JG. 8 (2026), H. 1
DOI: 10.18716/OJS/MIDU/2026.1.9
VERÖFFENTLICHT AM 27.07.2026

M i D U

Medien im Deutschunterricht

WIE INTERPRETIEREN STUDIERENDE COMPUTERSPIELE MIT KÜNSTLICHER INTELLIGENZ?

Florian Hesse
Friedrich-Schiller-Universität Jena | florian.hesse@uni-jena.de

Gerrit Helm
Friedrich-Schiller-Universität Jena | gerrit.helm@uni-jena.de

Maximilian Arend
Friedrich-Schiller-Universität Jena | maximilian.arend@uni-jena.de

ABSTRACT

(Angehende) Lehrkräfte haben empirischen Studien zufolge Probleme, literarische Texte zu verstehen. KI wird hier unterstützendes Potenzial zugeschrieben, wenngleich Untersuchungen zur tatsächlichen Nutzung noch ausstehen. In dieser Studie wird daher betrachtet, wie Lehramtsstudierende (*N* = 34) Computerspiele mit Hilfe von KI interpretieren. Geprüft wird dabei, inwiefern sich die Deutungen der Studierenden und der KI hinsichtlich ihrer Plausibilität und Elaboriertheit in Abhängigkeit von Studienfortschritt und zu interpretierendem Computerspiel unterscheiden. Zudem wird kodiert, wie die Studierenden KI nutzen. Dabei zeigt sich, dass fortgeschrittene Studierende die Spiele nicht signifikant plausibler und elaborierter interpretieren. Wohl aber lassen sich signifikante Unterschiede in Abhängigkeit vom Spiel nachweisen. Die KI-Deutungen sind den Studierendendeutungen signifikant überlegen und über beide Spiele konstant. Trotz dieses Potenzials zeigt sich in der Nutzung, dass die Studierenden die Deutungen der KI nur selten in eigene Deutungen integrieren. Außerdem treten mehr problematische als produktive Nutzungsweisen auf. Der Beitrag diskutiert daraus folgende Implikationen für Forschung und Lehre.

SCHLAGWÖRTER

KI UND UNTERRICHTSPLANUNG — KI UND LITERATURUNTERRICHT — KI UND TEXTVERSTEHEN — KI UND LEHRAMT — KI UND LITERARISCHE TEXTE

JG. 8 (2026), H. 1
DOI: 10.18716/OJS/MIDU/2026.1.9
VERÖFFENTLICHT AM 27.07.2026

M i D U

Medien im Deutschunterricht

ABSTRACT (ENGLISH)

HOW DO STUDENT TEACHERS INTERPRET VIDEO GAMES WITH ARTIFICIAL INTELLIGENCE?

According to empirical studies, (prospective) teachers have problems understanding literary texts. AI is considered to have supportive potential here, although studies on its actual use are still missing. This study therefore analyses how student teachers (N = 34) interpret video games with AI. It examines the extent to which the interpretations of the students and the AI differ in terms of their plausibility and elaboration depending on the progress of the study and the computer game to be interpreted. In addition, it is coded how the students use AI. The results show that advanced students do not interpret the games significantly more plausibly or elaborately. However, there are significant differences with regard to the game. The AI interpretations are significantly superior to the student interpretations and are constant across both games. Despite this potential, the findings show that students rarely integrate the AI's interpretations into their own interpretations. In addition, there are more problematic than productive ways of use. The article discusses the following implications for research and teaching.

KEYWORDS

— AI AND LESSON PLANNING — AI AND LITERATURE TEACHING — AI AND TEXT COMPREHENSION — AI AND TEACHING — AI AND LITERARY TEXTS

1 — EINLEITUNG

Sichtet man aktuelle empirische Studien zur deutschdidaktischen Professionsforschung, lässt sich wiederkehrend die bisweilen ernüchternde Beobachtung machen, dass sich Lehramtsstudierende damit schwertun, die im Unterricht zu verhandelnden literarischen Gegenstände selbst zu verstehen (vgl. Hesse / Magirius / Heins 2025). Dies stellt für die Vorbereitung von Unterricht eine Herausforderung dar, insofern ein Verständnis des Unterrichtsgegenstandes eine notwendige Voraussetzung für die Gestaltung qualitätsvoller und mithin adaptiver Lehr-Lernsettings darstellt (vgl. Hesse 2024).

Seit dem Aufkommen von ChatGPT und vergleichbaren KI-basierten Chatbots wird vor diesem Hintergrund in der Deutschdidaktik darüber diskutiert, inwiefern KI als Verstehenshilfe nicht nur für Schüler:innen (vgl. hierzu im Überblick Magirius et al. 2024), sondern auch für angehende und erfahrene Lehrpersonen fungieren könnte (vgl. Magirius / Scherf 2023). So liegt die besondere Attraktivität der KI darin, dass sie Impulse für das Textverstehen von lokalen Verständnisfragen bis hin zu umfassenden Interpretationen liefern (vgl. für die Schule z. B. Führer / Nix, 2023) und darüber hinaus auf mögliche Verstehensanforderungen hinweisen oder Aufgabenstellungen samt Erwartungshorizont generieren kann. Welche Qualität die so generierten Informationen aufweisen und wie angehende Lehrkräfte diese für ihren Verstehens- und Planungsprozess nutzen, ist in der Deutschdidaktik bislang allerdings noch nicht umfassend erforscht worden.

Mit dem vorliegenden Artikel möchten wir zur Schließung dieser Forschungslücke beitragen, indem wir untersuchen, wie angehende Lehrpersonen literarisch anspruchsvolle Gegenstände (hier: Computerspiele) mit und ohne Unterstützung von KI interpretieren und welche Rolle dabei der Studienfortschritt und der betrachtete Gegenstand spielen. Dazu gibt der Beitrag zunächst einen Überblick über die aktuelle Forschung, ehe in einem zweiten Schritt das methodische Vorgehen erläutert wird. Darauf aufbauend werden die Ergebnisse unserer Untersuchung vorgestellt und in den Forschungsstand eingeordnet.

2 — FORSCHUNGSSTAND

2.1 ZUR ROLLE PROFESSIONELLER KOMPETENZEN VON LEHRKRÄFTEN

Der vorliegende Beitrag untersucht die Fähigkeiten von Studierenden, sich mit literarischen Texten interpretierend auseinanderzusetzen, und lässt sich damit dem Forschungsfeld der literaturdidaktischen Professionalisierungsforschung zuordnen. Eine Kernannahme dieses Forschungsfeldes besteht darin, dass professionelles Lehrkräftehandeln entscheidend für die Durchführung eines qualitätsvollen Unterrichts ist, der wiederum kognitive, emotionale oder motivationale Outcomes von Lernenden bedingt. Diese Annahme wird im Kaskaden-Modell professioneller Lehrkräftekompetenz von Krauss et al. (2020) illustriert, das den Weg von Dispositionen der Lehrkraft bis hin zum Lernen der Schüler:innen heuristisch nachzeichnet (Abb. 1).

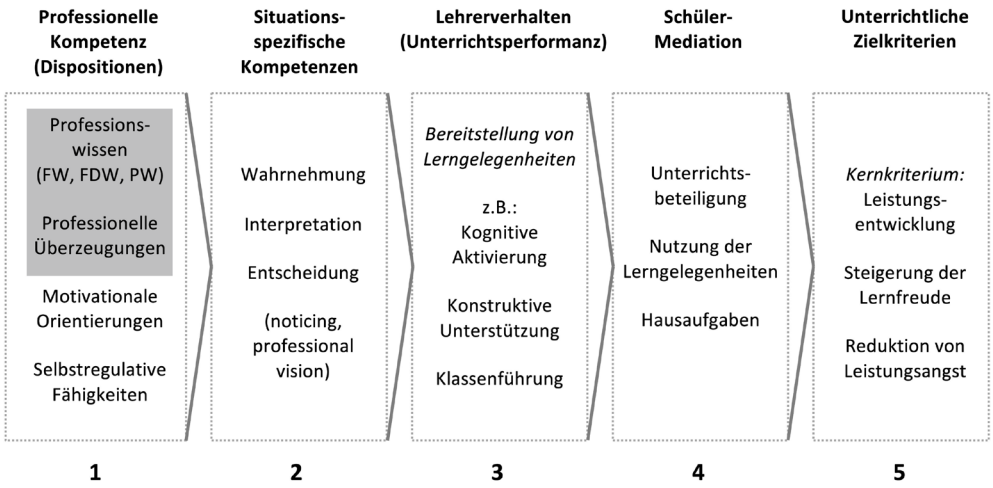

Abb. 1: Das Kaskadenmodell als Integration von Modellen professioneller Lehrkraft-Kompetenzen, Unterrichtsqualitätsmodellen sowie Paradigmen der Forschung über Lehrkräfte (nach Krauss et al., 2020)

Dass die Leistungsentwicklung von Lernenden (Säule 5) mit Blick auf das Lesen (und vermutlich auch für das literarische Verstehen) nicht den Erwartungen entspricht, führen uns nationale und internationale Vergleichsstudien regelmäßig vor Augen (vgl. z.B. PISA: Heine et al., 2023). Ursächlich hierfür dürften neben sozio-ökonomischen Voraussetzungen der Kinder vor allem Merkmale des Unterrichts sein. So zeigen etwa die Studien von Lotz (2016), Winkler (2017, 2020), Tengberg et al. (2021), Schilcher / Glondys / Wild (2023) oder Hesse (2024), dass die Lerngelegenheiten im alltäglichen erstsprachlichen Unterricht (Säule 3) zwar eine anregende Arbeitsatmosphäre bieten, darüber hinaus aber wenig kognitiv-emotional aktivierend sind sowie fachlich-inhaltliche Probleme aufweisen. Um Defizite in der angebotsseitigen Unterrichtsqualität erklären zu können, kann die literaturdidaktische Professionsforschung mittlerweile auf zahlreiche Studien verweisen, die Überzeugungen, Wissensbestände (Säule 1) oder situationsspezifische Fähigkeiten (Säule 2) von (angehenden) Lehrkräften untersucht haben (vgl. im Überblick: Hesse / Magirius / Heins 2025). Diese Erkenntnisse sind insofern wichtig, als sie (in Ansätzen) Auskunft darüber geben, welche Stellschrauben relevant sind, um in der Lehrkräfteaus- und -weiterbildung Grundlagen für eine spätere qualitätsvolle Unterrichtsgestaltung zu legen. Im folgenden Abschnitt werden vor diesem Hintergrund und mit Blick auf die Fragestellung des Beitrags vor allem solche Studien besprochen, die sich dezidiert mit der Fähigkeit von Studierenden oder Lehrkräften auseinandersetzen, literarische Gegenstände zu interpretieren.

