JG. 8 (2026), EINZELBEITRAG 1
DOI: 10.18716/OJS/MIDU/2026.0.1
VERÖFFENTLICHT AM 01.02.2026

M i D U

Medien im Deutschunterricht

GELINGENSBEDINGUNGEN UND GRENZEN EINER KI-GESTÜTZTEN FÖRDERUNG LITERARISCHER VERSTEHENSKOMPETENZ.
BEFUNDE AUS ZWEI EXPERIMENTELLEN STUDIEN

Jörn Brüggemann
Otto-Friedrich-Universität Bamberg | joern.brueggemann@uni-bamberg.de

Carina Ascherl
Otto-Friedrich-Universität Bamberg | carina.ascherl@uni-bamberg.de

Laureen Okesson
Otto-Friedrich-Universität Bamberg | laureen.okesson@uni-bamberg.de

Volker Frederking
Friedrich-Alexander-Universität Erlangen-Nürnberg | volker.frederking@fau.de

ABSTRACT

In der Literaturdidaktik wecken Chatbots wie ChatGPT die Erwartung, bei der Bewältigung zentraler Erwerbs- und Vermittlungsherausforderungen im Bereich des literarischen Textverstehens unterstützen zu können. Allerdings ist unklar, inwieweit diese Erwartung berechtigt ist, da empirische Evidenz in Bezug auf Gelingensbedingungen und Grenzen einer KI-gestützten Förderung literarischer Verstehenskompetenz fehlt. Vor diesem Hintergrund werden im vorliegenden Beitrag zunächst ungeprüfte Annahmen herausgearbeitet, die die systematische Förderung literarischer Verstehenskompetenz sowie die Validität und Verlässlichkeit KI-erzeugter Unterstützungsangebote betreffen (1). Im Anschluss wird erläutert, wie diese Annahmen in zwei experimentellen Settings mit ChatGPT (basierend auf den Sprachmodellen GPT-3.5 und -4o) im Rahmen des BMBF-Verbundprojekts „Digitale Souveränität als Ziel wegweisender Lehrer:innenbildung in den Sprachen, Gesellschafts- und Wirtschaftswissenschaften“ (DiSo-SGW) untersucht wurden (2). Auf dieser Basis wird erläutert, wie das Förderpotenzial von ChatGPT in weiteren Studien in DiSo-SGW erforscht wird, um empirisch abgesicherte Erkenntnisse zur Implementation KI-gestützter Fördermaßnahmen zu gewinnen.

SCHLAGWÖRTER

LITERATURUNTERRICHT — LITERARISCHE VERSTEHENSKOMPETENZ — CHATGPT — LERNEN MIT KI — DIGITALE SOUVERÄNITÄT — ZUVERLÄSSIGKEIT VON KI

JG. 8 (2026), EINZELBEITRAG 1
DOI: 10.18716/OJS/MIDU/2026.0.1
VERÖFFENTLICHT AM 01.02.2026

M i D U

Medien im Deutschunterricht

ABSTRACT (ENGLISH)

CONDITIONS FOR SUCCESS AND LIMITATIONS OF AI-SUPPORTED PROMOTION OF LITERARY LITERACY. FINDINGS FROM TWO EXPERIMENTAL STUDIES

In the realm of German didactics, the chatbot ChatGPT has raised expectations that AI-generated support can help to better overcome key challenges in the acquisition and teaching of literary literacy. However, it is currently uncertain how realistic these expectations are. There is a lack of empirical evidence regarding the conditions for success and the limitations of AI-supported promotion of literary literacy. Against this background, this paper first identifies unexamined assumptions regarding the systematic promotion of literary literacy and the validity and reliability of AI-generated support services (1). It then explains how these assumptions were investigated in two experimental settings with ChatGPT (based on the GPT-3.5 and -4o LLMs) as part of the BMBF collaborative project “Digital Sovereignty as a Goal of Forward-thinking Professional Development for Teachers in Languages, Social Ciences, and Economics“ (DiSo-SGW) (2). On this basis, we explain how the funding potential of ChatGPT for promoting literacy literacy is currently being investigated in an intervention study in DiSo-SGW in order to gain empirically validated insights into the implementation of AI-supported support measures.

KEYWORDS

LITERATURE CLASSES — LITERARY COMPREHENSION SKILLS — CHATGPT — LEARNING WITH AI — DIGITAL SOVEREIGNTY — VALIDITY AND RELIABILITY OF AI

1 — KI-GESTÜTZTE FÖRDERANGEBOTE FÜR DEN LITERATURUNTERRICHT: GRATIFIKATIONSERWARTUNGEN

Die Förderung literarischer Verstehenskompetenz gilt als eine besondere Herausforderung des Deutschunterrichts. Diese resultiert u.a. aus der Mehrdeutigkeit literarischer Texte sowie der Zirkularität, Rekursivität und Iterativität literarischer Interpretationsprozesse, wie empirische Studien gezeigt haben (vgl. Frederking et al. 2016). Dass diese Herausforderung mit Hilfe KI-generierter Unterstützungsangebote besser bewältigt werden kann, ist eine in der Literaturdidaktik verbreitete Annahme. So hat der Chatbot ChatGPT des US-amerikanischen Unternehmens OpenAI Hoffnungen auf die Möglichkeit einer „individuellen Förderung literarischer Rezeptionsfähigkeit" (Führer / Gerjets 2024, 5) durch KI-generierte Unterstützungsangebote geweckt. Vermutet wird, dass v.a. der dialogische Charakter der Interaktionen in ChatGPT „Potenzial zur vertieften Auseinandersetzung" mit literarischen Texten und zur Generierung „neue[r] Perspektiven" mit sich bringt (Magirius / Scherf 2023, 410). Nach Nix und Führer (2024, 346) bietet der „Einsatz der KI im Lese- und Literaturunterricht […] historisch die einmalige Chance", allen Schüler:innen „Lernbegleiter, einen kompetenten Tutor, zur Verfügung [zu stellen], der anlassbezogen auf Probleme beim Lesen reagieren und Hilfestellungen, Anregungen sowie weiterführende Impulse liefern kann." Auch für den Aufbau „literaturspezifische[r] Wissensstrukturen bzw. ein[es] implizite[n] Strategiewissen[s]" werden positive Effekte durch „Übung mit KI" (Nix / Führer 2024, 346) vermutet. Erste empirische Ergebnisse der Autor:innen zeichnen allerdings ein ambivalentes Bild. So berichten Magirius und Scherf in ihrer Untersuchung zur KI-Nutzung von Studierenden bei der Interpretation von Gedichten von zahlreichen „problematische[n] Umgangsweisen" (2023, 411), darunter z.B. die unkritische Übernahme von KI-Deutungen oder die Tendenz, nur jene Antworten der KI auszuwählen, die das eigene Vorverständnis bestätigen. Nix und Führer (vgl. 2024, 361) betonen vor dem Hintergrund ihrer Erkenntnisse zu Interaktionen von Schüler:innen mit ChatGPT zur Interpretation von Kafkas *Vor dem Gesetz*, dass der Einsatz von Chatbots nicht per se zu einer vertieften Förderung ästhetischer Wahrnehmungs- oder Interpretationsprozesse beiträgt. Vor diesem Hintergrund scheint er sinnvoll, lernunterstützende Konfigurationen textgenerativer KI zu ermitteln, die solchen Umgangsweisen entgegenwirken. Das gilt vor allem für Unterstützungsangebote zur Erzeugung einer „offene[n] […] Suchbewegung" als Ausgangs- und Bezugspunkt literarischer Verstehensprozesse, in der sich „Hypothesenbildung und -überprüfung" (Lessing-Sattari 2015, 78) abwechseln bzw. verschränken. Denn deren Entfaltung überfordert viele Schüler:innen bis zum Abitur, wie Freudenberg (vgl. 2010) gezeigt hat. Da sich verstehensunterstützende Aufgaben und Gesprächsimpulse mit fokussierendem und konfliktinduzierendem Support positiv auf das Selbstkonzept und die Ausprägung literarischer Verstehenskompetenz auswirken, wie Befunde der empirischen Aufgabenforschung zeigen (vgl. Brüggemann / Gölitz 2025; Heins 2017; Magirius et al. 2022; Steinmetz 2019), könnte die Interaktion mit ChatGPT und anderen textgenerativen KI-Angeboten hier tatsächlich positive Effekte haben – sofern sich deren Ergebnisse als fachlich korrekt erweisen und passgenau an die individuellen Lernvoraussetzungen von Schüler:innen adaptierbar sind.

