

DER DIGITALE FEDERKIEL STUDIERENDENSEITIGE NUTZUNG EINES TUTORIELLEN LLMS UND DIE NUTZUNGS- PROGRESSION IM VERLAUF EINES SEMESTERS. EINE QUALITATIV-QUANTITATIVE STUDIE

Sebastian Bernhardt

Universität Münster | sebastian.bernhardt@uni-muenster.de

Friedrich Bach

Universität Münster | friedrich.bach@uni-muenster.de

Ann-Kathrin Koppenhöfer

Universität Münster | akoppenh@uni-muenster.de

Ninja Schmiedgen

Universität Münster | ninja.schmiedgen@uni-muenster.de

ABSTRACT

Dieser Beitrag untersucht, wie Studierende in fünf germanistischen literatur- und mediendidaktischen Seminaren ein tutorielles Large Language Model (*Tutor.AI*) über den Verlauf eines Semesters hinweg nutzten. Auf der Grundlage von 2.277 protokollierten Prompts von 129 Studierenden wurden Nutzungskategorien entwickelt und in Beziehung zu den seminarplanerischen Phasen gesetzt. Erstens zeigen die Ergebnisse eine grobe Orientierung der Nutzung an der Seminararchitektur. Zweitens steigt die Nutzungsintensität in Test- und Prüfungswochen signifikant an und lässt ein konsistentes prüfungsorientiertes Nutzungsmuster erkennen. Drittens weisen Studierende mit hoher individueller Nutzungsintensität eine stärkere Progression von begriffsbezogenen Anfragen hin zu analytischen und anwendungsbezogenen Fragestellungen auf. Die Studie liefert empirische Hinweise darauf, wie tutoriellen LLM-Systemen im deutschdidaktischen Kontext begegnet wird und welche Potenziale und Grenzen sich im Hinblick auf Verständnisunterstützung und selbstreguliertes Lernen abzeichnen.

SCHLAGWÖRTER

— LLM — AI — RAG — LITERARISCHES LERNEN — LEHRARCHITEKTONIK —
EVIDENZGESTÜTZTE AUSWERTUNG — LANGFRISTAUSWERTUNG — LITERATUR- UND
MEDIENDIDAKTIK

ABSTRACT (ENGLISH)

THE DIGITAL QUILL STUDENT-SIDE USE OF A TUTORIAL LLM AND THE USAGE PROGRESSION OVER THE COURSE OF A SEMESTER. A QUALITATIVE-QUANTITATIVE STUDY

This article examines how students in five German literature and media didactics seminars used a tutorial large language model (*Tutor.AI*) over the course of a semester. Based on 2.277 logged prompts from 129 students, categories of use were developed and related to the phases of seminar planning. First, the results show a rough orientation of use towards the seminar architecture. Second, the intensity of use increases significantly during test and exam weeks, revealing a consistent exam-oriented usage pattern. Thirdly, students with high individual intensity of use show a stronger progression from concept-related queries to analytical and application-related questions. The study provides empirical evidence on how tutorial LLM systems are encountered in the context of German language teaching and what potential and limitations are emerging regarding comprehension support and self-regulated learning.

KEYWORDS

— LLM — AI — RAG — LITERARY LEARNING — LEHRARCHITEKTONIK —
EVIDENZGESTÜTZTE AUSWERTUNG — LANGFRISTAUSWERTUNG — LITERATUR- UND
MEDIENDIDAKTIK

1 — EINLEITUNG

Der Einsatz von Large Language Models (LLMs) gewinnt in der Hochschullehre zunehmend an Bedeutung. Während institutionelle und hochschuldidaktische Debatten häufig die Potenziale generativer AI für Wissensarbeit, Textproduktion oder formative Unterstützung adressieren, ist bislang nur wenig darüber bekannt, wie Studierende LLM-basierte tutorielle Systeme tatsächlich in Lehrveranstaltungen nutzen (vgl. Bond et al. 2024, 2). Viele Annahmen über Chancen und Risiken beruhen auf theoretischen Überlegungen oder kontrollierten Laborszenarien; empirische Einsichten in Nutzungsmuster unter authentischen Studienbedingungen liegen dagegen kaum vor (vgl. Zawacki-Richter et al. 2019; Bond et al. 2024). Die Frage, wie Studierende ein AI-basiertes tutorielles System über die Dauer eines Kurses hinweg nutzen – etwa entlang der Seminarstruktur, in prüfungsnahen Phasen oder in Abhängigkeit von individuellen Lernstrategien –, ist bislang weder in der deutschdidaktischen Forschung noch in der AI-bezogenen Hochschulforschung systematisch empirisch untersucht worden (vgl. Zawacki-Richter et al. 2019; Bond et al. 2024).

Die vorliegende Studie adressiert diese Forschungslücke, indem sie 2.277 User:innenprompts von 129 Studierenden aus fünf Lehrveranstaltungen analysiert und die Nutzung in Bezug auf (a) Seminarphasen, (b) Prüfungsnähe und (c) individuelle Lernverläufe untersucht. Ziel ist es, zu verstehen, welche Muster sich über das Semester hinweg herausbilden, welche Faktoren die Nutzung bestimmen und inwiefern sich lernlogische Progressionen im Umgang mit dem tutoriellen LLM erkennen lassen.

Diese Ausgangslage wirft auch zentrale deutschdidaktische Fragen auf:

— Welche Funktion übernehmen tutorielle AI-Systeme im Prozess literarästhetischen Lernens? Unterstützen sie vor allem die Festigung fachlicher Grundlagen oder werden sie auch für vertiefende analytische und didaktische Anwendungen genutzt?

— Die Studie knüpft damit sowohl an aktuelle Arbeiten zu AI-basierten Lernsettings als auch an grundsätzliche Überlegungen der germanistischen Literatur- und Mediendidaktik an und verortet sich an der Schnittstelle von Lernprozessforschung, digitaler Bildung und hochschuldidaktischer Seminarentwicklung.