2.2 STUDIEN UNTER LEHRAMTSSTUDIERENDEN ZUR INTERPRETATION LITERARISCHER TEXTE

Die Fähigkeit von (angehenden) Deutschlehrkräften, selbständig diejenigen Texte zu verstehen, die sie auch im Unterricht vermitteln müssen, wird selten eigenständig untersucht. Dennoch scheinen in umfangreicheren Studien, die das Textverstehen

mindestens als Teilleistung professionellen Lehrkräftehandelns mitdenken, mit großer Übereinstimmung Defizite im Textverstehen auf (vgl. im Überblick auch Hesse et al. 2025).

— Anselm (2011) kann in einer Studie zur professionellen Kompetenz unter (angehenden) Lehrpersonen (*N* = 169) zeigen, dass insbesondere unter den von ihr untersuchten Studienanfänger:innen (*N* = 41) nur 30 % der Proband:innen in der Lage waren, zwei oder mehr Deutungen zu einem epischen Kurztext (*Das Märchen vom kleinen Hans, der eine Glatze kriegte,* W. Biermann) anzugeben; etwa 50 % fanden keinen interpretativen Zugang (vgl. Anselm 2011, S. 179).

— Pissarek / Schilcher (2017) untersuchten im Rahmen der FALKO-Studie die Fähigkeit von (angehenden) Lehrkräften (*N* = 271), Potenziale literarischer Texte einschätzen zu können. Studierende bzw. Lehrkräfte erreichen hier im Mittel nur 2.9 bzw. 3.1 von insgesamt 7 möglichen Punkten und unterscheiden sich darüber hinaus in dieser Dimension *nicht* signifikant. Letzterer Befund deutet darauf hin, dass Fähigkeiten in diesem Bereich durch bloße Berufserfahrung nicht automatisch erworben werden (vgl. Pissarek / Schilcher 2017, S. 93).

— Im Rahmen des Praxissemester-Projekts OVID-PRAX konnten Winkler / Seeber (2020) zeigen, dass rund ein Drittel der Studierenden (*N* = 127) z. T. erhebliche Probleme damit hatte, die an Schüler:innen gerichteten Textverstehensaufgaben zur Fabel *Wolf und Lamm* (Äsop) selbst zu lösen. Ebenso verschlechterte sich im Verlauf der Praxisphase die Fähigkeit der Studierenden, die Anforderungen des Textes differenziert zu bestimmen (vgl. Winkler / Seeber 2020, S. 43).

— Dick (2024) kann in einer Interventionsstudie zum vernetzten Professionswissen zeigen, dass die Studierenden sowohl im Prä- als auch im Posttest (deutlich) weniger als die Hälfte der möglichen Punkte im Textverstehen erreichen und damit hinter normativen Erwartungen zurückbleiben (vgl. Dick 2024, S. 315).

— Winkler / Wieser (2025) untersuchen Fähigkeiten von fortgeschrittenen Studierenden im Bereich der Erzähltextanalyse in Abhängigkeit von verschiedenen Unterstützungsformen. Hierbei zeigt sich, dass die Studierenden den vorgegebenen literarischen Text (Kurzgeschichte *Nacht* von Sibylle Berg) generell nicht zufriedenstellend verstehen und dass selbst unter den Studierenden mit „hohem Support“ die Zahl der Probanden mit „vollständiger Lösung“ bei den Analyseaufgaben nur rund 30 % beträgt (Winkler / Wieser 2025, S. 16).

Wenngleich die unterschiedlichen Studien aufgrund der variierenden methodischen Zugriffsweisen, Standorte, literarischen Texte oder abweichenden Studienfortschritte der Proband:innen nur bedingt vergleichbar sind, zeichnen sich über alle Studien hinweg Gemeinsamkeiten hinsichtlich mangelnder interpretativer Fähigkeiten ab. Diese lassen sich auch durch Studien zur Unterrichtsplanung (vgl. Masanek / Doll 2020) oder professionellen Wahrnehmung (vgl. Heins 2022; Hesse / Seeber 2025) stützen, in denen Probleme hinsichtlich der Wahrnehmung des Unterrichtsgegenstandes sichtbar werden. Besonders frappierend sind diese Befunde deshalb, weil die Studierenden trotz Absolvierung zahlreicher literaturwissenschaftlicher Studienmodule Texte offensichtlich nicht selbstständig erschließen können. Da erste

Untersuchungen zeigen, dass insbesondere „[d]ie interpretatorische und analytische Kompetenz der KI […] beeindruckende, jedenfalls schulisch und speziell prüfungstechnisch wohl satisfaktionsfähige Ergebnisse“ hervorbringt (vgl. Susteck / Perder 2023, S. 11), liegt es nahe, KI Potenziale zur Unterstützung von Verstehensprozessen im Rahmen der Unterrichtsvorbereitung zuzuschreiben. Einschränkend muss gleichwohl anmerkt werden, dass die Leistungsfähigkeit der KI nicht pauschal bewertet werden kann, da diese vom genutzten Modell sowie der Kanonizität der zu analysierenden Texte (und den im Trainingssatz verfügbaren Daten zum Text) abhängen dürfte. So stellen Susteck und Perder (2023, S. 11) in ihrer Untersuchung ebenso fest, dass „Verknüpfungen, Kontexte und Erläuterungen […] nur bedingt funktionierten.“

2.3 STUDIEN ZUR NUTZUNG VON KI FÜR DAS INTERPRETIEREN LITERARISCHER TEXTE BEI STUDIERENDEN

Zur Frage, inwiefern Studierende KI zielführend für die Interpretation literarischer Texte sowie für weitere Planungsaufgaben – z.B. die Generierung von Lernzielen – nutzen können, liegt bislang nur eine explorative Untersuchung von Magirius / Scherf (vgl. 2023) vor. Die beiden Autoren ließen 21 Studierende der Universitäten Heidelberg und Berlin u.a. Interpretationen zum expressionistischen Gedicht *Weltende* (Jakob van Hoddis) erarbeiten. Hierbei wurden sowohl problematische als auch produktive Nutzungsweisen sichtbar.

Als produktive Umgangsweisen kodierten die Autoren Fälle, in denen die Studierenden durch die KI Impulse für ihre eigene Interpretationshypothese erhielten, mit der KI dialogisch interagierten oder mit Hilfe der KI eigene Deutungshypothesen auf ihre Plausibilität prüften. Als problematisch wurden indes Umgangsweisen gekennzeichnet, die sich durch eine (vollständige) Übernahme fehlerhafter Deutungsvorschläge auszeichneten oder in denen KI-Antworten als selektive Bestätigung eigener unzureichender Deutungen genutzt wurden.

Hinsichtlich des quantitativen Verhältnisses produktiver versus problematischer Umgangsweisen lässt sich aufgrund der kleinen Stichprobe bislang kein abschließendes Urteil fällen. Allerdings formulieren die Autoren die Hypothese, dass „problematische Umgangsweisen mit der KI auf Unsicherheiten im Umgang mit den zu deutenden literarischen Texten zurückzuführen sind“ (Magirius / Scherf 2023, S. 413). Dies wiederum lege einen Matthäus-Effekt nahe, demzufolge Studierende mit guten Voraussetzungen eher von der KI profitieren werden, während Studierende mit weniger guten Voraussetzungen die Potenziale nicht ausreizen werden (vgl. ebd., S. 414). Abschließend weisen die Autoren darauf hin, dass weitere Studien notwendig seien, die den Einfluss von „Personen- und Textvariablen“ auf das „Deutungsgeschehen mit KI-Nutzung“ untersuchen (ebd., S. 413).

Die bei Magirius und Scherf ermittelten Ergebnissen lassen sich teilweise durch Studien zur Unterrichtsplanung mit KI decken, die nicht spezifisch auf das Interpretieren literarischer Texte zielen. Neumann und Steensen (in Vorb.) zeigen auf Basis einer quantitativen Befragung von 92 Deutschlehramtsstudierenden nach dem Praxissemester, dass die Nutzung von ChatGPT in der Unterrichtsplanung bereits

weit verbreitet ist – vor allem zur Materialerstellung, Differenzierung und Formulierung von Lernzielen. Als Vorteile der KI werden Zeitersparnis und kreative Impulse genannt, während Fehleranfälligkeit, Oberflächlichkeit und mangelnde fachdidaktische Fundierung als Herausforderungen aufscheinen. Gabriel (2025) untersuchte in leitfadengestützten Interviews mit 21 Studierenden des Primarstufenlehramts, wie diese Unterrichtsentwürfe mit und ohne KI-Unterstützung bewerten. Die Ergebnisse zeigen, dass Studierende den Einsatz von ChatGPT ähnlich wie die vorangehenden Studien insbesondere zur Ideengenerierung und Zeitersparnis schätzen, zugleich aber fehlende Quellenangaben, mangelnde Individualisierung und die Notwendigkeit kritischer Prüfung der KI-Ergebnisse problematisieren. Treß (2024) beschreibt schließlich aus musikdidaktischer Perspektive vier typische Anwendungssituationen von ChatGPT: als Assistent für die Unterrichtsplanung, Assoziations- und Ideengeber, Hilfsmittel zur Materialerstellung sowie Wissensquelle. Während sich der Einsatz in frühen Planungsphasen teils als produktiv erwies, wurde die kleinschrittige Arbeit mit der KI von einer Lehrkraft sogar als „Zeitfresser" bezeichnet (Treß 2024, S. 129). Fachpraktische und implizite Wissensbestände erwiesen sich häufig als ungenau oder fehlerhaft, wurden aber selten kritisch überprüft, weshalb Treß die Entwicklung einer fachspezifischen AI-Literacy fordert.

