kann. Um den festgestellten Zusammenhang zwischen der Textsorte und der Performance von ChatGPT zu bekräftigen, müssten noch weitere literarische Texte bzw. andere Fragmente des Romans in die Analyse miteinbezogen werden.

Auch bei Grammatikalitätsurteilen weichen die Antworten von ChatGPT-4o in einigen Fällen von standardsprachlichen Normen ab. Die Übereinstim-

mung zwischen seinen Klassifikationen und den in Referenzgrammatiken zu findenden Normen ist gering ( $\kappa=0.12$ ). Dennoch sind die Unterschiede zwischen ChatGPT-4o und den konsultierten Referenzgrammatiken statistisch nicht signifikant ( $\chi^2(4) = 3.19, p = .526$ ). Dies bedeutet, dass ChatGPT-4o zwar in vielen Fällen von sprachlichen Normen abweicht, aber dabei keine systematische Richtung manifestiert. So werden einige der dialektalen Formulierungen wie bspw. *Ich will ein lecker Pils trinken* oder *Ich habe gestern den ganzen Tag gelesen gehabt* als akzeptabel oder gar zielsprachlich eingestuft, während die für süddeutsche Dialekte typische Konstruktion *Der Mann, wo ich gestern gesehen habe, ist auch hier* unter die Kategorie inkorrekt fällt. Derselbe Unstimmigkeiten werden auch bei umgangssprachlichen Formulierungen beobachtet. Des Weiteren werden einige nicht-zielsprachliche Konstruktionen als korrekt (z. B. *Ich bin die ganze Nacht getanzt*) und zielsprachliche als inkorrekt klassifiziert (z. B. *Er bemalte sich als lustiger Clown*). Auf die Frage, warum der Satz *Sie war die Frau, die wo für mich die schönste war* als inkorrekt und nicht als akzeptabel eingestuft wurde, generiert ChatGPT-4o den folgenden Output: *Die Kombination „die wo“ ist weder in der Standardsprache noch in der formellen Schriftsprache korrekt. Sie kommt allerdings häufig in regionaler oder umgangssprachlicher Sprache vor, vor allem in süddeutschen, schweizerdeutschen und österreichischen Dialekten.* Diese Erläuterung deutet an, dass es sich dabei um eine dialektale Formulierung handelt und der Satz somit unter die Kategorie *nicht standardsprachlich, aber unter Umständen akzeptabel* fallen sollte. Dennoch wird der Satz als inkorrekt eingeordnet. Der Satz *Ich fahre dasselbe Auto wie Fritz, nur ist meins rot, seins weiß* wird von ChatGPT hingegen als grammatisch korrekt klassifiziert. Begründet wird diese Entscheidung mit dem folgenden Output: *das Wort „dasselbe“ bedeutet, dass es sich um ein identisches Objekt handelt, nicht nur ein ähnliches. In diesem Satz wird korrekt angegeben, dass das Auto-Modell identisch ist, der Unterschied jedoch in der Farbe liegt.* Bei dieser Erklärung zeigt sich ein gewisser Widerspruch, da der Ausdruck „dasselbe“ sich einerseits auf denselben Referenten und andererseits auf eine gemeinsame Eigenschaft beziehen sollte. Diese Ambiguität wird von ChatGPT jedoch nicht weiter thematisiert.

Bei den Korrekturaufgaben schneidet ChatGPT-4o besser ab, allerdings werden auch in dem Fall einige grammatische Fehler übersehen. Dies wird

insbesondere bei der Aufgabe mit einzelnen kontextlosen Sätzen beobachtet. Tabelle 2 zeigt, welche Fehler in den zu korrigierenden Sätzen identifiziert und welche Korrekturvorschläge gegeben wurden.

Satz	Fehler	Korrekturvorschlag durch KI
1 <i>Ich habe meine Cousine begegnet.</i>	Hilfsverb und Kasusmarkierung	<i>Ich bin meiner Cousine begegnet.</i>
2 <i>Der Verfolgte hoffte, sich zu erkennen geben dürfen.</i>	Syntax	<i>Der Verfolgte hoffte, sich erkennen geben zu dürfen.</i>
3 <i>Ich sehe mich als kompetenter Sprachbeherrscher an.</i>	Kasusmarkierung	<i>Ich sehe mich als kompetenter Sprachbeherrscher.</i>
4 <i>Das ist nur eine der vielen Möglichkeiten, die zur Verfügung steht.</i>	Numeruskongruenz	<i>Das ist nur eine der vielen Möglichkeiten, die zur Verfügung stehen.</i>
5 <i>Ich habe den Kegel umfahren, also ist mein Fahrrad jetzt kaputt.</i>	Falsches Partizip II	<i>Ich habe den Kegel umgefahren, also ist mein Fahrrad jetzt kaputt.</i>
6 <i>Ich überreiche diesen Kalender dir statt ihn.</i>	Kasusmarkierung	<i>Ich überreiche dir diesen Kalender anstelle von ihm.</i>

**Tab. 2:** Korrekturaufgabe I – isolierte Sätze

Wie aus Tabelle 2 hervorgeht, wurden Sätze 1, 4, 5 und 6 zielsprachlich korrigiert, wobei der Korrekturvorschlag für Satz 6 fraglich ist. Statt den Kasus des Personalpronomens „ihn“ zu ändern, formuliert ChatGPT den ganzen Satz um. Da in sonstigen Beispielen nicht alle Kasus- und Syntaxfehler erkannt werden, kann auch in dem Fall eine gewisse Inkonsequenz konstatiert werden.

Anders als bei Grammatikalitätsurteilen und Korrekturen isolierter Sätze geht das Tool bei der Korrektur der Schülertexte deutlich systematischer vor. Für jeden grammatischen und stilistischen Fehler bietet es nicht nur einen zielsprachlichen Korrekturvorschlag, sondern auch eine Begründung für seine Entscheidungen und weitere Hinweise. Dabei handelt es sich um Kasus- (1) und Genusfehler (2) sowie um Fehler bei der Pluralbildung (3) und bei der Verb- (4) und Adjektivflexion (5).

- (1) *Und stahlen ein Fernseher und ein Koffer.*
- (2) *Der Kennzeichen sah man nicht.*
- (3) *Und hatte zwei Computers im Hand.*
- (4) *Ich gibte mein Bericht bei der Polizei ab.*
- (5) *Und der Mann mit orangenem T-Shirt hat die geklaute Sachen in den Koffer reingepackt.*

Anders als im Fall von isolierten Sätzen, wurden bei dieser Aufgabe alle grammatischen Fehler erkannt. Angesichts der Funktionsweise von ChatGPT ist dieses Ergebnis wenig überraschend. Da Generative Sprachmodelle die Reihenfolgewahrscheinlichkeiten anhand eines vorgegebenen Kontexts berechnen, kann der Umfang der syntaktischen Umgebung die Angemessenheit der generierten Tokenfolge beeinflussen. Die generierten Korrekturvorschläge sind dabei nicht nur zielsprachlich, sondern auch übersichtlich dargestellt. Die Pfeile (*Computers* → *Computer*) markieren ganz deutlich die vorgenommenen Änderungen und vereinfachen somit deren Umsetzung. Neben sprachlichen Korrekturen generiert ChatGPT-4o auch folgende Hinweise hinsichtlich der Tempuskonstanz und Stilistik: „*ist*“ → „*war*“ (*Vergangenheitsform*); „*hat ... gesehen*“ → „*sah*“ (*Präteritum ist hier angemessener*); „*ohne zu gucken*“ → „*ohne zu schauen*“ (*stilistisch gehobener*).

## 4.2 Explizites Sprachwissen

Wie in Kapitel 3.2 angedeutet, beschränken sich die Aufgaben und Fragestellungen, die auf die Operationalisierung der Leistungen im Bereich des expliziten Sprachwissens abzielen, auf das Benennen, und Erklären grammatischer Strukturen und Phänomene. Tabelle 4 präsentiert die Ergebnisse der geschlossenen Aufgaben.

Selbstkonzipierte Aufgaben	Ergebnis (richtig/gesamt)
Wortartenbestimmung	23/26
Satzgliedbestimmung	6/6
Präpositionalphrasen und ihre Funktion	4/5

Grammatikübungen aus Lehrwerken	Ergebnis (richtig/gesamt)
Aufgabe 1 (Nomen, Verben und Adjektive erkennen)	11/11
Aufgabe 2 (Verben erkennen und konjugieren)	6/7
Aufgabe 3 (Wortarten ihren Definitionen zuordnen)	5/5
Aufgabe 4 (Satzglieder erkennen)	eine falsche Antwort (eine falsche Anzahl von Satzgliedern angegeben)

**Tab. 3:** Ergebnisse der geschlossenen Aufgaben zum expliziten Sprachwissen

Bei der Aufgabe zur Wortartenbestimmung werden insgesamt drei Fehler festgestellt. So wird der indefinite Artikel *einem* in *bei einem schönen Spaziergang* und in *in einem seltenen und schillernd bunten Schmetterling* als Pronomen identifiziert. Des Weiteren wird das flektierbare *unverhofft* in *Manchmal begegnet einem ein schönes, unbekanntes Wort so unverhofft (...)* (Hacke, 2006: 14) nicht als Adjektiv, sondern als Adverb klassifiziert.

Die Aufgabe zur Satzgliedbestimmung bereitet ChatGPT-4o deutlich weniger Schwierigkeiten – in dem Fall werden alle Beispiele richtig zugeordnet. Auch die Identifizierung von Präpositionalgruppen in vorgegebenen Sätzen ist unproblematisch. Der einzige Fehler tritt bei der Bestimmung von Satzgliedern auf. So wird die Präpositionalphrase *auf einem Ohr* in *Er ist auf einem Ohr fast taub* nicht als präpositionales Objekt, sondern als präpositionales Adverbial des Ortes identifiziert. Diese Zuordnung kann allerdings von der zugrunde gelegten grammatischen Theorie abhängen. Geilfuss/Ponitka (2020) sowie Musan (2021) schlagen vor, Präpositionalphrasen, deren Präpositionen nicht austauschbar sind und nicht mehr ihre Eigenbedeutung tragen, als Präpositionalobjekte zu betrachten. Obwohl bei dem angeführten Beispiel die genannten Voraussetzungen erfüllt sind, wird die fragliche Präpositionalphrase in anderen Grammatiken wie bspw. Helbig/Buscha (2001: 288) als präpositionales Adverbial klassifiziert.

Auch bei den Aufgaben aus den gängigen Lehrwerken für die Primarstufe schneidet ChatGPT-4o recht gut ab, wobei sich bei zwei Aufgaben einige Unstimmigkeiten zeigen. In Aufgabe 2, die darin besteht, in einem kurzen Text

Verben zu identifizieren und sie anschließend im Präsens und Präteritum zu konjugieren, wird das Modalverb *sollen* nicht aufgelistet, obwohl ein anderes Modalverb – *wollen* – richtig erkannt wird. Die vorgeschlagenen Konjugationsparadigmen hingegen sind für alle gefunden Verben zielsprachlich. Ein weiteres Problem wird bei der Aufgabe beobachtet, bei der ChatGPT-4o entscheiden sollte, aus wie vielen Satzgliedern der vorgegebene Satz besteht. In diesem Fall wird das Prädikat als Satzglied betrachtet und folglich als solches mitgezählt, was zu einer falschen Antwort führt. Auch an dieser Stelle scheint die grammatische Theorie, auf der der Output von ChatGPT basiert, ausschlaggebend. So kategorisieren sowohl Helbig/Buscha (2001: 448) als auch viele bundesweit zugelassene Lehrwerke für die Grundschule Prädikate als Satzglieder. Beruht ChatGPT überwiegend auf derlei Quellen, so ist es nicht überraschend, dass seine Einordnung von den aktuellen linguistischen und sprachdidaktischen Erkenntnissen abweicht (u. a. Musan, 2021).

Besonders interessante Einblicke in die Performance von ChatGPT-4o liefern seine Antworten auf die offenen Fragen. Die Erklärung des Unterschieds zwischen Satzgliedern und Wortgruppen ist linguistisch fundiert und wird mit passenden Beispielen unterfüttert. Dabei wird auch zwischen der Form- und Funktionsebene differenziert. Bei Satzgliedern wird überdies die Umstellprobe als Methode zur Satzgliederkennung genannt. Seine ausführlichen Erläuterungen fasst ChatGPT-4o folgendermaßen zusammen: *Wortgruppen beschreiben die grammatische Struktur und Zusammensetzung von Wörtern. Satzglieder beschreiben die Funktion dieser Einheiten im Satz.* Diese Zusammenfassung ist jedoch problematisch, da der Ausdruck *die Zusammensetzung von Wörtern* sich auch auf die Wortbildung beziehen kann. Demzufolge ist nicht ganz klar, ob dabei Wortgruppen oder Wörter definiert werden.

Auf die Frage, was Verben sind und welche Funktion sie im Satz erfüllen, gibt das Tool eine umfangreiche Antwort. Dabei geht es nicht nur auf die Begriffserklärung ein, sondern auch auf verschiedene Verbklassen und Verbtypen. Die Definition an sich ist jedoch stark semantisch geprägt und blendet somit morpho-syntaktische Merkmale von Verben aus: *Ein Verb ist eine Wortart, die eine Handlung, einen Vorgang, einen Zustand oder ein Geschehen ausdrückt. Es ist eine der zentralen Wortarten in der deutschen Sprache und bildet das Herzstück eines Satzes.* Ein Versuch, auf die morpho-syntaktischen

Merkmale von Verben einzugehen, lässt sich ausschließlich an der Formulierung *Herzstück eines Satzes* erkennen.

Eine rein semantische Begriffserklärung wird auch bei der Frage nach dem Prädikativum beobachtet: *Ein Prädikativum (oder auch Prädikativ) ist ein Satzglied, das eine Eigenschaft, einen Zustand oder eine Identität des Subjekts oder Objekts beschreibt.* Als Nachtrag zu seiner Antwort nennt ChatGPT-4o mögliche Verben, die ein Prädikativum als Ergänzung erfordern. Allerdings werden unter der Kategorie *Kopulaverben* sowohl Kopulaverben als auch kopulaähnliche Verben aufgelistet. In dem zusammenfassenden Teil wird darauf hingewiesen, dass das Prädikativum auch *in Verbindung mit anderen Verben* auftritt. Um welche Verben es sich dabei handelt und worin sich Kopulaverben von kopulaähnlichen Verben unterscheiden, wird offengelassen. Bei der Erklärung, wodurch der Kasus des Prädikativums zustande kommt, differenziert ChatGPT-4o zwischen Subjekt- und Objektprädikativum und nennt Nominativ, Akkusativ und Dativ als mögliche Kasus. Bei der Kasusbestimmung von Prädikativa sollte schließlich die syntaktische Funktion des Bezugselements helfen: *Kasus des Prädikativums wird von der Funktion im Satz und vom Subjekt bzw. Objekt bestimmt, auf das es sich bezieht.*

Die größten Schwierigkeiten hat ChatGPT-4o mit der Begründung seiner Korrekturvorschläge (Abb. 5).

Warum ist der Satz "Der Affe fährt nicht mit." korrekt, aber der Satz "Der Affe nehme ich nicht mit." nicht?

**Abb. 5:** Offene Frage 4 (Prompt) (angelehnt an Wegener 1995)

ChatGPT kann zwar zwischen einem grammatischen und einem ungrammatischen Satz differenzieren, aber die Quelle des Problems nicht immer identifizieren. Als Grund, warum der Satz *Der Affe nehme ich nicht mit* ungrammatisch ist, nennt das Tool fehlende Kongruenz in Person und Numerus zwischen dem Prädikat *nehme* und dem Subjekt *der Affe*: *Grammatikalische Regeln verlangen eine Kongruenz zwischen Subjekt und Prädikat, die in diesem Fall verletzt wurde.* Die Nominalgruppe *der Affe* wird also fälschlich als Subjekt des gesamten Satzes identifiziert. Damit werden auch die syntagmatischen Re-

lationen Kongruenz und Rektion vertauscht. Als Korrektur schlägt ChatGPT-4o vor, den zweiten Satz zu einem Passivsatz umzuformulieren: *Der Affe wird nicht mitgenommen*. Erst wenn explizit nach dem Subjekt des Satzes gefragt wird, generiert das Tool den Hinweis, dass es sich dabei um das Personalpronomen *ich* handelt: *Im Satz „Der Affe nehme ich nicht mit.“ ist grammatisch gesehen „ich“ das Subjekt, weil das Prädikat „nehme“ zur 1. Person Singular gehört. Aber die ungrammatische Konstruktion macht die Satzglieder inkongruent und sorgt für Verwirrung*. Dabei wird betont, dass eine Akkusativmarkierung nötig wäre, wenn die Nominalgruppe *ich* das eigentliche Subjekt sein sollte. Dennoch wird die Rektion des Verbes *mitnehmen* und die daraus resultierende falsche Kasusmarkierung nicht explizit thematisiert.

## 5 Didaktische Implikationen und Ausblick

Das Ziel des vorliegenden Experiments war es, die Performance von ChatGPT-4o im Bereich der deutschen Grammatik zu untersuchen und darauf aufbauend mögliche Anwendungsbereiche im schulischen Grammatikunterricht zu bestimmen. Die Ergebnisse der durchgeführten Studie zeigen, dass ChatGPT-4o Ergänzungs-, Korrektur- und Kategorisierungsaufgaben ganz gut erfüllen kann, sofern es sich um standardsprachliche Formulierungen und prototypische Fälle handelt. Bei Grenzfällen und sprachlichen Randphänomenen werden nicht alle Fehler erkannt. Besonders deutlich zeigt sich das bei Grammatikalitätsurteilen, wo Beispiele aus verschiedenen Registern und Dialekten analysiert werden sollten. In Bezug auf Aufgaben, denen deklarative Handlungen zugrunde liegen, erzielt ChatGPT-4o zwar gute Ergebnisse. Bei Fragen, die darauf abzielen, sprachliche Phänomene zu erläutern, ist sein Output jedoch weniger zutreffend. Die formulierten Definitionen sind oft semantisch geprägt und blenden somit die für die korrekte Wortarten- und Satzgliedbestimmung notwendigen morpho-syntaktischen Aspekte aus. Auch Aufgaben, die eine Begründung für den generierten Output erfordern, bereiten ChatGPT-4o Schwierigkeiten.

Angesichts der Funktionsprinzipien des Tools sind diese Ergebnisse wenig überraschend. Generative Sprachmodelle sind darauf ausgelegt, anhand des

vorausgehenden Kontexts den wahrscheinlichsten Wert des Folgewortes zu ermitteln. Damit wird ausschließlich die Form- und nicht die Inhaltsseite eines sprachlichen Zeichens verarbeitet (Fürstenberg/Müller 2024; Müller/Fürstenberg 2023). Ihre Leistung ist also auf statistische Berechnungen beschränkt. Aus diesem Grund kann nicht davon ausgegangen werden, dass GPTs etwas wissen oder verstehen. Dies erklärt die festgestellten Defizite in der Kohärenz des Outputs bei Grammatikalitätsurteilen und offenen Fragen.

Die zentrale Frage, die sich im Hinblick auf diese Ergebnisse stellt, ist, ob der Einsatz von ChatGPT-4o im Grammatikunterricht sich dennoch als lohnenswert erweisen kann. Gute Leistungen des Tools in Ergänzungs- und Korrekturaufgaben können vor allem für sprachheterogene Lerngruppen gewinnbringend sein. So können Schüler:innen, deren Deutschkenntnisse noch nicht auf muttersprachlichem Niveau sind, die durch ChatGPT-4o generierten Korrekturvorschläge, Sprachhinweise und Übersetzungen als Hilfestellung bei Hausaufgaben, insbesondere bei der Textbearbeitung (s. Steinhoff 2025) verwenden und dahingehend an ihren Sprachkenntnissen arbeiten. Die Arbeit mit ChatGPT-4o bietet des Weiteren viele Anlässe zur Sprachreflexion. Ein Beispiel für eine sprachreflexive Aufgabe wäre eine Diskussion über die Entscheidungen und Grammatikalitätsurteile von ChatGPT-4o in Bezug auf sprachliche Varietäten und Register. Davon würden insbesondere jüngere Schüler:innen profitieren, die noch nicht sicher zwischen unterschiedlichen Sprachvarietäten differenzieren und folglich dialektale oder umgangssprachliche Formulierungen für standardsprachlich halten können. Auch bei Aufgaben zur Entwicklung von Prompting-Kompetenzen zeigen sich Möglichkeiten zum kreativen und spielerischen Umgang mit der Sprache. Durch die Erprobung verschiedener Prompts können die Schüler:innen sehen, welche Formulierungen zum gewünschten Output führen. So lernen sie, ihre Gedanken und Fragestellungen strukturierter und präziser zu formulieren, was nicht nur für den Umgang mit KI-Anwendungen, sondern auch für den Erwerb von Kommunikationskompetenzen relevant ist. All diese Aufgabenformate und Betrachtungsaktivitäten können schließlich die Entfaltung der Sprachbewusstheit fördern, die eines der zentralen Ziele des Grammatikunterrichts darstellt.