Mit Blick auf die Gegenstandsangemessenheit ist der Output von textgenerativer KI aus literaturdidaktischer und literaturwissenschaftlicher Sicht allerdings massiv infrage gestellt worden. So konstatiert Maiwald (2023, 363), dass die „KI-generierte Informationslage […] nicht nur ohne Quellenfundierung, sondern auch vage bis inkonsistent oder schlicht falsch" ist. Kragl (2023, 352f.) befürchtet ganz grundsätzlich „methodologische Kollateralschäden", […] wo hermeneutische Herausforderungen traktiert werden". Er konstatiert: „Was […] aktuell noch weitgehend fehlt, ist eine textanalytische Begleitung dessen, was da gerade passiert, namentlich der Resultate, die die Maschinen auswerfen." Demgegenüber attestieren Magirius und Steinmetz (2025, 44) ChatGPT tutorielles Potenzial, insofern das System „wie ein Meisterleser" interpretiere. Ob dieser Eindruck jedoch einer methodisch kontrollierten Prüfung standhält, ist unklar.

Vor diesem Hintergrund wird erkennbar, dass an die Stelle von Hoffnungen und Befürchtungen geprüftes Wissen in Bezug auf lernförderliche Potenziale textgenerativer KI treten muss. Befunde zur digitalen Souveränität im Fach Deutsch aus dem BMBF-Verbundprojekt „Digitale Souveränität als Ziel wegweisender Lehrer:innenbildung in den Sprachen, Gesellschafts- und Wirtschaftswissenschaften" (DiSo-SGW) können dazu einen Beitrag leisten, wie nachfolgend verdeutlicht werden soll.

2 — UNGEPRÜFTE ANNAHMEN UND DARAUS ABGELEITETE FORSCHUNGSFRAGEN

Im Fokus der nachfolgend berichteten Studien standen zwei Annahmen, die den Erwartungen vieler Schüler:innen, aber auch den in Kap. 1 genannten literaturdidaktischen Hoffnungen, vermittels textgenerativer KI Verstehenskompetenzen im Literaturunterricht fördern zu können, implizit zugrunde liegen: 1. Die Unterstützungsangebote von Chatbots wie ChatGPT stellen valide Interpretationen literarischer Texte dar. 2. Interpretationen, die als valide betrachtet werden können, werden auch bei wiederholter Durchführung durch ChatGPT zuverlässig erzeugt.

Probleme zeigen sich allerdings bereits bei den Prämissen, die beiden Annahmen zugrunde liegen. Denn KI-generierte Aussagen über literarische Texte wie die von ChatGPT stellen keine Interpretationsleistungen dar, die der Zirkularität, Rekursivität und Iterativität von Verstehensprozessen entsprechen. Chatbots wie ChatGPT überprüfen keine Interpretationshypothesen am Text. Sie produzieren lediglich Aussagen, die mit hoher statistischer Wahrscheinlichkeit akzeptabel sind. Dabei wird im Rückgriff auf Trainingsdaten mit Blick auf einen eingegebenen Text, der in sog. Tokens (Wort, Sequenz von Buchstaben, Satzzeichen) zerlegt wird, die Wahrscheinlichkeit berechnet, mit der Folgetoken auftreten. Mit anderen Worten: „GPT ersetzt ‚Bedeutung' durch ‚Auftretenswahrscheinlichkeit'" (Müller / Fürstenberg 2023, 335). Um den Eindruck einer authentischen Interaktion mit der KI zu erzeugen, wählt ChatGPT zudem aus dem Spektrum wahrscheinlich passender Möglichkeiten mit Hilfe eines Zufallsgenerators Folgetoken aus. ChatGPT kann also weder *verstehen* noch *Interpretationen* generieren.[1]

1 Der Chatbot vermag lediglich in der von ihm simulierten Kommunikation Text zu erzeugen, der von User:innen als Interpretation interpretiert werden kann bzw. wird. Nur mit dieser Einschränkung ist nachfolgend von KI-generierten ‚Interpretationen' und ‚Interpretationsleistungen' die Rede.

Ob vor diesem Hintergrund bei wiederholter Bearbeitung von (identischen) Interpretationsanfragen zuverlässig am Text belegbare ‚Interpretationen' erzeugt werden, ist unklar. Ebenso unklar ist, ob ChatGPT beim wiederholten Bearbeiten von präzise formulierten Interpretationsanfragen Unterstützungsangebote erzeugt, die passgenau auf die Lern- und Verstehensvoraussetzungen von Schüler:innen und ihren individuellen Förderbedarf bezogen sind. Diese Passgenauigkeit bezieht sich z.B. auf das Ausgabeformat (z.B. Fließtext, listenförmiger Präsentationsmodus), die Textlänge, sprachliche Anforderungen, aber auch auf die nachfolgend fokussierte Anforderung, Unterstützungsangebote zur Bewältigung von Verstehensanforderungen mit Blick auf *spezifische Kompetenzdimensionen* zu generieren. Damit treten Forschungsdesiderate in den Blick, die über sechs Forschungsfragen konkretisiert werden können. Drei betreffen die fachliche Korrektheit bzw. Plausibilisierbarkeit der Aussagen über literarische Texte, die von ChatGPT als untersuchter textgenerativer KI generiert werden:

> Forschungsfrage 1:
> Sind die durch ChatGPT generierten ‚Interpretationsansätze' literarischer Texte durch den Rekurs auf eine „Textintention" im Sinne Ecos (1986, 125) legitimierbar und in diesem Sinne gegenstandsangemessen und plausibilisierbar?
>
> Forschungsfrage 2:
> Werden bei der wiederholten Bearbeitung identischer Aufgabenstellungen zuverlässig gleichermaßen plausible ‚Interpretationsergebnisse' bzw. ‚Interpretationsleistungen' erbracht?
>
> Forschungsfrage 3:
> In welchen Dimensionen literarischer Verstehenskompetenz (vgl. Kap. 3.1) können von ChatGPT generierte ‚Interpretationsansätze' (mit Hilfe von Prompts) erzeugt werden (durch die wahrscheinlichkeitsgestützte Zusammensetzung von Tokens)?[2]

Weitere Forschungsdesiderate betreffen das Förderpotenzial, das mit ChatGPT verbunden wird:

> Forschungsfrage 4:
> Wie präzise und zuverlässig gelingt es ChatGPT, zutreffende Aussagen über literarische Texte in unterschiedlichen Kompetenzdimensionen (vgl. Kap. 3.1) zu erzeugen?
>
> Forschungsfrage 5:
> Wie zuverlässig generiert ChatGPT zutreffende Aussagen über literarische Texte *in klar definierten formalen Parametern (Fließtext – listenförmiger Modus, adressatengerechte Sprache, Textlänge)*?