Insgesamt zeigt sich über die genannten Studien hinweg ein ähnliches Muster: Studierende erkennen die Potenziale generativer KI vor allem in der Ideenfindung, Strukturierung und Effizienzsteigerung didaktischer Planungsprozesse, stoßen jedoch zugleich auf Grenzen hinsichtlich fachlicher Präzision, kritischer Bewertung und reflektierter Nutzung. Damit scheint sich auch im Kontext der Unterrichtsplanung die von Magirius und Scherf formulierte Annahme zu bestätigen, dass der Ertrag der KI-Nutzung von den fachlichen und reflexiven Kompetenzen der Studierenden abhängt.

3 — METHODIK

3.1 FRAGESTELLUNGEN UND HYPOTHESEN

Genau an diesem Desiderat setzt die vorliegende Studie an, wenn sie Umgangsweisen mit KI bei der Interpretation zweier literarisch anspruchsvoller Computerspiele untersucht und dabei Studierende mit unterschiedlichem Studienfortschritt berücksichtigt. Dabei resultiert der Fokus auf Computerspielen neben ihren vielfältigen Potenzialen für das literarische Lernen (z. B. Boelmann 2023, Kepser 2023) daraus, dass bisherige Studien ausschließlich literarische Texte i. e. S. untersuchen und ein Desiderat in der Erforschung anderer literarischer Medien liegt. Unsere Forschungsfragen lauten wie folgt:

1. Unterscheiden sich Plausibilität und Elaboriertheit der Interpretationen in Abhängigkeit vom Studienfortschritt? Unterscheiden sich diesbezüglich menschliche und KI-generierte Deutungen?

2. Unterscheiden sich Plausibilität und Elaboriertheit der Interpretationen in Abhängigkeit vom untersuchten Spiel? Unterscheiden sich diesbezüglich menschliche und KI-generierte Deutungen?
3. Wie gehen Studierende konkret mit den Deutungen der KI um?

Hinsichtlich der ersten Fragestellung wäre aus normativer Sicht zu erwarten, dass Studierende am Ende ihres Studiums (im Vergleich zu Beginn) in der Lage sind, „Methoden der Textanalyse/Textinterpretation" (KMK 2024, S. 27) anzuwenden. Gleichwohl deuten beispielsweise die o. g. Befunde aus FALKO darauf hin, dass zwischen Studierenden und erfahrenen Lehrpersonen keine signifikanten Unterschiede bestehen, ebenso wie die Befunde von OVID-PRAX nahelegen, dass die Fähigkeiten zur Analyse und Interpretation während der Praxisphase sogar abnehmen. Insofern kann hier keine an normativen Erwartungen orientierte Hypothese formuliert werden. Bezüglich der KI kann in Anlehnung an die Befunde bei Magirius / Scherf (2023) vage vermutet werden, dass fortgeschrittene Studierende tendenziell eher in der Lage sind, plausible und elaborierte Deutungen der KI zu elizitieren, als Studierende, die noch am Anfang ihres Studiums stehen und (mutmaßlich) über weniger ausgereifte interpretatorische bzw. textanalytische Fähigkeiten verfügen.

Dies gilt auch für die zweite Fragestellung, da die gewählten Computerspiele (s. Abschn. 3.3, unten) zwar beide anspruchsvoll sind, ihre Verstehensanforderungen aber auf unterschiedlichen Ebenen liegen. Zudem wäre es angesichts der bislang vorliegenden Studien denkbar, dass sich Unterschiede zwischen den Spielen statistisch nur schwer ausmachen lassen, weil die Studierenden in beiden Spielen gleichermaßen Verstehensprobleme zeigen.

Hinsichtlich der dritten, eher rekonstruktiven Fragestellung gehen wir davon aus, dass sich wie bei Magirius und Scherf (2023) auch in unserer Studie sowohl problematische als auch produktive Nutzungsweisen der KI unterscheiden lassen. Dabei kann einerseits angenommen werden, dass produktive Nutzungsweisen bei Studierenden im fortgeschrittenen Semester häufiger auftreten, da diese die KI-Outputs und literarischen Texte möglicherweise aufgrund eines ausgeprägteren Fachwissens bzw. KI-Wissens besser in Beziehung setzen können. Andererseits kann vermutet werden, dass das Fachwissen gar nicht unbedingt in Beziehung zum Studienfortschritt steht (siehe oben) und zum Erhebungszeitpunkt KI-spezifische Lerngelegenheiten im Studium allenfalls rudimentär implementiert waren, weshalb nicht zwingend mit Unterschieden im KI-bezogenen Wissen zu rechnen ist.

3.2 STICHPROBE

An der Studie nahmen insgesamt 34 Studierende teil, davon 16 im Bachelor-Studium (meist 3.–5. Fachsemester) an der Universität Wuppertal und 18 im Hauptstudium eines Staatsexamensstudiengangs (meist 8.-10. Fachsemester) an der Universität Jena. Diese Zusammensetzung resultiert aus den jeweiligen Lehrverpflichtungen der beiden Erstautoren zum Erhebungszeitpunkt.[1] Beim Vergleich der beiden Kohorten

1 Dass an dieser Stelle zwei verschiedene Studienstrukturen vermischt werden, ist einerseits nicht optimal und als limitierender Faktor zu berücksichtigen. Andererseits kann relativierend angemerkt werden, dass gerade mit Blick auf die BA-Studierenden die Studiengangszugehörigkeit noch insofern weniger relevant ist, als hier bei beiden Standorten allenfalls Überblicksveranstaltungen im Bereich der Germanistik und der Deutschdidaktik besucht wurden, die keinen spezifischen Wissensvorsprung in für die Studie relevanten Bereichen der einen Gruppe gegenüber der anderen vermuten lassen.

muss insofern limitierend berücksichtigt werden, dass es sich lediglich um einen Quasi-Längsschnitt handelt, der zukünftig durch echte Längsschnitte validiert werden muss.

Beide Studierendengruppen spielten das Spiel *Every day the same dream* (Molleindustria 2009). Die Gruppe der fortgeschrittenen Studierenden spielte eine Woche nach der ersten Erhebung zusätzlich das Spiel *The Plan* (Krillbite 2013), um mit Blick auf Forschungsfrage 2 Aussagen darüber treffen zu können, ob die Qualität der Interpretationen gegenstandsabhängig ist.

Bei beiden Kohorten handelt es sich jeweils um eine Vollerhebung der zum Zeitpunkt der Erhebung anwesenden Studierenden in deutschdidaktischen Seminaren mit mediendidaktischem Schwerpunkt, die von den ersten beiden Autoren des Beitrags geleitet wurden. Somit kann ausgeschlossen werden, dass es sich bei der Probandengruppe um eine Positivselektion besonders interessierter Studierender handelt.

3.3 DIE SPIELE

Every Day the Same Dream ist ein kurzes 2D-Browserspiel, das durch seine repetitive Struktur und minimalistische Gestaltung die Entfremdung und Verweigerung von Arbeit in einem kapitalistischen System thematisiert. Spieler:innen übernehmen in einer schwarz-weiß gehaltenen Welt die Rolle eines anonymen Büroarbeiters, der jeden Tag denselben Ablauf erlebt: Aufstehen, zur Arbeit fahren, im Großraumbüro unter Dutzenden gleich aussehenden Menschen arbeiten.

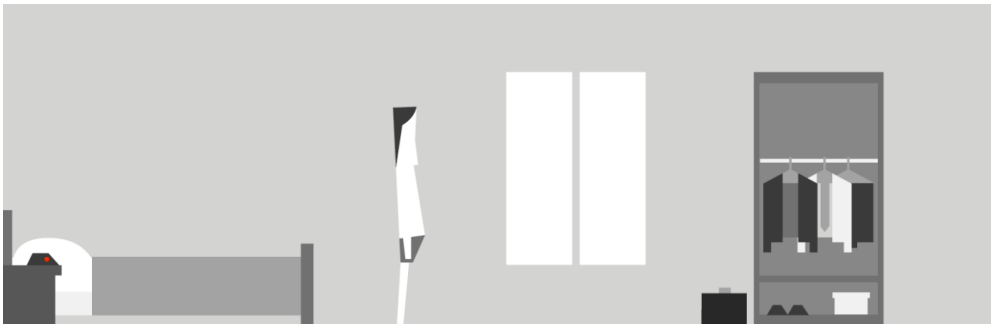

Abb. 2: Screenshot: Every day the Same Dream (Molleindustria 2009)

Die Spielmechanik basiert auf einfachen Steuerungselementen (Pfeiltasten und Interaktionstaste) und einem Loop-System, bei dem der Tag immer wieder von vorn beginnt. Das Ziel des Spiels wird durch eine alte Frau in einem Fahrstuhl vorgegeben: Es seien fünf Schritte zu absolvieren, um ein besserer Mensch zu werden. Diese Schritte erweisen sich als auffällig willkürlich (ein heruntergefallenes Blatt aufheben, eine Kuh streicheln, unbekleidet zur Arbeit kommen …) und tragen in keiner Weise dazu bei, den oben angesprochenen Konflikt aufzulösen. Das Spiel endet damit, dass man einer Figur, die genauso aussieht wie die eigene, dabei zusieht, wie sie vom Hochhausdach springt.

Interpretativ lässt sich das Spiel in jedem Fall als Kritik an einer kapitalistischen Arbeitswelt lesen. Gleichwohl würde es zu kurz greifen, das Spiel als Aufforderung zum Eskapismus zu begreifen, führen doch die vermeintlichen Selbstoptimierungsstipps alle ins Leere. Vielmehr lässt sich mit Kepser (2025) argumentieren, dass das Spiel Anleihen am Brecht'schen Theater nimmt, also auf Verfremdung und Aufklärung statt auf Immersion und Unterhaltung setzt. Dies zeigt sich neben der anonymen Gestaltung des eigenen Avatars insbesondere am Schluss, der die Aufgabe, Lösungen für das dargestellte Problem zu finden, den Spielern aufbürdet. Eine solche Deutung passe laut Kepser auch gut in das Gesamtwerk des Entwicklers Paolo Pedercini, der für seine systemkritischen Spiele bekannt ist.