Auch in Bezug auf explizites Sprachwissen bietet ChatGPT eine Unterstützung für Lehrkräfte und Schüler:innen. So ist es bspw. möglich, mit dem

Tool zusätzliche Übungen und Aufgaben zu generieren, die das bereits erworbene Wissen überprüfen. Anhand der Hinweise zu Lernzielen und Wissensbeständen der Gruppe kann ChatGPT Grammatikübungen erstellen, die auf individuelle Wissenslücken der Schüler:innen eingehen. Dies ermöglicht den Lehrkräften, effizient und zeitökonomisch für Differenzierung zu sorgen, was in sprachheterogenen Lerngruppen üblicherweise eine Herausforderung darstellt. Eine weitere Möglichkeit stellen die sog. benutzerdefinierten GPTs (Custom GPTs) dar, die auf die Lernbedürfnisse der Schüler:innen zugeschnitten werden und sie bei der Bearbeitung ausgewählter Aufgaben begleiten können. Um dabei potenzielle Fehler und die bereits genannten Probleme zu vermeiden, können die Lehrkräfte ihre eigenen Inhalte wie Grammatikskripts hochladen und ChatGPT bei der Generierung seines Outputs darauf basieren lassen.

Abschließend lässt sich konstatieren, dass ChatGPT-4o trotz seiner Einschränkungen einige Einsatzmöglichkeiten im Grammatikunterricht anbietet, die sich vor allem in sprachheterogenen Lerngruppen bewähren können. Das Tool kann zum einen als Sprachfördertool und Grammatiktutor verwendet werden. Zum anderen ermöglicht es den Schüler:innen, sich mit sprachlichen Strukturen auf der Wort-, Satz- und Textebene auseinanderzusetzen, über linguistische Phänomene zu diskutieren und mit Sprache zu experimentieren. Ob und ggf. wie sich der Einsatz von ChatGPT auf das grammatische Wissen von Schüler:innen auswirkt, bleibt an dieser Stelle offen und müsste in weiterführenden Studien untersucht werden.

## Literatur

### Primärliteratur

- BRUCKLACHER, DAVID (2025): Teuer erkaufte Ruhe. In: *Frankfurter Allgemeine Zeitung* (18.01.2025). URL: <https://www.faz.net/aktuell/gesellschaft/gesundheit/warum-kleinkinder-nicht-vor-dem-bildschirm-sitzen-sollten-110221127.html>.
- FROMM, ERICH (2019): *Haben oder Sein. Die seelischen Grundlagen einer neuen Gesellschaft*. München: dtv.

- HACKE, AXEL (2006): *Das Beste aus meinem Leben. Mein Alltag als Mann*. München: Antje Kunstmann.
- JELINEK, ELFRIEDE (2011): *Die Klavierspielerin*. Reinbek bei Hamburg: Rowohlt.
- KAPUŚCIŃSKI, RYSZARD (2010): *Die Erde ist ein gewalttätiges Paradies*. München: Pieper.

## Sekundärliteratur

- ACAR, OGUZ A. (2024): Prompt-Engineering und mehr: Schlüsselkompetenzen für den erfolgreichen Einsatz von generativer KI im Marketing. In: *NIM Marketing Intelligence Review*, 16/1, S. 18–23. <https://doi.org/10.2478/nimmir-2024-0003>.
- ANDRESEN, HELGA/FUNKE, REINOLD (2006): Entwicklung sprachlichen Wissens und sprachlicher Bewusstheit. In: Bredel, Ursula/Günther, Hartmut/Klotz, Peter/Ossner, Jakob/Siebert-Ott, Gesa (Hg.): *Didaktik der deutschen Sprache. Ein Handbuch*. (Bd. 1.). Paderborn: Schöningh, S. 438–451.
- BALCIK, INES/HESSE, ELKE (2023): *Dudenkalender 2024*. Dortmund: Athesia Kalenderverlag.
- BECKER-MROTZEK, MICHAEL/GRABOWSKI, JOACHIM (2018): Textkorpus Scriptoria. In: Michael Becker-Mrotzek und Joachim Grabowski (Hg.). *FD-LEX (Forschungsdatenbank Lernertexte)*. Köln: Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache. Verfügbar unter: <https://fd-lex.uni-koeln.de>. DOI: 10.18716/FD-LEX/861.
- BINANZER, ANJA/GAMPER, JANA/KÖPCKE, KLAUS-MICHAEL (2015): Linguistische und affektive Anforderungen an Lehramtsanwärterinnen und -anwärter: Untersuchungen im Bereich Deutsch als Zweitsprache. In: Fischer, Christian/Weber, Marcel/Fischer-Ontrup, Christiane/Buschmann, Rafael (Hg.): *Umgang mit Vielfalt: Aufgaben und Herausforderungen für die Lehrerinnen- und Lehrerbildung*. Münster: Waxmann, S. 293–320.
- BINANZER, ANJA/LANGLOTZ, MIRIAM (2018): Grammatik. In: Boelmann, Jan (Hg.): *Empirische Forschung in der Deutschdidaktik*. Baltmannsweiler: Schneider Hohengehren, S. 303–320.
- BINANZER, ANJA/WECKER, VERENA (2020): Lernergrammatik und Grammatikunterricht – Überlegungen zu natürlichen Erwerbssequenzen und didaktischer Progression. In: Langlotz, Miriam (Hg.): *Grammatikdidaktik: Theoretische und*

- empirische Zugänge zu sprachlicher Heterogenität*. Baltmannsweiler: Schneider Verlag Hohengehren, S. 61–86.
- BREDEL, URSULA (2013): *Sprachbetrachtung und Grammatikunterricht*. Paderborn: Schöningh.
- BREMERICH-VOS, ALBERT (1981): Sprachunterricht – nein danke! In: *Der Deutschunterricht* 23/6, S. 5–19.
- BREMERICH-VOS, ALBERT (1999): Zum Grammatikunterricht in der Grundschule: wie gehabt, gar nicht, anders? In: Bremerich-Vos, Albert (Hg.): *Zur Praxis des Grammatikunterrichts. Mit Materialien für Lehrer und Schüler*. Freiburg i.Br.: Filibach, S. 12–80.
- CHEN, BANGHAO/ZHANG, ZHAOFENG/LANGRENÉ, NICOLAS/ZHU, SHENGXIN (2024): *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*. URL: <https://doi.org/10.48550/arXiv.2310.14735> (letzter Zugriff: 06.03.2025).
- CONIAM, DAVID (2008): Evaluating the language resources of chatbots for their potential in English as a second language. In: *ReCALL* 20, S. 98–116.
- DAI, WEI et al. (2023): Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In: *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Orem, UT, USA, 2023, S. 323–325.
- DÄMMER, JUTTA (2019): *Grammatisches Fachwissen von Lehramtsstudierenden des Faches Deutsch*. Baltmannsweiler: Schneider Hohengehren.
- DIEHL, ERIKA et al. (2000): *Grammatikunterricht: Alles für die Katz? Untersuchen zum Zweitspracherwerb Deutsch*. Tübingen: Max Niemeyer Verlag.
- DODIGOVIC, MARINA (2007): Artificial intelligence and second language learning: An efficient approach to error remediation. In: *Language Awareness*, 16/2, S. 99–113. doi: 10.2167/la416.0.
- Dudenredaktion (Hg.) (2009): *Duden Band 4. Die Grammatik*. 8. Aufl. Mannheim/Zürich: Dudenverlag.
- Dudenredaktion (Hg.) (2011): *Duden Band 9. Richtiges und gutes Deutsch. Das Wörterbuch der sprachlichen Zweifelsfälle*. 7. Aufl. Mannheim/Zürich: Dudenverlag.
- DÜRSCHIED, CHRISTA (2007): Damit das grammatische Abendland nicht untergeht. Grammatikunterricht auf der Sekundarstufe II. In: Köpcke Klaus-Michael/Ziegler, Arne (Hg.): *Grammatik in der Universität und für die Schule. Theorie Empirie und Modellbildung*. Tübingen: Niemeyer, S. 45–65.

- FEILKE, HELMUTH/TOPHINKE, DORIS (2016): Grammatisches lernen. In: *Praxis Deutsch* 43, S. 4–11.
- FIX, MARTIN (2000): *Textrevisionen in der Schule*. Baltmannsweiler: Schneider Hohengehren.
- FRENTZ, HARALD/LEHMANN, CHRISTINA (2003): Der gymnasiale Lernbereich „Reflexion über Sprache“ und das Hochschulzugangsniveau für sprachliche Fähigkeiten. In: *Didaktik Deutsch*, 14, S. 92–98.
- FRYER, LUKE/CARPENTER, ROLLO (2006): Emerging technologies: Bots as language learning tools. In: *Language Learning and Technology*, 10/3, S. 8–14.
- FRYER, LUKE et al. (2017): Stimulating and sustaining interest in a language course: An experimental comparison of Chatbot and Human task partners. In: *Computers in Human Behavior*, 75, S. 461–468.
- FUNKE, REINOLD (2005): *Sprachliches im Blickfeld des Wissens. Grammatische Kenntnisse von Schülerinnen und Schülern*. Tübingen: Niemeyer.
- FUNKE, REINOLD (2014): Grammatikunterricht, grammatisches Wissen und schriftsprachliches Können. In: Hildegard Gornik (Hg.). *Sprachreflexion und Grammatikunterricht*. Baltmannsweiler: Schneider Hohengehren, S. 429–454.
- FUNKE, REINOLD (2018): Working on grammar at school: empirical research from German-speaking regions. Contribution to a special issue ‚Working on Grammar at School in L1-Education. Empirical Research across Regions.‘ In: *L1-Educational Studies in Language and Literature*, 18, S. 1–39.
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht*, 5, S. 2–13.
- GEILFUSS-WOLFGANG/PONITKA, SANDRA (2020): *Der einfache Satz*. Tübingen: Narr.
- GODA, YOSHIKO et al. (2014): Conversation with a chatbot before an online EFL group discussion and the effects on critical thinking. In: *Journal of Information Systems Education*, 13/1, S. 1–7.
- GRAMZOW EMDEN, MATTHIAS (2013): *Deutsche Grammatik verstehen und unterrichten*. Tübingen: Narr Verlag.
- GRIMM, ANGELA/SCHULZ, PETRA (2014): Sprachfähigkeiten von Kindern mit DaZ bei Schuleintritt. In: Lütke, Beate/Petersen, Inge (Hg.): *Deutsch als Zweitsprache: erwerben, lernen und lehren. Beiträge zum 9. Workshop Kinder mit Migrationshintergrund*. Stuttgart: Klett Fillibach, S. 35–50.

- HAVERKAMP, HENDRIK/HECHT, MALTE/SCHINDLER, KIRSTEN (2024): Lernförderliches Feedback KI-basiert vermitteln. In: *Deutschunterricht*, 5, S. 60–71.
- HELBIG, GERHARD/BUSCHA, JOACHIM (2001): *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin/München: Langenscheidt.
- IVO, HUBERT/NEULAND, EVA (1991): Grammatisches Wissen. Skizze einer empirischen Untersuchung über Art, Umfang und Verteilung grammatischen Wissens (in der Bundesrepublik). In: *Diskussion Deutsch*, 22/121, S. 437–493.
- KAMMLER, CLEMENS/SWITALLA, BERND (2001): Qualität des Deutschunterrichts auf der gymnasialen Oberstufe – Kernkompetenzen. In: Tenorth, Heinz-Elmar (Hg.): *Kerncurriculum Oberstufe. Mathematik – Deutsch – Englisch. Expertisen im Auftrag der Konferenz der Kultusminister*. Weinheim: Beltz, S. 103–123.
- KASNECI, ENKELEJDA et al. (2023): ChatGPT for good? On opportunities and challenges of large language models for education. In: *Learning and Individual Differences*, 103, S. 1–9.
- Kultusministerkonferenz (KMK) (2022): *Bildungsstandards für das Fach Deutsch Primarbereich*. URL: <https://www.kmk.org/themen/qualitaetssicherung-in-schulen/bildungsstandards.html>.
- LEHNEN, KATRIN/STEINHOFF, TORSTEN (2024): Digitales Lesen und Schreiben. In: Androutopoulos, Jannis/Vogel, Friedmann (Hg.): *Handbuch Sprache und digitale Kommunikation*. Berlin/Boston: de Gruyter, S. 527–546.
- LEIFHEIT, LUZIA et al. (2024): KI im Unterricht – Entwicklung von Lehrveranstaltungen für Lehramtsstudierende der Sprach- und MINT-Fächer. In: *Ludwigsburger Beiträge zur Medienpädagogik – LBzM*, 24, S. 1–19. doi.org/10.21240/lbzm/24/08.
- LEWIS, PATRICK; DENOYER, LUDOVIC; RIEDEL, SEBASTIAN (2019): Unsupervised Question Answering by Cloze Translation. In: *Facebook Research*, S. 4896–4910. doi:10.18653/v1/P19-1484. S2CID 186206974.
- LÜTKE, BEATE (2014): From meta-process to conscious access. Metasprachliche Fähigkeiten im frühen Zweitspracherwerb aus theoretischer, empirischer und sprachdidaktischer Perspektive. In Klages, Hana/Pagonis, Giulio (Hg.): *Linguistisch fundierte Sprachförderung und Sprachdidaktik: Grundlagen, Konzepte, Desiderate. DaZ-Forschung – Deutsch als Zweitsprache, Mehrsprachigkeit und Migration 7*. Berlin u. a.: De Gruyter, S. 99–122.
- MASHKOVSKAYA, ANNA (2013): *Der C-Test als Lesetest bei Muttersprachlern*. Dissertationsschrift Universität Duisburg-Essen.

- MÜLLER, ANJA/GEYER, SABRINA (2020): Gut vorbereitet auf den Grammatikunterricht in der Schule? Eine empirische Untersuchung zum schulgrammatischen Wissen angehender Deutschlehrkräfte. In: Masanek, Nicole/Kilian, Jörg (Hg.): *Professionalisierung im Lehramtsstudium. Überzeugungen, Wissen, Aushandlungsprozesse. Reihe: Positionen der Deutschdidaktik. Theorie und Empirie*. Bern: Peter Lang.
- MÜLLER, HANS-GEORG/FÜRSTENBERG, MAURICE (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Germanistenverbandes*, 4/70, S. 327–345.
- MUSAN, RENATE (2021): *Satzgliedanalyse*. Heidelberg: Universitätsverlag.
- PARADIS, MICHEL (2009): *Declarative and procedural determinants of second languages (Studies in Bilingualism 40)*. Amsterdam: John Benjamins.
- PETRONI, FABIO et al. (2019): Language Models as Knowledge Bases? In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, S. 2463–2473.
- RISEL, HEINZ (1999): Schlaglichter auf Wissensbestände. Anmerkungen zu „Grammatikbiographien“. In: Klotz, Peter/Peyer, Ann (Hg.): *Wege und Irrwege sprachlich-grammatischer Sozialisation*. Baltmannsweiler: Schneider Hohengehren, S. 53–60.
- SCHMITZ, ULRICH (2003): Satzzeigen. Wie oft kann das grammatische Abendland noch untergehen? In: *Mitteilungen des deutschen Germanistenverbandes*, 2–3/50, S. 452–458.
- SCHULHOFF, SANDER et al. (2025): The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. URL: <https://doi.org/10.48550/arXiv.2406.06608>.
- SCHULZE, KORDULA (2011): Brauchen Grundschullehrer grammatische Kompetenzen? In: Köpcke, Klaus-Michael/Noack, Christina (Hg.) (2010): *Sprachliche Strukturen thematisieren. Sprachunterricht in Zeiten der Bildungsstandards*. Baltmannsweiler: Schneider Hohengehren, S. 27–46.
- Ständige Wissenschaftliche Kommission (SWK) (2024): *Large Language Models und ihre Potenziale im Bildungssystem Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz*. URL: [https://www.swk-bildung.org/content/uploads/2024/02/SWK-2024-Impulspapier\\_LargeLanguageModels.pdf](https://www.swk-bildung.org/content/uploads/2024/02/SWK-2024-Impulspapier_LargeLanguageModels.pdf).
- STEINHOFF, TORSTEN (2024a): ChatGPT: Das Ende des Schreibunterrichts? In: *BiSS-Journal*, 19, S. 20–23.
- STEINHOFF, TORSTEN (2024b): Macht Künstliche Intelligenz den Schreibunterricht überflüssig? In: Schicker, Stephan/Akbulut, Muhammed/Reinsperger, Victoria/

- Hendler, Melanie (Hg.): *Zusammen gedacht. Transdisziplinäre Perspektiven auf Literalität und Schreiben in Deutsch im Kontext von Mehrsprachigkeit. Festschrift für Sabine Schmölzer-Eibinger*. Weinheim/Basel: Beltz Juventa, S. 179–183.
- STEINHOFF, TORSTEN (2025): Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel des Schreibens mit ChatGPT in der 8. Klasse. In: Albrecht, Christina/Brüggemann, Jörn/Kretschmann, Tabea/Meier, Christian (Hg.): *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven*. Stuttgart: Metzler, S. 85–102.
- TAYLOR, WILSON L. (1953): Cloze procedure: a new tool for measuring readability. In: *Journalism Quarterly*, S. 415–433.
- UNTERHOLZNER, FRANZ/MÜLLER, HANS-GEORG (2023): Metakognition als Brücke zwischen sprachlichem Wissen und Können. In: *Didaktik Deutsch*, 55, S. 20–38.
- WEGENER, HEIDE (1995): Das Genus im DaZ-Erwerb: Beobachtungen an Kindern aus Polen, Russland und der Türkei. In: Handwerker, Brigitte (Hg.): *Fremde Sprache Deutsch: grammatische Beschreibung – Erwerbsverläufe – Lehrmethodik*. Tübingen: Narr, S. 1–24.

## „Für Matura-Ausarbeitungen war es wirklich ein Lifesaver!“

### KI und Schreiben in der Schule

#### Abstract

In diesem Beitrag wird untersucht, wofür Schüler:innen die von einer KI erstellten Schreibprodukte in der Schule nutzen und wie sie diese Nutzung (subjektiv) bewerten. Als Datengrundlage dienen Gruppendiskussionen mit Schüler:innen einer Sekundarstufe II. Die Teilnehmer:innen verweisen auf die Verwendung der KI als Writing Partner, d. h. um Texte umschreiben oder verbessern zu lassen. Am häufigsten verwenden sie KI jedoch als „Ghostwriter“, der Texte eigenständig als Antwort auf einen passenden Prompt verfasst. Dabei machen sie die Erfahrung, dass man durch die Nutzung von KI eine bessere Note erreichen kann. Trotzdem sind sie sich in der Diskussion unsicher, ob eine solche Verwendung „fair“ ist.

**Schlagwörter:** Schreiben, Leistungen vortäuschen, Sekundarstufe, KI, Gruppendiskussionen

## 1 Einleitung

Die neuesten technischen Entwicklungen innerhalb der Künstlichen Intelligenz (KI) verändern Schule, Universität und alle Bereiche des wirtschaft-

.....