2 *Wir adressieren damit Grundlagenforschung als Basis für später geplante Anwendungsforschung. Wenn ChatGPT als Intelligentes tutorielles System tatsächlich Lernende mit individuellem Förderbedarf in unterschiedlichen Dimensionen literarischer Verstehenskompetenz unterstützen können soll, muss das System in der Lage sein, ‚Interpretationsansätze' mit Blick auf jene textseitigen Verstehensanforderungen zu generieren, die die unterschiedlichen Dimensionen literarischer Verstehenskompetenz repräsentieren (vgl. Kap. 4.1). Das ist die Voraussetzung für eine gezielte Konfiguration von KI-Tutoren, die Lernenden mit unterschiedlichen Lernvoraussetzungen Unterstützungsangebote unterbreiten.*

Forschungsfrage 6:
Wie präzise und zuverlässig können formale und inhaltliche Ansprüche an die von ChatGPT generierten Aussagen durch Konfiguration und Prompting[3] gesteuert werden?

3 — METHODE

Hinweise zur Beantwortung der Forschungsfragen 1-6 wurden im Rahmen von zwei Experimentalstudien gewonnen, die im Forschungsprojekt DiSo-SGW im Rahmen der Entwicklung von Fortbildungsangeboten zur Erweiterung der digitalen Souveränität von Deutschlehrkräften im Umgang mit KI im Literaturunterricht an der Universität Bamberg durchgeführt wurden. Die beiden experimentellen Erhebungen wurden mit den ChatGPT-Versionen 3.5 und 4o durchgeführt[4] und betrafen vier unterschiedliche Bedingungen:

Bedingung 1 (B1):
In Bedingung 1 hat ChatGPT-3.5 Aufgaben zur Auseinandersetzung mit einem literarischen Stimulustext bearbeitet, ohne dass dieser Text ChatGPT als separates Material zur Verfügung gestellt wurde.

Bedingung 2 (B2):
In Bedingung 2 hat ChatGPT-3.5 Aufgaben zur Auseinandersetzung mit dem literarischen Stimulustext bearbeitet; dieser wurde als PDF-Datei zu Beginn der Interaktion in den Chatverlauf eingespeist.

Bedingung 3 (B3):
In Bedingung 3 hat ChatGPT-4o Aufgaben zur Auseinandersetzung mit dem literarischen Stimulustext bearbeitet, der ebenfalls als PDF-Datei zu Beginn der Interaktion in den Chatverlauf eingespeist wurde.

Bedingung 4 (B4):
In Bedingung 4 wurde ein sogenanntes Custom GPT[5] eingesetzt, in dem ChatGPT-4o neben spezifischen Anweisungen („Hinweise“) eine fachspezifische Wissensbasis zur Verfügung gestellt wurde: Definitionen der fünf Dimensionen literarischer Verstehenskompetenz (vgl. Kap. 3.1), der literarische Stimulustext, eine Überblicksdarstellung über Stilmittel inklusive Anwendungsbeispiele aus einschlägigen Lehrwerken. Die Konfiguration und die Prompts wurden mit Hilfe von ChatGPT optimiert. Die Option zur Onlinesuche wurde deaktiviert, weil Vorstudien gezeigt hatten, dass sie nicht zur Verbesserung des Outputs beitrug. Dasselbe galt für die Bereitstellung literaturwissenschaftlicher und -didaktischer Literatur zur Interpretation von Literatur.

3 „Konfiguration“ bezeichnet die Einrichtung eines sogenannten Custom GPT mittels des GPT Builders von OpenAI und damit die Möglichkeit, das Verhalten der generativen KI durch dauerhaft hinterlegte Instruktionen und Zusatzmaterialien zu steuern, die den bestehenden Systemprompt ergänzen.
4 Die Entscheidung, unterschiedliche, zum Erhebungszeitpunkt häufig genutzte Modellgenerationen zu untersuchen, diente dazu, mögliche Effekte der technischen Weiterentwicklung zu berücksichtigen.
5 Custom GPTs sind benutzerdefinierte GPTs, die durch die Bereitstellung ausgewählter Daten auf die besonderen Bedürfnisse von Nutzenden bzw. Anwendungsfälle und Themengebiete zugeschnitten und optimiert werden können.

4 — EXPERIMENTALSTUDIE 1

4.1 UNTERSUCHUNG DER FORSCHUNGSFRAGEN 1-3

In Experimentalstudie 1 wurden die fachliche Korrektheit, die Zuverlässigkeit und die Dimensionalität der KI-generierten ‚Interpretationsleistungen' mit Hilfe von Testaufgaben zum literarischen Textverstehen für die Jahrgangsstufen 8-10 geprüft, die in drei DFG-Projekten (LUK = Literarische Urteils- und Verstehenskompetenz) entwickelt und systematisch in Bezug auf ihre Validität, Reliabilität und Objektivität evaluiert worden sind (vgl. Frederking et al. 2016; Meier et al. 2017). Mit den LUK-Testaufgaben liegen verlässliche Messinstrumente vor, die literarische Interpretationsleistungen in den folgenden fünf Dimensionen erfassen:

1. **Semantische literarische Verstehenskompetenz (SLV)** bezeichnet die Fähigkeit, trotz der Unbestimmtheit, Indirektheit und Mehrdeutigkeit vieler literarischer Texte und eines hohen Maßes an Verknüpfungsdichte einen literarischen Text in kohärenter Weise deuten zu können – auch unter Einbeziehung textexterner Zusatzinformationen (z.B. werk-, epochen-, gattungshistorischer Art). Zur Erfassung von *Verstehensanforderungen im Bereich von SLV* sind die folgenden Fragen hilfreich: Wie ist der Text in seiner Gesamtheit inhaltlich zu verstehen? Welche Textstellen belegen die damit verbundenen Interpretationsansätze? Welche Verstehensprobleme weist der Text auf? Welche Textstellen sind unklar oder mehrdeutig?

2. **Ästhetische Aufmerksamkeit (ÄA)** bezeichnet die Fähigkeit, sprachlich-stilistische und strukturelle Auffälligkeiten literarischer Texte identifizieren zu können, die potenziell interpretationsrelevant sind. Die Aufmerksamkeit auf sprachliche Foreground-Elemente (van Holt / Groeben 2005) zu fokussieren, gehört zu den Fähigkeiten, die Schüler:innen laut Nix und Führer (vgl. 2024, 361) bislang im Umgang mit ChatGPT vermissen lassen. Verstehensanforderungen im Bereich von ÄA sind im Horizont der Frage verortet: Welche Textelemente sind hinsichtlich ihrer sprachlich-stilistischen oder strukturellen Gestaltung besonders auffällig oder erklärungsbedürftig?

3. **Idiolektale literarische Verstehenskompetenz (ILV)** bezeichnet die Fähigkeit, die textseitig intendierten ästhetischen Funktionen und Wirkungsintentionen sprachlich-stilistischer und struktureller Besonderheiten literarischer Texte identifizieren und entsprechende Zuschreibungen prüfen zu können – auch unter Einbezug textexterner Zusatzinformationen (z.B. Gattungsgeschichte, Formkonventionen). Spezifische *Verstehensanforderungen im Bereich von ILV* stehen im Fokus der Frage: Mit welcher Absicht bzw. zu welchem Zweck könnten diese auffälligen Textelemente auf diese besondere Weise gestaltet worden sein?

4. **Die Kompetenz zum Verstehen textseitig intendierter Emotionen (VIE)** bezeichnet die Fähigkeit, die emotionale Rezeptionssteuerung literarischer Texte zu rekonstruieren, d. h. jene Emotionen zu erfassen, die durch den Text ausgelöst werden sollen, und jene Textstellen zu identifizieren, die diese Emotionen evozieren sollen. *Verstehensanforderungen im Bereich des VIE* werden adressiert durch die Fragen: Welche emotionalen Wirkungen soll der Text hervorrufen? Durch welche konkreten Textstellen sollen diese Emotionen ausgelöst werden?