2 — FORSCHUNGSSTAND

Schulen und Universitäten stehen vor der Frage, welche Konsequenzen die mit der fortschreitenden Etablierung von LLMs und AI in den Alltag einhergehende Automatisierung für Vermittlungssituationen haben wird (vgl. Simbeck 2025, 21–23). Steinhoff und Lehnen arbeiten heraus, dass in den von ihnen beobachteten typischen Nutzungsszenarien agentieller LLMs oftmals ein Chat-to-Create oder Chat-to-Generate-Verhalten an den Tag gelegt wird (vgl. Steinhoff / Lehnen 2025, 4): Die Nutzer:innen greifen auf das LLM zur Ideenfindung und -strukturierung oder für die Produktion von Textbausteinen zurück. Erste Erkenntnisse aus der Kognitionsforschung legen nahe, dass die Nutzung von LLMs im Sinne eines Chat-to-Generate oder Chat-to-Create der kognitiven Verarbeitungstiefe abträglich ist (vgl. Kosmyrna et al. 2025). Diese Erkenntnis empfiehlt, dass ein anderes Nutzungsszenario zu favorisieren ist,

nämlich ein tutorielles Lernbegleitungssystem, das als dialogischer Verstehenssupport angelegt und nicht vornehmlich als Hilfe zur Generierung von Textbausteinen oder zur Entlastung des Rückmeldeaufwandes durch die Lehrenden konzipiert ist. Deziert geht es auch nicht um ein Rückmeldesystem zu eingegebenen Texten, sondern um eine Chat-Situation, die für Rückfragen, zur Hilfe bei Verständnisschwierigkeiten und zur Entwicklung von Ideen in Bezug auf Analysen und didaktische Perspektiven dient, weniger im Sinne der Hilfe zur Erstellung eines Texts als eher zur Begleitung bei der Vorbereitung auf Seminarsituationen und Seminargespräche. Ziel ist das Ausloten von Möglichkeiten, den Studierenden einen Verstehens- und Lernsupport zu bieten.

Ein entsprechend auf Verständnisunterstützung ausgelegtes agentielles System zu implementieren, stellt sich als eine Herausforderung sowohl organisatorischer- als auch technischerseits dar; zudem lassen sich die tatsächlichen Nutzungsszenarien schwerlich evaluieren. Corvacho del Toro, Fultrott und Steinhoff (2025) entwickeln beispielsweise für ihr Projekt KIMADU (Künstliche Intelligenz im Mathematik- und Deutschunterricht) eine Anwendungsumgebung für ein solches agentielles System, das mithilfe eines Megaprompts die Kommunikation mit dem System filtert und zudem weitere datenschutzfilternde Einstellungen vornimmt. Plattformen wie *fobizz* nutzen derartige Möglichkeiten in Bildungskontexten (vgl. ebd., 71). Bei diesem Use-Case bleibt es allerdings dabei, dass das LLM selbst im Sinne Bajohrs einer Black-Box gleichkommt (vgl. Bajohr 2022, 153), deren genaue Funktionsweise nicht steuerbar ist.

3 — FRAGESTELLUNG UND HYPOTHESEN

Die zunehmende Integration textgenerierender AI in universitäre Lehr-Lernkontexte eröffnet neue Möglichkeiten der Lernunterstützung, gleichzeitig liegen bislang jedoch nur begrenzte empirische Erkenntnisse zum tatsächlichen Nutzungsverhalten von Studierenden vor (vgl. Bond et al. 2024). Da belastbare Daten zu Prompting-Mustern, Lernprogressionen und Nutzungsweisen im Kontext der literatur- und medien-didaktischen Hochschullehre bisher fehlen, verfolgt die vorliegende Studie das Ziel, die alltagsnahe Nutzung eines tutoriellen LLMs durch Lehramtsstudierende systematisch zu untersuchen. Die Datengrundlage bilden 2.277 User:innen-Prompts von 129 Studierenden aus fünf literatur- und mediendidaktischen Seminaren des Sommersemesters 2024.

Ziel ist es, Nutzungsmuster im Semesterverlauf sichtbar zu machen, diese an die Seminararchitektur rückzubinden und erste Hypothesen über typische Formen der Nutzung sowie deren Veränderung im Lernprozess abzuleiten.

Ausgehend von dieser Zielsetzung leiten sich folgende Forschungsfragen ab:

- a. Nutzung im Semesterverlauf: Wie verteilt sich die Nutzung des tutoriellen LLMs über die Vorlesungszeit hinweg und inwiefern spiegeln sich die seminarinternen Phasen (Theorie – Reaktivierung – Anwendung) in den Nutzungsmustern wider?

- b. Einfluss von Leistungsanforderungen: In welchem Verhältnis stehen Nutzungsspitzen zur zeitlichen Nähe von Tests, Leistungsnachweisen und Prüfungsanforderungen?
- c. Individuelle Nutzungsintensität: Wie verändert sich das Prompting-Verhalten in Abhängigkeit von der Anzahl der Nachrichten pro Person?

Die Studie ist primär deskriptiv angelegt und zielt auf eine systematische Beschreibung von Nutzungsmustern eines tutoriellen LLMs im Seminarverlauf. Auf Grundlage der beobachteten Regelmäßigkeiten im Datenmaterial werden Annahmen formuliert, die im Rahmen der vorliegenden Daten empirisch weiter plausibilisiert werden können. Daraus ergeben sich die folgenden Erwartungshypothesen:

EH1 – Phasenorientierte Nutzung:

Es wird erwartet, dass die Nutzung der Studierenden sich an den durch die Seminarplanung vorgegebenen Phasen orientiert.

EH2 – Prüfungsnähe-Effekt:

Zudem wird erwartet, dass die Gesamtzahl der Prompts in den Wochen vor Tests und Leistungsnachweisen ansteigt.

EH3 – Nutzungsintensität und lernlogische Progression:

Schließlich wird erwartet, dass Studierende mit hoher individueller Nutzungsintensität eine stärkere Progression zu Analyse- und Anwendungsfragen zeigen als Studierende mit geringer Nutzung.