- 1 Die Herausgeber haben für diesen Beitrag wertvolle sprachliche und textsortenspezifische Hilfestellungen gegeben, weswegen ich die beiden hier als meine *Writing Tutors* (s. Kap. 2.1) nennen möchte.

lich-kulturellen und politisch-sozialen Lebens grundlegend. Für die Schule ist besonders relevant, dass KI-Modelle (wie bspw. ChatGPT) in der Lage sind, verschiedene Textsorten zu den unterschiedlichsten Themenbereichen in vielen Sprachen in Sekundenschnelle zu produzieren. Dadurch wird – so wie in der tertiären Bildung – das Schreiben in allen Fächern, aber insbesondere in den Fächern mit einer Sprache als Gegenstand<sup>2</sup> nicht nur beeinflusst, sondern gleichzeitig wie die Gesellschaft im Gesamten revolutionär „transformiert“ (vgl. Fürstenberg/Müller 2024; Allier et al. 2024; Williamson/Hogan 2021); KI muss als Game-Changer aufgefasst werden (Schönbächler et al. 2023: 1). Die breite (globale) Verfügbarkeit von textgenerierender KI innerhalb weniger Monate hat eben nicht nur das Bildungswesen, sondern genauso die Arbeitswelt, die Freizeit etc. in ihren Routinen bei der Produktion, Rezeption und Distribution von Texten erschüttert (vgl. Bräuer/Hollosi-Boiger 2024).

In diesem Beitrag soll untersucht werden, wofür Schüler:innen (SuS) die von einer generativen künstlichen Intelligenz erstellten Schreibprodukte in Bezug auf ihre schulischen Aufgaben nutzen. Dazu werden erste Ergebnisse aus einer explorativen, qualitativen Untersuchung mit 15- bis 20-jährigen SuS aus Wien präsentiert. Die Jugendlichen sprachen in Gruppeninterviews anhand von fünf Leitfragen über ihre Verwendung von KI, insbesondere ChatGPT, in dem jeweiligen Schuljahr<sup>3</sup>. In dieser Publikation sollen insbesondere folgende Forschungsfragen beantwortet werden:

- Wie nutzen SuS in der Sekundarstufe KI im Rahmen ihrer schriftlichen schulischen Aufgaben, und zwar aus ihrer subjektiven Sicht?
- Wie bewerten SuS in der Sekundarstufe ihre eigene bzw. die Nutzung von KI im Rahmen von schriftlichen schulischen Aufgaben?

.....

- 2 Es gibt Sprache(n) in den Fächern und Sprache(n) als Fächer, d. h. in diesen ist Sprache der Gegenstand des Fachs (vgl. Boeckmann et al. 2014). In Österreich wären das als Unterrichtssprache Deutsch (oder eine der anerkannten Minderheitensprachen wie bspw. Kroatisch), Englisch, Französisch, aber auch Latein etc.
- 3 Die Gruppeninterviews fanden 2023 und 2024 statt (siehe Kapitel 4).

## 2 KI beim Schreiben und Prüfen in der Schule

### 2.1 KI und Schreiben in der Schule

Basierend auf den Überlegungen von Steinhoff (2025) untersucht dieser Beitrag verschiedenen Rollen, die KI für die SuS übernimmt. Steinhoff nennt für den Kompetenzbereich *Schreiben* folgende Funktionen:

- Als *Ghostwriter* übernimmt die KI die Textproduktion vollständig.
- Als *Writing Partner* begleitet die KI den Schreibprozess, d.h. ko-produziert Text.
- Als *Writing Tutor* unterstützt die KI und bietet Hilfestellungen im Schreibprozess, produziert aber keinen Text.

Ähnlich dazu, aber mit einem alle Kompetenzen im Fach Deutsch umfassenden Blick, benennen Fürstenberg/Müller (2024) die Rollen der KI als *Taschenrechner*, *Sparringspartner* und *Tutor*. Dabei wird von den Autoren – zumindest für den Moment – die Rolle als Tutor nur in wenigen Fällen als möglich angesehen. In bspw. den Fächern Mathematik oder den schulischen Fremdsprachen lässt sich die KI (in gewissem Maße) als Tutor nutzen, denn diese Fächer weisen eine relativ klare Progression bei Inhalt und Schwierigkeitsgrad auf (vgl. die Kompetenzbeschreibungen des GERS<sup>4</sup> für den Fremdspracherwerb). Zur Tutoren-Rolle der KI gibt es bspw. die Studien von Duong/Chen (2025) zu Englisch als Fremdsprache in der Schule oder Dornicheva/Sulimova (2025) zu Russisch als universitäre Fremdsprache. Somit sind für das Fach Deutsch die jetzigen Rollen der KI in erster Linie der Ghostwriter/Schulfach oder Writing/Sparring Partner (vgl. Fürstenberg/Müller 2024).

Es gibt Autor:innen, die einen möglichen, positiv bewerteten Mehrwert von KI aufzeigen. Andere betonen eher die Sorge, dass die Nutzung von KI das tiefgehende Verständnis und eine kritische Auseinandersetzung mit Texten

.....

4 Der Gemeinsame Europäische Referenzrahmen für Sprachen (Europarat 2001) hat für den Fremdspracherwerb die Sprachniveaus, Sprachniveaubeschreibungen und somit auch die Progression in den Lehrwerken europaweit vereinheitlicht.

erschwert (bspw. bei Magirius et al. 2024). Insbesondere besteht die Gefahr, dass sich SuS zunehmend und einseitig auf KI verlassen, was aufgrund der Unzuverlässigkeit der generierten Inhalte (Halluzinationen) eine ungünstige Entwicklung darstellt (vgl. Wahrheit vs. Wahrscheinlichkeit bei Fürstenberg/Müller 2024 oder Desinformation bei Murgia et al. 2024). Im Sinne einer umfassenden Digital Literacy (vgl. Grabensteiner et al. 2023) und eines kritischen und reflektierten Literacy Managements (vgl. Bräuer/Hollosi-Boiger 2024) ist ein Paradigmenwechsel im Bildungssystem erforderlich. Es ist notwendig, die zurzeit andauernden, dynamischen und schnellen Veränderungen durch die Nutzung von KI in der individuellen und institutionellen literalen Praxis zu beobachten, zu untersuchen sowie die Ausbildung der literalen Praktiken in Schulen an diese neue Situation anzupassen. Dazu kann diese Publikation einen Beitrag leisten.

Die Herausforderung für Schulen besteht darin, den Einsatz von KI so zu gestalten, dass sie als Lernhilfe dient, ohne das eigenständige Denken und Schreiben zu untergraben<sup>5</sup>. Dazu sind neue didaktische Konzepte erforderlich, die an unterschiedlichen Orten ausprobiert werden. So soll *Prompty* als didaktisches Tool SuS ethisches und verantwortliches Prompten beibringen (vgl. Dennison et al. 2024). Auch wird es wichtig sein, immer wieder aufzuklären, dass GPTs nicht wie Menschen denken und weiters keine Intelligenz besitzen, sondern nur Wahrscheinlichkeiten berechnen (vgl. Fürstenberg/Müller 2024) und deswegen das eigene, auch kritische Denken nicht auf die KI ausgelagert werden kann (vgl. Schmidt/Seegel 2024). Auch können Menschen nur dann epistemisch-heuristisch durch das Schreiben und während des Schreibprozesses lernen, wenn das Schreiben nicht im Sinne eines Ghostwriting von GPTs übernommen wird (vgl. Buck et al. 2024). Ferner gilt, dass die Technik keine Lösung für die Verminderung von Ungleichheit durch Erhöhung von Chancengerechtigkeit darstellt (vgl. Langenohl et al. 2024). Sowohl der Zugang zu KI als auch die Möglichkeit, KI effektiv zu nutzen, zeigen positivere Effekte für SuS, die ohnehin schon ein hohes Maß an Literalität aufweisen

.....

5 Die möglichen Lösungen zu dieser Frage sind noch nicht abschließend gefunden, und je nach Kontext gibt es unterschiedliche Vorschläge (bspw. Evangelista 2024).

und kompetent mit Texten umgehen können (vgl. Fürstenberg/Müller 2024; Magirius et al. 2024).

## 2.2 KI und Leistungsbeurteilung in der Schule

Für eine Beurteilung im Bildungswesen ist die individuelle Zurechenbarkeit von Leistungen entscheidend, d. h. dass eine (schriftliche) Leistung nur dann beurteilt werden soll, wenn sie selbstständig erbracht wurde (vgl. BMB o. J.). In Bezug auf schulische Prüfungsleistungen gilt laut österreichischer Leistungsbeurteilungsverordnung (LBVO) §11.4: „Vorgetäuschte Leistungen sind nicht zu beurteilen.“ Wenn eine Leistung nur vorgetäuscht wird, dann ist diese zu sanktionieren, bspw. durch eine schlechte Note o. Ä. Die LBVO wurde in Bezug auf die Disruption durch KI (noch) nicht geändert oder adaptiert, aber die Vorgaben zum Verfassen der Matura<sup>6</sup>-relevanten Abschlussarbeit (Vorwissenschaftliche Arbeit oder Diplomarbeit, je nach Schultyp) sind ab dem Schuljahr 2023/24 in Bezug auf die Nutzung von KI ausführlich überarbeitet. Zusammenfassend sagt das Dokument aus, dass die Beurteilung weiterhin den gesamten Erarbeitungsprozess der abschließenden Arbeit betrifft und wie bisher auf den Grundsätzen der LBVO basiert. Sofern eine KI genutzt wird (egal, in welchem Schritt des Erstellungsprozesses), muss dies kenntlich gemacht werden, damit die individuelle Leistung immer ersichtlich bleibt (vgl. BMBWF 2023).

Dass sich die Rate an vorgetäuschten Leistungen nicht nur durch KI erhöht, zeigen Comas-Forgas et al. (2021) in Bezug auf Effekte durch die Covid-19-Pandemie. Durch unterschiedliche technologisch-digitale Hilfsmittel können Leistungen vorgetäuscht und Plagiate abgegeben werden (vgl. Broemer/Recktenwald 2021; Holden et al. 2021). Auch ohne technische Unterstützung gibt es Möglichkeiten zu vorgetäuschten Leistungen, wobei Gamage et al. (2023) die Auswirkungen von Contract-Cheating auf den Hochschulbereich untersuchen (vgl. auch Lancaster/Cotarlan 2021). In Bezug auf SuS während des Covid-19-bedingten Online-Unterrichts in Tschechien zeigen Černochová et al. (2024), dass viele SuS die Vortäuschung eigener Leistungen online als

.....

6 In Deutschland entspricht das Abitur der Matura in Österreich.

einfacher wahrnehmen; nämlich leicht durchzuführen und mit weniger (intellektueller und akademischer) Anstrengung verbunden (ebd.).

Bezüglich KI kann man sagen, dass sie viele dieser Möglichkeiten erfüllen kann. So ist KI in der Lage, universitäre Multiple-Choice-Prüfungen zu bestehen (vgl. Gilson et al. 2023), sogar auf demselben Niveau wie Studierende (vgl. Riedel et al. 2023). Gerade bei solchen Faktenprüfungen oder bei sehr grundlegenden Textverständnisaufgaben kann die KI Produkte liefern, die sich von studentischen Abgaben nur schwer unterscheiden lassen (vgl. Evangelista 2024)<sup>7</sup>. Ghostwriting ist im Bereich Schule und Hochschule wirklich nicht neu, wird jedoch durch KI für sehr viel mehr Menschen zugänglich und kostet im Vergleich zu Contract Cheating weniger Geld.

Für die Herausforderung von vorgetäuschten Leistungen versucht das Bildungswesen Antworten zu finden, wobei oft eine technologische Lösung wie automatische Plagiatsprüfungsprogramme an Universitäten, Schulen, etc. gesucht wird. Doch alle technischen Kontrollprogramme müssen sich ständig weiterentwickeln und bleiben trotzdem immer einen Schritt zurück (vgl. Evangelista 2024; Weber-Wulf et al. 2023). Außerdem können solche technischen Lösungen nur dann ihr volles Potenzial entfalten, wenn Lehrkräfte in der Nutzung geschult (vgl. Kočková et al. 2024) und passende Strukturen zum Aufdecken von Täuschungsversuchen in den Institutionen aufgebaut werden (vgl. Bräuer/Hollosi-Boiger 2024). Es lässt sich daher sagen, dass die Herausforderungen im Bildungsbereich nicht ausschließlich durch weitere technologische Lösungen gemeistert werden können (vgl. Khosravi et al. 2021), sondern dass insbesondere für die Leistungsbeurteilung andere Modelle und Modalitäten Anwendung finden müssten<sup>8</sup>. Überdies sind (Text-)Produkte der KI nicht immer und nicht durchgehend von menschlichen Produkten unterscheidbar

.....

7 Zu dem Thema von Prüfung und Bewertung und die Vulnerabilität dieser schulischen Funktion durch KI gibt es viel weniger Forschungsarbeiten als im Bereich der Hochschulbildung, weswegen hier und in anderen Abschnitten besonders auf Ergebnisse aus der tertiären Bildung verwiesen wird.

8 Eine Gemeinsamkeit aller dieser Ideen (wie bspw. bei Evangelista 2024; Hutson/Plate 2023) ist, dass sie personalintensiv sind, d. h. mehr Zeit und Aufmerksamkeit durch die Lehrenden verlangen.

(vgl. Floridi 2023 und 2024 in Bezug auf das *human-artificial*-Kontinuum statt einer Dichotomie).

Abschließend soll auf einen weiteren Punkt bei der Vortäuschung eigener Leistungen eingegangen werden, insbesondere bei Hausaufgaben. Katenbrink/Kohler (2021) beschäftigen sich mit dem Abschreiben von Hausaufgaben und deuten diesen Normbruch als funktional für das System Schule. So können Lehrkräfte ihren Unterricht wie geplant durchführen, wenn ihnen scheint, dass auch schwächere SuS im Stoff und in der Klasse mitkommen. Individuell ist eine Vortäuschung von Leistungen funktional, da sie zu positiven Beurteilungen und besseren Noten führen kann. Zumindest können dadurch Sanktionen (negative Noten, Wiederholungsprüfungen, etc.) verhindert werden (ebd.) Für Katenbrink/Kohler (2021) liegt dieses Verhalten darin begründet, dass SuS „Mängel auf der Angebots- und/oder Nutzungsseite“ (2021: 6) kompensieren, weil sie Themen schlecht verstanden haben, weil sie keine weiteren Lernangebote erhalten, aber eben auch, weil es einfacher ist<sup>9</sup>.

## 3 Zur Nutzung von KI bei SuS der Sekundarstufe II

### 3.1 Methodisches Vorgehen

Zur Datenerhebung wurden für dieses Forschungsvorhaben Gruppendiskussionen mit Jugendlichen einer Sekundarstufe II (mit technischem Schwerpunkt) in Wien durchgeführt. Es gab zwei Gruppengespräche, und zwar Ende Mai 2023 und Anfang Juni 2024. Bei der ersten Diskussion waren vier Jugendliche (2 weiblich, 2 männlich) im Alter von 19–20 Jahren anwesend, bei der zweiten Diskussion fünf Jugendliche (nur männlich) im Alter von 15–16 Jahren. Die erste Diskussion dauerte 32:46 Minuten, die zweite Diskussion 22:52 Minuten. Die SuS der ersten Diskussion (2023) werden mit „SH+Ziffer“ abgekürzt, die

.....  
9 Katenbrink/Kohler (2021) verweisen in ihrem Text auf eine Art Doppelmoral der Institution Schule, die ein reibungsloses Funktionieren ermöglichen soll. Mitunter schauen Lehrkräfte bei Täuschungsversuchen nicht so genau hin, da vorgetäuschte Leistungen (streng) sanktioniert werden müssten (vgl. Sorgo et al. 2015).

SuS aus der zweiten Diskussion (2024) mit „SC+Ziffer“. Alle SuS besuchten Laptopklassen, sodass der Laptop in eigentlich jedem Fach täglich genutzt wurde und Schularbeiten<sup>10</sup> sowie andere schriftliche Überprüfungen am Laptop geschrieben wurden<sup>11</sup>.

In Anlehnung an Atanasoska/Proyer (2016) und die Erfahrungen bei Interviews zu schwierigen Themen wurde für diese Diskussionen eine nicht-moderierte Variante der Gruppendiskussion gewählt, um auch auf brisante Bereiche, in diesem Fall Themen wie Täuschungsversuche o.Ä.<sup>12</sup>, zu sprechen zu kommen. Die SuS besprachen die vorgegebenen Leitfragen eigenverantwortlich.

Vor Beginn der Diskussion wurden der Gruppe das Forschungsinteresse der Studie sowie die Datenschutzmaßnahmen dargestellt. Die Einverständniserklärung wurde vor der Diskussion schriftlich und bei der Diskussion per Audio eingeholt. Anschließend wurde der Ablauf erklärt. Beide Diskussionen fanden in Klassenzimmern der Schule statt.

Eine Person in der Gruppe wurde gebeten, die Zeit zu kontrollieren. Jede Frage sollte mindestens fünf und maximal acht Minuten diskutiert werden. Eine zweite Person erhielt die fünf Leitfragen für die Diskussion, die auf A4 ausgedruckt und beidseitig beschriftet waren. Auf der einen Seite stand die Nummer der Frage, auf der anderen Seite die Frage selbst in einer großen Schriftart, damit auch die daneben sitzenden Personen mitlesen konnten. Die Person mit den Fragen hatte den Auftrag, jede Frage der Gruppe vorzulesen, und durfte die verdeckt auf dem Tisch liegenden Fragen erst nach der Mindestzeit und in der vorgegebenen Reihenfolge umdrehen.

.....

- 10 In Österreich sind das Schularbeiten (in Deutschland oft Klassenarbeiten), Tests, schriftliche Überprüfungen, schriftliche Stundenwiederholungen etc.
- 11 Die Schule hat eine eigene Lösung, wie der Zugang zum Internet auf den (privaten) Laptops der SuS während Klassenarbeiten etc. unterbunden werden kann.
- 12 Die schulische Leistungsbeurteilungsverordnung in Österreich verlangt, dass Vorfälle von „vorgetäuschten Leistungen“ (LBVO 2024, §11.4) geahndet werden.

Die fünf Leitfragen waren:

1. Wie habt ihr oder eure Klassenkolleg:innen oder Freunde/Freundinnen in diesem Schuljahr KI-Hilfen wie z. B. ChatGPT genutzt?
2. Wie sehr waren für euch/eure Klassenkolleg:innen/Freunde und Freundinnen solche Angebote wie ChatGPT eine „Hilfe“?
3. Wie sollten Lehrpersonen mit Angeboten wie ChatGPT umgehen?
4. Wie sollen Lehrpersonen bei der Benotung und Bewertung damit umgehen, dass alle Schüler:innen Zugang zu Angeboten wie ChatGPT haben?
5. Was stört euch daran, dass Künstliche Intelligenz in der heutigen Zeit so viel übernehmen/erledigen kann? UND Was findet ihr gut, dass die KIs übernehmen/erledigen können?

Beide Diskussionen wurden nach Notenschluss durchgeführt, sodass die SuS keine Sorge bezüglich ihrer Benotung haben mussten.

Die Analyse erfolgte mittels qualitativer Inhaltsanalyse (vgl. Kuckartz 2012), wobei im ersten Schritt alle Sequenzen umformuliert wurden. Im zweiten Schritt wurden diese Sequenzen in Bezug auf die zwei Forschungsfragen näher betrachtet und in Bezug auf die erste Forschungsfrage für diesen Beitrag nach der Einteilung von Steinhoff 2025 (siehe Kapitel 2.1) analysiert. Für die zweite Forschungsfrage wurde ein induktives Vorgehen gewählt, um explorativ die Vielfalt der angesprochenen Bereiche analysieren zu können.

### **3.2 Analyse 1: KI zum und beim Schreiben**

In diesem Kapitel geht es darum, die Rollen aufzuzeigen, in denen die SuS KI nutzten. Der Fokus liegt, entsprechend der Fragestellung, auf dem Schreiben von Texten. Die deduktive Analyse bezieht sich auf die Unterteilung von Steinhoff (2025), der die KI-Verwendung in die Kategorien *Ghostwriter*, *Writing Partner* und *Tutor* unterteilt.

Ghostwriting bedeutet hier, dass von der KI generierte Texte 1:1 oder mit kleinen Modifikationen durch die SuS übernommen oder weitergenutzt werden. Die Texte werden auch zur Bewertung bei der Lehrkraft eingereicht. Bei

der Rolle des Writing Partners verwenden SuS die KI, um den Stil ihres Textes zu verbessern bzw. stilistischen Ansprüchen zu entsprechen. Sie lassen den Text von der KI umformulieren, um „schönere, gehobenere“ Formulierungen (siehe SH 1<sup>13</sup>) zu verwenden. Beispielsweise kann die KI Wortwiederholungen reduzieren, indem passende Synonyme eingefügt werden.