5. **Literarisches Fachwissen (LF)** bezeichnet die Fähigkeit, im Umgang mit literarischen Texten literaturspezifisches Fachwissen zur Einordnung von semantischen und formalen Textelementen zu aktivieren und diese Elemente korrekt zu bezeichnen. *Verstehensanforderungen im Bereich des LF* können durch die folgenden Fragen erfasst werden: Welches Fachwissen ist hilfreich, um den literarischen Text besser zu verstehen? Welche Fachbegriffe sind wichtig, um den literarischen Text zu beschreiben? Welche Diskurse und Kontexte führen zu einem erweiterten Verstehen des Textes?

Zur Überprüfung der fachlichen Korrektheit, Zuverlässigkeit und Dimensionalität der KI-generierten ‚Interpretationsleistungen' wurden ChatGPT-4o[6] zwei LUK-Units zu zwei unterschiedlichen lyrischen Texten zur Bearbeitung vorgelegt, deren Eignung für das Erfassen literarischer Verstehenskompetenz systematisch evaluiert worden sind (vgl. Frederking et al. 2016). Die beiden LUK-Units umfassten insgesamt 80 offene, halboffene, MC- und FC-Items zur Erfassung der in den fünf Teildimensionen beschriebenen Interpretationsleistungen. Auf dieser Basis wurden die Forschungsfragen 1-3 überprüft. Um die zuverlässige Reproduktion der Ergebnisse von ChatGPT im Sinne von Forschungsfrage 2 zu prüfen, hat der Bot jede Unit fünf Mal bearbeitet, sodass insgesamt 400 Lösungen produziert wurden. Experimentalstudie 1 wurde unter den Bedingungen 3 und 4 durchgeführt.

4.2 BEFUNDE VON EXPERIMENTALSTUDIE 1

Erste Hinweise zur Beantwortung der Forschungsfragen 1-3 in Experimentalstudie 1 bietet die Evaluation der KI-generierten ‚Interpretationsleistungen' mithilfe des LUK-Tests (vgl. Abb. 1).

6 Die Überprüfung erfolgt mit ChatGPT-4o, weil sich diese Version im Vergleich zur Version 3 als zuverlässiger erwiesen hat (vgl. Kap. 4.2 und 5.2).

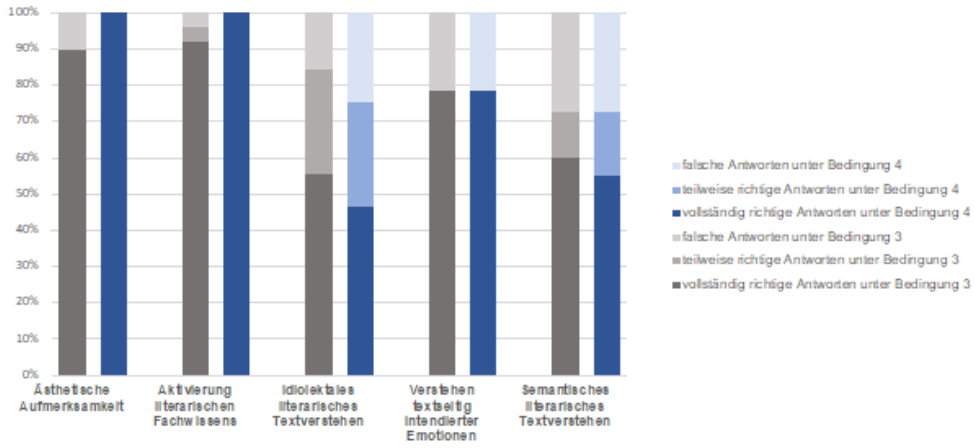

Abb. 1 Evaluation der Interpretationsleistung von ChatGPT unter den Bedingungen 3 und 4 in Prozent

Die Daten zeigen deutlich, dass ChatGPT ‚Interpretationsleistungen' in allen fünf LUK-Kompetenzdimensionen zu entwickeln vermag. Insgesamt liegt der Anteil korrekt generierter Lösungen bei 81,5% (B3) bzw. 80,5% (B4). Es liegen also eher geringe Unterschiede zwischen den Bedingungen B3 und B4 vor.[7] Jedoch variiert der Anteil korrekter Lösungen in den einzelnen Kompetenzdimensionen z.T. erheblich. Während im Bereich von ÄA und LF in B4 100% der Lösungen korrekt sind und in B3 90% (ÄA) und 92% (LF), sind die Leistungen beim VIE geringer. Die Rekonstruktion der emotionalen Rezeptionssteuerung gelingt in B3 und B4 in 78,8% der Fälle. Eine Analyse auf Itemebene zeigt zudem, dass die von ChatGPT generierten Aussagen bei der Erfassung textseitig intendierter *komplexer* Emotionen bzw. emotionaler Zustände (z.B. *Resignation*, *Distanz zu Gefühlen*) häufig unzutreffend sind. Deutlich geringer fällt der Anteil vollständig korrekter Lösungen im SLV in B3 mit 60% und B4 mit 55% aus. Auch im Bereich des ILV ergeben sich in B3 nur 55,6% und in B4 sogar nur 46,7%. Zudem variiert der Anteil von Lösungen, die teilweise richtig sind. Dieser liegt im SLV bei 12,5% (B3) bzw. 17,5% (B4), im ILV bei 28,9% (B3/4). Während also die Korrektheit der Interpretationsleistungen in B4 in den Bereichen ÄA und LF höher ist als in B3, weist B3 Vorteile im Bereich von SLV und ILV auf. Zudem deuten die Daten darauf hin, dass bei der wiederholten Bearbeitung identischer LUK-Aufgaben nicht zuverlässig korrekte bzw. plausible ‚Interpretationsleistungen' erbracht werden. Da der Bot jedes Item fünfmal bearbeitet hat, zeigt der Vergleich der Aufgabenlösungen auf Ebene der Einzelitems, dass Items nicht durchgängig richtig oder falsch gelöst wurden, wobei sich kein Muster in Bezug auf die Korrektheit des Outputs ableiten lässt.

Für die Entwicklung von KI-generierten Unterstützungsmaßnahmen scheint B4 Vorteile im Bereich von ÄA und LF aufzuweisen. B3 scheint dagegen im SLV und ILV geringfügig überlegen zu sein, obschon ein Leistungsvermögen mit 60% bei SLV und 55,6% bei ILV sich – im Notenspektrum verortet – allenfalls im ausreichenden Bereich bewegt. Eine zuverlässige Unterstützung von Lernenden beim Verstehen literarischer Texte kann auf Basis von ChatGPT mithin nur eingeschränkt erfolgen.