Weniger systemkritisch, dafür aber philosophisch aufgeladen ist das zweite Spiel, *The Plan*, vom Krillbite Studio aus dem Jahr 2013. Es handelt sich um ein kurzes, experimentelles Spiel, das aus der Top-Down-Perspektive den vertikalen Aufstieg einer Fliege begleitet (Gesamtspielzeit: ca. 5 Minuten). Die Spielmechanik beschränkt sich auf die einfache Steuerung der Fliege (Flugbewegung mit den Pfeiltasten), wobei Hindernisse wie Windböen oder fallendes Laub den Aufstieg der Fliege stören. Mit zunehmender Höhe wird die Fliege immer kleiner (Zoom-Out); der nahezu lautlose Aufstieg wird kurz vor Schluss von einer eindringlichen musikalischen Untermalung (Edvard Griegs *Der Tod des Åses*) begleitet und endet nach Durchquerung einer quasi-göttlichen Sphäre abrupt, wenn die Fliege unerwartet an einer riesigen Glühbirne verglüht (Abb. 3).

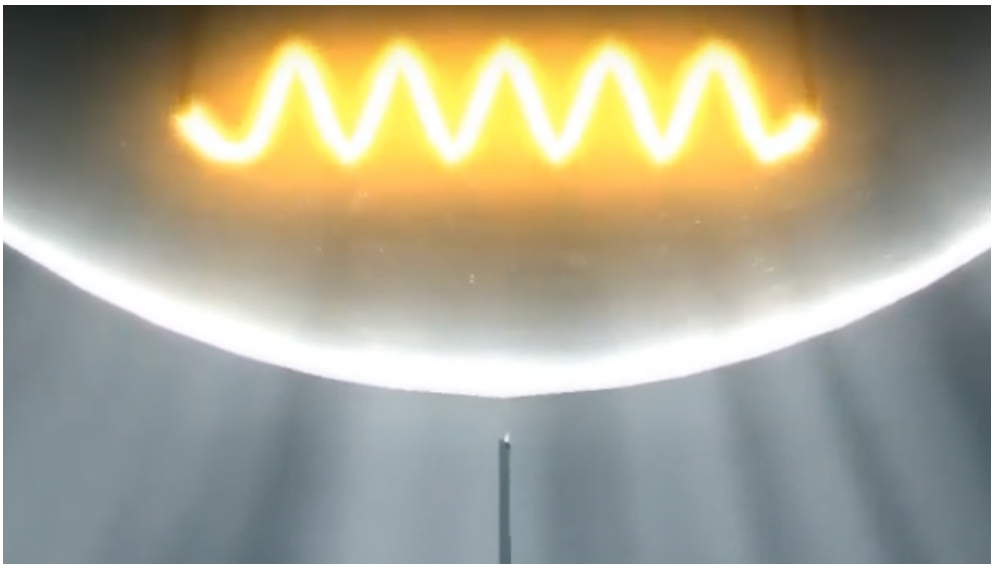

Abb. 3: Screenshot aus The Plan (Krillbite 2013)

Insbesondere durch den überraschenden Schluss eröffnet das Spiel Raum für vielfältige Deutungen. Schöffmann / Hoiß (2023) arbeiten drei davon heraus. In einer anthropozentrischen Lesart könnte man das Spiel als Frage nach dem freien Willen lesen, da die Fliege am Ende nicht mehr steuerbar und der Glühbirne ausgeliefert ist; in einer deterministischen und zum Titel passenden Lesart zeigt das Spiel auf, dass

das Leben ungeachtet der Frage, ob man auf dem Lebensweg Hindernisse überwindet oder nicht, mit dem Tod endet; eine metaphysische Lesart kann schließlich grundlegende Fragen des Seins im Zusammenhang mit dem Aufstieg der Fliege zum Licht sowie den religiösen Elementen fokussieren.

3.4 ERHEBUNGSMETHODIK

Die Erhebung fand Ende Januar/Anfang Februar 2024 statt. Die Studierenden sollten das Spiel *Every day the same dream* spielen und anschließend verschiedene Aufgaben schriftlich in einem Online-Fragebogen beantworten. Die fortgeschrittenen Studierenden spielten darüber hinaus mit Blick auf Forschungsfrage 2 eine Woche später das Spiel *The Plan* und bearbeiteten erneut die Aufgaben im Fragebogen. Da die Studierenden das Spiel bzw. die beiden Spiele zuhause spielten, gab es keine Begrenzung der Spielzeit (wiederholte Spieldurchläufe waren möglich) und der Bearbeitungszeit der Aufgaben.

Um eine möglichst hohe Anschlussfähigkeit an den bisherigen Diskurs zu ermöglichen, wurde die Erhebung nah an das Setting von Magirius / Scherf (2023) angelehnt. Dies schließt ein, dass die verwendeten Aufgabenstellungen fast wörtlich übernommen und nur punktuell Anpassungen mit Blick auf die eingesetzten Untersuchungsgegenstände vorgenommen wurden. Wie in Tab. 1 deutlich wird, basiert das Untersuchungsdesign auf der Idee, den Studierenden zunächst eine eigenständige Auseinandersetzung mit dem Spiel zu ermöglichen (Schritte 0+1), die dann durch die Perspektive der KI erweitert wird (Schritt 2)[2]. Abschließend konnten die Studierenden nach eigenem Ermessen die selbstständig entwickelte Deutung überarbeiten (Schritt 3).

Schritt	Aufgabe	Konkrete Aufgabenstellung
0	(mehrfaches) Spielen des Spiels/der Spiele	Spielen Sie das Spiel *Every day the same dream* [alle Studierende] und *The plan* [nur Fortgeschrittene].. (Anmerkung: Das Spiel konnte natürlich auch während der Aufgabenbearbeitung wiederholt gespielt werden.)
1	Eigene Deutung (ohne KI)	Schreiben Sie Interpretationsansätze zum Spiel auf. Wie lässt sich das Spiel **unter Berücksichtigung von Darstellungs- und Spielelementen** deuten? (Sie können beginnen mit ‚Das Spiel könnte aussagen, dass...' oder ‚Das Spiel zeigt mir, dass...').
2	Deutung mit KI	Nutzen Sie ChatGPT oder Google Bard, um eine Deutungshypothese zu formulieren. (Anmerkung: Die KI-Logs wurden von uns in einem separaten Textfeld zur Verfügung gestellt.)
3	Ggf. Überarbeitung der eigenen Deutung	Würden Sie Ihre eigene Deutung nach der Sichtung des Ergebnisses von ChatGPT oder Bard überarbeiten wollen? Schreiben Sie ggf. die ‚neue' Deutung auf.

Tab. 1: Ablauf der Erhebung

2 *Wie die Studierenden die KI in diesem Schritt ansteuerten, war ihnen freigestellt. Im Vorfeld der Erhebung hat sich für beide Spiele gezeigt, dass die KI schon bei einem simplen Interpretationsauftrag und Angabe des Spiels (z.B. „Interpretiere das Spiel ‚Every day the Same Dream'.") brauchbare Deutungen liefert. Eine Sichtung der Prompts der Studierenden zeigte, dass diese ebenfalls mit kurzen Prompts arbeiteten. Anzunehmen ist folglich, dass die KI bei der Darbietung von Interpretationshypothesen nicht das Spiel an sich interpretiert hat, sondern im Internet befindliche Überlegungen zum Spiel (aus Rezensionen, Blogs, Foren etc.) aus dem Trainingsdatensatz der KI.*

3.5 AUSWERTUNGSMETHODIK

3.5.1 RATING DER INTERPRETATIONEN

Um mit Blick auf die Forschungsfragen 1 und 2 die Qualität der Deutungen in Abhängigkeit von Studienfortschritt und Spiel vergleichen zu können, war es notwendig, ein Ratingverfahren zu entwickeln. Dieses musste einerseits abstrakt genug sein, um die beiden unterschiedlichen Spiele gleichermaßen berücksichtigen zu können. Andererseits musste es zulassen, qualitative Unterschiede zwischen den verschiedenen Deutungshypothesen der Studierenden sowie der KI sichtbar zu machen.

Bei einer ersten Sichtung der von den Studierenden sowie von den KI-Chatbots generierten Deutungshypothesen fielen zwei Aspekte unmittelbar auf, hinsichtlich derer sich die Antworten stark unterschieden: Zum einen differierten die Deutungen hinsichtlich ihrer Plausibilität und mithin der Frage, wie stimmig die vorgeschlagene Deutung in Bezug auf das jeweilige Spiel ist. Zum anderen variierten die Deutungen erheblich im Grad der Elaboriertheit. Damit ist gemeint, dass die Deutungshypothesen trotz einer durch Fettdruck hervorgehobenen Aufforderung, die Hypothesen auf konkrete Darstellungs- und Spielelemente zu beziehen, mitunter gar nicht, zum Teil aber auch sehr ausführlich begründet wurden.

Für die Auswertung der Daten wurden folglich sowohl die Deutungen der Studierenden als auch die uns vorliegenden Deutungen der KI jeweils dreistufig hinsichtlich Plausibilität und Elaboriertheit eingeschätzt. Die Stufen lauteten dabei:

	Plausibilität	**Elaboriertheit**
1.	plausibel (Score: 3)	elaboriert (Score: 3)
2.	teilweise plausibel (Score: 2)	teilweise elaboriert (Score: 2)
3.	gar nicht plausibel (Score: 1)	gar nicht elaboriert (Score: 1)

Lagen innerhalb einer Studierendenantwort bzw. eines KI-Outputs mehrere Deutungen vor, wurde jede Deutung separat eingeschätzt und die Punktwerte gemittelt. Alle Deutungshypothesen wurden durch den dritten Autor kodiert, zusätzlich wurden ca. 30 % der Daten durch die beiden anderen Autoren geprüft. Unstimmigkeiten bei der Einschätzung wie auch Zweifelsfälle wurden in der Autorengruppe diskutiert und bei Bedarf in einem iterativen Prozess erneut geratet.

Um die Ratings zu veranschaulichen, stellen wir im Folgenden zwei exemplarische Deutungshypothesen zum Spiel *The Plan* vor, die sich im Grad der Plausibilität und Elaboriertheit unterscheiden.