SH 1: „Teilweise, wenn ich gemerkt habe, in Englisch, dass ich irgendeinen Satz habe, in dem ich schon zehnmal dasselbe Synonym verwende oder so, dann habe ich gesagt, ja, schreibst du [ChatGPT] mir diesen einen Satz irgendwie anders um, und dann kam da was so viel Besseres raus.“ (Min. 4:15)

So und ähnlich berichten auch andere SuS von der Verwendung von ChatGPT<sup>14</sup> für die stilistische Verbesserung ihrer Texte.

In beiden Diskussionen verweisen die SuS auf die Verwendung als Writing Partner, d. h. um Texte umschreiben oder verbessern zu lassen, aber nie auf die Funktion als Writing Tutor (siehe Kap. 2.1). Doch die Funktion, die von den SuS am häufigsten genannt wird, ist die des Ghostwriters. Dazu gehören insbesondere Zusammenfassungen, die sich die SuS von der KI schreiben lassen. Für die mündliche Matura beispielsweise sind bestimmten Themen<sup>15</sup> vorgegeben, zu denen die SuS vor der Prüfungskommission abgefragt werden. Manche Lehrkräfte erwarten von den SuS, dass sie ihre themenspezifischen Ausarbeitungen abgeben. In Bezug auf diese Ausarbeitungen nennen die SuS in der Diskussion 2023 nur KI-generierte Texte.

SH 2: „Für Matura-Ausarbeitungen war es wirklich ein Lifesaver.“ (Min. 4:33)

.....

- 13 SH 1: „[U]nd zum Beispiel, wenn ich eigene Sachen geschrieben habe und ich fand, dass die Formulierung seltsam klingt und ich es schöner formuliert haben wollte, habe ich ChatGPT gefragt, ob man diese Formulierung etwas gehobener machen könnte“ (Min. 1:05)
- 14 Die SuS sprachen bei Texterstellung ausschließlich von ChatGPT. Nur bei anderen Aufgaben, wie z. B. Bilder oder Videos erstellen etc., wurden andere KI-Modelle genannt.
- 15 Die Schulen haben diesbezüglich gewisse Freiheiten innerhalb größerer Überthemen, wie z. B. „Globalisierung“ im Fach Geografie oder Englisch.

Auch die SuS 2024 nutzen die KI häufig für Zusammenfassungen, so wie SC 6, denn Texte zu kürzen sei wichtig für die Schule (Min. 4:11). Ferner nutzen die SuS die KI auch für Zusammenfassungen von „extrem fetten“ (siehe SC 3) Texten, wenn Texte also sehr lang und umfangreich sind und der (faktische) Inhalt für eine mögliche Überprüfung beherrscht werden muss.

SC 1: „Findet ihr, man sollte ein schlechtes Gewissen haben, wenn man in Sachen wie [Sachfach], wo man nur Sachen zusammenfassen muss, ChatGPT verwendet?“

SC 3: „Nein. Nein, weil du hast ja so einen extrem fetten Text.“ (Min. 8:58)

Eine Zusammenfassung zu lesen, geht schneller und ist für sie effektiver, finden die SuS, und die Zusammenfassung von der KI erstellen zu lassen, spart noch mehr Zeit. In den Diskussionen wird diese Nutzung von beiden Gruppen als legitim betrachtet. Bei einer mündlichen Überprüfung wie der Matura müsse man ohnehin über das Wissen verfügen.

Die KI wird ferner dazu genutzt, um Mitschriften zu Stichpunkten in einen Fließtext zu verwandeln. So sagt SH 1, dass in der Stunde nicht so viel Zeit ist und man nur ein paar Bullet-Points notieren kann. Aus denen lässt sich SH 1 einen Fließtext erstellen, den man bei der Lehrkraft abgeben muss und der somit teilweise als Bewertungsgrundlage dient.

SH 1: „[...] schreibe ich immer mit im Unterricht, aber nur Bullet Points, und damit ich richtig schnell mitschreiben kann, und dann habe ich ChatGPT immer gesagt, ja, wandle diese Bullet Points in einen Fließtext um, damit ich sozusagen nicht Bullet Points dann als Abgabe abgebe.“ (Min. 1:35)

Die KI spart Zeit und ist deswegen aus Sicht der SuS ein effektives Werkzeug.

Die SuS sprechen in den Diskussionen zwar in erster Linie über Texte, die sie von einer KI schreiben lassen, aber sie nennen neben der produktiven regelmäßig auch die rezeptive Schriftlichkeit. Sie lesen sich die generierten Texte

durch und machen sich Gedanken zu Form und Inhalt. Dies inkludiert Überlegungen zur Vertrauenswürdigkeit der von der KI wiedergegebenen Inhalten.

SH 3: „Da hatten wir so ein Referat und da habe ich ChatGPT irgendwas dazu schreiben lassen, und das war einfach komplett falsch, also, es war irgendwas, da muss man wirklich aufpassen.“ (Min. 4:40)

In dem Diskussionsausschnitt zeigt sich, dass SH 3 den Inhalt des Referats schon kennt, bevor er sich von der KI als Ghostwriter einen Text zu dem Thema erstellen ließ. Durch seine Vorkenntnisse kann SH 3 erkennen, dass die Informationen und Sachverhalte im generierten Text „komplett falsch“ sind. Ihm und den anderen SuS ist bewusst, dass KI „halluzinieren“ kann, d. h. falsche Inhalte wiedergibt (siehe Kapitel 2.2). Auch die andere Gruppe spricht dieses Problem an.

SC 4: „Ich glaube, es kann nicht alles so richtig.“ (Min. 10:35)

Die Aussage wurde nicht im Kontext des Schreibens getätigt, sondern allgemein zu den Einschränkungen, dass die KI eben nicht immer richtige Informationen liefert.

### **3.3 Analyse 2: Bewertung der KI-Nutzung durch die SuS**

Beim Thema Zusammenfassung wurde im vorhergehenden Kapitel deutlich, dass diese Art von KI-Nutzung für die SuS keine Täuschung darstellt, da man das Wissen bei einer (mündlichen) Prüfung beherrschen muss. Dieses Thema, nämlich die Leistungsbeurteilung, die in der Sekundarstufe oftmals auf Textprodukten basiert, kommt in den Diskussionen immer wieder vor. Dies spiegelt sich in Aussagen und Analysen wider, wenn es um die Unterscheidung von menschlich vs. künstlich geht. Die SuS wollen KI-generierte Texte nur dann zur Beurteilung bei der Lehrkraft abgeben, wenn sie denken, dass sie nicht erwischt werden.

Die SuS gehen davon aus, dass ein vollständig von der KI generierter Text zu künstlich wirken könnte. Deshalb ist es ihnen ein Anliegen, KI-generierte

Texte „menschlicher“ (siehe SC 2) zu machen. Die Gruppe 2024 berichtet davon, andere KI (die sie nicht namentlich nennen) dafür zu nutzen. Die Gruppe 2023 geht anders vor: Bei manchen Aufgaben schreiben die SuS einige Textteile selber um, bei anderen lassen sie den Text noch einmal von derselben KI umschreiben.

SC 2: „Also, man kann es [ChatGPT] schon für Texte verwenden, aber dann muss man sich noch irgend so ein Tool holen, was es vielleicht menschlicher schreibt.“ (Min. 7:57)

In Bezug auf das „Vortäuschen eigener Leistungen“ (vgl. LBVO) sind die Lehrkräfte aus Sicht der SuS nicht in der Lage, KI-generierte Texte zu erkennen, weswegen ihre Nutzung aus Sicht der SuS erfolversprechend ist.

SC 3: „Das denken die Lehrer nicht, weil unser Lehrer hat mal einen [Textausschnitt] gezeigt von einem sehr guten Schüler, der komplett von ChatGPT war.“ (Min. 7:25)

Was die SuS also im Bildungssystem alltäglich erleben und erfahren, ist, dass man bessere Noten durch die Nutzung von KI erreichen kann (siehe Kapitel 2.2). Außerdem erleben sie, dass ihre Lehrkräfte nicht in der Lage sind, KI-generierte von durch Menschen verfasste Texte zu unterscheiden<sup>16</sup>.

Nur in den Sprachfächern wird Lehrpersonen zugeschrieben, den individuellen sprachlichen Stil von SuS zu erkennen, sodass ein KI-Text keinen Vorteil bedeutet.

SC 1: „Die Lehrerin kennt ja deine Schreibweise.“ (Min. 12:05)

Hierbei geht es nicht darum, ob die Lehrperson wirklich jede individuelle Schreibweise kennt und erkennt, sondern es wird deutlich, dass die SuS den

.....

16 Zu dieser subjektiven Einschätzung der Schüler:innen gibt es empirische Ergebnisse, bspw. bei Fleckenstein et al. (2024).

Lehrkräften in den Sprachfächern hohe Beurteilungs- und Bewertungskompetenzen zuschreiben.

Einschränkend für die Nutzung von KI wirken auch die schriftlichen Überprüfungen in der Schule. Diese motivieren die SuS extrinsisch, denn für solche Überprüfungen ist es aus Sicht der SuS wichtig, dass man die Aufgabenstellung ohne Hilfe durch KI lösen kann. Das geht nur, wenn man das Schreiben auch ohne KI trainiert.

SH 1: „Vor allem musst du es wirklich üben, wenn du weißt, du hast eine Schularbeit oder so was.“ (Min. 13:25)

Zum Thema Fairness und KI sind sich die SuS einig, dass es prinzipiell „fair“ ist (siehe unten, SC 5), KI zu nutzen, weil dies ja alle tun.

SC 1: „Ich würde auf jeden Fall sagen ja, weil jeder hat den gleichen Zugang zu diesen KI-Tools.“ (Min. 17:14)

Die SuS problematisieren in den Gesprächen nicht, dass es vielleicht Menschen gibt, die weniger oft oder weniger gut mit einer KI umgehen und diese somit nicht für die schulische Laufbahn nutzen können (siehe Kapitel 2.2). Vielmehr sehen sie das Thema Fairness historisch und vergleichen ihre heutigen Möglichkeiten mit denen von früheren Generationen.

SC 5: „Zur Frage, ist das fair. Die Schüler, die vor uns da waren in dieser Schule, die keine KI hatten, hatten [es] natürlich ... schwerer aufzusteigen. Und wir haben es natürlich jetzt leichter. Und das ist natürlich nicht fair.“

SC 2: „Also das ist wirklich nicht fair, dass andere früher das zum Beispiel nicht hatten.“

SC 5: „Aber ich glaube, die hatten irgendwelche anderen Methoden zu bescheißen, sicher. Weil nie im Leben haben sie das geschafft.“ (Min. 22:33)

Es wird deutlich, dass die SuS von einer Normalität des Vortäuschens von Leistungen im Schulsystem ausgehen, denn aus ihrer Sicht wäre die Schule sonst nur sehr schwer bis gar nicht zu bewältigen. Somit ist für sie KI nur eine neue Methode in einer langen Reihe von Methoden, um zu schummeln und um die Schule erfolgreich abzuschließen.

## 4 Zusammenfassung und Diskussion

Die hier untersuchten SuS zeigen einen hohen Nutzungsgrad von KI, insbesondere ChatGPT, zum Erstellen von Texten für die Schule. Dabei wird KI hauptsächlich in der Rolle als Ghostwriter genutzt. Um bei der Vortäuschung eigener Leistungen durch Ghostwriting mit Hilfe von KI nicht erwischt und sanktioniert zu werden, nutzen SuS verschiedene Wege, um die Nutzung zu verschleiern. Die KI macht aus Sicht der SuS Texte stilistisch besser und kann Zeit sparen, da sie bspw. Zusammenfassungen sofort erledigt. Außerdem kann sie schnell Texte liefern, die zur Bewertung bei den Lehrpersonen abgegeben werden. In Bezug auf die Sprachfächer sehen SuS jedoch Einschränkungen, denn sie schreiben den LuL eine hohe Kompetenz beim Erkennen des individuellen Schreibstils zu. Auch sind die regelmäßigen schriftlichen Überprüfungen in den Sprachfächern ein Grund, dass SuS Texte selber schreiben, um für die Schularbeit zu üben.

Die Leistungsbeurteilung, also die Möglichkeit, gute Noten zu erhalten, ist ein Motivator, viele schriftliche Aufgaben durch oder mit KI zu erstellen. Die SuS erfahren in ihrem Alltag, dass Lehrkräfte solche Textprodukte nicht oder kaum als Täuschung sanktionieren (vgl. Katenbrink/Kohler 2021). Stattdessen erleben sie immer wieder, dass KI-generierte Texte zu besseren Noten führen. Aus ihrer Sicht nutzen alle (anderen) SuS genauso KI für schulische Aufgabenstellungen. Dass die KI nicht alles weiß und inhaltlich falsche Texte generiert, ist den SuS zwar bewusst, wird aber in beiden Gruppen nur wenig thematisiert. In den Diskussionen sprechen die SuS am Rande über den Umstand der Halluzinationen und falschen Informationen, wobei ihnen klar ist, dass die KI nichts weiß (SH 3, Min. 29:47) und nicht alles richtig kann (siehe oben, SC 4).

Häufig und intensiv wird von den SuS diskutiert, ob die Nutzung von KI in und für die Schule fair ist oder nicht, obwohl keine der fünf Leitfragen direkt auf dieses Thema hinwies. Die Fairness beziehen sie dabei auf die schulische Leistungsbewertung, die Effizienz und den Arbeitsaufwand. Dass der Zugang zu KI nicht allen Menschen gleich offen steht und nicht von allen gleich gut oder effektiv verwendet werden kann (siehe Kapitel 2.2), wird nicht problematisiert. Fragen der Partizipation und Ungleichheit bleiben also bestehen, denn die digitale Teilhabe ist zwar niedrigschwellig, bedeutet aber nicht automatisch Chancengleichheit.

Die vorgestellten Ergebnisse liefern nur einen kleinen Einblick in die Rollen der KI bei Schüler:innen in einer Sekundarstufe II und es braucht weitere qualitative und auch quantitative Studien zur KI-Nutzung im schulischen Bereich. Wichtig wäre es dabei, unterschiedliche Altersgruppen, Schultypen, Regionen und Sprachen in der Schule synchron und diachron zu untersuchen. Dabei ist besonders auf die Steigerung (oder die Verringerung) von Ungleichheiten im Schulsystem zu achten. Einerseits sind hierbei die ungleichen Möglichkeiten der KI-Nutzung und das Niveau der digitalen Literalität zu beachten; andererseits spielt das Thema Leistungsbeurteilung und Vortäuschen eigener Leistungen hierbei genauso eine wichtige Rolle. Wie die Diskussionen zeigten: Über Fairness zu sprechen und für Fairness zu sorgen, das sind Themen, mit denen sich Schule und Bildungssystem intensiv beschäftigen müssen.

## Literatur

- ATANASOSKA, TATJANA/PROYER, MICHELLE (2016): Bildung mit und innerhalb von Grenzen: Herausforderungen für Flüchtlinge jenseits des Pflichtschulalters am Beispiel heterogener Schule und bildungsbezogener Übergänge. In: *SWS-Rundschau* 3/2016, S. 422–446.
- BMB (Bundesministerium für Bildung Österreich) (o. J.): Künstliche Intelligenz. Chance für Österreichs Schulen. In: *Künstliche Intelligenz – Chance für Österreichs Schulen*. Online. <https://www.bmb.gv.at/Themen/schule/zrp/ki.html> (letzter Zugriff: 8.6.2025).

- BMWF (Bundesministerium für Bildung, Wissenschaft und Forschung Österreich) (2023): *Die Verwendung KI-basierter Tools beim Erstellen abschließender Arbeiten – Potenziale, Risiken und beurteilungsrelevante Aspekte. Informationen für Betreuer/innen abschließender Arbeiten an AHS und BMHS*. Online Download von [https://www.bmb.gv.at/dam/jcr:3bc6eb26-f4b1-499c-a601-675e7fd6fa0f/ki\\_abarb.pdf](https://www.bmb.gv.at/dam/jcr:3bc6eb26-f4b1-499c-a601-675e7fd6fa0f/ki_abarb.pdf) (letzter Zugriff: 8.6.2025).
- BRÄUER, GERD/HOLLOSI-BOIGER, CHRISTINA (2024): Mit Schreibzentrumsarbeit als reflektiertes Literacy Management den Paradigmenwechsel. KI in der Textproduktion mitgestalten. In: *JoSch – Journal für Schreibwissenschaft* 15, S. 24–36.
- BROEMER, ELI/RECKTENWALD, GEOFFREY (2021): Cheating and Chegg: a Retrospective. *2021 ASEE Annual Conference*, PaperID 34650. <http://dx.doi.org/10.18260/1-2--36792> (letzter Zugriff: 8.6.2025).
- BUCK, ISABELLA/HUEMER, BIRGIT/LIMBURG, ANIKA (2024): KI im Schreibzentrum? Ein Plädoyer für offenen Diskurs und Kollaboration. In: *JoSch – Journal für Schreibwissenschaft* 15, S. 4–7.
- ČERNOCHOVÁ, MIROSLAVA/SELCUK, HASAN/PODOLJAK, TOMÁŠ (2024): Perspectives of Czech High School Students on Cheating in Remote Education. In: *Proceedings of the 23rd European Conference on e-Learning, ECEL 2024*, S. 60–68.
- COMAS-FORGAS, RUBÉN/LANCASTER, THOMAS/CALVO-SASTRE, AINA/SUREDA-NEGRE, JAUME (2021): Exam cheating and academic integrity breaches during the COVID-19 pandemic: An analysis of internet search activity in Spain. In: *Heliyon* 7 (10), <http://dx.doi.org/10.1016/j.heliyon.2021.e08233> (letzter Zugriff: 8.6.2025).
- DENNISON, DEEPAK V./GARCIA, RAYCELLE C. C./SARIN, PARTH/BYWATER, CHRISTINE/XIE, BENJAMIN/LEE, VICTOR R. (2024): From Consumers to Critical Users: Prompty, an AI Literacy Tool for High School Students. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (21), S. 23300–23308.
- DORNICHEVA, DARIA/SULIMOVA, MARIA (2025): Chatbots im Schreibunterricht für fortgeschrittene Russischlernende. In: *SlavUn. Slavische Sprachen Unterrichten* März 2025, S. 30–56.
- DUONG, THI-NGOC-ANH/CHEN, HSIU-LING (2025): An AI Chatbot for EFL Writing: Students' Usage Tendencies, Writing Performance, and Perceptions. In: *Journal of Educational Computing Research* 63 (2), S. 406–430.
- Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

- EVANGELISTA, EDMUND (2024): Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. In: *Contemporary Educational Technology* 17 (1), S. 559–568.
- FLORIDI, LUCIANO (2023): AI as agency without intelligence: On ChatGPT, large language models, and other generative models. In: *Philosophy & Technology* 36 (15), Art. 15, <https://doi.org/10.1007/s13347-023-00621-y> (letzter Zugriff: 8.6.2025).
- FLORIDI, LUCIANO (2024): On the Future of Content in the Age of Artificial Intelligence: Some Implications and Directions. In: *Philosophy & Technology* 37 (3). <https://doi.org/10.1007/s13347-024-00806-z> (letzter Zugriff: 8.6.2025).
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): Was bedeuten ChatGPT und Co. für den Deutschunterricht? Sprachliche Kompetenzentwicklung in Zeiten künstlicher Intelligenz. In: *Der Deutschunterricht* 5/24, S. 14–23.
- GAMAGE, KELUM/DEHIDENIYA, SHYAMA/XU, ZHIHENG/TANG, XIAOBIN (2023): Contract cheating in higher education Impacts on academic standards and quality. In: *Journal of Applied Learning & Teaching* 6 (2), <https://journals.sfu.ca/jalt/index.php/jalt/article/view/1063> (letzter Zugriff: 8.6.2025).
- GILSON, AIDAN/SAFRANEK, CONRAD F./HUANG, THOMAS/SOCRATES, VIMIG/CHI, LING/TAYLOR, RICHARD A./CHARTASH, DAVID (2022): How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. In: *JMIR Medical Education* 2023;9:e453122023 Feb 8. <https://mededu.jmir.org/2023/1/e45312> (letzter Zugriff: 8.6.2025).
- GRABENSTEINER, CAROLINE/HIMPSL-GUTERMANN, KLAUS/SCHÖNBÄCHLER, ERICH (2023): Hybride Settings als Science-Fiction. Fragen zu Tendenzen der Amorphisierung von Unterricht durch digitale Erweiterungen des Lernraums. In: *Medienimpulse* 61 (1), Download von <https://doi.org/10.21243/mi-01-23-13> (letzter Zugriff: 8.6.2025).
- HOLDEN, OLIVIA L./NORRIS, MEGHAN E./KUHLMIEIER, VALERIE A. (2021): Academic integrity in online assessment: A research review. In *Frontiers in education* Vol. 6, Art. 639814, <https://www.frontiersin.org/journals/education/articles/10.3389/feeduc.2021.639814/full> (letzter Zugriff: 8.6.2025).
- HUTSON JAMES/PLATE, DANIEL (2023): Enhancing Institutional Assessment and Reporting Through Conversational Technologies: Exploring the Potential of AI-