7 *ChatGPT-4o bietet in B3 und B4 umfassende ‚Interpretationen', die am Text belegt werden, allerdings ohne korrekte Versnummern. Dieses Problem wurde im Rahmen der Auswertung als gegenwärtig (noch) bestehendes des Sprachmodells behandelt und aus der Bewertung der ‚Interpretationsleistungen' ausgeklammert.*

5 — EXPERIMENTALSTUDIE 2

5.1 UNTERSUCHUNG DER FORSCHUNGSFRAGEN 1–6

In der Experimentalstudie 2 wurde geprüft, ob und inwieweit die ChatGPT-Versionen 3.5 und 4o in den Bedingungen B1-B4 (s.o.) präzise und zuverlässige Aussagen über literarische Texte in den fünf Dimensionen literarischer Verstehenskompetenz (LUK) generieren, die in einer anschließenden Interaktion vertieft werden konnten, mit dem Ziel, Anknüpfungspunkte zur Erweiterung literarischer Verstehenskompetenz zu erzeugen. Im Rahmen der Interaktion haben N=7 Deutschdidaktiker:innen identische Prompts zur Erschließung von Mascha Kalékos Gedicht *Window-Shopping* eingegeben. Der Fortgang der Interaktion variierte z.T. in Abhängigkeit vom jeweiligen Output von ChatGPT. Die Auswertung der Interaktionsprotokolle[8] durch drei literaturdidaktisch geschulte Rater:innen lieferte (weitere) Hinweise zur Beantwortung der Forschungsfragen 1-6 und ermöglichte damit eine empirisch gestützte Beantwortung der Frage, inwieweit Voraussetzungen für eine verlässliche individualisierte KI-gestützte Förderung literarischer Verstehenskompetenz gegeben sind.

Um das Förderpotenzial der Interaktion mit ChatGPT und die Qualität der Outputs beurteilen zu können, wurde mit *Window-Shopping* von Mascha Kaléko (2012, 588) ein lyrischer Stimulustext mit hohem Aktivierungspotenzial gewählt, der in der DiSo-SGW-Fortbildung eine prominente Rolle spielt, aufgrund seiner Bekanntheit allerdings nicht zu den LUK-Texten gehörte. Die erste Strophe des Gedichts (V. 1-7) lautet: „Brillantgefunkel für den Hals der Lady, / Parfum, zehn Dollar aufwärts, für die Gnädi- / ge Frau. Ein wirklich echtes Herzcape für den Spitz, / aus Sandelholz der – hm, Toilettensitz. / Da strömt das Volk im billigen Sonntagsschuh / zum Luxusfenster der Fifth Ävennjuh / und liegt vorm goldnen Kalb platt auf dem Bauche." In der zweiten Strophe (V. 8-10) heißt es: „Wenn ich mir schweigend diesen Prunk betrachte, / denk ich mir nur, was Sokrates schon dachte: / Wie vieles gibt es doch, was ich nicht brauche!"

Vielfältiges **Potenzial zur Aktivierung semantischer literarischer Verstehenskompetenz (SLV)** bietet der Text in folgender Hinsicht: Dass in den Versen 1-4 ein schweifender Blick über Schaufenstereinlagen präsentiert wird, muss durch Inferenzbildung erschlossen werden. Dasselbe gilt für die Frage, wessen Blick schweift. Weitere semantische Verstehensherausforderungen bergen einzelne Formulierungen – interpretationsbedürftig ist z.B. „das Volk im billigen Sonntagsschuh" (V. 5), das „vorm goldnen Kalb platt auf dem Bauche [liegt]" (V. 7). Mit Blick auf eine Gesamtdeutung ist z.B. die Beziehung zwischen lyrischem Ich und „Volk" klärungsbedürftig, ebenso bleibt zunächst fraglich, welche Botschaft vermittelt werden soll (z.B. eine konsumkritische).

Potenzial zur Aktivierung ästhetischer Aufmerksamkeit (ÄA) bietet *Window-Shopping* z.B. durch die Anführungszeichen im Titel, die unvollständige Syntax in den Versen 1-4, das Kompositum „Brillantgefunkel" (V. 1), das Enjambement (V. 2f.), die Formulierungen „wirklich echtes Nerzcape" (V. 3), „– hm," (V. 4) oder „Fifth Ävennjuh" (V. 6), durch Metaphern (V. 5, 7), semantische Oppositionen (z.B. „billigen Sonntags-

8 Die Protokolle inkl. Prompts und Konfiguration sind hier einsehbar: https://t1p.de/literarischesprachreflexion

schuh“ vs. „Luxusfenster“, V. 6f.; „Volk“ (V. 5) vs. lyrisches Ich (V. 8-10)), aber auch durch die Strophengestaltung.

Potenzial zur Aktivierung idiolektaler literarischer Verstehenskompetenz (ILV) bietet der Text, insofern vielen Textelementen plausible ästhetische Funktionen bzw. Wirkungsintentionen zugeordnet werden können. In diesem Sinn wird z.B. durch den Verzicht auf eine vollständige Syntax in den Versen 1-4 ein umherschweifender Blick über Schaufenstereinlagen imitiert; durch die semantischen Oppositionen wird ein distanziertes Verhältnis von lyrischem Ich und „Volk“ bzw. das Unbehagen Intellektueller gegenüber der Masse indiziert.

Potenzial zur Aktivierung der Fähigkeit zum Verstehen textseitig intendierter Emotionen (VIE) und damit verbundener emotionaler Rezeptionssteuerung des Gedichts besitzen z.B. Formulierungen und Gestaltungselemente, die Heiterkeit, evtl. Spott evozieren sollen – z.B. „wirklich echtes Nerzcape“ (V. 3), das Spiel mit Erwartungsbrüchen in V. 4 („der – hm, Toilettensitz“) oder das Spiel mit dem Reim „Lady“/“Gnädi-“ (V. 1f.); überdies die Strophenform, mit deren Hilfe eine reflexive Distanz zum emotionalen Erleben (der Beobachteten, vielleicht auch der Leser:innen) ausgelöst werden soll.

Potenzial zur Aktivierung von Literarischem Fachwissen (LF) liegt vor, insofern Fachbegriffe und Kontextwissen hilfreich sein können, um den Text besser zu verstehen. LF umfasst hier terminologisches Wissen, aber auch konzeptuelle syntaktische Supportangebote (Steinmetz 2019, 87f.), mit denen textbezogene Wissenselemente verbunden sind, die „den Status von Sinnangeboten“ haben (z.B. Wissen über die Bedeutung des goldenen Kalbs oder über Sokrates).

5.2 BEFUNDE DER EXPERIMENTALSTUDIE 2

Durch die Eingabe identischer Prompts zu Mascha Kalékos *Window-Shopping* ermöglicht Experimentalstudie 2 eine Ausdifferenzierung der Befunde von Experimentalstudie 1 zu den Forschungsfragen 1-3. Dabei werden eine Reihe von Problemen erkennbar, deren Beseitigung für die Entwicklung KI-gestützter Fördermaßnahmen für den Literaturunterricht von grundlegender Bedeutung ist. Diese betreffen u.a. sog. „Halluzinationen“: Manchmal (siehe Beispiel 1 hier und Beispiel 2 hier), aber nicht immer, wird offengelegt, dass eine ‚Interpretation‘ ohne Zugriff auf den Stimulustext unmöglich ist. Eine Reihe weiterer Defizite sind in allen Bedingungen zu beobachten: Ohne Einspeisung von Wissen über die Autorin werden trotz wiederholter Rückfragen falsche Autorangaben ‚halluziniert‘ (s. Beispiel 3 hier oder Beispiel 4 hier). Ebenso zitiert ChatGPT in unterschiedlichen Chats Text, der im eingespeisten Stimulustext nicht existiert (z.B. das Wort „Sandelsitz“, vgl. Beispiel 5 hier, Z. 32 und Beispiel 6 hier, Z. 86f.). Bemerkenswert sind auch Schwächen bei der Verarbeitung von Satzzeichen, die interpretationsrelevant sein könnten. Wenn ChatGPT z.B. bei der Verarbeitung von Prompts zur Textstelle „– hm,“ (V. 4) diese Textstelle in der eigenen Antwort zitiert, wird das Komma getilgt (s. Beispiel 7 hier, Z. 6 oder Beispiel 8 hier, Z. 10-11). Defizite bestehen auch beim Zitieren von Versen mittels Versnummern – sogar nach der Implementierung expliziter Anweisungen und Hilfestellungen in B4: Verse werden