Deutungshypothese	Kommentiertes Rating
Das Spiel zeigt mir, dass egal, wie ungünstig die persönlichen Voraussetzungen sind, man alles schaffen kann. (S12)	**Plausibilität: 1 | Elaboriertheit: 1** Die Antwort ist nicht plausibel, da im Spiel weder unterschiedliche Voraussetzungen dargestellt werden noch die Antwort zum Ende des Spiels passt (etwas „schaffen" können, passt nicht zum Verglühen der Fliege; Stufe 1). Zudem wird die Deutungshypothese nicht weiter begründet oder an konkreten Spielelementen festgemacht (Stufe 1).
Das Spiel könnte aussagen, dass alle Menschen im Leben beim gleichen Ausgangspunkt starten (Geburt) und am gleichen Endpunkt ankommen (Tod). Im Verlauf des Lebens haben alle die gleichen Handlungsoptionen (hoch, runter, links, rechts), mit denen sie die gleichen Hindernisse (Sturm, Spinnennetz, fallende Blätter) umgehen. (S16)	**Plausibilität: 2 | Elaboriertheit: 3** Diese deterministische Deutung mag vom Zielpunkt her einleuchten (Tod als unvermeidbarer Endpunkt des Lebens), jedoch geht das Spiel an keiner Stelle auf „gleiche[] Ausgangspunkt[e]" ein oder führt diese vor. Insofern ist die Deutung nur teilweise plausibel (Stufe 2). Die Elaboriertheit ist hingegen als hoch zu bewerten, da konkrete Spielelemente und Gestaltungsaspekte zur Begründung herangezogen werden (Stufe 3).

Tab. 2: Beispielratings

3.5.2 KODIERUNG DER UMGANGSWEISEN

Um mit Blick auf Forschungsfrage 3 weiterführend zu kodieren, wie die Studierenden mit den KI-generierten Deutungshypothesen umgegangen sind, erfolgte eine deduktive Kodierung der Daten. Diese orientierte sich an den bei Magirius / Scherf (2023) gefundenen Umgangsweisen, legt diese jedoch in vier Dimensionen aus (Addition, Modifikation, Affirmation und Rejektion), wie Tab. 3 zeigt.

Umgangsweise	produktiv	problematisch
(1) Addition	Studierende fügen ihrer eigenen Deutung einen *zusätzlichen* Aspekt hinzu, der zuvor übersehen wurde und der plausibel ist.	Studierende fügen ihrer eigenen Deutung einen *zusätzlichen* Aspekt hinzu, der unplausibel ist.
(2) Modifikation	Studierende verändern ihre Deutung aufgrund der KI-Deutung oder ersetzen diese vollständig. Dies führt zu einer *Zunahme* der Plausibilität oder Elaboriertheit.	Studierende verändern ihre Deutung aufgrund der KI-Deutung oder ersetzen diese vollständig. Dies führt zu einer *Abnahme* der Plausibilität oder Elaboriertheit.
(3) Affirmation	Studierende fühlen sich durch die KI-Deutung zurecht in ihrer eigenen Deutung bestätigt (und verändern zurecht nichts an ihrer Deutung).	Studierende fühlen sich durch die KI-Deutung zu Unrecht in ihrer eigenen Deutung bestätigt oder bestätigen ihre Deutung bloß selektiv (und verändern nichts an ihrer Deutung, obwohl sie es hätten tun sollen).
(4) Rejektion	Die Studierenden weisen unplausible oder wenig(er) elaborierte KI-Deutungen (zurecht) zurück.	Die Studierenden weisen eine KI-Deutung zurück, obwohl diese einen höheren Grad der Plausibilität oder Elaboriertheit aufweist als ihre eigene.

Tab. 3: Umgangsweisen mit KI

Die vier Umgangsweisen wurden analog zu den Ratings durch den dritten Autor vollständig kodiert, 30% der Daten wurden zudem durch die beiden anderen Autoren unabhängig kodiert. Zweifelsfälle oder Unstimmigkeiten wurden daraufhin konsensuell geklärt.

3.5.3 STATISTISCHE AUSWERTUNG

Für den Vergleich der beiden Studierendengruppen mit unterschiedlichem Studienfortschritt (Forschungsfrage 1) wurden Kruskal-Wallis-Tests für unabhängige, nicht normalverteilte Stichproben mit Bonferroni-korrigierten paarweisen post-hoc-Tests gerechnet. Um zu prüfen, ob die fortgeschrittene Studierendengruppe in Abhängigkeit vom Spiel unterschiedlich abschnitt (Forschungsfrage 2), wurden Wilcoxon-Rang-Vorzeichen-Tests für abhängige, nicht normalverteilte Stichproben berechnet. Die Kodierungen der Umgangsweisen mit KI-Deutung (Forschungsfrage 3) wurden deskriptiv-statistisch ausgewertet. Alle Auswertungen erfolgten mit SPSS (IBM Corp 2023).

4 — ERGEBNISSE

Forschungsfrage 1: Unterscheiden sich Plausibilität und Elaboriertheit der Interpretationen in Abhängigkeit vom Studienfortschritt? Unterscheiden sich diesbezüglich menschliche und KI-generierte Deutungen?

Vergleicht man die Deutungen von Studierenden zu Beginn (Universität Wuppertal) und gegen Ende ihres Studiums (Universität Jena), zeigt sich in den Mittelwerten zwar ein leichter Vorteil der fortgeschrittenen Studierenden gegenüber den Studierenden am Studienbeginn (Abb. 4). Dieser Vorteil wird aber weder im Bereich der Plausibilität ($z = 1.33$, $p = 1.0$) noch im Bereich der Elaboriertheit ($z = 1.37$, $p = 1.0$) im Kruskal-Wallis-Test signifikant. Anders ist dies beim Vergleich der von den Studierenden und von der KI generierten Deutungen: Hier wird deutlich, dass die KI in beiden Studienabschnitten überlegen ist, also sowohl die fortgeschrittenen Studierenden signifikant übertrifft ($z = -3.41$, $p = .004$, $r = .57$) als auch die weniger fortgeschrittenen Studierenden ($z = -2.87$, $p = .024$, $r = .51$). Auffällig ist schließlich, dass die KI-Deutungen der fortgeschrittenen Studierenden im Rating besser bewertet werden als die der Studienanfänger, wobei der Unterschied im Falle der Elaboriertheit sogar signifikant ist ($z = 2.692$, $p = .043$, $r = 0.63$).

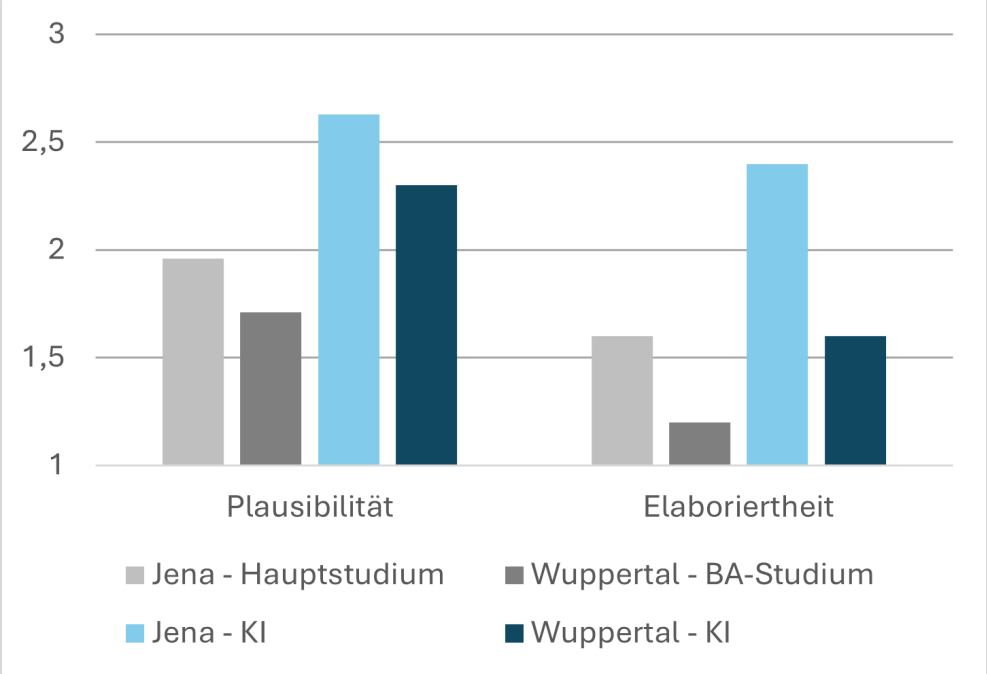

Abb. 4: Ratings der Deutungen in Abhängigkeit vom Studienfortschritt

Forschungsfrage 2: Unterscheiden sich Plausibilität und Elaboriertheit der Interpretationen in Abhängigkeit vom untersuchten Spiel? Unterscheiden sich diesbezüglich menschliche und KI-generierte Deutungen?

Um herauszufinden, ob die Qualität der studentischen und der KI-Deutungen in Abhängigkeit vom Spiel variiert, wurde in der fortgeschrittenen Gruppe mit *The Plan* ein weiteres Spiel gespielt und mit KI interpretiert. Wilcoxon-Tests für abhängige Stichproben (Tab. 4) können hier verdeutlichen, dass sich die Qualität der Deutungen von Studierenden signifikant und mit mittlerer Effektstärke unterscheiden, insofern die Auseinandersetzung mit *The Plan* offensichtlich deutlich schwerer fiel als mit *Every day the same dream* – ungeachtet der Tatsache, dass die Studierenden in beiden Spielen Medianwerte unter dem theoretischen Skalenmittelwert (2.0) erreichten. Bei der KI ist dies anders: Sie erzielte in beiden Spielen sowohl mit Blick auf die Plausibilität als auch die Elaboriertheit konstant hohe Werte, die sich nicht signifikant unterschieden.

	Mdn TP	**Mdn E**	**Mdn Diff**	**Z**	p	R
Plausibilität: Studierende	1.25	2.00	.75	2.287	.02	0.54
Elaboriertheit: Studierende	1.00	1.00	0.00	1.963	.05[3]	0.46
Plausibilität: KI	2.33	2.66	.33	1.149	.251	–
Elaboriertheit: KI	2.50	2.37	.23	-0.45	.964	–

Tab. 4: Wilcoxon-Tests für den Vergleich von The Plan (TP) und Every day the same dream (E)

Forschungsfrage 3: Wie gehen Studierende mit den Deutungen der KI um?