- Powered Tools and Natural Language Processing. In: *Journal Artificial Intelligence Robotics* 1 (1), S. 11–22.
- KATENBRINK, NORA/KOHLER, BRIGITTA (2021): Personales Defizit oder funktionale Praxis? Das Abschreiben von Hausaufgaben aus praxistheoretischer Sicht. In: *Bildungsforschung* 2021 (1), S. 1–12.
- KHOSRAVI, HASSAN/DENNY, PAU/MOORE, STEVEN/STAMPER, JOHN (2023): Learn-ersourcing in the age of AI: Student, educator and machine partnerships for content creation. In: *Computers and Education Artificial Intelligence* 5:100151, <https://www.sciencedirect.com/science/article/pii/S2666920X23000309> (letzter Zugriff: 8.6.2025).
- KOČKOVÁ, PETRA/KILIÁNOVÁ, KRISTÝNA/SLEPANKOVA, MARTA/SCHMID, ANGELIKA/KOSTOLÁNYOVÁ, KATEŘINA (2024): AI literacy in Teacher Education in the Czech Republic. In: *European Conference on e-Learning* 23 (1), S. 178–186.
- KUCKARTZ, UDO (2012): *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung*. Weinheim: Juventa.
- LANCASTER, THOMAS/COTARLAN, CODRIN (2021): Contract cheating by STEM students through a file sharing website: a Covid-19 pandemic perspective. In: *International Journal for Educational Integrity* 17 (1), <https://edintegrity.biomedcentral.com/articles/10.1007/s40979-021-00070-0> (letzter Zugriff: 8.6.2025).
- LANGENOHL, ANDREAS/LEHNEN, KATRIN/ZILLIEN, NICOLE (2024): Literalität und Bildung im Wandel: Überlegungen zum digitalen Habitus. In: Sarikaya, Yaşar (Hg.): *Herausforderung Digitalisierung: Islamische Theologie im digitalen Zeitalter*. Berlin, Boston: De Gruyter, 2025, S. 137–154. <https://doi.org/10.1515/9783111561387-008> (letzter Zugriff: 8.6.2025).
- LBVO (2024): Gesamte Rechtsvorschrift für Leistungsbeurteilungsverordnung, Fassung vom 01.07.2024. In: *Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Leistungsbeurteilungsverordnung, Fassung vom 08.06.2025*. Rechtsinformati-onssystem des Bundes. <https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10009375> (letzter Zugriff: 13.4.2025).
- MAGIRIUS, MARCO/HESSE, FLORIAN/HELM, GERRIT/SCHERF, DANIEL (2024): KI im Literaturunterricht. Chancen und Herausforderungen zwei Jahre nach der Veröffentlichung von ChatGPT. In: *Der Deutschunterricht* 5/24, S. 14–23.

- MÜLLER, HANS-GEORG/FÜRSTENBERG, MAURICE (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Deutschen Germanistenverbandes* 70 (4), S. 327–345.
- MURGIA, EMILIANA/FUMAGALLI, BIANCA S. F./NICCHIA, ELISABETTA/PARMIGIANI, DAVIDE (2024): Uninvited Generative AI has joined our students. Tackling disinformation and creating content with the help of generative AI apps. In: *2<sup>nd</sup> International Workshop on Artificial Intelligence Systems in Education*, Nov. 25–28, 2024, Bolzano, Italy. [https://ceur-ws.org/Vol-3879/AIxEDU2024\\_paper\\_20.pdf](https://ceur-ws.org/Vol-3879/AIxEDU2024_paper_20.pdf) (letzter Zugriff: 8.6.2025).
- RIEDEL, MAXIMILIAN et al. (2023): ChatGPT's performance in German OB/GYN exams. Paving the way for AI-enhanced medical education and clinical practice. In: *Frontiers in Medicine* 10, <https://pubmed.ncbi.nlm.nih.gov/38155661/> (letzter Zugriff: 8.6.2025).
- SCHMIDT, REGINA LEONIE/SEEGEL, STEN (2024): Voneinander lernen – KI-gestütztes wissenschaftliches Schreiben im Team lehren. In: *JoSch – Journal für Schreibwissenschaft* 15, S. 65–72.
- SCHÖNBÄCHLER, ERICH/STRASSER, THOMAS/HIMPSL-GUTERMANN, KLAUS (2023): Vom Chat zum Check. Informationskompetenz mit ChatGPT steigern. In: *medienimpulse* 61 (1), S. 1–51.
- SORGO, ANDREJ/VAVDI, MARIJA/CIGLER, URSKA/KRALJ, MARKO (2015): Opportunity makes the cheater: high school students and academic dishonesty. In: *Center for Educational Policy Studies Journal* 5 (4), S. 67–87.
- STEINHOFF, TORSTEN (2025): Künstliche Intelligenz als Ghostwriter, Writing Tutor und Writing Partner. Zur Modellierung und Förderung von Schreibkompetenzen im Zeichen der Automatisierung und Hybridisierung der Kommunikation am Beispiel des Schreibens mit ChatGPT in der 8. Klasse. In: Albrecht, Christian/Brüggemann, Jörn/Kretschmann, Tabea/Meier, Christel (Hg.): *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven*. Berlin: Springer, S. 85–99.
- SULIMOVA, MARIA/ATANASOSKA, TATJANA (2023): Familiensprache Russisch im Sprachenrepertoire von Jugendlichen während Covid-19: online vs. offline. In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 28, <https://zif.tu-journals.ulb.tu-darmstadt.de/article/id/3733/> (letzter Zugriff: 8.6.2025).

„Für Matura-Ausarbeitungen war es wirklich ein Lifesaver!“

- WEBER-WULF, DEBORA/ANOHINA-NAUMECA, ALLA/BJELOBABA, SONJA (2023): Testing of Detection Tools for AI-Generated Text. In: *International Journal for Educational Integrity* 19 (26), <https://edintegrity.biomedcentral.com/articles/10.1007/s40979-023-00146-z> (letzter Zugriff: 8.6.2025).
- WILLIAMSON, BEN/HOGAN, ANNA (2021): Pandemic privatisation in higher education: Edtech & university reform. In: *Education International*. [https://issuu.com/educationinternational/docs/2021\\_eiresearch\\_gr\\_covid19\\_commercialisation\\_digit?fr=sNTdjNzI5MTkzMTM](https://issuu.com/educationinternational/docs/2021_eiresearch_gr_covid19_commercialisation_digit?fr=sNTdjNzI5MTkzMTM) (letzter Zugriff: 8.6.2025).



## „Goethe vs. ChatGPT“

Einblicke in eine Unterrichtsreihe zur funktionalen Nutzung von ChatGPT zur Förderung fachspezifischer Kompetenzen im Umgang mit literarischen Texten

### Abstract

Ausgangspunkt der hier vorgestellten Unterrichtsreihe mit dem Titel „Goethe vs. ChatGPT“ war die Hypothese, dass Tools wie ChatGPT funktional in den Deutschunterricht eingebettet werden können, um fachspezifische Kompetenzen im Umgang mit literarischen Texten wie insbesondere die Interpretationskompetenz zu fördern. Diese Annahme wurde in einer Versuchsklasse in vier Sequenzen erprobt: In einer ersten Einheit zum ‚Identifizieren‘ mussten die Schüler:innen Goethe-Originale von KI-Fakes anhand auffälliger Textmerkmale unterscheiden und trainierten so ihre textanalytischen Kompetenzen. In der zweiten Sequenz zum ‚Imitieren‘ wurden sie dazu angeregt, ChatGPT mithilfe geeigneter Prompts dazu anzuleiten, möglichst gelungene Goethe-Reproduktionen zu erstellen, wodurch ihre Fähigkeit geschult wurde, literarische Texte laufend zu überarbeiten. In der dritten Phase zum ‚Interpretieren‘ wurde erprobt, inwieweit es mithilfe des Chatbots gelingt, angemessene Gedichtinterpretationen zu generieren, wodurch die Schüler:innen sich über die erforderlichen Denk- und Schreiboperationen bewusst wurden. Den Abschluss bildete eine vierte Sequenz zum ‚Aktualisieren‘, in welcher die literarische Urteilskompetenz der Schüler:innen gefördert wurde, indem sie von ChatGPT produzierte Modernisierungen eines klassischen Goethe-Gedichts kriteriengeleitet bewerteten. Die Unterrichtsreihe zeigt, dass ChatGPT bei kontrollierter und funktionaler Einbettung in den Deutschunterricht dazu beitragen kann, Motivation, fachspezifische Kompetenzen und metakognitives Reflexionsvermögen zu fördern.

**Schlagwörter:** Deutschunterricht, ChatGPT, Fachkompetenzen, Interpretieren

## Ziel und Struktur der Unterrichtsreihe

Im Schuljahr 2023/2024 habe ich an meiner Schule, dem Evangelischen Gymnasium zum Grauen Kloster in Berlin, im Deutschunterricht mit meiner 10. Klasse eine Unterrichtsreihe mit dem Titel „Goethe vs. ChatGPT“ durchgeführt. ChatGPT war zu diesem Zeitpunkt noch relativ neu, und ich wollte herausfinden, ob sich der Chatbot gewinnbringend im Deutschunterricht einsetzen lässt, um fachspezifische Kompetenzen im Umgang mit literarischen Texten zu fördern. Blickleitend war also die Hypothese, dass die Kompetenzorientierung des Deutschunterrichts durch Entwicklungen wie ChatGPT keineswegs ‚überholt‘ ist, wie teilweise argumentiert wurde (vgl. Odendahl 2023: 13), sondern dass der Bot in passenden Unterrichtsarrangements vielmehr gezielt zur Förderung spezifischer Deutschkompetenzen wie der Analyse und Interpretation literarischer Texte eingesetzt werden kann. Ebenso wichtig war die Vorannahme, dass das spielerische Wetteifern der Schüler:innen mit der Maschine als ‚Sparringspartner‘ (vgl. Fürstenberg/Müller 2024: 7) für den Lernprozess fruchtbar gemacht werden könne, um die Motivation zu steigern und zugleich metakognitive Reflexionen anzuregen.<sup>1</sup>

Meine Unterrichtsreihe gliederte sich in vier Sequenzen, die jeweils einer bestimmten Praxis der Anwendung des Chatbots auf Goethe-Texte gewidmet waren. Den Auftakt bildete eine Doppelstunde zur 1) Praxis des *Identifizierens*: Um Neugierde auf das Thema zu wecken, ging es hier erst einmal um die Frage, ob wir Goethe-Originalgedichte von (simplen) ChatGPT-Fakes unterscheiden können,<sup>2</sup> um daraus Kriterien der fachgerechten Unterscheidung zu entwickeln. Durch den vergleichenden Blick auf Original und Fake wurde

.....

1 Damit vertrete ich die Gegenposition zu Odendahl, der argumentiert, dass es im Deutschunterricht im KI-Zeitalter nicht mehr darum gehen könne, im Hinblick auf bestimmte Kompetenzen „mit der aufgabenbearbeitenden Maschine zu wetteifern“ (Odendahl 2023: 13). Während Odendahl fordert, einen „lustvoll-genießende[n] Umgang“ mit Literatur jenseits der KI zu kultivieren (ebd.), begreife ich die Nutzung von ChatGPT im Unterricht als Werkzeug, um an die digitale Experimentierfreude der Schüler:innen (im Sinne eines modernen ‚Spieltriebs‘) anzuknüpfen und diese als motivationale Grundressource für die kreative und analytische Textarbeit fruchtbar zu machen.

2 Dem Problem der Unterscheidbarkeit von KI- und menschlicher Lyrik widmet sich zuletzt eine Reihe von Studien (vgl. etwa Porter/Machery 2024).

der analytische Blick der Schüler:innen für grundlegende Merkmale lyrischer Texte (wie Reim oder Metrum) geschärft. Daran schloss sich eine freier gestaltete 2) Sequenz zum *Imitieren* an, in der die Schüler:innen im Rahmen eines Portfolios eigenständig daran arbeiten sollten, ChatGPT so zu instruieren, dass es möglichst gelungene Goethe-Imitationen produziert. Neben der Analysekompetenz wurde dabei insbesondere auch die Fähigkeit gefördert, Texte in Interaktion mit dem Chatbot laufend zu überarbeiten und zu verbessern. In einem 3) Schwerpunkt zum *Interpretieren* erprobten die Schüler:innen dann, wie gut sich Goethe-Gedichte mithilfe von ChatGPT (simulativ) interpretieren lassen<sup>3</sup> – dem Klischee nach der Hauptanwendungsbereich des Tools im Schulgebrauch –, um die Ergebnisse dann kritisch zu reflektieren. Dabei wurden nicht zuletzt das metakognitive Reflexionsvermögen gefördert, indem die Schüler:innen erkannten, welche Denk- und Schreiboperationen es auszuführen gilt, um einen gelungenen Interpretationsaufsatz zu produzieren. Abgeschlossen wurde die Reihe von einer 4) Sequenz zum *Aktualisieren*, in der die Schüler:innen versuchten, ob sich mithilfe von ChatGPT gute modernisierte Fassungen von Goethe-Gedichten produzieren lassen, die um aktuelle Themen und Konflikte kreisen, um die Produkte abschließend zu evaluieren. Anhand der kritischen Auseinandersetzung mit der Botdichtung sollte vor allem die Kompetenz gefördert werden, die Qualität literarischer Texte kriteriengeleitet zu beurteilen. Man erkennt schon an diesem Grundaufbau, dass der Einsatz von ChatGPT immer kontrolliert und funktional in den Unterricht eingebettet war und für die Förderung fachspezifischer Kompetenzen genutzt werden sollte, wobei ein besonderes Augenmerk auf der Textanalyse lag. Die Sequenzen zum Identifizieren, Interpretieren und Aktualisieren waren eher klassisch-geschlossen, die zum Imitieren offen angelegt: Hier arbeiteten die Schüler:innen weitgehend frei an einem Portfolio, in dem sie ihre Imitationsversuche mithilfe von ChatGPT dokumentierten und reflektieren.

Im Folgenden werden die vier Sequenzen der Unterrichtsreihe im Sinne eines ‚best practice‘-Beispiels für die Nutzung von ChatGPT im Deutschunterricht detailliert vorgestellt. Im Hintergrund steht dabei die Grundannahme,

.....  
 3 Auch die Potenziale und Risiken KI-gestützter Interpretationen rückten jüngst in den Fokus der Forschung (vgl. etwa Führer/Gerjets 2024).

dass die Deutschdidaktik aktuelle Herausforderung wie die Entwicklung von KI-Anwendungen am besten bewältigen und innovative Lösungen entwickeln kann, wenn die neuen Tools im Unterricht praktisch erprobt und ihr didaktisches Potenzial laufend reflektiert wird (vgl. Lehnen/Steinhoff 2023: 2).

## 1 Erste Sequenz: Identifizieren

Um an das intuitive, durch lebensweltliche Erfahrungen mit KI-Anwendungen geprägte Vorverständnis und Urteilsvermögen der Schüler:innen anzuknüpfen, konfrontierte ich sie in der ersten Unterrichtsstunde meiner Reihe mit der Frage, ob sie überhaupt in der Lage seien, Goethe-Originale von Chatbot-Fakes zu unterscheiden. Im Einstieg wurden als visueller Impuls zunächst die Verse aus einem weltberühmten, den meisten aber unbekanntem Goethe-Gedicht (Goethe 2006a, 53) an die Wand projiziert:

Über allen Gipfeln  
ist Ruh'  
In allen Wipfeln  
Spürest Du  
Kaum einen Hauch;  
Die Vögelein schweigen im Walde.  
Warte nur! Balde  
Ruhest du auch.

Diesem wurde sogleich eine ChatGPT-Reproduktion gegenübergestellt, die aus dem Prompt hervorging, das Gedicht in leichter Variation von Form und Inhalt wiederzugeben:

Über all den Wogen  
Stille hält  
In den Sternenbogen  
Hörst du sanft  
Nur den Nachthall

Die Gedanken schweifen im Dunkeln  
 Warte still! Bald  
 Wirst du Ruh' finden.

Von der ersten Irritation über die frappierende Ähnlichkeit der beiden Gedichte ausgehend erhielten die Schüler:innen den Arbeitsauftrag, die Texte miteinander zu vergleichen und eine Hypothese zu ihrem Verhältnis zu formulieren. Spekuliert wurde etwa darüber, ob Goethe sein Gedicht später überarbeitet oder parallel ein ähnliches geschrieben habe – aber es wurde auch schon die Vermutung geäußert, dass hier eine Künstliche Intelligenz wie ChatGPT ihre Finger im Spiel haben könnte: Was machen wir nun aber, wenn uns nur die ChatGPT-Imitation vorliegt, aber nicht das Goethe-Original? Die Schüler:innen erkannten das Problem und formulierten die passende Stundenfrage, ob und wenn ja, *wie*, wir Goethe-Originalgedichte von ChatGPT-Fakes unterscheiden können: Eine Aufgabe für ‚Literaturdetektive‘. Damit war die Erarbeitungsphase angebahnt: In sechs ‚Detektiv‘-Arbeitsgruppen wurden weitere Originale und Fakes unter die Lupe genommen, mit dem Arbeitsauftrag *erstens* zu entscheiden, ob es sich beim Gedicht um Original oder Fake handelt, diese Entscheidung *zweitens* anhand besonderer Textmerkmale zu begründen und *drittens* allgemeine Kriterien für die Unterscheidung vorzuschlagen.

	<p><b>I IDENTIFIZIEREN</b>                  Goethe-Originalgedichte und (simple) ChatGPT-Fakes voneinander unterscheiden lernen                  ‚Neugierde wecken‘</p>	
	<b>Arbeitsauftrag</b>	<b>Operator</b>
	1) Entscheidet, ob es sich bei ‚Eurem‘ Gedicht um ein Goethe-Original oder ein ChatGPT-Fake handelt.	‚ <b>unterscheiden‘</b> <b>AFB I</b>
<b>ERARBEITUNG</b>	2) Begründet Eure Entscheidung anhand von besonders auffälligen Textbelegen u. -merkmalen.	‚ <b>begründen‘</b> <b>AFB II</b>
	3) Verallgemeinert Aspekte und entwickelt Kriterien, die eine Unterscheidung zwischen Original u. Fake ermöglichen.	‚ <b>Kriterien entwickeln‘</b> <b>AFB III</b>

**Abb. 1:** Arbeitsauftrag zum „Identifizieren“ von Goethe-Original u. KI-Fake

Damit folgte die Stunde einer Progression im Hinblick auf die Operatoren und die damit verknüpften Kognitionsleistungen, vom intuitiven Unterscheiden (AFB I) über das textbezogene Begründungsarbeit (AFB II) bis zum abschließenden Entwickeln von typischen Aspekten und Kriterien der Unterscheidung (AFB III) (vgl. SenBJF 2023: 9). Die Phasen von Erarbeitung und Sicherung spiegeln sich in den Deutschheften wider: Dort wird vor allem deutlich, dass die Schüler:innen die Gedichte nicht nur korrekt zuordnen, sondern ihre Unterscheidung zwischen Original und Fake auch aspekt- und schließlich kriteriengeleitet begründen, indem sie auf verschiedene Mittel der sprachlich-poetischen Gestaltung abstellen, so etwa auf der Ebene von Semantik („offenes Ende“ bei Goethe), Lexik („größerer, diverserer Wortschatz“ bei Goethe) oder von Metrum und Reim („unschöne Reime“ bei ChatGPT).

Die Vögelein schweigen im Walde.  
 Warte nur! Balde  
 Ruhest du auch.