4 — FORSCHUNGSDESIGN UND DATENERHEBUNG

4.1 SEMINARKONTEXT UND STRUKTUR DER LEHRVERANSTALTUNGEN

Obwohl sich die fünf untersuchten literatur- und mediendidaktischen Seminare hinsichtlich ihrer thematischen Ausrichtung unterschieden, ähnelten sie sich im Verlauf der lernbezogenen Anforderungen sowie in der Zielsetzung, fachliche Begriffe zu klären, analytische Schritte nachzuvollziehen und in späteren Wochen eigenständig anzuwenden. Damit boten alle Lehrveranstaltungen vergleichbare Gelegenheitsstrukturen für den Einsatz des tutoriellen LLMs und erlauben eine gemeinsame Analyse der zeitlichen Nutzungsmuster und individuellen Lernverläufe.

Die Seminare folgten einer vergleichbaren Grundstruktur, die sich in drei inhaltlichen Phasen beschreiben lässt: In einer einführenden Grundlagenvermittlungsphase (Phase 1) wurden zentrale Begriffe, Konzepte und Theorien eingeführt. Daran schloss eine Phase zur Vermittlung literatur- und medienwissenschaftlicher Arbeitstechniken (Phase 2) an, in der die Studierenden mit ausgewählten Texten arbeiteten und die zuvor eingeführten Konzepte auf Beispiele aus Literatur und Sprachgebrauch anwendeten. In der Anwendungs- und Transferphase (Phase 3) standen vertiefende Aufgabenformate, eigenständige Analysen und Prüfungsvorbereitung im Zentrum.

Alle Seminare wiesen feste zeitliche Zäsuren auf, etwa die Lernstandüberprüfung durch Tests, die Abgabe schriftlicher Leistungsnachweise oder mündliche Prüfungen. Diese Prüfungsformate beeinflussten maßgeblich die Lernaktivität innerhalb des Semesters und bilden einen zentralen Bezugspunkt für die Analyse der Nutzungsspitzen von *Tutor.AI*.

4.2 ENTWICKLUNG EINES TUTORIELLEN LLMS AUF UNIVERSITÄTSEIGENEN RESSOURCEN

In Kooperation mit dem Startup colloc.AI und dem Gründungszentrum REACH auf Ressourcen der Universität Münster wurde eine textgenerierende tutorielle AI-Umgebung auf universitätseigenen Servern geschaffen, die den Vorteil der wirtschaftlichen Unabhängigkeit von externen Anbieter:innen bietet und eine datenschutzkonforme Nutzung ermöglicht (vgl. dazu Bach et al. 2025, 92). Um die Wissensbasis der textgenerierenden AI steuern zu können, sodass keine beliebigen und nicht geprüften Quellen als Grundlage der jeweiligen Antworten dienen, sondern die Möglichkeit besteht, die Eingaben in der hochschuldidaktischen Anwendung gezielt zu beeinflussen (vgl. Walter 2024), ist die Retrieval Augmented Generation (RAG)-Methode für solche wissensintensiven Anwendungen zweckmäßig (vgl. Lewis et al. 2020) und wurde entsprechend umgesetzt. Da das Projekt auf eigenen Ressourcen betrieben wird, können zudem Nutzungsverhalten und Eingaben eingesehen und Anpassungen am System vorgenommen werden. Dies bietet die Möglichkeit, eine Langzeitbetrachtung des Nutzungsverhaltens anzustellen (vgl. Bernhardt / Koppenhöfer / Schmiedgen, angenommen).

4.3 SEMINARKONZEPTION UND PHASIERUNG

Die Seminarkonzeption ist jeweils folgendermaßen angelegt:

Phase 1: Grundlagenvermittlung (3 Sitzungen)

Grundlagenvermittlung von Theorien literarischen Lernens, ästhetischer Erfahrung und Literaturvermittlung im Unterricht

Phase 2: Einüben literatur- und medienwissenschaftlicher Arbeitstechniken (2 Sitzungen)

Die Studierenden werden mit Narratologie, Analyseverfahren für Bilderbücher, Romane, Hörspiele, Filme und deren Anwendung vertraut gemacht.

Phase 3: Anwendung und Transfer: konkrete Anwendung auf literarästhetische Gegenstände (mind. 7 Sitzungen)

Es werden innerhalb der Seminare jeweils zwei literarästhetische Gegenstände über mehrere Sitzungen hinweg behandelt.

3a: *Analyse*: Es werden als Grundlage für die später erfolgende Didaktisierung Schwerpunkte zur Analyse literarästhetischer Gegenstände gelegt und Analysen angestellt.

3b: Didaktisierung: Die erarbeiteten Analyseergebnisse werden didaktisch aufbereitet: Die Befunde werden zielgerichtet auf Modelle literarischen Lernens/literarästhetischer Erfahrungen angewendet und entsprechend in Bezug auf die Vermittlungssituation perspektiviert.

3c: Methodischer Ausblick: Es werden Diskussionen zur konkreten schulpraktischen Verwertbarkeit angestellt und Aufgaben zur Methodisierung bearbeitet, wobei jeweils auch ein Schwerpunkt auf den Einsatz von KI im Unterricht angeboten wird.

Durch diese Seminarkonzeption wird angestrebt, dass die Studierenden basierend auf den theoretischen Vorüberlegungen zu einer Handlungsfähigkeit im Umgang mit konkreten Gegenständen gelangen.

4.4 DATENERHEBUNG

Die Erhebung fand im Sommersemester 2024 statt. Die Studierenden wurden mit der Nutzung von *Tutor.AI* vertraut gemacht und darüber informiert, dass die Daten anonymisiert genutzt werden; die Nutzung war freiwillig. Bei der ersten Datenerhebung konnten 2.277 Prompts von 129 Studierenden mit folgenden Informationen erhoben werden: (1) Eingaben bzw. User:innen-Prompts, (2) Ausgaben des LLMs, (3) Seminarzuordnung, und (4) Datum/Uhrzeit.