Insbesondere mit Blick auf die Studie von Magirius / Scherf (2023) war abschließend zu fragen, wie die Studierenden KI genutzt haben. Bezogen auf die von uns differenzierten Nutzungsarten ist in Abb. 5 zunächst auffällig, dass das Hinzufügen einzelner Deutungen (Addition) sowie die Ablehnung von Deutungen (Rejektion) die insgesamt am häufigsten kodierten Fälle darstellen. Weniger frequent waren dagegen Modifikationen und Affirmationen. Lediglich die BA-Studierenden neigten häufiger zu Modifikationen, während die Studierenden im Hauptstudium häufiger als die BA-Studierenden zu Affirmationen tendierten.

3 *Der Wilcoxon-Test kann trotz identischer Medianwerte signifikante Unterschiede hervorbringen, da sich der Test nicht direkt auf Unterschiede der Mediane, sondern auf systematische Unterschiede in den gepaarten Rängen der Differenzen (bzw. der Werte zwischen zwei Gruppen) bezieht.*

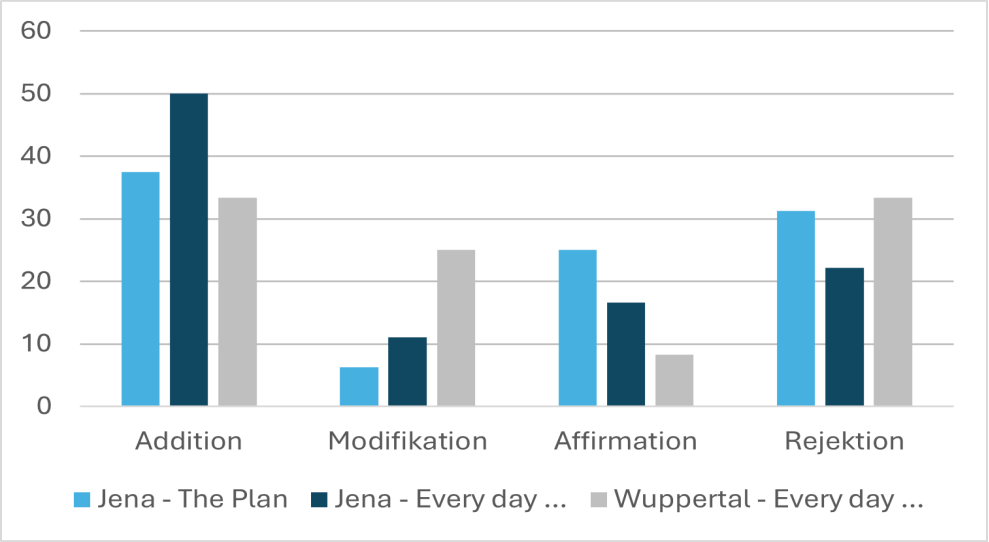

Abb. 5: Nutzungsweisen von KI

Fragt man darüber hinaus nach dem Anteil produktiver versus problematischer Nutzungsweisen – hier aus Gründen der Übersichtlichkeit für alle Kategorien übergreifend dargestellt (Abb. 6) –, zeigt sich ein klarer Trend, demzufolge die problematischen Nutzungsweisen die produktiven deutlich übertreffen. Auffällig ist auch, dass die fortgeschrittenen Studierenden tendenziell produktiver mit der KI umgehen (42.86 % bzw. 33.33 %) als die Studienanfänger:innen (25.00 %), wenngleich auch hier gegenstandsspezifische Abweichungen zu bestehen scheinen. So ist der Unterschied zwischen *Every day the same dream zwischen* den beiden Kohorten fast ebenso groß wie der Unterschied zwischen *Every day the same dream* und *The Plan* innerhalb der Kohorte der fortgeschrittenen Studierenden.

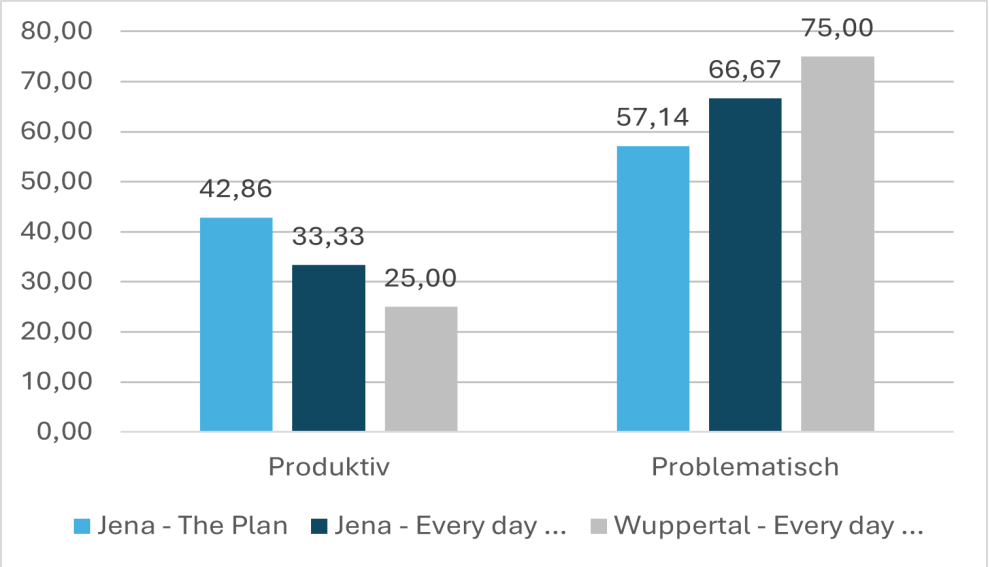

Abb. 6: Produktive vs. problematische Nutzungsweisen

5 — DISKUSSION

Die vorliegende Studie ging von der empirischen Beobachtung aus, dass Lehramtsstudierende Probleme mit der Interpretation literarischer Texte haben. KI wird in diesem Zusammenhang konzeptionell als Unterstützungsmöglichkeit diskutiert (vgl. Magirius et al. 2024; Magirius / Scherf 2023), wenngleich empirische Studien zum tatsächlichen Einsatz zu diesem Zweck bislang noch rar sind. Der Beitrag setzte an diesem Desiderat mit einer explorativen Untersuchung an, die aufgrund ihrer kleinen Stichprobengröße, dem fehlenden echten Längsschnitt sowie ihrer Standortgebundenheit in ihrem Aussagewert notwendigerweise begrenzt und zukünftig durch umfangreichere Studien zu ergänzen ist.

Die Studie liefert Ergebnisse, die sich gut in den bisherigen Diskurs einfügen und mithin eine hohe Konvergenzvalidität aufweisen (vgl. Carl / Scherf 2022, S. 4). So wird in Übereinstimmung mit anderen Untersuchungen deutlich, dass sich Studierende des Lehramts Deutsch damit schwertun, literarisch anspruchsvolle Gegenstände zu verstehen. Hat die bislang vorliegende Forschung dies überwiegend für schriftlich fixierte epische und lyrische Texte zeigen können, liefert die hier dargestellte Untersuchung Evidenz dafür, dass auch Computerspiele hohe Verstehensanforderungen bereithalten können: Ungeachtet des Studienfortschritts und des betrachteten Spiels erreichen die Studierenden Durchschnittswerte (weit) unterhalb des theoretischen Skalenmittelwerts. Dieses Ergebnis ist auch deshalb bemerkenswert, weil für den Arbeitsauftrag (als Hausaufgabe) keine Zeitbeschränkung gesetzt wurde und mehrfaches Spielen möglich war. Eventuell hat aber gerade das Heimarbeitssetting dazu verleitet, die Aufgabe nur oberflächlich zu lesen und nicht in einen vertieften Verstehensprozess einzutauchen. Ebenso ist denkbar, dass einige Studierende mit Computerspielen nicht vertraut waren (hierzu liegen leider keine Daten vor), was eine tiefgründige Einlassung verhindert haben könnte. Zugleich darf dieser mögliche ‚Mediumseffekt' aber insofern nicht überbewertet werden, als andere Studien vergleichbare Verstehensschwierigkeiten auch für anspruchsvolle literarische Texte i. e. S. zeigen konnten (s. Abschn. 2.2).

Sichtet man die Ergebnisse der vorliegenden Studie entlang der drei Forschungsfragen, konnte hinsichtlich **Forschungsfrage 1** gezeigt werden, dass Studierende im fortgeschrittenen Studium zwar leicht, aber nicht signifikant besser abschneiden als Studierende im BA-Studium. Auch dieses Ergebnis ist nicht gänzlich neu, insofern bereits die FALKO-Befunde zeigen, dass sich erfahrene Lehrkräfte bzgl. der Analyse des Potenzials von Texten nicht signifikant von Studierenden unterscheiden (vgl. Pissarek / Schilcher 2017). Ebenso deuten die Ergebnisse von Winkler / Wieser (2025, S. 18) mit Blick auf erzählanalytisches Wissen darauf hin, dass fortgeschrittene Studierende im Umgang mit Texten zwar im Studium erworbene Kategorien deklarativ benennen, diese aber oft nicht für die Textanalyse fruchtbar machen können. Die in der Literaturwissenschaft und -didaktik traditionsreiche und zugleich schwierige Frage, inwiefern das Interpretieren von Texten gezielt (im Studium) lehr- und lernbar ist (vgl. Carl 2022), erfährt vor diesem Hintergrund neue Relevanz – gerade auch mit Blick auf literarische Texte i. w. S. Gewinnbringend schiene es hier, an die aktuell neu aufkeimende Debatte zur Rolle des Übens anzuknüpfen (vgl. Heins et al., 2022; Heins

/ Wieser, 2025) und hier nicht nur für den schulischen Unterricht, sondern auch für die Lehrkräftebildung ggf. spiralcurricular organisierte und intermedial ausgerichtete Übe-Konzepte zum Verstehen literarischer Gegenstände zu entwickeln, zu implementieren und zu evaluieren.

Ein weiterer auffälliger Nebenbefund besteht darin, dass sich zwar die studentischen Interpretationen nicht signifikant unterscheiden, wohl aber die KI-generierten Interpretationen der fortgeschrittenen Studierenden die KI-generierten Interpretationen der BA-Studierenden übertreffen – in der Dimension Elaboriertheit sogar signifikant. Dieser Befund kann möglicherweise dahingehend gedeutet werden, dass fortgeschrittene Studierende über bessere Promptingstrategien verfügen als BA-Studierende. Zukünftige Studien sollten diesem Befund durch qualitative Analysen von Logfiles weiter nachgehen.