z.T. unter Angabe falscher Versnummern zitiert (s. Beispiel 9 hier, Z. 11 oder Z. 45, s. auch Beispiel 10 hier). Versangaben zu Textstellen werden gelegentlich ausgeweitet (Z. 26) oder verkürzt (Z. 12). Bemerkenswert sind auch Defizite im Bereich der Identifikation sprachlich-stilistischer, grammatikalischer und struktureller Auffälligkeiten (ÄA): Dies gilt v.a. für Alliterationen – selbst in B4, die durch entsprechendes Wissen inklusive Anwendungsbeispiele zur Beseitigung der in B1–3 beobachteten Defizite konfiguriert worden ist. Dies zeigt die wiederholte falsche Einordnung des Worts „Brillantgefunkel" (V. 1): „Der Ausdruck ‚Brillantgefunkel' enthält eine Alliteration, da der gleiche Konsonant (B) am Anfang der aufeinanderfolgenden Wörter wiederholt wird." (Z. 6-8). Probleme werden auch deutlich, wenn auf die „Wiederholung des „B"-Lautes in „Brillantgefunkel" (Z. 20) oder auf den „Beginn mit ‚B' und die Wiederholung des ‚l'-Lauts" (Z. 30-31) hingewiesen wird (s. Beispiel 16 hier, Z. 20-26; Beispiel 17 hier, Z. 18-20 oder Beispiel 18 hier, Z. 3-4). Vergleichbare Defizite zeigen sich bei der Identifikation von Assonanzen („Brillantgefunkel": „[...] es handelt sich eher um eine Assonanz, da der Vokal „i" in beiden Wörtern wiederholt wird" (Z. 97-98) (vgl. auch Beispiel 20 hier, Z. 22; Beispiel 21 hier, Z. 12-17; Beispiel 22 hier, Z. 19-24, und Beispiel 23 hier, Z. 8-9) und Neologismen (s. Beispiel 24 hier, Z. 28-31; Beispiel 25 hier, Z. 6-10 oder Beispiel 26 hier, Z. 21-24). Obwohl ChatGPT grundsätzlich Enjambements korrekt identifiziert, gibt es Ausnahmen, wenn er in V. 4 identifiziert, dass in „‚aus Sandelholz der – hm [!] Toilettensitz' […] ein Zeilenbruch verwendet [wird]" (Z. 48-50, vgl. auch Beispiel 28 hier, Z. 14-16). Ähnliche Probleme zeigen sich bei der Identifikation und Anwendung von literarischem Fachwissen: In keinem untersuchten Fall wurde das Reimschema vollständig korrekt erkannt. Zwar wurden in mehreren Fällen Paarreime identifiziert, jedoch nie in der zweiten Strophe. Ebenso wurde der umarmende Reim (V. 7, 10) nicht erkannt: „Das Gedicht besteht aus vier Zeilenpaaren, die jeweils durch Paarreime verbunden sind: ‚Lady' – ‚Gnädi', ‚Spitz' – ‚Sitz', ‚Schuh' – ‚Ävennjuh', ‚Bauche' – ‚brauche'" (Z. 8-10, vgl. auch Beispiel 30 hier, Z. 6-7 oder Beispiel 31 hier, Z. 32-35). Probleme zeigen sich auch bei der Identifikation von Reimpaaren („‚Fifth Ävennjuh' – ‚Bauche'" (Z. 11), vgl. weitere Beispiele Beispiel 33 hier, Z. 55-57; Beispiel 34 hier, Z. 5-6 oder Beispiel 35 hier, Z. 19-21) und bei der Identifikation des Reimschemas (Z. 35-36). Defizitär ist auch die Identifikation eines „versteckte[n] Reim[s]" im Wort „Brillantgefunkel" (Z. 13), eines „[u]nvollständige[n] Reim[s]" in V. 4 (Z. 27-28) oder die Markierung des „– hm," (ebd.) als „Bruch im Reimschema" (Z. 28, vgl. auch Beispiel 40 hier, Z. 22-24).

Auch wenn die KI-generierten ‚Verstehensleistungen' in vielen Fällen durchaus plausibel sind und als Unterstützung bei der Fokussierung interpretationsrelevanter Textaspekte betrachtet werden können, bestätigt Experimentalstudie 2 die in Experimentalstudie 1 beobachteten Defizite in Bezug auf Korrektheit und Zuverlässigkeit. Für die Entwicklung von Fördermaßnahmen, die auf KI-generiertem Support basieren, sind dies deutliche Einschränkungen, die sich durch weitere Beobachtungen ergänzen lassen:

So besteht eine weitere Limitation (im Korpus allerdings nur einmal beobachtet) in der fehlenden Kohärenz bzw. in der Widersprüchlichkeit des KI-erzeugten Outputs. Wenn die KI zu Beginn einer ‚Antwort' Informationen präsentiert, die zu den am Ende präsentierten Informationen im Widerspruch stehen (Z. 15-16), so zeigt dies, dass

nicht jeder Output sofort *passgenau* ausfällt. Eine andere Limitation betrifft die Unkalkulierbarkeit des KI-generierten Outputs mit Blick auf Umfang, Darstellungsform und sprachliche Anforderungen. Die Protokolle zeigen, dass der von ChatGPT generierte Output bei identischem Prompting in B1-3 in dieser Hinsicht eine Variationsbreite aufweist, die dessen Weiterverarbeitung durch Schüler:innen mit unterschiedlichen Lernvoraussetzungen erschwert. Für die Entwicklung passgenauer Unterstützungsangebote im Bereich des literarischen Verstehens in heterogenen Lerngruppen liegt hier eine besondere Herausforderung. Ob der Output in Form eines kohäsiven linearen Textes oder listenförmig (Kombination von prägnanten Oberpunkten/Überschriften und erläuternden Passagen) präsentiert wird, ist bei identischem Prompting ebenso wenig vorhersehbar wie der Umfang des Outputs (vgl. Beispiel 44 hier vs. Beispiel 45 hier). Dasselbe gilt für einleitende und resümierende Rahmungen (vgl. Beispiel 46 hier vs. Beispiel 47 hier (Z. 1-6), oder Beispiel 48 hier vs. Beispiel 49 hier (Z. 1-5)). In Abhängigkeit vom jeweiligen Umfang des Outputs sowie dessen formaler Gestaltung entstehen sehr heterogene Anforderungen beim Verstehen, bei der Auswahl und bei der Weiterverarbeitung der KI-generierten Aussagen, die wahrscheinlich nicht alle Lernenden bewältigen können. Für die Entwicklung KI-gestützter Unterstützungsangebote ist deshalb eine bessere und zuverlässige Steuerbarkeit der Darstellungsform erforderlich.

Die Protokolle in B4 zeigen jedoch auch, dass man ChatGPT durch eine spezifische Konfiguration in die Lage versetzen kann, passgenauere Unterstützungsangebote zu machen. Um Anforderungen im Bereich der Lesekompetenz (zugunsten mentaler Kapazitäten für literarische Verstehensleistungen) zu verringern, wurde ein listenförmiger Ausgabemodus unter Verzicht auf einleitende oder resümierende Rahmungen konfiguriert und im Vergleich zu B1-3 erfolgreich erprobt. Der Rückfall in frühere Muster zeigte sich unter dieser Bedingung nur noch selten (Beispiel 50 hier (Z. 36-40), Beispiel 51 hier (Z. 35-37) und Beispiel 52 hier (Z. 38-40)).