4.5 KATEGORIENBILDUNG UND CODING-PROZESS

Für die Analyse der Nutzung wurden die User:innen-Prompts in einem mehrstufigen Kodiervorgang kategorisiert. In einem ersten Schritt sichteten drei Rater:innen ergebnisoffen unabhängig voneinander das vollständige Material aus allen fünf Seminaren. Auf dieser Grundlage wurden in einem trilateralen Abgleich vorläufige Kategorien definiert, die eine fachlich-inhaltliche Strukturierung der Prompts ermöglicht. Anschließend wurde das Material erneut unabhängig voneinander gesichtet, um diese Kategorien zu prüfen, Unschärfen zu identifizieren und erforderliche Modifikationen vorzunehmen. Die finalen acht Kategorien zielten darauf ab, die Prompts von den jeweiligen Seminarphasen zu abstrahieren und fachlich zu bündeln, sodass der Datensatz für zukünftige Analysen und potenzielle automatisierte Klassifikationen nutzbar bleibt. Zur Unterstützung des Kodiervorganges wurde ergänzend eine technische Vor-Klassifikation mittels Python implementiert. Dafür wurde für jede Kategorie eine Liste mit Schlüsselwörtern erstellt, die als linguistische Marker dienen. Die automatisiert generierten Zuordnungen wurden von allen Rater:innen einzeln überprüft und korrigiert, sodass maschinelle Vorschläge konsequent fachlich abgesichert wurden. Die Interrater:innen-Reliabilität wurde auf Basis der finalen händischen Kodierungen berechnet, wobei hier Fleiß Kappa bei 0,67, also nach Landis und Koch (1977) in einem soliden und zuverlässigen Bereich, liegt und somit eine hohe Übereinstimmung zwischen den Rater:innen zeigt.

Die so gewonnenen Kategorien bilden die Grundlage für die quantitative Auswertung. Sie ermöglichen es, Nutzungsschwerpunkte über den Semesterverlauf hinweg sichtbar zu machen, Unterschiede zwischen Nutzer:innengruppen herauszuarbeiten

und unseren Erwartungshypothesen zur Nutzung entlang von Seminarphasen, Prüfungsnähe und individueller Progression nachzugehen.

5 — ERGEBNISSE

Die deskriptive Auswertung der erhobenen Daten gibt einen Überblick über das Nutzungsverhalten der Studierenden während der gesamten Vorlesungszeit. Der Kodierprozess brachte folgende 8 Kategorien hervor:

a) Begriffsklärung & Definition:

Fragen nach der Bedeutung von Begriffen oder Konzepten und entsprechend basale Klärung terminologischer oder konzeptioneller Wissensbestände (n = 617 Eingabeprompts, 27,10%)

b) Theoriebezug und wissenschaftlicher Kontext:

Bezugnahmen auf konkrete theoretische Positionen, die in konkrete Anwendungsszenarien gestellt werden (n = 386, 16,95%)

c) Analyseaufgabe zu literarischen Texten:

Fragen zur Analyse eines literarästhetischen Gegenstandes, beispielsweise eine Anwendung von Analysewerkzeugen auf einen konkreten Gegenstand oder eine Szene (n = 371, 16,29%)

d) Didaktische Reflexion:

Anwendungen von Befunden auf Vermittlungskontexte, didaktische Perspektiven, die Anwendung von Modellen literarästhetischen Lernens auf bestimmte Befunde (n = 149, 6,54%)

e) Unterrichtspraktische Frage:

Fragen nach Methoden oder konkreten Perspektiven der Unterrichtsgestaltung (n = 171, 7,51 %)

f) Zusammenfassungen:

Konkrete Fragen nach der Zusammenfassung oder der Herausarbeitung der wichtigsten Gedanken aus einem oder mehreren Texten (n = 154, 6,76%)

g) Sonstiges / Unklar / Small Talk:

User:innenprompts, deren Kontext sich nicht erschließen ließ, etwa weil sie nur aus einem Wort bestanden oder deren Bezug nicht klar war (n = 298, 13,09%)

h) LLM:

User:innenprompts, die eine Übertragung ins Deutsche forderten, da das LLM gelegentlich ins Englische wechselt (n = 131, 5,75%)

Insgesamt dominieren Anfragen zur Begriffsklärungen und Definitionen (Kategorie a) mit 27,10 % aller Prompts. Darauf folgen Anfragen in Bezug auf die literatur- und medienwissenschaftlichen Analyseverfahren (Kategorie b; 16,95 %) sowie analytische Aufgabenstellungen zu literarischen Texten (Kategorie c; 16,29 %). Didaktische Reflexionen (Kategorie d) und unterrichtspraktische Fragen (Kategorie e) machen demgegenüber jeweils einen geringeren Anteil aus.

Das Erkenntnisinteresse unserer Studie besteht darin, das Nutzungsverhalten der Studierenden zu erarbeiten und herauszufinden, inwiefern sich die Konzeption der Seminare von der Grundlagenvermittlung hin zur eigenständigen Anwendung und Transferleistungen im Nutzungsverhalten widerspiegelt. Vor diesem Hintergrund lassen sich die von uns erarbeiteten Kategorien c, d und e aggregieren, da es sich hierbei um jeweils spezifische Anwendungs- und Transferabfragen handelt. Die Kategorien f–g stellen Sammelkategorien dar, die allesamt keine Rückschlüsse auf die Lernprogression zulassen.

Entsprechend lassen sich folgende Kategorisierungen aggregieren:

- I Begriffsklärung & Definition
- II Theoriebezug und wissenschaftlicher Kontext
- III Anwendungs- und Transferaufgaben (sowohl literatur-/medienwissenschaftliche Analysen als auch didaktische oder methodische Reflexionen).

5.1 PHASENORIENTIERTE NUTZUNG

Im Folgenden vergleichen wir die aggregierten Kategorien I–III mit den Seminarphasen 1–3. Rein quantitativ besteht keine vollständige Übereinstimmung zwischen Seminarverlauf und Prompting: Während Seminarphase 1 und 2 zusammen etwa ein Viertel der Veranstaltungszeit umfassen, entfällt rund drei Viertel des Semesters auf Phase 3, in der es um Anwendung und Transfer geht. Der Anteil der Prompts in der anwendungs- und transferbezogenen Kategorie III nimmt aber nicht proportional zu, sondern verbleibt auf einem relativ moderaten Niveau. Damit wird sichtbar, dass *Tutor.AI* durch alle Seminarphasen hinweg vor allem für inhaltliche Klärungen genutzt wird, weniger jedoch für Anwendungs- oder Transferszenarien.