Mit Blick auf **Forschungsfrage 2** zeigt sich, dass sich die Qualität der Deutungen in Abhängigkeit vom untersuchten Gegenstand bei den Studierenden signifikant unterscheidet. So haben die Studierenden zum Spiel *Every day the same dream* signifikant plausiblere und elaboriertere Deutungen verfasst als zum Spiel *The Plan*. Dies lässt sich möglicherweise über die unterschiedlichen Verstehensanforderungen beider Spiele begründen. Während bei *Every day the same dream* die Kapitalismuskritik durch das gesamte Setting des Spiels vielfältig angedeutet wird (sich wiederholende Arbeitstage, Berufsverkehr auf der Autobahn, Großraumbüros), ist bei *The Plan* eine größere Transferleistung notwendig. Hier wird in hohem Maße die Fähigkeit gefordert, komplexe und teils metaphysische Fragen, bspw. nach dem Sinn des Lebens, in einer einfachen Symbolik (Fliege, Spinnennetz, Glühbirne etc.) wiederzuerkennen. Hilfreich könnte somit sein, universitäre Lehrangebote gezielt auf Herausforderungen beim Verstehen solcher Merkmale literarischer Texte zu lenken und lehrpersonenseitige Umgangsweisen mit diesen gezielt in den Blick zu nehmen.

KI-Anwendungen werden bereits vor der öffentlichen Zugänglichmachung von *ChatGPT* als ein vielversprechendes Unterstützungsangebot für den Umgang mit literarischen Texten diskutiert (vgl. Balyan / McCarthy / McNamara 2017). Dies liegt zum einen daran, dass sie anders als traditionelle Strategieprogramme zum Textverstehen adaptiv am konkreten Text bzw. Spiel ansetzen. Andererseits kann die KI als Dialogpartner fungieren und ermöglicht es, konkrete Rückfragen zu stellen, wenn einzelne Aspekte unklar sind. Über das Rating der KI-generierten Deutungen in der vorliegenden Studie wurde ersichtlich, dass die KI für beide Spiele zuverlässig gute und dabei signifikant plausiblere und elaboriertere Deutungen liefert als die Studierenden – und das, obwohl die Studierenden vorher nicht im Prompten geschult wurden und die Computerspiele der KI nur als Name bzw. Spieltitel eingespeist wurden. Entsprechend ist davon auszugehen, dass die KI durchaus brauchbare Interpretationsangebote bereithält.

Mit Blick auf die in **Forschungsfrage 3** untersuchten Umgangsweisen der Studierenden mit KI wird allerdings deutlich, dass dieses Potenzial nur partiell abgerufen wird: Die Mehrheit der Studierenden nutzt die KI, um die eigene Deutung um zusätzliche Aspekte zu erweitern oder um die von ChatGPT angebotenen Deutungsvorschläge

als falsch oder nicht weiterführend zurückzuweisen. Weniger frequent sind indes Nutzungsweisen, die eine Modifikation der eigenen Deutung beinhalten (z.B. durch gezielte Überarbeitung einzelner Aussagen). In Kombination mit der Beobachtung, dass die problematischen Nutzungsweisen (z.B. Ablehnung einer richtigen Deutungshypothese, selektive Bestätigung der eigenen Deutung) die produktiven deutlich überwiegen, liegt der Verdacht nahe, dass die Studierenden den Output der KI insgesamt nur oberflächlich mit ihren eigenen Annahmen in Beziehung setzen.

Für dieses Ergebnis kommen verschiedene Erklärungen in Betracht. Denkbar wären beispielsweise motivationale Gründe: Das Modifizieren der eigenen Deutung ist zumeist kognitiv aufwendiger als das Hinzufügen oder Ablehnen von Deutungsvorschlägen, sodass diese ökonomische(re)n Varianten präferiert werden. Auch könnten defizitorientierte *Beliefs* zu KI bereits im Vorfeld eine tiefgründigere Einlassung verhindern. Zudem erfordert die Arbeit mit dem KI-Output Fähigkeiten, die bereits im Kontext des Lesens multipler Dokumente (vgl. Philipp 2020) oder des (argumentierenden) materialgestützten Schreibens (vgl. Feilke et al. 2016) als besonders anforderungsreich beschrieben wurden. Die Studierenden müssen nämlich nicht mehr nur den literarischen Text verstehen, sondern sich auch noch mit einer oder gar mehreren Deutungshypothesen der KI auseinandersetzen, diese dabei auf ihre Stimmigkeit prüfen und mit der eigenen Deutungshypothese abgleichen. Da dieser Prozess voraussetzt, dass die Studierenden bereits über ein elaboriertes Text-Welt-Modell des literarischen Textes verfügen, dürften – Magirius / Scherf (2023) folgend – vor allem diejenigen Studierenden von KI profitieren, die ohnehin über solide Fähigkeiten im Interpretieren verfügen. Studierende, die mit weniger ausgeprägten Fähigkeiten im Interpretieren in den Prozess starten, werden hingegen von der KI lediglich mit einem ‚weiteren' Text konfrontiert, der zusätzliche Verstehensanforderungen bereithält (Matthäus-Effekt, vgl. Steinmetz 2020).

Mit Blick auf die zukünftige Forschung scheint es vor diesem Hintergrund relevant, in Übereinstimmung zur Forderung von Magirius / Scherf (2023) die Zusammenhänge zwischen personenbezogenen Faktoren (Beliefs, Textverstehenskompetenz etc.) und tatsächlicher KI-Nutzung für professionelle Aufgaben anhand größerer und standortübergreifender Stichproben weiter zu erforschen. Hierzu können experimentelle Designs mit gezielter Kontrolle einzelner Faktoren ebenso erhellend sein wie Erhebungen in ökologisch validen Settings. Aus methodischer Sicht sollte dabei darauf geachtet werden, mehrere Texte einzubeziehen, da die Ergebnisse offensichtlich nicht nur aufgrund von lernenden-, sondern auch gegenstandsseitigen Faktoren variieren können.

Für die Lehrkräfteprofessionalisierung indizieren die in der Studie gewonnenen Befunde einmal mehr dringenden Handlungsbedarf, da Literaturunterricht zweifelsohne nur dort gelingen dürfte, wo Lehrpersonen die Texte, die sie unterrichten, auch selbst verstehen. Die Annahme, dass Studierende über den Erwerb des Abiturs die Fähigkeit zum Verstehen anspruchsvoller Texte mitbringen, scheint sich daher als Trugschluss zu erweisen. Dementsprechend ist Winkler / Wieser (2025, S. 19) zuzustimmen, „dass viele Studierende für ihr eigenes Verständnis literarischer Texte wohl oft mehr Support benötigen als gemeinhin angenommen." KI sollte in diesem Zusam-

menhang als *eine* Form des Supports mitgedacht werden, wobei sich angesichts der o. g. Befunde kritisch diskutieren ließe, für wen KI tatsächlich eine Unterstützung (*support*) und für wen nicht eher eine zusätzliche Anforderung (*demand*) darstellt. In jedem Fall scheint KI (bislang) den Erwerb einer (digitalen) Textsouveränität (Frederking 2023) nicht ersetzen zu können. Dass entsprechende Erwerbsprozesse trotz vieler Lehrveranstaltungen im Bereich der Literaturwissenschaft und -didaktik bzw. angesichts noch fehlender Lehrveranstaltungen im Bereich der Medienwissenschaft und -didaktik offenbar nicht (hinreichend) gelingen, sollte als Anregung betrachtet werden, über neue Lehrveranstaltungsformate nachzudenken.

QUELLENVERZEICHNIS

PRIMÄRQUELLEN

— *Every day the same dream.* 2009 (Molleindustria). — *The Plan.* 2013 (Krillbite).