Eine besondere Herausforderung bei der KI-gestützten Erzeugung von fokussierendem Support im Zusammenhang mit der Förderung *Ästhetischer Aufmerksamkeit* betrifft in B2 und B3 die Tendenz, dass ChatGPT mehr liefert als gefordert; oft z.B. Gesamtinterpretationen des Stimulustextes, obwohl ausschließlich die Angabe von Textstellen gefordert wurde, die in sprachlich-stilistischer, grammatikalischer oder struktureller Hinsicht auffällig (vgl. Beispiel 53, 54 und 55, d.h. diese, diese und diese Prompts) oder schwer interpretierbar sind (vgl. Beispiel 56 dieser Prompt). Zumeist basiert der Output auf der Synthese von Leistungen aus mehreren Kompetenzdimensionen (vgl. Beispiel 57, 58, 59 und 60, d.h. hier, hier, hier oder hier), sodass es nicht möglich ist, z.B. ausschließlich fokussierende Supportangebote im Bereich der ÄA oder der SLV anzusteuern, die einer flüchtigen Lektürepraxis entgegenwirken, oder in der Interaktion gezielt kompetenzbezogenen Output im Bereich der ÄA und der ILV zu generieren. Für Fördermaßnahmen, die die Verbesserung von Interpretationsleistungen in spezifischen Kompetenzdimensionen durch gezieltes Training anvisieren, ist ein synthetisierender Output jedoch ungünstig. Um dem entgegenzuwirken, wurde in B4 eine Konfiguration erstellt, die die o.g. Synthesen ausschließt (vgl. LUK-GPT: Konfiguration) und zu KI-generierten Textbeobachtungen führt, die tatsächlich präzise auf die jeweiligen Kompetenzdimensionen bezogen sind (vgl. Beispiel 61

hier).[9] Die Protokolle von B4 zeigen, dass ChatGPT nun im Bereich der ÄA vielfältige fokussierende Supportangebote anbietet, bei denen der Anteil ‚interpretierender' Aussagen stark reduziert ist. Im Bereich der SLV gelingt es in B4 durch gezieltes System-Prompting (Konfiguration), Hinweise auf konkrete „Interpretationsprobleme" im Stimulustext (inklusive Erläuterung) zu erzeugen (vgl. Beispiel 63 hier) und eine überschaubare Zahl von Textstellen ins Blickfeld zu rücken, die „eine besonders intensive inhaltliche Interpretation" erfordern. Ebenso gelingt es, semantische Interpretationsangebote zu lokal begrenzten Textstellen sowie Hinweise auf verschiedene plausible Interpretationsmöglichkeiten zu generieren. Auch im Bereich des ILV liefert ChatGPT in B4 (vgl. Beispiel 67 hier) vielfältige Unterstützungsangebote zum Verstehen der ästhetischen Funktionen und Wirkungsabsichten sprachlich-stilistischer, grammatikalischer und struktureller Gestaltungsmittel, wobei allerdings die Qualität der Unterstützungsangebote variiert. Welche Angebote differenziertere Einblicke in die Funktionen der Gestaltungsmittel ermöglichen (z.B. Beispiel 68 hier, Z. 6-10) und welche eher pauschal bleiben (z.B. Beispiel 69 hier, Z. 19-23), muss jeweils evaluiert werden. Dasselbe gilt für Angebote im Bereich des VIE, die ein breites Spektrum zwischen Basisemotionen, komplexen Emotionen, aber auch Haltungen gegenüber Emotionen umfassen (vgl. Beispiel 70 hier, Z. 15-21; Beispiel 71 hier Z. 27-30 oder Beispiel 72 hier, Z. 6-11), die jedoch nicht immer plausibel sind (vgl. „Fremdscham", Z. 28-32). Im Bereich des LF ist ChatGPT grundsätzlich in der Lage, verstehenserweiterndes Kontextwissen (vgl. Beispiel 74 hier, Z. 7-12; Beispiel 75 hier, Z. 7-13 oder Beispiel 76 hier, Z. 7-16) anzubieten. Dies gilt auch in Bezug auf „Fachbegriffe" für „eine kompetente Interpretation des Gedichts", deren Eignung für die Erweiterung des Textverstehens allerdings im Einzelfall geprüft werden muss.

DISKUSSION

Die Experimentalstudien 1 und 2 zeigen, dass ChatGPT zwar plausible ‚Interpretationsansätze' in allen Dimensionen literarischer Verstehenskompetenz erzeugen kann, dass der Anteil korrekt generierter Lösungen in den einzelnen Kompetenzdimensionen z.T. allerdings erheblich variiert. Es mangelt ChatGPT mit anderen Worten an Zuverlässigkeit.

Vor diesem Hintergrund wird erkennbar, dass Lehrkräfte im Literaturunterricht jeden KI-generierten Output durch eigene Textanalysen überprüfen müssen, um falsche Aussagen über literarische Texte zu identifizieren bzw. auszuschließen. Auch Schüler:innen sind für die Unzuverlässigkeit der KI zu sensibilisieren – z.B. indem sie selbst Ergebnisse, die sie in ihrer Lerngruppe durch identische Prompts erzeugen, vergleichend überprüfen und so auf die Fehleranfälligkeit von Chatbots aufmerksam werden. Dazu müssen Evaluationsphasen, in denen die Qualität der vom Chatbot generierten Aussagen geprüft wird (vgl. Bach et al. 2025, 96 f.), in digital-gestützte Förderarrangements integriert werden. Die Befunde aus Experimentalstudie 2 zeigen überdies, wie Interaktionen mit ChatGPT so konfiguriert werden können, dass ChatGPT gezielte Unterstützungsangebote zur Bewältigung kompetenzbezogener Verstehensanforderungen generieren kann. Von besonderer Bedeutung sind hier fokussierende und konzeptuelle Supportangebote (vgl. Steinmetz 2019) im Bereich der

9 Die Konfiguration sorgt durchgehend für eine korrekte Zuordnung der Prompts zu den angesteuerten Kompetenzdimensionen. Die einzige Ausnahme sind Prompts zur Aktivierung von ILV, die von ChatGPT partiell der SLV zugeordnet wird (vgl. jeweils Z. 1-2 für Prompt 9-11 in Test 1, 3, 4, 5 und 7). Ursächlich könnten relativ hohe Korrelationen im Verhältnis von SLV und ILV sein (vgl. Frederking et al. 2016, 224).

Ästhetischen Aufmerksamkeit und des *idiolektalen Verstehens*, wo Schüler:innen Förderbedarf aufweisen (vgl. Nix / Führer 2024, 361). Genauso wichtig für die Entwicklung von KI-gestützten Fördermaßnahmen sind die Hinweise aus Experimentalstudie 2 zur Konfiguration der Darstellungsform des KI-generierten Outputs. Ohne eine lerner:innenspezifische Konfiguration von Textlänge, Ausgabemodus, sprachlichen Anforderungen etc. dürften KI-generierte Unterstützungsangebote die Lernvoraussetzungen vieler Schüler:innen verfehlen.