Zur statistischen Absicherung dieser Beobachtungen wurden die Prompts mit ermittelten Kategorien anhand ihrer Zeitstempel den jeweiligen Seminarphasen zugeordnet. Die weitere Aggregation der Kategorien in die Prompt-Kategorien I–III und deren Projektion auf die drei Seminarphasen ergibt die in Tabelle 1 dargestellte Kontingenztafel.

	Seminarphase 1	Seminarphase 2	Seminarphase 3	Summe
Promptkategorie I	321	40	256	617
Promptkategorie II	211	45	130	386
Promptkategorie III	180	100	411	691
Summe	712	185	797	1.694

Tabelle 1: Aggregation der Promptkategorien auf Seminarphasen

Hier lassen sich die Phasen signifikant unterscheiden ($\chi^2 = 132,02$; $p < 0,001$). Während die Zuordnung der Kategorien in den Seminarphasen 1 und 3 überwiegend den erwarteten Mustern folgt, verteilen sich die Prompts der Kategorie 2 über alle Phasen hinweg und lassen sich nicht eindeutig einer einzelnen Seminarphase zuordnen.

5.2 NUTZUNG IM SEMESTERVERLAUF

Die zeitliche Entwicklung der Nutzung von *Tutor.AI* im Semesterverlauf ist in Abbildung 1 für alle fünf untersuchten Seminare (A–E) dargestellt. Die Abbildung zeigt die Anzahl der User:innen-Prompts pro Kalenderwoche, differenziert nach den drei aggregierten Promptkategorien I–III.

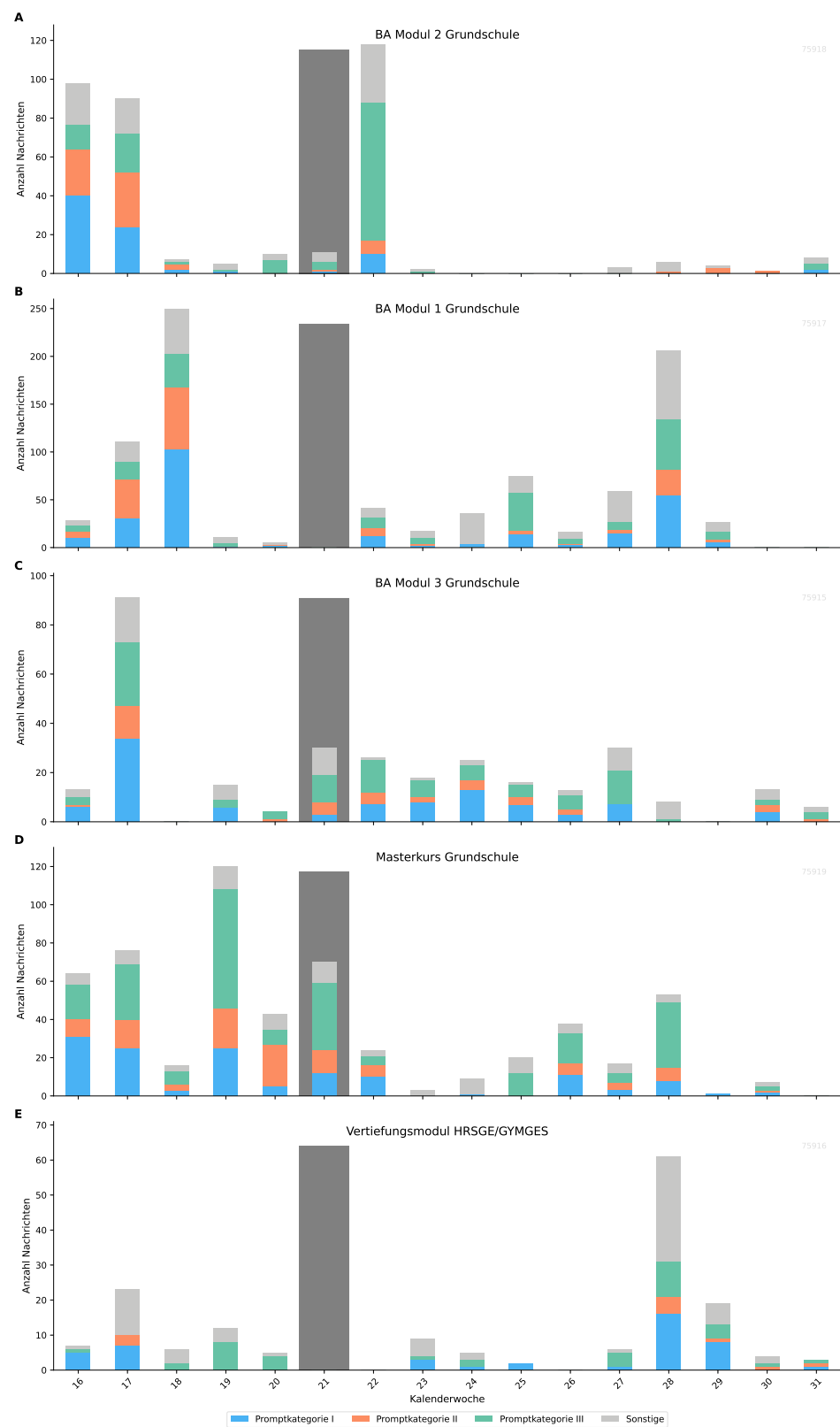


Abb. 1: Nutzungshäufigkeiten von *Tutor.AI* pro Kalenderwoche, differenziert nach Seminar (A–E) und Prompt-Kategorien (I–III), grau hinterlegt ist die Pfingstwoche

Über alle Seminare hinweg zeigen sich deutliche Nutzungsspitzen in test- oder prüfungsnahen Wochen. Zwischen diesen Phasen ist die Nutzung deutlich geringer. Dieses Muster tritt unabhängig von Seminarformat, Studiengang oder Modulzuordnung auf. Zugleich lassen sich kursspezifische Unterschiede erkennen: Während einzelne Seminare eine eher kontinuierliche Nutzung aufweisen, konzentriert sich die Interaktion in anderen stark auf einzelne Prüfungszeiträume. Inhaltlich dominieren zu Beginn des Semesters Eingaben der Kategorie I. Im weiteren Verlauf verschiebt sich der Schwerpunkt in Richtung anwendungs- und transferbezogener User:innen-Prompts (Kategorie III). Diese Verschiebung verläuft jedoch nicht linear, sondern wird durch die Prüfungsstruktur der Seminare überlagert.