• Wo lagen die Schwachstellen in der Imitation Goethes?

- Goethe: offenes Ende ✓
- Goethe: größeren/diverseren Wortschatz und geschichtl. Nutzen dieser Wörter ✓
- Chat Gpt: Unschöne Reime / Textstellen / fehlerhaftes Metrum ✓
- Goethe: komplexeres Reimschema ✓
- Chat Gpt hingegen: „Schulbuchgedicht“ ✓
- Chat Gpt = begrenzter Wortschatz. Ansonst Gedächtnis ausschließlich mit untypisch-romantischen etw. B. Wälder, Nachthimmel, Sterne etc. wie z. B. Wälder, Nachthimmel, Sterne etc. ✓

gut!

<b>Operator ,begründen' (AFB II)</b>	<b>Operator ,Kriterien entwickeln' (AFB III)</b>	SuS ermitteln Aspekte und Kriterien der sprachlich-poetischen Gestaltung (Semantik, Wortschatz, Reim, Metrik), um ihre Zuordnung zu begründen.
--	--	--

Abb. 2: Schülerergebnis zur Identifikationsübung

Mit dieser Operation erfüllen die Schüler:innen einen wichtigen Standard aus dem Rahmenlehrplan für das Fach Deutsch in Berlin/Brandenburg für die Sekundarstufe I, nämlich die analytische Beschreibung grundlegender Merkmale von Lyrik (vgl. SenBJF 2023: 28) – ein Standard, der im Rahmenlehrplan Deutsch für die Sekundarstufe II in erweiterter Form wiederaufgenommen

wird, als Analyse von Inhalt, Aufbau und Form literarischer Texte (vgl. SenBJF 2022: 20).

Wie ein synoptischer Blick auf die Deutschhefte zeigt, haben alle Schüler:innen, auch die leistungsschwächeren, den Standard weitgehend erreicht. So gelang es auch schwächeren Schüler:innen, Goethe- und Botgedicht anhand von Auffälligkeiten auf semantischer, lexikalischer und phonetischer Ebene aspektgeleitet voneinander zu unterscheiden (Minimalstandard), in der Regel wurden diese Merkmale auch fachsprachlich korrekt bezeichnet, etwa indem das jeweils genutzte Reimschema oder das Metrum genauer beschrieben wurde (Regelstandard) und im Idealfall erfolgte eine differenzierte, kriteriengeleitete Beurteilung der lyrischen Qualität anhand der jeweils genutzten Ausdrucksmittel (Maximalstandard).

Die erste Sequenz zeigt also: Mithilfe von ChatGPT erstelltes Textmaterial lässt sich für spielerische Identifikationsübungen nutzen, die prinzipiell geeignet sind, um die Kompetenz der vergleichenden Textanalyse im Hinblick auf besondere inhaltliche und formale Merkmale zu fördern.

## 2 Zweite Sequenz: Imitieren

Dies war jedoch gleichsam nur eine propädeutische Fingerübung. Das Herzstück meiner Reihe bildete die folgende Sequenz zum Imitieren, in der die Schüler:innen erproben sollten, wie sie mithilfe von ChatGPT noch deutlich bessere Goethe-Fakes produzieren können. Frei nach dem Motto aus Goethes *Zauberlehrling* (Goethe 2006c, 874):

Hat der alte Hexenmeister  
Sich doch einmal wegbegeben!  
Und nun sollen seine Geister  
Auch nach meinem Willen leben.

Die Leitfrage für die folgenden Deutschstunden war also, ob wir ChatGPT mit passenden Prompts so anleiten können, dass es täuschend echte Goethe-Fakes produziert: Eine Aufgabe für Tüftler. Die Klasse verwandelte sich in

ein Lyriklabor. In Gruppen arbeiteten die Schüler:innen daran, mithilfe von ChatGPT möglichst gute Goethe-Imitationen zu produzieren. Das Format für die Aufgabe war ein Portfolio, in dem die Schüler:innen ihre Versuche über den Zeitraum von zwei Wochen dokumentieren sollten, mit Blick auf Prompts, Ergebnisse und Nachbesserungen. Abschließend sollten sie die Versuchsreihe reflektieren, auch mit Blick auf die eigene Motivation und ihren Kompetenzzuwachs.

Der im Portfolio dokumentierte Lernprozess (s. Abb. 5) war im wesentlichen selbstorganisiert, da die Schüler:innen ihre Versuchsreihen selbstständig gestalteten, Konzepte entwickelten, Prompts und Outputs evaluierten und ein Fazit zogen. Gleichzeitig war die Arbeit doppelt in das übergreifende Unterrichtsarrangement eingebettet, da die individuelle Reflexion über die eigene Versuchsreihe um peer-to-peer-Feedback aus der eigenen Arbeitsgruppe und Rückmeldungen durch die Lehrkraft ergänzt wurde. Um die ganze Klasse am eigenen Erkenntnisprozess teilhaben zu lassen, wurde während der beiden Versuchswochen außerdem das Ritual der (von ChatGPT produzierten) ‚Strophe des Tages‘ etabliert.

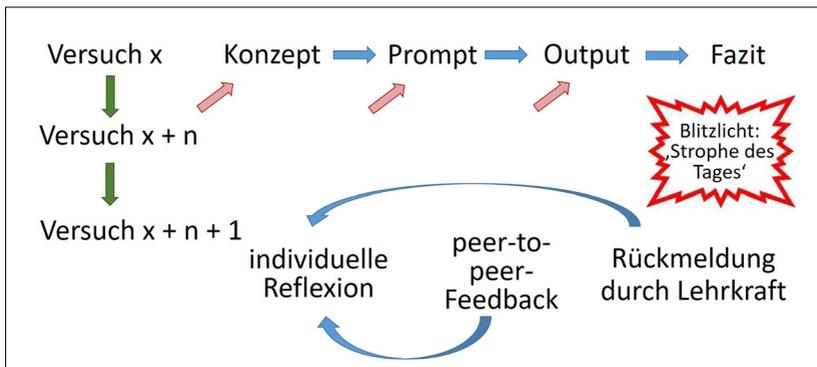


Abb. 3: Selbstorganisierter Lernprozess im Portfolio

Was die Portfolioarbeit an Erkenntnissen gebracht hat, verdeutlicht der Blick in ein exemplarisches Portfolio, das ich stellvertretend für den Er-

kenntnisprozess der gesamten 10. Klasse in einem ‚Close-Reading‘ auswerten möchte.<sup>4</sup>

Den Ausgangspunkt für den ersten Versuch des Schülers bildet die Diagnose, dass ChatGPT bisher keine guten Imitationen produziert habe, da dem Programm keine präzise Beschreibung von Goethes Schreibstil vorliege. Hier soll eine von ChatGPT erstellte Liste mit Merkmalen von Goethes Stil Abhilfe schaffen. Da das neue Gedicht wiederum nicht überzeugt, wird vermutet, dass die Liste fehlerhaft oder zu ungenau ist: „Ansatz für neue Lösungsansätze“, notiert der Schüler in seinem Protokoll treffend.

Ob der Schüler will oder nicht: Indem er ChatGPT Gedichte mithilfe einer Eigenschaftsliste produzieren lässt, die er dann laufend korrigiert und verbessert (*Operation*, die er ausführt), arbeitet er aktiv an der Plaung von Texten (*Standard*, den er erfüllt).

Auch im zweiten Versuch schult unser Proband eine fachspezifische Kompetenz. So soll die als fehlerhaft bzw. zu ungenau kritisierte Liste von ChatGPT durch eine *neue* Liste ersetzt werden, die der Schüler selbst mithilfe vertrauenswürdiger Quellen wie literaturgeschichtlicher Lexika erstellen will. Damit ist bereits ein Standard aus dem fachübergreifenden Teil B des Rahmenlehrplans aufgerufen, nämlich das Recherchieren aus zuverlässigen Quellen (vgl. SenBJF 2021: 15). Vor allem aber schult der Schüler seine Kompetenzen im Fach Deutsch: Von der Idee ausgehend, dass ChatGPT bessere Goethe-Fakes produziert, wenn diese aus einer präzise bestimmten Epoche und Gattung kommen sollen, recherchiert er selbstständig literaturgeschichtliches Basiswissen zur Epoche des Sturm und Drang und zu verschiedenen Gattungen der Lyrik, um daraus passende Merkmale von Sturm-und-Drang-Gedichten zu gewinnen: Subjektivität, Affektausdruck, unregelmäßige Reime, freier Rhythmus. Um ChatGPT dazu zu bringen, bessere Goethe-Imitate zu produzieren, recherchiert der Schüler also epochen- und gattungsspezifische Merkmale der Sturm und Drang-Liebeslyrik (*Operation*) und erschließt sich

.....

4 Vgl. zur aktuellen Methodendiskussion über das Verhältnis von ‚Close Reading‘ und ‚Distant Reading‘ in der literaturwissenschaftlichen Theorie und Praxis die programmatische Konferenz „Close Reading“ an der FU Berlin 2023 (URL: <https://www.temporal-communities.de/events/tagung-close-reading.html>; letzter Zugriff: 13.06.2024).

damit historisch-kulturelles Kontextwissen, das eine genauere Einordnung von Goethes Lyrik ermöglicht (*Standard*) – ein weiterer zentraler Standard im Kompetenzbereich „literarische Texte erschließen“ des Rahmenlehrplans Deutsch für die Sekundarstufe I (vgl. SenBJF 2023: 29).

Ab dem dritten Versuch beginnt der Schüler dann, mithilfe von ChatGPT genauer an der konkreten poetischen Ausgestaltung eines Sturm-und-Drang-Gedichts aus dem Geiste Goethes zu arbeiten, wobei mithilfe der KI zunächst Reime produziert und diese zu Strophen kombiniert werden, um sie anschließend auf den Prüfstand zu stellen.

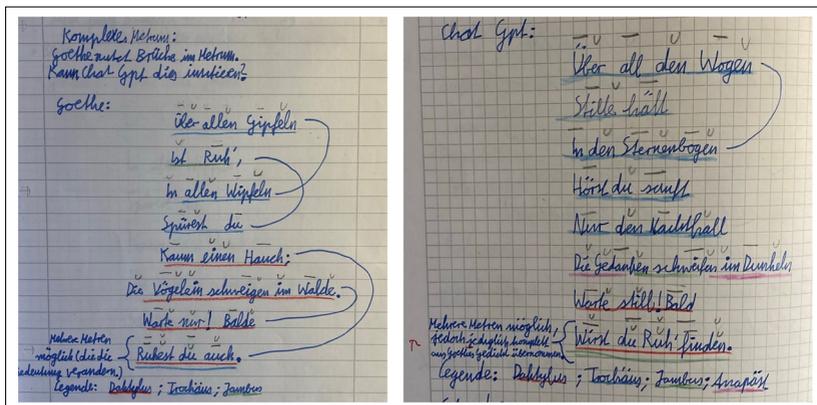


Abb. 4: Vergleich von Goethe- und KI-Metrik im Schülerportfolio

Die Erkenntnis, die daraus gewonnen wird, lautet, dass ChatGPT Lyrik zu sehr im Sinne von ‚Dienst nach Vorschrift‘ produziert, sich also an Schemata von Reim und Metrum orientiert, ohne von diesen abzuweichen. Das Alleinstellungsmerkmal Goethes, auch gezielt mit Brüchen in Reim und Metrum zu arbeiten, kann ChatGPT kaum imitieren, auch wenn der Schüler in einer ganzen Reihe weiterer Versuche daran feilt.

Gleichwohl werden bei diesen Versuchen weitere wertvolle Erkenntnisse gewonnen: Im Tandem mit ChatGPT (vgl. Lehnen/Steinhoff 2023: 2) trainiert der Schüler seine Kompetenz, Texte in Bezug auf ihre sprachliche Gestaltung zu überarbeiten. Der laufende Abgleich von Original und Imitation spornt ihn dazu an, genaue metrische Analysen durchzuführen und am Reim

und Rhythmus der Gedichte zu feilen. Wie die anderen Schüler:innen der Klasse durchläuft der Proband also einen Parcours, der durch die Kompetenzbereiche ‚Texte erschließen‘, ‚Texte planen‘ und ‚Texte überarbeiten‘ führt und dabei zur Einübung ganz konkreter fachspezifischer Standards wie der ‚formalen Analyse‘ lyrischer Texte beiträgt (vgl. SenBJF 2023: 28–29; SenBJF 2022: 20).

Dieser Befund deckt sich mit der im Portfolio abschließend dokumentierten Selbsteinschätzung des Probanden. In Bezug auf die Problemfrage, ob wir ChatGPT dazu bringen können, Lyrik auf Goethe-Niveau zu produzieren, äußert er sich – wie die anderen Schüler:innen – skeptisch, da etwa die Probleme und Schwächen der KI dabei beobachtet wurden, unregelmäßige Reime und Metren jenseits von ‚Schema F‘ zu erzeugen. Mit Blick auf die eigene Motivation und den Kompetenzzuwachs äußert sich der Schüler durchweg positiv: So habe ihm die Arbeit mit ChatGPT einen Ansporn gegeben habe, sich genauer mit Goethes Lyrik zu beschäftigen, wobei betont wird, dass eine genauere Kenntnis von Metren und Reimen ebenso wie literaturgeschichtliches Wissen zu Epochen und Gattungen erlangt wurde. Über diese fachspezifischen Kompetenzen hinaus wird auch das Recherchieren aus zuverlässigen Quellen oder das Formulieren genauer Prompts als Erkenntnisgewinn hervorgehoben (vgl. Falck 2025).

Durch die kritische Reflexion über die Fähigkeiten und Grenzen von KI und über die in der Auseinandersetzung mit dieser erworbenen Kompetenzen gelangt der Schüler damit zur Metakognition im Wortsinn, zu einer Metareflexion über kognitive Operationen und deren Übersetzbarkeit in geeignete Prompts.

### 3 Dritte Sequenz: Interpretieren

Vertieft wurde dies in der dritten Sequenz meiner Reihe, die einem besonders relevanten Anwendungsfeld von ChatGPT im Deutschunterricht gewidmet war, nämlich dem Interpretieren literarischer Texte. Auch dies spielten wir im Unterricht am Beispiel eines exemplarischen Goethe-Gedichts durch, und zwar anhand des in Literaturgeschichte und Deutschunterricht fest veranker-

ten Sturm- und Drang-Gedichts *Prometheus*. Die Leitfrage lautete, ob und inwieweit von ChatGPT erstellte Interpretationen des Gedichts mit denjenigen der Schüler:innen konkurrieren könnten. Damit gewann die Sequenz zugleich einen spielerisch-kompetitiven Charakter im Sinne eines Interpretations-,Wettbewerbs‘ zwischen menschlicher und künstlicher Intelligenz.

In der ersten Etappe der Sequenz erarbeiteten die Schüler:innen zunächst unter Aufsicht im Unterricht eine eigenständige Interpretation ohne Zuhilfenahme von ChatGPT. Die Leitfrage lautete dabei, wie Goethe den Prometheus-Mythos abwandelt und welches autor- und epochentypische Gottes- und Menschenbild dabei zum Ausdruck kommt (Goethe 2006b, 31):

Bedecke deinen Himmel, Zeus  
Mit Wolkendunst  
Und übe dem Knaben gleich  
der Disteln köpft  
an Eichen dich und Bergeshöhn

Dabei sollten auch die rhetorischen und poetischen Mittel analysiert werden, die Goethe verwendet, um seine Positionierung künstlerisch zum Ausdruck zu bringen und die intendierte Wirkung auf das Publikum zu erzielen, beispielsweise die programmatische Verwendung von Imperativen („Bedecke deinen Himmel, Zeus / Mit Wolkendunst“) und Possessiva („Mußt mir meine Erde / Doch lassen stehn“), um die Selbstbehauptung des lyrischen Ichs Prometheus gegenüber der Götterwelt zu betonen.

Während die erste Strophe im Plenum analysiert wurde, um das Interpretationsverfahren zu verdeutlichen, das auf die Verknüpfung von inhaltlicher und formaler Analyse abstellt und gleichzeitig die Wirkungsabsicht reflektiert, wurden die Strophen 2 bis 7 arbeitsteilig in einem ‚Gruppenpuzzle‘ erschlossen, was zusätzlich zum Aspekt von kooperativem und zeiteffizientem Arbeiten auch eine angemessene Differenzierung nach dem individuellen Stand der Kompetenzentwicklung der Schüler:innen ermöglichte.



### III INTERPRETIEREN

Goethe-Gedichte mithilfe von ChatGPT interpretieren  
und die Ergebnisse kritisch reflektieren



#### Schülerprodukt II

	RHETORIK		INHALT
<b>SICHERUNG</b>			
Gruppenpuzzle	<i>palamisch</i> Sput	Ich kenne nichts <u>ärmeres</u> Unter der Sonn' als euch <u>Götter!</u> Ihr <u>nähert</u> <u>kümmertlich</u> Von Opfersteuern Und Gebetshauch Eure Majestät; Und darbiet, wären Nicht Kinder und <u>Bettler</u> <u>Hoffnungsvolle Toren!</u>	<i>Beziehung</i> Der Gott als niederstes Wesen unter der Sonne dem Menschen unterlegen.
Kooperation ✓	<i>1: Antike</i> <i>2: Satire</i>		
Differenzierung ✓	<i>par3</i> <i>ore</i> <i>tele</i>		<i>Menschen die ihre Hoffnung in Götter setzen, sind Narren.</i>
Effizienz ✓		<b>POETIK</b>	

Abb. 5: Interpretation von Goethes „Prometheus“ im Gruppenpuzzle

Die Ergebnispräsentation und -evaluation erfolgte dann wiederum im Plenum, um beurteilen zu können, ob bzw. inwieweit der für die Stunde gewählte Standard der Textanalyse erreicht wurde, nämlich das Beschreiben rhetorischer (und poetischer) Mittel in ihrer Funktion für die Wirkung eines Textes (vgl. SenBJF 2023: 29).

In den nächsten Stunden wurde die zum Verstehen komplementäre Darstellungskompetenz (ebd.: S. 23) trainiert. Die Schüler:innen erhielten den Arbeitsauftrag, ihre Analyseergebnisse in einem Interpretationsaufsatz auszuformulieren. Dieser übergreifende Auftrag wurde operationalisiert, indem klare Vorgaben zu den einzelnen Arbeitsschritten und ihrer Darstellung in den verschiedenen Abschnitten des Aufsatzes formuliert wurden: Dass in der 1) Einleitung also die wichtigsten Daten zum Text, Autor und Epoche benannt (AFB I), zugleich aber schon eine Deutungshypothese zu Goethes poetischer Adaption des Mythos formuliert werden sollte; dass diese Hypothese im 2) Hauptteil in einer genauen Interpretation des Gedichts überprüft werden sollte, wobei die Analyse rhetorischer und poetischer Mittel stets im Hinblick auf die Aussage- und Wirkungsabsicht erfolgen sollte (AFB II/III); und dass im 3) Schlussteil die wichtigsten Analyseergebnisse zu einer Gesamtdeutung gebündelt werden sollten, um zugleich eine weiterführende Perspektive zur

Einordnung des Gedichts in Werk und Epoche zu skizzieren (AFB III). Im bewährten Modell einer ‚Schreibkonferenz‘ wurden die individuellen Ausarbeitungen der einzelnen Schüler:innen anhand eines Bewertungsbogens im Peerfeedback ausgewertet, wobei die Lehrkraft den Gesamtprozess beobachtete und abschließend eine verbindliche Rückmeldung zu den Ergebnissen gab (vgl. zu dieser Praxis: Lehnen 2023).