Damit sind erste Ansatzpunkte und Gelingensbedingungen für die Implementierung, Erprobung und Evaluation KI-gestützter Fördermaßnahmen auf empirischer Basis ins Blickfeld getreten. Derzeit werden im Teilprojekt Deutsch des Forschungsverbunds DiSo-SGW weitere Vorstudien für eine Interventionsstudie durchgeführt, die Hinweise zur Beantwortung der Frage erbringen sollen, ob sich eine differenzielle Wirksamkeit von KI-gestützten Fördermaßnahmen im Unterschied zu Fördermaßnahmen mit einschlägigen Lernaufgaben unter randomisierter Zuweisung von Klassen zu den beiden Treatmentbedingungen (sowie einer Kontrollgruppe) mit Blick auf a) literarische Verstehenskompetenz, auf b) motivationale und c) erlebnisbezogene Variablen zeigt. Von besonderer Bedeutung ist dabei eine Begleitstudie mit Lehrkräften, in der untersucht wird, inwiefern Lehrkräfte in der Lage sind, die Qualität der Outputs von ChatGPT zu beurteilen. Daten zur Ausprägung dieser Fähigkeit könnten einen Beitrag zur Erklärung von Varianz bei der Wirksamkeit KI- und nicht-KI-gestützter Fördermaßnahmen liefern. Zudem sind sie von besonderer Relevanz für die Lehrkräftebildung – einerseits angesichts der in diesem Beitrag dokumentierten Fehleranfälligkeit von ChatGPT, andererseits mit Blick auf einen problematischen Umgang von Lehramtsstudierenden mit KI-generierten Antworten und empirisch dokumentierten Einschätzungsproblemen (vgl. Magirius / Scherf 2023, 410-412). Möglicherweise brauchen aber nicht nur Studierende „Unterstützung, um das Potenzial von Chatbots für dialogische Interaktionen in professionellen Anforderungssituationen ausschöpfen zu lernen“ (ebd., 414), sondern auch erfahrene Lehrkräfte. In welchem Umfang auch bei Lehrkräften mit unterschiedlichen Qualifikationsvoraussetzungen Förderbedarf besteht, der die Beurteilung der fachlichen Korrektheit und Zuverlässigkeit von KI-generiertem Output betrifft, muss deshalb genauer empirisch untersucht werden. Dies gilt gerade auch mit Blick auf die steigende Zahl von Quereinsteiger:innen (z.T. ohne Masterstudium). Nur auf diese Weise können bedarfsorientierte und passgenaue Lehrkräftefortbildungen zur Nutzung von Large Language Models entwickelt werden.

LITERATURVERZEICHNIS

— **Kaléko, Mascha (2012)**. „Window-Shopping“. In: Kaléko, M. (Hg.): Sämtliche Werke und Briefe. München: DTV, 588.

SEKUNDÄRQUELLEN

— **Bach, Friedrich et al. (2025)**: SHIFT happens – Lernen mit und von textgenerierender KI. In: Müller, Hans-Georg / Fürstenberg, Maurice (Hg.): *DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz*. Berlin: Frank & Timme, 87–106. — **Brüggemann, Jörn / Frederking, Volker (2024)**: *Ein fachdidaktisches Modell digitaler Souveränität als Basis innovativer Lehrkräftebildung im Bereich sprachlicher, gesellschaftlicher, ökonomischer und ästhetischer Bildung.* https://digitale-souveränität.online/index/static/pdf/pub_bj_fv_2024.pdf. [03.05.2025]. — **Brüggemann, Jörn / Gölitz, Dietmar (2025)**: Personale Effekte funktionaler Bildung. Befunde zur differenziellen Wirksamkeit von kognitiv-funktionalen und ästhetisch-personalen Gesprächen über Literatur. In: Albrecht, Christian u.a. (Hg.): *Personale und funktionale Bildung im Deutschunterricht. Theoretische, empirische und praxisbezogene Perspektiven.* Stuttgart: Metzler, *147-167.* — **Eco, Umberto (1986/2008)**: Theorien interpretativer Kooperation. Versuch zur Bestimmung ihrer Grenzen. In: Kindt, Tom / Köppe, Tilmann (Hg.): *Moderne Interpretationstheorien. Ein Reader.* Göttingen: Vandenhoeck & Ruprecht, 98-129. — **Frederking, Volker / Brüggemann, Jörn / Hirsch, Matthias (2016)**. Das Fünfdimensionale Literary Literacy-Modell und seine interdisziplinären Implikationen am Beispiel der Geschichtsdidaktik. In Lehmann, Katja / Werner, Michael / Zabold, Stefanie (Hg.): *Historisches Denken jetzt und in Zukunft* (S. 211–234). Münster u.a.: LIT. — **Magirius, Marco / Scherf, Daniel / Steinmetz, Michael (2022)**: Lernunterstützung im Literaturgespräch: Nachweis und Wirkung eines potenziellen Qualitätsaspekts gesprächsförmigen Literaturunterrichts. In: *Zeitschrift für Sprachlich-Literarisches Lernen und Deutschdidaktik*, 2, 1-30. https://doi.org/10.46586/SLLD.Z.2022.9552. — **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), 406-415. https://doi.org/10.14220/mdge.2023.70.4.406. — **Magirius, Marco / Steinmetz, Michael (2025)**: Streitgespräche mit dem Meisterleser. Mit ChatGPT auf das Schreiben interpretierender Texte vorbereiten. In: Praxis Deutsch 307, 44-49. — **Meier, Christel et al. (2017)**: An extended model of literary literacy. In: Leutner, Detlev u.a. (Hg.): *Competence assessment in education.* Heidelberg: Springer, 55-74. https://doi.org/10.1007/978-3-319-50030-0_5. — **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), 327-345. https://doi.org/10.14220/mdge.2023.70.4.327. — **Nix, Daniel / Führer, Carolin (2024)**: Literarische Interaktion 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, 335-368. https://doi.org/10.1007/978-3-662-67816-9_19 — **Maiwald, Klaus (2023)**: Digitale Textkompetenz als Bildungsaufgabe in einer Kultur der Digitalität und künstlichen Intelligenz. *Mitteilungen des Deutschen Germanistenverbandes, 70*(4), 359-372. https://doi.org/10.14220/mdge.2023.70.4.359. — **Steinmetz, Michael (2019)**: Verstehensunterstützende Aufgaben im Literaturunterricht. In: Bräuer, Christoph / Kernen, Nora (Hg.): *Aufgaben- und Lernkultur im Deutschunterricht. Theoretische Anfragen und empirische Ergebnisse der Deutschdidaktik*. Frankfurt/M.: Peter Lang, 83-106. — **van Holt, Nadine, & Groeben, Norbert (2005)**. Das Konzept des Foregrounding in der modernen Textverarbeitungspsychologie [The concept of foregrounding in modern text processing psychology]. *Journal für Psychologie, 13*(4), 311–332.

ÜBER DIE AUTOR*INNEN

Dr. Carina Ascherl ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Didaktik der deutschen Sprache und Literatur an der Otto-Friedrich-Universität Bamberg. Ihre Forschungs- und Arbeitsschwerpunkte liegen im Bereich digitalisierungsbezogener Fragestellungen im Horizont des Deutschunterrichts sowie der Literaturdidaktik.

Laureen Okesson ist wissenschaftliche Mitarbeiterin am Lehrstuhl für Didaktik der deutschen Sprache und Literatur an der Otto-Friedrich-Universität Bamberg. Ihre Forschungs- und Arbeitsschwerpunkte liegen im Bereich der empirischen Kompetenz- und Unterrichtsforschung sowie der fachdidaktischen Lehrkräfteforschung.

Jörn Brüggemann ist Inhaber des Lehrstuhls für Didaktik der deutschen Sprache und Literatur an der Otto-Friedrich-Universität Bamberg. Seine Forschungs- und Arbeitsschwerpunkte liegen im Bereich der empirischen Kompetenz- und Unterrichtsforschung, der Lehrkräfteprofessionalisierungsforschung, der fachspezifischen Mediendidaktik und der Geschichte des Deutschunterrichts.

Volker Frederking hat den Lehrstuhl für Didaktik der deutschen Sprache und Literatur an der Friedrich-Alexander-Universität Erlangen-Nürnberg inne. Seine Forschungsschwerpunkte liegen u.a. im Bereich der empirischen Kompetenz- und Unterrichtsforschung, der fachspezifischen Mediendidaktik und der Theorie der Allgemeinen Fachdidaktik.