Zur statistischen Überprüfung des vermuteten Prüfungsnähe-Effekts wurden alle Prompts nach Test- und Nicht-Testwochen getrennt betrachtet. Tabelle 2 zeigt deskriptiv die aggregierte Aufteilung der Prompts in diesen beiden Zeiträumen für jedes Seminar und die zugehörige Testfrequenz.

Seminar	Testfrequenz	Nicht-Testwochen	Testwochen
A	2 / 16	155	208
B	4 / 16	310	573
C	4 / 16	145	163
D	5 / 16	265	296
E	1 / 16	101	61
Gesamt		976	1301

Tabelle 2: Anzahl von Prompts in Test- und Nicht-Testwochen pro Seminar

Es wurde ein lineares Fixed Effects Modell bestimmt, das kursbedingte Idiosynkrasien kontrolliert und die jeweiligen Testwochen mit einem Dummy markiert. Diese Regression zeigt, dass die Anzahl der Prompts in Testwochen im Mittel um 64 Prompts signifikant höher liegt als in Nicht-Testwochen ($\beta = 63,71$; $t = 6,54$; $p < 0,001$; $R^2 = 0,45$), unabhängig von kursbezogenen Basisunterschieden. Damit lässt sich plausibel annehmen, dass die Nutzung von *Tutor.AI* im Semesterverlauf maßgeblich durch Prüfungs- und Leistungsanforderungen strukturiert wird.

5.3 INDIVIDUELLE NUTZUNGSINTENSITÄT / PROGRESSION

Zur Analyse der individuellen Nutzung wurden die User:innen anhand der jeweiligen bisherigen Anzahl ihrer Prompts und in Hinblick auf die abzulesende Progression, betrachtet. Zunächst zeigt sich eine stark ungleiche Verteilung der Nutzung: Die größte Gruppe mit 56 Nutzer:innen (≤ 10 Prompts) umfasst knapp die Hälfte aller Studierenden, erzeugt jedoch nur einen geringen Anteil des gesamten Prompt-Volumens (14,7 %). Demgegenüber verantworten nur 22 intensiv nutzende Studierende (≥ 31 Prompts) einen überproportional hohen Anteil der Gesamtinteraktionen (44,4 %). Die kategoriale Progression der durch die jeweils verbleibenden Nutzenden ermittelten anteiligen Verteilung der Prompt-Kategorien wird in Abbildung 2 visualisiert.

Die x-Achse zeigt die kumulative Nutzung, die Prompt-Progression, die y-Achse die aggregierte Verteilung nach ermittelten Kategorien der Eingaben der verbliebenen Nutzer:innen.

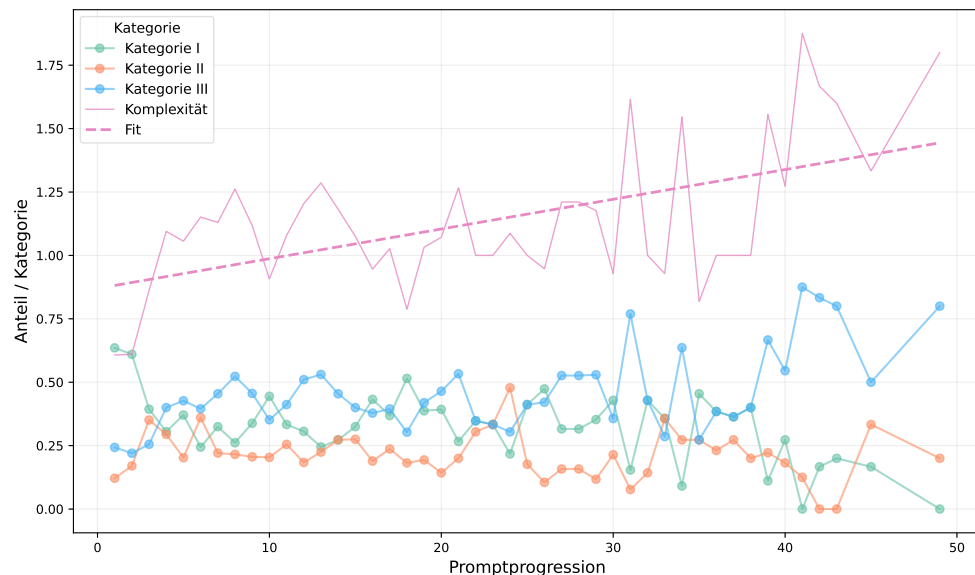


Abb. 2: Kumulative Nutzung und kategoriale Verteilung nach Prompt-Progression

Mit zunehmender Nutzung verlagert sich der Anteil der Prompts von Kategorie I und II hin zu Kategorie III. Zur statistischen Absicherung dieser Beobachtung wurde die individuelle Prompt-Progression mit einer simplen gewichteten Komplexitätskennzahl ($\text{Komplexität} = 1 \times \text{Kategorie II} + 2 \times \text{Kategorie III}$) in Beziehung gesetzt. Die Rangkorrelation zeigt einen signifikant positiven Zusammenhang zwischen Progression und Komplexität der Prompts (Spearman $\rho = 0,39$; $p = 0,008$). Eine lineare Regression bestätigt diesen Befund mit einem signifikanten Anstieg der Komplexität um $\beta = 0,012$ pro zusätzlichem Prompt ($t = 4,43$; $p < 0,001$; $R^2 = 0,31$).

Zusammenfassend lässt sich festhalten, dass eine höhere individuelle Nutzungsintensität mit einer stärkeren Verlagerung hin zu analytischen, didaktischen, oder unterrichtspraktischen Fragestellungen der Kategorie III einhergeht.

5.4 RÜCKBINDUNG AN DIE ERWARTUNGSHYPOTHESEN

Die Ergebnisse erlauben eine evidenzgestützte Rückkoppelung an unsere oben formulierten Erwartungshypothesen.