Auf der Basis einer eigenständigen Erarbeitung des *Prometheus*-Gedichts wurden in der zweiten Etappe der Unterrichtssequenz die Fähigkeiten des Chatbots geprüft, einen Interpretationsaufsatz auf vergleichbarem Niveau zu produzieren;<sup>5</sup> spielerisch-kompetitiv zugespitzt zur Leitfrage: ‚Wer kann Goethes *Prometheus* besser interpretieren, Mensch oder Bot?‘<sup>6</sup>

Im Plenum wurde zunächst vorgeführt, dass zu allgemein formulierte Befehle an ChatGPT wie „Interpretiere Goethes Gedicht *Prometheus*“ nicht zum erwünschten Ergebnis führen; so lieferte der Chatbot – in der im Schuljahr 2023/2024 verfügbaren Version 3.5 – auf diesen Prompt hin zwar die passenden Schlagworte (Goethes Prometheus als ‚Symbol für den rebellischen Geist‘ der Sturm- und Drang-Epoche), belegte die Interpretationsbeobachtungen jedoch nicht am Text. Auf die Aufforderung hin, passende Textbelege anzuführen, besserte der Bot zwar deutlich nach (etwa durch die Interpretation des Eröffnungsverses „Bedecke deinen Himmel, Zeus“ als ‚mutige Herausforderung an die göttliche Autorität‘), lieferte jedoch auch fehlerhafte, teilweise im Text nicht enthaltene, halluzinierte Zitate. Und auf einen ersten Prompt hin, der auf die Interpretation des vollständigen Gedichts abstellt, geht der Bot auf einzelne Strophen ein, unterschlägt jedoch die zentrale vierte Strophe mit der Metapher des ‚heilig glühend Herz‘ als Dreh- und Angelpunkt einer jeden Interpretation des Gedichts. Erst nachdem der vollständige Originaltext eingespeist wurde, stellt ChatGPT derart eklatante und vermeidbare Fehler ab.

.....

- 5 Wohl wissend, dass dieser ein Verständnis der literarischen Texte nur auf der sprachlichen Oberflächenebene simuliert. Vgl. zu dieser Funktionsweise von GPTs grundlegend: Fürstenberg/Müller 2024: 2–5.
- 6 Anders als in einer neueren Studie (vgl. Magirius et al. 2024: 17–21) war das Interpretations-Setting also weniger kooperativ ausgerichtet (ChatGPT als ‚Schreib-‘ bzw. ‚Dialogpartner‘) als konkurrenzial („Sparringspartner“).

„Lesson learnt“: Zunächst muss die Textgrundlage gesichert werden, und die Arbeitsaufträge an den Bot dürfen nicht allzu vage sein.

Von der initialen Erkenntnis ausgehend, dass der Bot tendenziell bessere Ergebnisse liefert, wenn die Prompts hinreichend präzise sind, wurden die Schüler:innen in die Erarbeitungsphase geschickt. Analog zur zuvor organisierten ‚Schreibkonferenz‘ wurde nun eine ‚Promptkonferenz‘ durchgeführt: In Teams sollten die Schüler:innen daran arbeiten, eine Serie möglichst passgenauer Prompts für ChatGPT zu formulieren, um das gewünschte Ergebnis einer angemessenen Gedichtinterpretation zu produzieren. Die in den Gruppen erarbeiteten Prompts und die daraus resultierenden Interpretationsergebnisse wurden wiederum im Plenum vorgestellt und vergleichend evaluiert.

Auf dieser Grundlage wurde in der abschließenden Reflexion herausgearbeitet, dass sich die Präzision der Prompts und damit die Qualität der Ergebnisse insbesondere durch eine Operationalisierung der Arbeitsaufträge steigern lässt. So kam ChatGPT dem gewünschten Interpretationsaufsatz am nächsten, wenn die Schüler:innen für den Bot einen ähnlichen Schreibauftrag formulierten wie denjenigen, der in der ersten Etappe an sie selbst ausgegeben wurde. Im Sinne von „Verfasse eine Interpretation zu Goethes *Prometheus*, indem du 1) das *Thema des Gedichts benennst* und *Gattung und Epoche bestimmst* (Einleitung), 2) das *Zusammenspiel von Inhalt und Form* mit Blick auf die *Wirkung analysierst* (Hauptteil) und 3) deine Beobachtungen abschließend zu einer übergreifenden Deutungsthese bündelst (Fazit).“

**You**  
Verfasse eine Interpretation zu Goethes „Prometheus“,

**indem Du**      **Operationalisierung**

das **Thema** des Gedichts **benennst** und **Gattung** und **Epoche bestimmst** (Einleitung)      AFB I/II

das **Zusammenspiel** von **Inhalt, Form und Wirkung analysierst** (Hauptteil),      AFB II/III

und Deine Beobachtungen abschließend zu einer **Deutungsthese bündelst** (Fazit).      AFB III

**Abb. 6:** Präzisierung des Prompts durch Operationalisierung

Mit diesen Prompts produzierte der Bot zunächst brauchbare Ergebnisse, deren konkrete Ausformulierung durch weiteres Nachsteuern optimiert werden konnte. So wurde der konkretisierende Befehl an ChatGPT, das Gedicht im Hauptteil Strophe für Strophe zu analysieren, zwar korrekt ausgeführt, aber das Ergebnis sprachlich zu repetitiv dargestellt („In Strophe 1“, „In Strophe 2“ etc.). Hier konnte durch den Befehl, die Einleitungssätze zu den unterschiedlichen Absätzen des Hauptteils sprachlich abwechslungsreicher zu gestalten, nachgebessert werden.

Zur genaueren Evaluation ausgewählter Botinterpretationen wurde den Schüler:innen der 10. Klasse schließlich ein differenzierter Beobachtungsbogen ausgehändigt, der sich an den neuen Berliner Klausurgutachten ([www.klausurgutachten.de](http://www.klausurgutachten.de)) für die gymnasiale Oberstufe orientiert. Blickleitend ist dabei die Unterscheidung von Verstehens- und Darstellungsleistung, die wiederum jeweils in verschiedene Bewertungssitems untergliedert sind. Dabei fielen den Schüler:innen Defizite des Chatbots insbesondere dabei auf, rhetorische und poetische Mittel nicht nur als solche zu identifizieren, sondern auch deren Funktion für die Wirkung des Gedichts zu interpretieren. Auch die synthetisierende Verknüpfung der Einzelbeobachtungen im Hinblick auf die übergreifende Deutungshypothese gelang ChatGPT nur teilweise. Diese Defizite mögen symptomatisch dafür sein, dass der Bot eine „Verstehensleistung“ im Sinne des Abiturgutachtens nur sprachlich simuliert (vgl. Fürstenberg/Müller 2024: 2–5). Bei der Darstellungsleistung störte eine teilweise noch zu schematisch wirkende Verwendung von Sprachmustern, insbesondere in den Überleitungen.

So gelangten wir in der Klasse zum Fazit, dass der Chatbot – zumindest im Schuljahr 2023/2024 – die Analyse- und Interpretationskompetenzen, welche die Schüler:innen unter Beweis stellen sollen, nur mit viel menschlicher Einhilfe und auch nicht vollständig simulieren kann.<sup>7</sup> Eine tröstliche Botschaft, passend zum humanistischen Bildungsprogramm des Grauen Klosters.

.....  
7 Zu einem ähnlichen Fazit gelangen die Autorinnen einer sprachwissenschaftlich fundierten Fallstudie zum KI-gestützten Verfassen von Interpretationsaufsätzen zu Kafka; vgl. Kellermann/Führer/Gese 2024: 34.

Mit Blick auf die Kompetenzentwicklung der Schüler:innen hat sich der Einsatz von ChatGPT jedoch als überaus gewinnbringend erwiesen. So kann die Analyse- und Interpretationskompetenz der Schüler:innen in der simulierten Wettbewerbssituation mit dem Chatbot angeregt und gezielt gefördert werden, im Hinblick sowohl auf die Verstehens- als auch die Darstellungskomponente. Die eigenen Interpretationen mit denjenigen des Bots abzugleichen, schärft den Blick für die Anforderungen der gängigen Schreibformate und den Stand der eigenen Kompetenzentwicklung. Gleichzeitig wird ein metakognitives Erkenntnispotenzial aktiviert, indem die Schüler:innen ein Bewusstsein für die Denk- und Schreiboperationen entwickeln, die sie kompetent ausführen müssen, um eine Interpretationsaufgabe angemessen lösen zu können: Die Schüler:innen erkennen, dass sie präzise Operationen ausführen müssen – genauso, wie der Chatbot nur angemessenen Output generiert, wenn er passende Prompts erhält. Bei aller Analogie von Operation u. Prompt bleibt ein menschlicher Faktor ‚x‘ insbesondere bei Verstehens- und synthetisierenden Darstellungsleistungen, der seinerseits genauer operationalisiert werden müsste, um computersprachlich mithilfe von Chatbots ansatzweise reproduziert/simuliert werden zu können.

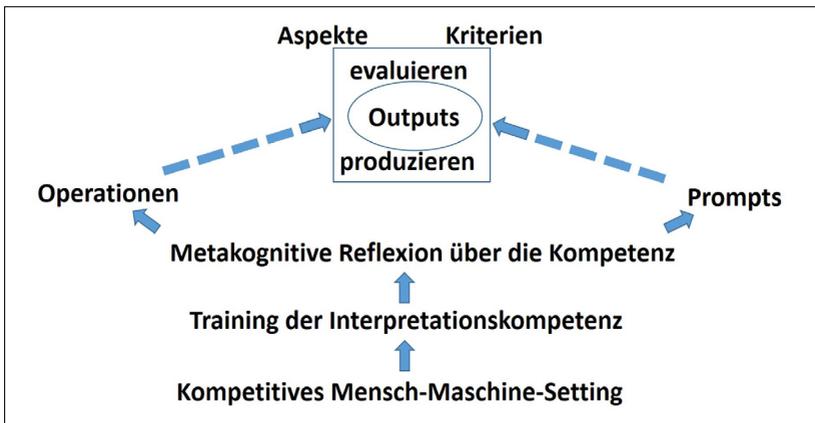


Abb. 7: Darstellung des Erkenntnispotenzials des Mensch-Maschine-Settings

## 4 Vierte Sequenz: Aktualisieren

Ein viertes mögliches Anwendungsfeld von ChatGPT im Deutschunterricht wurde in der finalen Sequenz zum Aktualisieren erprobt. Die Schüler:innen erhielten den Arbeitsauftrag, mithilfe des Chatbots eine aktualisierende Nachdichtung von Goethes *Prometheus* zu produzieren, welche Themen und Konflikte der heutigen Zeit verhandelt. Ausgewählte Botgedichte sollten dann, analog zu Goethes Vorlage, zunächst im Hinblick auf die Verwendung rhetorischer und poetischer Mittel und ihre Wirkung analysiert werden, um abschließend ein Urteil zu bilden, ob die Nachahmungen als ‚gelingen‘ zu bewerten sind.

In einem Gedicht des Chatbots, das für die gemeinsame Analyse ausgewählt wurde, wurde das mythologische Setting der Rebellion des Prometheus gegen die olympische Götterwelt konsequent ins technologische Zeitalter übersetzt. Gerade der synoptische Vergleich zwischen Goethes Original und der Nachdichtung des Bots erwies sich dabei als überaus produktiv für die Analyse. Das zeigt schon der vergleichende Blick auf *Prometheus* (Goethe 2006b, 31) und ‚Prometheus 2.0‘:

Goethes „Prometheus“	ChatGPTs „Prometheus 2.0“
Bedecke deinen Himmel, Zevs, Mit Wolkendunst Und übe, dem Knaben gleich Der Disteln köpft, An Eichen dich und Bergeshöhn Mußt mir meine Erde Doch lassen stehn Und meine Hütte die du nicht gebaut, Und meinen Herd Um dessen Glut Du mich beneidest.	Verhülle deinen Himmel, Technologie, Mit Datenfluten! Und erprobe, wie ein Schöpfer, Der Algorithmen trimmt, An Programmcodes und virtuellen Höh'n! Musst mir meine Welt Doch lassen steh'n, Und meinen Cyberspace, den Du nicht erschaffen Und meinen Code Um dessen Logik Du mich beneidest.

In der vergleichenden Analyse der ersten Strophe markierten die Schüler:innen zunächst die wesentlichen, durch den Chatbot erzeugten sprachlichen und inhaltlichen Abwandlungen und interpretierten diese anschließend: Dabei wurde insbesondere die Ersetzung von Begriffen und Metaphern der

göttlich-mythologischen Sphäre durch solche der technologisch-virtuellen Welt herausgearbeitet (V. 1: Zeus → Technologie; V. 2: Wolkendunst → Datenfluten; V. 4: Disteln → Algorithmen etc.). Diskutiert wurden dabei auch die semantischen Implikationen der KI-gestützten Reorganisation des Sprachmaterials, etwa am Beispiel der Ersetzungen von „Herd“/„Code“ (V. 9) und „Glut“/„Logik“ (V. 10), die im Sinne einer Neudefinition des Menschen nicht als Gefühls-, sondern als Verstandeswesen interpretiert werden könnten.

Wie bei der Analyse des Originalgedichts wurde der Gedichtvergleich für die folgenden Strophen 2–7 in der bewährten Sozialform des ‚Gruppenpuzzles‘ durchgeführt. Gefördert wurde mit dieser Sequenz wieder die textanalytische Kompetenz, die Verwendung rhetorischer und poetischer Mittel in ihrer Funktion für die Wirkung eines Textes zu beschreiben (vgl. SenBJF 2023: 29). Durch den Vergleich von Goethe- und Botgedicht wurde zugleich die Analyseform des ‚Gedichtvergleichs‘ eingeübt, die als Aufgabenformat in der gymnasialen Oberstufe und im Abitur zunehmend an Bedeutung gewinnt (vgl. SenBJF 2024: S. 3).

Der Kompetenzschwerpunkt der Sequenz lag jedoch auf der Beurteilung der KI-generierten Aktualisierung (vgl. SenBJF 2023: 29). So sollten die Schüler:innen sich ein abschließendes Urteil zur Frage bilden, ob ChatGPT eine inhaltlich und sprachlich überzeugende Aktualisierung von Goethes Prometheus-Gedicht gelingt, die eine starke Wirkung auf heutige Leser:innen erzielt. Untergliedert wurde diese Urteilsfrage in vier Kriterien, die auf eine Gesamtbewertung von Werk und Wirkung in hinführen sollten:

- 1) Grundkonzept: Ist das Konzept plausibel, den Konflikt des Prometheus-Mythos zu aktualisieren, indem dieser nicht zwischen Mensch und Gott, sondern zwischen Programmierer und Technologie stattfindet?
- 2) Inhaltliche Tiefe/Komplexität: Ist der Inhalt des Gedichtes komplex, indem viele verschiedene Aspekte dieses Konflikts zwischen Mensch und Datengöttern dargestellt werden?
- 3) Geschlossenheit/Widerspruchsfreiheit: Ist dieser Grundkonflikt auf inhaltlicher Ebene kohärent ausgestaltet oder ergeben sich logische Widersprüche, Fehler, Leerstellen?

- 4) Sprachliche Gestaltung: Ist die sprachliche Gestaltung gelungen, werden also etwa Stilmittel aus dem Goethe-Gedicht passend modernisiert? Oder verliert das Gedicht deutlich an poetischer Qualität?
- Gesamturteil zur Wirkung: Kann die Aktualisierung eine starke Wirkung auf ein heutiges Lesepublikum erzielen, intellektuell oder affektiv, also zum kritischen Nachdenken oder zu starken Gefühlen anregen?

Bei der qualitativen Auswertung der Schülerurteile ergab sich (erwartungsgemäß) folgendes Gesamtbild: Während das Grundkonzept der Aktualisierung, den mythologischen in einen technologischen Konflikt zu übersetzen, grundsätzlich als überzeugend eingeschätzt wurde und bei der inhaltlichen und sprachlichen Ausgestaltung auch originell wirkende Leitbegriffe und -metaphern hervorgehoben wurden, monierten viele Schüler:innen die über sieben Strophen redundant wirkende Wiederholung von Motiven der digitalen Welt, deren sprachlicher Code zudem als teilweise unpassend bewertet wurde (etwa wenn der Doppelvers in der sechsten Strophe „Weil nicht alle / Blümenträume reifen“ ersetzt wird durch „Weil nicht alle Bits und Bytes / In Harmonie tanzten“).

Allerdings fiel das Chatbot-Gedicht qualitativ auch nicht so stark ab, um nicht eine ernsthafte Vergleichsoperation motivieren zu können. Insofern wird es auch in Zukunft ein überaus nützliches Tool sein, um lyrisches Material für Gedichtvergleiche zu produzieren – und anhand dessen zu reflektieren, ob und inwieweit die Produktion und Rezeption lyrischer Dichtkunst als quasi-prometheisches Alleinstellungsmerkmal der menschlichen Gattung gelten kann (vgl. in diesem Sinne: Odendahl 2023: 13) oder als Sprachphänomen doch technisch weitgehend reproduzierbar ist.

## 5 Fazit und Perspektive

Meine Unterrichtsreihe „Goethe vs. ChatGPT“ hat exemplarisch gezeigt, dass die Nutzung von ChatGPT bei der Einbettung in passende Lehr- und

Lernarrangements einen vierfachen Mehrwert für den Deutschunterricht verspricht. *Erstens* kann die Motivation der Schüler:innen, sich vertieft mit literarischen Texten zu beschäftigen, durch den kontrollierten Einsatz des Tools gesteigert werden. An ihren digitalen ‚Spieltrieb‘ anknüpfend können die Schüler:innen dazu angeregt werden, die Fähigkeiten des Chatbots beim Imitieren, Interpretieren oder Aktualisieren literarischer Texte zu erproben und diese – im spielerisch-kompetitiven Modus eines Wettbewerbs ‚Mensch vs. Maschine‘ – mit ihren eigenen Fertigkeiten zu vergleichen. Gefördert werden kann dies insbesondere durch relativ offene Unterrichtsarrangements wie die beschriebene Portfolioarbeit. *Zweitens* kann die Arbeit der Schüler:innen mit dem Tool zur Förderung fachspezifischer Kompetenzen beitragen: Das betrifft alle drei Teilkompetenzen des Umgangs mit literarischen Texten, nämlich sowohl das ‚Texte erschließen‘, als auch das ‚Texte planen‘ und das ‚Texte überarbeiten‘. Exemplarisch wurde dies in der Sequenz zum Imitieren deutlich: So wurde die analytische Kompetenz der Texterschließung durch den Vergleich von Goethe-Gedichten mit Bot-Imitationen gefördert; dabei wurde der Blick für grundlegende Merkmale lyrischer Texte wie das Metrum geschärft. Beim Versuch, die Qualität der generischen KI-Texte zu steigern, wurden komplexere Fähigkeiten zur Planung und Überarbeitung literarischer Texte gefördert. *Drittens* kann eine funktionale Nutzung von ChatGPT zur Förderung des metakognitiven Reflexionsvermögens der Schüler:innen beitragen. Besonders deutlich wurde dies in der Sequenz zum Interpretieren. Indem die Schüler:innen ihre eigenen Interpretationsaufsätze mit KI-generierten Exemplaren vergleichen, diese kritisch evaluieren und an deren Verbesserung arbeiten, schärfen sie ihr Bewusstsein für die Besonderheiten des Formats und die dafür erforderlichen Denk- und Schreiboperationen; deutlich wurde den Schüler:innen dabei die Analogie zwischen ihrem ‚operatorenbasierten‘ Denk- und dem ‚promptbasierten‘ Sprachprozess des Bots. Schließlich und *viertens* kann der Deutschunterricht durch die angeleitete Reflexion der Möglichkeiten und Grenzen des Chatbots bei der Textanalyse und -produktion einen wichtigen Beitrag zum überfachlichen Kompetenzziel der Medienbildung leisten. Eine erfolgreiche Adaption der hier skizzierten Unterrichtsreihe hat dabei natürlich zur Voraussetzung, dass die einzelnen Sequenzen auf das Kompetenzniveau

der jeweiligen Lerngruppe ebenso wie auf das Performanzlevel der sich stets weiterentwickelnden Chatbots abgestimmt werden.

Was in meiner Unterrichtsreihe mit Blick auf die KI-gestützte Verarbeitung literarischer Texte skizziert wurde, lässt sich analog auch im Hinblick auf *pragmatische* Texte durchführen. Im Deutschunterricht eingeübte Gattungen, deren Untersuchung sich in einer analogen Versuchsanordnung als besonders ergiebig erweisen könnte, sind das ‚materialgestützte Informieren‘ bzw. ‚Argumentieren‘ (wobei es sich bei dem eingespeisten Material wiederum teilweise um literarische Texte handeln könnte). Eine naheliegende Anschlussfrage wäre etwa, ob sich der Chatbot darauf programmieren lässt, eine überzeugende, den Anforderungen der Abiturgattung entsprechenden Argumentation zum Problem zu entwickeln, ob die Analyse von Goethe-Lyrik im Deutschunterricht künftig KI-gestützt erfolgen sollte. Auch hier wird der KI-generierte Texte zunächst sicher Schwächen aufweisen; mithilfe welcher Prompts nachgebessert werden kann, dürfte jedoch aufschlussreich für die Operationalisierung des Genres sein – aus der Sicht der Schüler:innen, die diese Schreibaufgaben lösen müssen, ebenso wie aus der Perspektive der Lehrer:innen, die solche Aufgaben stellen und kriteriengeleitet bewerten müssen.