SEKUNDÄRQUELLEN

— **Anselm, Sabine** (2011): *Kompetenzentwicklung in der Deutschlehrerbildung: Modellierung und Diskussion eines fachdidaktischen Analyseverfahrens zur empiriegestützten Wirkungsforschung.* Frankfurt am Main u.a.: Peter Lang. — **Balyan, Renu / McCarthy, Kathryn / McNamara, Danielle** (2017): *Combining Machine Learning and Natural Language Processing to Assess Literary Text Comprehension.* In: Hu, Xiangen / Barnes, Tiffany /Hershkovitz, Arnon / Paquette, Luc (Hg.): Proceedings of the 10th International Conference on Educational Data Mining. 244-249. https://www.researchgate.net/publication/318909202 — **Boelmann, Jan M**. (2024): Lernförderliche Aspekte von Computerspielen im Unterricht. In: Der Deutschunterricht, H. 4 (2023), 2-10. — **Carl, Mark-Oliver** (2022): *Kann man Interpretieren gezielt lernen?: Eine aktuelle Kontroverse mit langer Vorgeschichte.* In: Bernhardt, Sebastian / Hardtke, Thomas (Hg.): Interpretation – *Literaturdidaktische Perspektiven*. Berlin: Frank & Timme, 25-52. https://doi.org/10.57088/978-3-7329-9143-3_2. — **Carl, Mark-Oliver / Scherf, Daniel** (2022): *Über den angemessenen Umgang mit empirischen Erkenntnissen in der Literaturdidaktik.* In: Zeitschrift für Sprachlich-Literarisches Lernen und Deutschdidaktik, H. 2 (2022), 1-25. https://doi.org/10.46586/SLLD.Z.2022.8894. — **Dick, Mirjam** (2024): *Vernetzung statt Addition: Eine Treatmentstudie in der de-fragmentierenden Deutschlehrerbildung am Beispiel Textverstehen und Aufgabenkonstruktion.* Herne: Gabriele Schaefer Verlag. — **Feilke, Helmuth / Lehnen, Katrin / Rezat, Sara / Steinmetz, Michael** (2016): *Materialgestütztes Schreiben lernen: Grundlagen - Aufgaben - Materialien* (Sekundarstufen I und II, Druck A). Braunschweig: Schroedel/Westermann. — **Frederking, Volker** (2023): *Von Fake News bis ChatGPT. Digitale Textsouveränität als ethisch-politische Bildungsaufgabe für Deutschdidaktik und Deutschunterricht in der digitalen Welt.* In: MiDU, H. 2 (2023). https://doi.org/10.18716/OJS/MIDU/2023.2.4. — **Führer, Carolin / Nix, Daniel** (2023): *Anschlusskommunikationen mit ChatGPT: Kann die Interaktion mit Künstlicher Intelligenz (KI) Schülerinnen und Schüler beim Verstehen literarischer Texte unterstützen?* In: leseforum.ch. https://doi.org/10.58098/LFFL/2023/3/805. — **Gabriel, Sonja** (2025): *„ChatGPT kennt halt die Schüler:innen nicht".* In: Medienimpulse, H. 1 (2025), 1-39. https://doi.org/10.21243/MI-01-25-09. — **Heine, Jörg-Henrik / Heinle, Martina / Hahnel, Carolin / Lewalter, Doris / Becker-Mrotzek, Michael** (2023): *Lesekompetenz in PISA 2022: Ergebnisse, Veränderungen und Perspektiven.* In: Lewalter / Doris / Diedrich, Jennifer / Goldhammer, Frank / Köller, Olaf / Reiss, Kristina (Hg.), *PISA 2022: Analyse der Bildungsergebnisse in Deutschland.* Münster/New York: Waxmann, 139-162. https://doi.org/10.31244/9783830998488. — **Heins, Jochen** (2022): *Unterrichtswahrnehmung von Literaturunterricht: Befunde zur Wahrnehmung von Unterrichtssituationen mit unterschiedlichem Anforderungsniveau.* SLLD, Bd. 2, 1-31. https://doi.org/10.46586/SLLD.Z.2022.9649. — **Heins, Jochen / Kleinschmidt-Schinke, Katrin / Wieser, Dorothee / Wiesner, Esther** (Hg.) (2022): *Üben. Theoretische und empirische Perspektiven in der Deutschdidaktik.* SLLD-B, Bd. 5. https://doi.org/10.46586/SLLD.248. — **Heins, Jochen / Wieser, Dorothee** (2025): *Unterstützung des konsolidierenden Übens im Deutschunterricht: Konzeptionelle Überlegungen, empirische Befunde und Forschungsdesiderata.* In: Unterrichtswissenschaft, H. 53 (2025), 205-225. https://doi.org/10.1007/s42010-025-00225-9. — **Hesse, Florian** (2024): *Qualitäten von Literaturunterricht. Eine Videostudie im Praxissemester.* Stuttgart u.a.: Springer / J. B. Metzler. — **Hesse, Florian / Magirius, Marco / Heins, Jochen** (2025): *Literaturdidaktische Lehrer*innenforschung – Ein Scoping Review.* In: SLLD-Z, H. 4 (2025), 1-29. https://doi.org/10.46586/SLLD.Z.2025.11220. — **Hesse, Florian / Seeber, Anna** (2025): *Professionelle Unterrichtswahrnehmung im Praxissemester. Analysen von videobasierten Peer-Feedbacks zu Literaturstunden im Fach Deutsch.* In: Didaktik Deutsch, H. 58 (2025), S. 97-123. — IBM Corp. (2023): *SPSS Statistics for Windows [Software].* Armonk: IBM Corp. — **Kepser, Matthis** (2023): *Gaming. Sprachlich-literarästhetisches Lernen im kulturellen Handlungsfeld digitaler Spiele.* In: Praxis Deutsch 298, 4-13. — **Kepser, Matthis** (2025): *Das ist unser Leben, das ist auch mein Leben. Sprachlich-literarisches Lernen mit dem digitalen Spiel „Every Day the same Dream".* In: MiDU, H. 2 (2025), 1-24. https://doi.org/10.18716/OJS/MIDU/2024.2.6. — KMK (2024): *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung.* Beschluss der Kultusministerkonferenz vom 16.10.2008 i. D. F. vom 08.02.2024. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildung.pdf. — **Krauss, Stefan / Bruckmaier, Georg / Lindl, Alfred / Hilbert, Sven / Birkner, Karin / Steib, Nicole / Blum, Werner** (2020): *Competence as a continuum in the COACTIV study: The "cascade model".* In: ZDM Mathematics Education, 52(2), 311-327. https://doi.org/10.1007/s11858-020-01151-z. — **Lotz, Miriam** (2016): *Kognitive Aktivierung im Leseunterricht der Grundschule: Eine Videostudie zur Gestaltung und Qualität von Leseübungen im ersten Schuljahr.* Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-10436-8. — **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, H. 5, 14-23. — **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. — **Masanek, Nicole / Doll, Jörg** (2020): *Vernetzung ja, aber ohne Fachwissenschaft? - Zur Nutzung fachlichen, fachdidaktischen und pädagogischen Wissens durch Lehramtsstudierende im Bachelorstudium.* In: Didaktik Deutsch, 25(48), 36-54. — **Neumann, Anne / Steensen, Gönke C**. (in Vorb.): *Zur Rolle Künstlicher Intelligenz in der Unterrichtsplanung von Deutschlehrkräften in Zeiten des Lehrkräftemangels.* In: Hesse, Florian (Hg.): Deutschlehrkräftebildung in Zeiten des Lehrkräftemangels. Systemische, empirische und konzeptionelle Perspektiven. — Philipp, Maik (2020): *Multiple Dokumente verstehen: Theoretische und empirische Perspektiven auf Prozesse und Produkte des Lesens mehrerer Dokumente.* Weinheim / Basel: Beltz Juventa. — **Pissarek, Markus / Schilcher, Anita** (2017): *FALKO-D: Die Untersuchung des Professionswissens von Deutschlehrenden: Entwicklung eines Messinstruments zur fachspezifischen Lehrerkompetenz und Ergebnisse zu dessen Validierung.* In: Krauss, Stefan / Lindl, Alfred / Schilcher, Anita / Fricke, Michael / Göhring, Anja / Hofmann, Bernhard / Kirchhoff, Petra / Mulder, Regina (Hg.): *FALKO: Fachspezifische Lehrerkompetenzen: Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik*. Münster / New York: Waxmann, S. 68-109. — **Schilcher, Anita / Glondys, Manuel / Wild, Johannes** (2023): *Leseunterricht in den Grundschulen in Deutschland.* In: McElvany, Nele / Lorenz, Ramona / Frey, Andreas / Goldhammer, Frank / Schilcher, Anita / Stubbe, Tobias (Hg.): IGLU 2021. Lesekompetenz von Grundschulkindern im internationalen Vergleich und im Trend über 20 Jahre. Münster / New York: Waxmann, 179-196. https://doi.org/10.31244/9783830997009. — **Schöffmann, Andreas / Hoiß, Christian** (2023): *Von Fliegen und Menschen. Macht und Ohnmacht im Computerspiel The Plan.* In: Praxis Deutsch, 50(298), 47-55. — **Steinmetz, Michael** (2020): *Verstehenssupport im Literaturunterricht: Theoretische und empirische Fundierung einer literaturdidaktischen Aufgabenorientierung.* Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-28378-0. — **Susteck, Sebastian / Perder, Christoph** (2023): Schreiben durch Künstliche Intelligenz: ChatGPT und automatisierte Lyrikanalysen. MiDU, H. 2 (2023), 1-20. https://doi.org/10.18716/OJS/MIDU/2023.0.2. — **Treß, Johannes** (2024): *Argumentierte Unterrichtsplanung auf dem Prüfstand. Eine explorative Studie mit Musiklehrkräften zur KI-gestützten Unterrichtsplanung.* In: Clausen, Bernd / Ehninger, Julia / Sachsse, Malte (Hg.): 45. Jahresband des Arbeitskreises Musikpädagogische Forschung. Münster / New York: Waxmann, 121-133. https://10.31244/9783830999188.10 — **Tengberg, Michael / van Bommel, Jorryt / Nilsberth, Marie / Walkert, Michael / Nissen, Anna** (2021): *The Quality of Instruction in Swedish Lower Secondary Language Arts and Mathematics.* In: Scandinavian Journal of Educational Research, 66(5) (2021), 760-777. https://doi.org/10.1080/00313831.2021.1910564. — **Winkler, Iris** (2017): *Potenzial zu kognitiver Aktivierung im Literaturunterricht: Fachspezifische Profilierung eines prominenten Konstrukts der Unterrichtsforschung.* In: Didaktik Deutsch, 22(43), 78-97. https://doi.org/10.25656/01:16157. — **Winkler, Iris** (2020): *Cognitive Activation in L1 Literature Classes: A content-specific framework for the description of teaching quality.* In: L1 – Educational Studies in Language and Literature, H. 20 (2020), 1-32. https://doi.org/10.17239/L1ESLL-2020.20.01.03. — **Winkler, Iris / Seeber, Anna** (2020): *Facetten literaturdidaktischer Kompetenz bei Deutschstudierenden vor und nach dem Praxissemester. Eine Interventionsstudie zur Wirksamkeit videobasierter Lernbegleitung.* In: Didaktik Deutsch, 25(49), 23-49. https://doi.org/10.25656/01:22286. — Winkler, Iris / Wieser, Dorothee (2025): *Fähigkeiten zur Erzähltextanalyse in Abhängigkeit vom Aufgabensupport: Eine Studie mit Lehramtsstudierenden des Faches Deutsch*. In: Zeitschrift für Sprachlich-Literarisches Lernen und Deutschdidaktik, H. 5 (2025), 1-23. https://doi.org/10.46586/SLLD.Z.2025.12112.

ÜBER DIE AUTOREN

Dr. Florian Hesse ist wissenschaftlicher Mitarbeiter (Postdoc) an der Friedrich-Schiller-Universität Jena. Seine Forschungsschwerpunkte liegen im Bereich der Unterrichts- und Professionsforschung sowie dem Einsatz von KI in Deutschunterricht und Lehrkräfteprofessionalisierung.

Dr. Gerrit Helm ist wissenschaftlicher Mitarbeiter (Postdoc) an der Friedrich-Schiller-Universität Jena. Seine Forschungsschwerpunkte liegen im hierarchieniedrigen Lesen und Schreiben sowie im Einsatz vom KI im Deutschunterricht.

Maximilian Arend ist wissenschaftlicher Mitarbeiter an der Friedrich-Schiller-Universität Jena. Sein Forschungsschwerpunkt liegt in der literaturdidaktischen Unterrichts- und Professionsforschung.