EH1 – Phasenorientierte Nutzung: Die Analyse der Kategorienverteilungen zeigt über alle Seminare hinweg eine grobe Orientierung am inhaltslogischen Aufbau der Veranstaltungen. Zu Beginn dominieren Anfragen der Kategorie I (Begriffsklärung & Definition), während im weiteren Verlauf verstärkt anwendungs- und transferbezogene Kategorien der Phase III auftreten. Dieses erwartbare Muster wird jedoch durch prü-

fungsnähe Nutzungsspitzen überlagert, die zu deutlichen Abweichungen vom linearen inhaltlogischen Verlauf führen. H1 lässt sich daher als teilweise gestützt ansehen: Eine grundlegende fachlogische Entwicklung ist erkennbar, allerdings erfolgt sie nicht kontinuierlich entlang der inhaltlichen Seminararchitektur, sondern wird durch die formale Ausrichtung (Prüfungsarchitektur) ebenfalls beeinflusst.

EH2 – Prüfungsnähe-Effekt: In allen fünf Seminaren steigt die Nutzung von *Tutor.AI* in den Wochen mit Leistungsnachweisen deutlich an. Dieses Muster wiederholt sich sowohl bei schriftlichen Tests als auch bei mündlichen Prüfungen, unabhängig von thematischen Schwerpunkten oder Kursformaten. *Tutor.AI* wird damit offenbar vor allem als lernunterstützendes Instrument in prüfungsrelevanten Phasen genutzt.

EH3 – Nutzungsintensität und lernlogische Progression: Studierende mit hoher individueller Nutzungsintensität verwenden *Tutor.AI* nicht nur häufiger, sondern auch qualitativ anders. Mit zunehmender Nutzung verschiebt sich der Schwerpunkt von Theoriefragen mit basalen Grundlagen (Kategorie I) hin zu anwendungs- und transferbezogenen Prompts (Kategorie III). Während sich diese Entwicklung im Gesamtkollektiv nur abgeschwächt zeigt, tritt sie in den individuellen Lernverläufen intensiver Nutzer:innen deutlich hervor. Eine höhere Nutzung geht tendenziell mit komplexeren und stärker anwendungsorientierten Prompts einher.

5.5 EINORDNUNG UND DISKUSSION DER BEFUNDE

Die Ergebnisse der Hypothesenprüfung zeigen, dass das Nutzungsverhalten der Studierenden sowohl durch strukturelle Elemente der Lehrveranstaltungen als auch durch individuelle Lernstrategien geprägt ist. Auffällig ist insbesondere die starke Prüfungsorientierung, die sich seminarübergreifend wiederholt und die intendierte Progression innerhalb der Seminare teilweise überlagert. Dieser Befund deckt sich mit Erkenntnissen aus der empirischen Lehr-Lernforschung, die testbezogene Aktivitätspeaks als typisches Muster studierendenbezogener Steuerung von Lernaufwand beschreibt (vgl. Gibbs / Simpson 2004). Im Kontext der Deutschdidaktik und insbesondere der Literatur- und Mediendidaktik wirft dies die Frage auf, inwiefern tutorielle Unterstützungssysteme tatsächlich zur kontinuierlichen Auseinandersetzung mit komplexen fachlichen und fachdidaktischen Inhalten beitragen oder ob sie primär als kurzfristige Bewältigungsressource in belastungsintensiven Phasen genutzt werden.

Die geringen Anteile an didaktischen und unterrichtspraktischen Prompts sind ebenfalls erklärungsbedürftig. Trotz der seminaristischen Schwerpunktsetzung auf literarästhetisches Lernen und die Anwendung theoretischer Modelle auf konkrete Vermittlungssituationen zeigt sich, dass Studierende das tutorielle System kaum als Unterstützung für fachdidaktische Reflexionen nutzen. Dies könnte darauf hinweisen, dass Studierende die Kompetenz des Systems in diesen Bereichen als begrenzt wahrnehmen oder dass didaktische Aufgabenstellungen stärker an dialogisches Arbeiten, Peer-Diskussionen oder angeleitete Präsenzphasen gebunden sind. Hier zeigt sich ein bedeutsamer Anschluss an die aktuelle Diskussion zur „AI Literacy“ in der Lehrer:innenbildung (vgl. Müller / Fürstenberg 2023): Wenn Studierende genera-

tive AI primär als Instrument zur Wissensrekonstruktion nutzen, bleibt ihr Potenzial zur Unterstützung komplexerer reflexiver Lernprozesse unausgeschöpft. Zudem legt die Analyse der individuellen Nutzungsverläufe nahe, dass Progressionen weniger sequenziell entstehen, sondern sich vor allem aus wiederholter Nutzung ergeben. Intensivere Nutzer:innen zeigen eine Verschiebung hin zu analytischen Kategorien, während die Mehrheit der Studierenden bei basalen Nachfragen verbleibt. Daraus lässt sich schließen, dass vertiefte kognitive Prozesse häufig erst durch wiederholte Interaktion entstehen und nicht allein durch die Struktur der Lernumgebung initiiert werden. Für die literatur- und mediendidaktische Arbeit bedeutet dies, dass tutoriell unterstützte Lernprozesse stärker darauf angewiesen sind, dass Studierende eigeninitiativ in längere Interaktionssequenzen eintreten.

Schließlich verdeutlichen die Ergebnisse, dass tutorielle Systeme wie *Tutor.AI* eine ambivalente Rolle im akademischen Lernen einnehmen: Sie bieten niedrighschwellige Unterstützung bei der Klärung fachlicher Grundlagen, werden jedoch nur begrenzt als Ressource für komplexe Transferleistungen genutzt. Die Befunde weisen damit zugleich auf das Potenzial und die Herausforderungen beim Einsatz generativer AI in der universitären Lehre hin. Eine gezielte didaktische Einbettung – insbesondere eine systematische Förderung des reflektierten und strategischen Prompts – erscheint als notwendig, um das Potenzial solcher Systeme für literatur- und mediendidaktisches Lernen umfassender nutzbar zu machen.