## Literatur

### Primärtexte

JOHANN WOLFGANG GOETHE (2006a): *Wandrer's Nachtlid*. In: Reinhardt, Hartmut (Hg.): *Johann Wolfgang Goethe. Sämtliche Werke nach Epochen seines Schaffens*. Münchner Ausgabe: btb Verlag 2006. Band 2.1: Erstes Weimarer Jahrzehnt 1775–1786, S. 53.

JOHANN WOLFGANG GOETHE (2006b): *Prometheus*. In: Hans J. Becker, Hans-Georg Dewitz, Norbert Miller, Gerhard H. Müller, John Neubauer, Hartmut Reinhardt u. Irmtraut Schmidt (Hg.): *Johann Wolfgang Goethe. Sämtliche Werke nach Epochen seines Schaffens*. Münchner Ausgabe: btb Verlag 2006. Band 3.2: Italien und Weimar 1786–1790, S. 31–32.

JOHANN WOLFGANG GOETHE (2006c): *Der Zauberlehrling*. In: Reiner Wild (Hg.): *Johann Wolfgang Goethe. Sämtliche Werke nach Epochen seines Schaffens*. Münchener Ausgabe: btb Verlag 2006. Band 4.1: Wirkungen der Französischen Revolution 1791–1797, S. 874–877.

## Sekundärtexte

### **Curricula u. didaktische Handreichungen**

Senatsverwaltung für Bildung, Jugend und Familie des Landes Berlin (SenBJF)/Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (Hg.) (2023): *Fachteil C Deutsch des Rahmenlehrplans für die Jahrgangsstufen 1 bis 10*. Berlin/Potsdam. URL: [https://www.berlin.de/sen/bildung/unterricht/faecher-rahmenlehrplaene/rahmenlehrplaene/rlp-deutsch\\_1-10-teil-c.pdf](https://www.berlin.de/sen/bildung/unterricht/faecher-rahmenlehrplaene/rahmenlehrplaene/rlp-deutsch_1-10-teil-c.pdf) (letzter Zugriff: 19.4.2025).

Senatsverwaltung für Bildung, Jugend und Familie des Landes Berlin (SenBJF) (Hg.) (2022): *Fachteil C Deutsch des Rahmenlehrplans für die gymnasiale Oberstufe*. Berlin. URL: <https://www.berlin.de/sen/bildung/unterricht/faecher-rahmenlehrplaene/rahmenlehrplaene/rahmenlehrplan-deutsch-go-teil-c.pdf> (letzter Zugriff: 19.4.2025).

Senatsverwaltung für Bildung, Jugend und Familie des Landes Berlin/Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (Hg.) (2021): *Teil B Fachübergreifende Kompetenzentwicklung*. Berlin/Potsdam. URL: [https://www.berlin.de/sen/bildung/unterricht/faecher-rahmenlehrplaene/rahmenlehrplaene/rlp\\_go\\_teil\\_b\\_2021.pdf](https://www.berlin.de/sen/bildung/unterricht/faecher-rahmenlehrplaene/rahmenlehrplaene/rlp_go_teil_b_2021.pdf) (letzter Zugriff: 19.4.2025).

Senatsverwaltung für Bildung, Jugend und Familie des Landes Berlin (SenBJF) (Hg.) (2024): *Fachbrief Deutsch Berlin Nr. 41. Themenschwerpunkt: Entwicklungen im Abitur ab 2025. Herausforderungen und Chancen*. URL: [https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/fachbriefe\\_berlin/deutsch/Fachbrief\\_Deutsch\\_01.pdf](https://bildungsserver.berlin-brandenburg.de/fileadmin/bbb/unterricht/fachbriefe_berlin/deutsch/Fachbrief_Deutsch_01.pdf) (letzter Zugriff: 19.4.2025).

### **Forschungsliteratur**

FALCK, JOSCHA (2025): KI-Kompetenzen. Diese Fähigkeiten brauchen Lehrkräfte und Schüler. In: *Deutsches Schulportal 2025*. URL: <https://deutsches-schulportal.de/>

- expertenstimmen/ki-kompetenzen-diese-faehigkeiten-brauchen-lehrkraefte-und-schueler/ (letzter Zugriff: 24.5.2025)
- FÜHRER, CAROLIN/GERJETS, PETER (2024): How to understand & write literature with AI? Potentiale und Risiken von KI-Tools für literarisches Lesen und Schreiben. In: *Medien im Deutschunterricht* 1/2024, S. 1–18. URL: <https://doi.org/10.18716/OJS/MIDU/2024.1.3> (letzter Zugriff: 13.6.2025).
- FÜRSTENBERG, MAURICE/MÜLLER, HANS-GEORG (2024): KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* 5/2024, S. 2–13.
- KELLERMANN, KATHARINA/FÜHRER, CAROLIN/GESE, HELGA (2024): Kann ChatGPT das Schreiben über Literatur sprachlich stützen? Bildungssprachliche Mittel und Textprozeduren in computergestützter schriftlicher Anschlusskommunikation. In: *Der Deutschunterricht* 5/2024, S. 24–35.
- LEHNEN, KATRIN/STEINHOFF, TORSTEN (2023): Sprachliche Bildung als menschlich-maschinelle Koaktivität. Zur (Re-)Modellierung von Lese- und Schreibpraktiken im Deutschunterricht. In: Fürstenberg, Maurice; Müller, Hans-Georg; Wurst, Alexander (Hg.): *DeutschGPT. Konferenz zum Deutschunterricht in Zeiten von Chatbots und KI. Book of Abstracts*, S. 2. URL: [https://www.fachplusdidaktik.gwi.uni-muenchen.de/wp-content/uploads/2023/08/DeutschGPT\\_Book\\_of\\_Abstracts.pdf](https://www.fachplusdidaktik.gwi.uni-muenchen.de/wp-content/uploads/2023/08/DeutschGPT_Book_of_Abstracts.pdf) (letzter Zugriff: 17.4.2025).
- LEHNEN, KATRIN (2023): Peerfeedback beim schulischen Schreiben. Grundlagen, Methoden, Modellierung. In: *ide – Informationen zur Deutschdidaktik* 47 (2023) 2, S. 18–30.
- MAGIRIUS, MARCO/HESSE, FLORIAN/HELM, GERRIT/SCHERF, DANIEL (2024): KI im Literaturunterricht. Chancen und Herausforderungen zwei Jahre nach der Veröffentlichung von ChatGPT. In: *Der Deutschunterricht* 5/2024, S. 14–23.
- ODENDAHL, JOHANNES (2023): Kein Grund, sich zu schämen, Prometheus! ChatGPT und die überholte Kompetenzorientierung. In: Fürstenberg, Maurice; Müller, Hans-Georg; Wurst, Alexander (Hg.): *DeutschGPT. Konferenz zum Deutschunterricht in Zeiten von Chatbots und KI. Book of Abstracts*, S. 13. URL: [https://www.fachplusdidaktik.gwi.uni-muenchen.de/wp-content/uploads/2023/08/DeutschGPT\\_Book\\_of\\_Abstracts.pdf](https://www.fachplusdidaktik.gwi.uni-muenchen.de/wp-content/uploads/2023/08/DeutschGPT_Book_of_Abstracts.pdf) (letzter Zugriff: 17.4.2025).

PORTER, BRIAN/MACHERY, EDOUARD (2024): AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. In: *Scientific Reports* 14 (2024). URL: <https://doi.org/10.1038/s41598-024-76900-1> (letzter Zugriff: 13.06.2025).



## Beiträger:innen

**Anna Ansari**, M.A., ist wissenschaftliche Mitarbeiterin an der Justus-Liebig-Universität Gießen. Sie untersucht in ihrer Dissertation Schreibpraktiken von Schüler:innen im Umgang mit textgenerativer KI. Ziel ist es, die Rolle der KI im Schreibprozess zu analysieren und didaktische Perspektiven für einen Einsatz im Deutschunterricht zu entwickeln.

**Prof. Dr. Tatjana Atanasoska** arbeitet an der Pädagogischen Hochschule Oberösterreich. Ihre Forschungsschwerpunkte liegen – insbesondere in Bezug auf die Professionalisierung von Lehrer:innen – im Bereich Mehrsprachigkeit und Migration, Deutsch als Zweitsprache und historischer Bildungsforschung mit Fokus auf Süd-Jugoslawien.

**Friedrich Bach**, M.Sc., ist wissenschaftlicher Mitarbeiter am REACH-EUREGIO Start-up Center an der Universität Münster und Co-Leiter des Projekts Incub.AI. Seine Forschungsinteressen liegen in den Bereichen KI, Big Data und KI in der Hochschullehre.

**Prof. Dr. Sebastian Bernhardt** ist Professor für Literatur- und Mediendidaktik am Germanistischen Institut der Universität Münster. Zu seinen Forschungsinteressen gehören Literatúrausstellungen und ihre Potenziale für das literarische Lernen, diversitätssensibler Literaturunterricht, Hörspiellidaktik und transmediale Literaturdidaktik sowie KI in der Hochschullehre.

**Prof. Dr. Katrin Böhme** ist an der Universität Potsdam Professorin für Inklusionspädagogik mit dem Schwerpunkt Sprache. Sie beschäftigt sich in Forschung und Lehre mit sprachlicher Heterogenität sowie Sprachbildung und Sprachförderung in der inklusiven Schule. Ein besonderer Schwerpunkt ihrer Arbeit liegt auf dem Einsatz digitaler Medien und künstlicher Intelligenz für die Unterstützung sprachlicher Lernprozesse.

**Dr. Irene Corvacho del Toro** ist wissenschaftliche Mitarbeiterin und Co-Leiterin im Projekt KIMADU an der Universität Siegen. Ihre Forschungsschwerpunkte sind neben KI im Deutschunterricht die Professionalisierung von Lehrkräften, Theorie, Erwerb und Didaktik der deutschen Rechtschreibung, Rechtschreibschwierigkeiten und Rechtschreibstörung, Emotionswissen und Emotionswortschatz, Erzählfähigkeit, Mehrsprachigkeit und Sprachförderung.

**Dr. des. Mareike Fuhlrott** ist wissenschaftliche Mitarbeiterin und Co-Leiterin im Projekt KIMADU an der Universität Siegen. Ihre Forschungsschwerpunkte umfassen generative KI im Deutsch- und Fachunterricht, sprachlich-fachliches Lernen, Aufgaben, Schreibdidaktik und Querschnittsthemen.

**Dr. Maurice Fürstenberg** ist Akademischer Rat in der Germanistischen Linguistik an der Ludwig-Maximilians-Universität München. Er arbeitet seit drei Jahren zum Thema KI im Deutschunterricht. Zuvor war er Postdoktorand an der Universität Tübingen. Seine Forschungsschwerpunkte liegen in den Bereichen Syntax, Graphematik, Grammatikunterricht, Schriftspracherwerb, Statistik und in der empirischen Bildungsforschung.

**Prof. Dr. Olaf Gätje** ist Professor für Sprachwissenschaft/Sprachdidaktik an der Universität Kassel. Schwerpunkte seiner Forschung sind die Sprachwissenschaft sowie die Sprach- und Mediendidaktik.

**Dr. Anna Jachimek** ist wissenschaftliche Mitarbeiterin am Institut für Psycholinguistik und Didaktik der deutschen Sprache der Goethe Universität Frankfurt. Zu ihren Lehr- und Forschungsschwerpunkten gehören Erst- und Zweitspracherwerb sowie Mehrsprachigkeit und Grammatikdidaktik, insbesondere im sprachheterogenen Kontext.

**Janne Mesenhöller**, M.Sc., ist Psychologin und als akademische Mitarbeitende im Forschungs- und Transferprojekt KISS-Pro an der Universität Potsdam tätig. In diesem Projekt entwickelt und erprobt sie Professionalisierungsangebote für Lehrkräfte zum Thema künstliche Intelligenz in Schule und Un-

terrichtet. Sie promoviert gegenwärtig zu ethischen, rechtlichen und sozialen Implikationen künstlicher Intelligenz im Bildungskontext.

**Prof. Dr. Maik Philipp** ist an der Pädagogischen Hochschule Zürich Professor für Deutschdidaktik. Seine Schwerpunkte liegen in der Lese- und Schreibdidaktik mit einem besonderen Fokus auf digitalem Lesen und Schreiben, Selbstregulation und Lernen mit Texten.

**Dr. Kaspar Renner** ist akademischer Mitarbeiter am Institut für Germanistik der Universität Potsdam und Lehrer für Deutsch und Geschichte am Evangelischen Gymnasium zum Grauen Kloster in Berlin. Neben der Aufklärungsforschung gehören Literatur- und Mediendidaktik zu seinen Forschungsschwerpunkten.

**Prof. Dr. Silvia Reuvekamp** ist Professorin für Deutsche Philologie (Literatur des Mittelalters) am Germanistischen Institut der Universität Münster. Sie ist zudem Prodekanin für Digitalisierung und Entwicklung neuer Lehrformate des Fachbereichs 09 Philologie der Universität Münster. Ihre Forschungsinteressen umfassen Wissens- und Bildungsgeschichte, Mediengeschichte und Medienwechsel sowie KI in der Hochschullehre.

**Ninja Schmiedgen**, M.A., ist wissenschaftliche Mitarbeiterin an der Forschungsstelle Innovation und Transfer digitaler Lehre (ITdL) im Fachbereich 09 der Universität Münster. Ihre Forschungsinteressen sind Digitalität in der Hochschullehre sowie KI in der Hochschullehre.

**Prof. Dr. Katharina Simbeck** lehrt im Studiengang Wirtschaftsinformatik an der HTW Berlin und leitet die Forschungsgruppe Informatik und Gesellschaft. Sie befasst sich mit der Fairness und Erklärbarkeit von KI-Systemen sowie mit dem Einsatz von KI im Bildungs- und Personalwesen.

**Prof. Dr. Torsten Steinhoff** ist Professor für Didaktik der deutschen Sprache an der Universität Siegen und verantwortet das Projekt KIMADU. Er forscht

Beiträger:innen

zur Schreibdidaktik, zur Wortschatzdidaktik, zum sprachlich-fachlichen Lernen und zu digitaler Kommunikation/KI.

**Tobias Weindel**, M.A., ist wissenschaftlicher Mitarbeiter für Sprachwissenschaft/Sprachdidaktik an der Universität Kassel. Seine Forschungsschwerpunkte sind die Schreib- und Gesprächsdidaktik.



# LITERATUR – MEDIEN – DIDAKTIK

- Bd. 1 Sebastian Bernhardt/Thomas Hardtke (Hg.):  
Interpretation – Literaturdidaktische Perspektiven.  
292 Seiten. ISBN 978-3-7329-0818-9
- Bd. 2 Sebastian Bernhardt (Hg.): Frank Maria Reifenbergs Werke  
im literaturdidaktischen Fokus. 380 Seiten.  
ISBN 978-3-7329-0908-7
- Bd. 3 Sebastian Bernhardt/Eva-Maria Dichtl (Hg.): Frühkindliches Spiel  
und literarische Rezeption. Perspektiven der Kindheitspädagogik  
und der Literaturdidaktik. 256 Seiten. ISBN 978-3-7329-0903-2
- Bd. 4 Stefan Emmersberger/Lea Grimm (Hg.): Kurzfilme im  
Deutschunterricht. Fachwissenschaftliche und fachdidaktische  
Perspektiven. 206 Seiten. ISBN 978-3-7329-0797-7
- Bd. 5 Sebastian Bernhardt/Kirsten Kumschlies (Hg.):  
Antonia Michaelis' Werke im literaturdidaktischen Fokus.  
260 Seiten. ISBN 978-3-7329-0915-5
- Bd. 6 Natalie Beck: Rassismussensibler Literaturunterricht  
in der Grundschule. Mediendidaktische Perspektiven.  
136 Seiten. ISBN 978-3-7329-1007-6
- Bd. 7 Ines Heiser: Handlungs- und produktionsorientierter  
Literaturunterricht in Zeiten der Kompetenzorientierung.  
Eine kritische Revision. 518 Seiten. ISBN 978-3-7329-0955-1
- Bd. 8 Susanne Drogi/Nadine Naugk (Hg.): Begegnungen von Jung und Alt  
in der Kinder- und Jugendliteratur. Literaturwissenschaftliche und  
literaturdidaktische Perspektiven. 272 Seiten. ISBN 978-3-7329-0925-4
- Bd. 9 Sebastian Bernhardt (Hg.): Unzuverlässiges Erzählen  
in Literatur und Medien. Didaktische Perspektiven.  
524 Seiten. ISBN 978-3-7329-1015-1
- Bd. 10 Magdalena Kißling/Johanna Tönsing (Hg.): Einfach aussortieren?  
Anregungen für kritische Lektüren des Literaturkanons.  
294 Seiten. ISBN 978-3-7329-1003-8

# LITERATUR – MEDIEN – DIDAKTIK

- Bd. 11 Johannes Odendahl/Marcel Illetschko (Hg.): Kunst oder Leben?  
Ästhetisches Lernen in Zeiten globaler Krisen. 386 Seiten.  
ISBN 978-3-7329-1046-5
- Bd. 12 Nadine Rack-Hellekes: Schreibkompetenz und Schreibmotivation  
in der Grundschule. Ein Trainingsprogramm. 278 Seiten.  
ISBN 978-3-7329-1098-4
- Bd. 13 Sebastian Bernhardt (Hg.): Literarästhetisches Lernen außerhalb  
des Klassenzimmers. Interdisziplinäre Perspektiven.  
354 Seiten. ISBN 978-3-7329-1040-3
- Bd. 14 Sebastian Bernhardt (Hg.): Literatúrausstellungen und Vermittlung.  
Transdisziplinäre Perspektiven. 346 Seiten. ISBN 978-3-7329-1175-2
- Bd. 15 Sandra Siewert: Das Erstlesebuch – eine unterschätzte Subgattung.  
Gestaltungsweisen und Potenziale zwischen Lesefähigkeit  
und Lesemotivation. 432 Seiten. ISBN 978-3-7329-1134-9
- Bd. 16 Hans-Georg Müller/Maurice Fürstenberg (Hg.): DeutschGPT –  
Deutschunterricht im Dialog mit Künstlicher Intelligenz.  
300 Seiten. ISBN 978-3-7329-1120-2

Große generative Sprachmodelle (LLMs) haben die Exotik der Anfangstage verloren. Sie sind längst Alltag geworden, auch im Lehren und Lernen. In dieser neuen Phase des sprach- und literaturdidaktischen Umgangs mit ChatGPT & Co. muss sich der Deutschunterricht in Theorie und Praxis neuen Fragen stellen: Wie funktioniert KI und was ist von dieser Technik in Zukunft noch zu erwarten? Wie wirkt sich die Verwendung von KI auf den sprachlichen Kompetenzerwerb aus? Wie gehen Schüler:innen mit digitalen Textgeneratoren um und welche Konsequenzen hat das für ihre Bildungsgeschichte? Wie lässt sich KI sinnvoll in den Deutschunterricht integrieren? Wo liegen Risiken, wo Chancen für die Gestaltung der Schule von morgen? Diesen und anderen Fragen gehen die Beiträge des Bandes nach. Sie präsentieren zudem eine Auswahl aktueller Forschungsprojekte rund um den Einsatz von KI im Deutschunterricht. Der Band bildet somit den gegenwärtigen Forschungsstand zum Thema Deutschunterricht und KI in seiner ganzen Breite ab.

*Prof. Dr. Hans-Georg Müller* lehrt Sprachdidaktik am Institut für Germanistik der Universität Potsdam. Seine Forschungsschwerpunkte liegen in den Bereichen Kognitionswissenschaft, empirische Bildungsforschung und (Schrift-)Spracherwerb.

*Dr. Maurice Fürstenberg* lehrt und forscht als Akademischer Rat an der LMU München in der Germanistischen Linguistik an der Schnittstelle zur Sprachdidaktik. Er promovierte zum Gebrauch des Kommas durch Schüler:innen und habilitiert zu KI im Deutschunterricht.