5.6 LIMITATIONEN

Die vorliegenden Befunde sind vor dem Hintergrund mehrerer Limitationen zu interpretieren. Erstens erlaubt das Studiendesign keine Aussagen darüber, wie sich das Lern- und Promptingverhalten ohne den Einsatz eines tutoriellen LLM-Systems entwickelt hätte. Da keine Vergleichsgruppe erhoben wurde, bleibt erstens offen, ob die beobachteten Muster – insbesondere die prüfungsnahen Nutzungsspitzen und die individuelle Progression intensiver Nutzer:innen – auf *Tutor.AI* selbst zurückzuführen sind oder lediglich bestehende Lernstrategien der Studierenden widerspiegeln. Zweitens erfolgte die Analyse auf Grundlage eines Datensatzes, der stark durch die Struktur der Leistungsanforderungen geprägt ist. Die ausgeprägten Nutzungsspitzen zu Tests und mündlichen Prüfungen zeigen, dass das Nutzungsgeschehen eng an prüfungsrelevante Situationen gekoppelt ist. Ob sich in anderen Prüfungsformaten oder in weniger stark testorientierten Lehrveranstaltungen ähnliche Muster ergeben würden, kann auf Grundlage der vorliegenden Daten nicht beurteilt werden. Drittens ist die Kategorisierung der Prompts trotz des triangulativ abgesicherten Kodiervorgangs nicht frei von Interpretationsspielräumen. Einige Kategorien weisen semantische Überschneidungen auf und die Zuordnung einzelner Prompts kann – trotz Konsensverfahrens – fachlich unterschiedlich bewertbar sein. Dies gilt insbesondere dort, wo Studierende mehrteilige und komplexe Anfragen formulierten. Viertens erlauben die Daten keine eindeutigen Rückschlüsse darauf, ob die Nutzung von *Tutor.AI* zu tatsächlichen Kompetenzzuwächsen geführt hat. Ohne begleitende Kompetenzmessungen bleibt unklar, ob der vermehrte Einsatz bestimmter Kategorien – etwa im anwendungs- und transferbezogenen Bereich – Ausdruck wachsender

Fähigkeiten ist oder lediglich einen Strategiewechsel im Umgang mit dem System darstellt. Schließlich bildet die Studie nur jene Studierenden ab, die *Tutor.AI* aktiv genutzt haben. Nutzer:innen mit geringer oder ausbleibender Nutzung erscheinen im Datensatz kaum und können damit hinsichtlich ihrer Lernstrategien, Schwierigkeiten oder potenziellen Barrieren nicht berücksichtigt werden. Dies schränkt die Generalisierbarkeit der Befunde ein und verweist auf die Notwendigkeit ergänzender Erhebungen, die auch nicht-nutzende Studierende systematisch berücksichtigen. Fünftens wurden die Daten im Kontext von fünf literatur- und mediendidaktischen Seminaren mit einer sehr ähnlichen Struktur erhoben. Inwiefern sich die Ergebnisse auch bei anderen Seminarkonzepten und auch in anderen Teildisziplinen analog gestalten würden, lässt sich nicht antizipieren.

Die dargestellten Befunde werden im folgenden Fazit mit Blick auf Implikationen für die Gestaltung universitärer Lehr-Lernsettings sowie auf weiteren Forschungsbedarf zusammengefasst.

6 — FAZIT

Die vorliegende Studie untersuchte die Nutzung des tutoriellen LLMs *Tutor.AI* in fünf germanistischen literatur- und mediendidaktischen Lehrveranstaltungen und gab Einblicke in tatsächliche Nutzungsmuster Studierender in den jeweiligen Seminarphasen, der Prüfungsnähe und bezüglich individueller Lernverläufe. Die Ergebnisse zeigen, dass *Tutor.AI* durchgängig primär als Instrument zur Begriffserklärung und zur Auseinandersetzung mit vermittelten Inhalten genutzt wird, während Transferaufgaben wie etwa didaktische und unterrichtspraktische Fragestellungen nur eine verhältnismäßig untergeordnete Rolle spielen. Damit bleibt die Nutzung deutlich hinter der intendierten lernlogischen Progression der Seminare zurück. Besonders prägend für das Nutzungsgeschehen ist die ausgeprägte Prüfungsorientierung. In allen Seminaren zeigen sich deutliche Nutzungsspitzen in Wochen mit Tests oder Prüfungsleistungen, während zwischen diesen Perioden eine vergleichsweise geringe Aktivität zu beobachten ist. Diese Befunde verdeutlichen, dass die Nutzung des tutoriellen Systems weniger kontinuierlich lernprozessbegleitend erfolgt, sondern stark an leistungsrelevante Anforderungen gekoppelt ist. Die Seminararchitektur und das Prüfungsdesign erweisen sich damit als zentrale Steuerungsgrößen des Nutzungsverhaltens. Gleichzeitig zeigen die Analysen individueller Nutzungsverläufe, dass Studierende mit hoher Nutzungsintensität *Tutor.AI* qualitativ anders einsetzen. In diesen Lernreisen ist eine Verschiebung von Begriffsanfragen hin zu anwendungs- und transferbezogenen Fragestellungen erkennbar, was auf eine lernlogische Progression im individuellen Umgang mit dem System hindeutet. Diese Entwicklung bleibt jedoch auf einen vergleichsweise kleinen Teil der Nutzer:innen beschränkt und prägt nicht das Gesamtbild.

Insgesamt legen diese Ergebnisse in literatur- und mediendidaktischen Seminaren nahe, dass tutorielle LLMs ihr Potenzial als Verstehens- und Lernsupport im universitären Kontext nur dann entfalten können, wenn ihre Nutzung stärker in die didaktische Gestaltung der Lehrveranstaltungen eingebettet wird. Insbesondere Prüfungsformate, Aufgabenstellungen und explizite Nutzungsszenarien scheinen

entscheidend dafür zu sein, ob Studierende das System über kurzfristige prüfungsbezogene Klärungen hinaus einsetzen.

Für die weitere Forschung ergibt sich daraus die Notwendigkeit, Nutzungsmuster systematisch mit unterschiedlichen Seminar- und Prüfungsdesigns zu vergleichen. Zudem bedarf es ergänzender Studien, die den Zusammenhang zwischen der Nutzung tutorieller LLMs und tatsächlichen Kompetenzentwicklungen untersuchen. Erst auf dieser Grundlage lässt sich beurteilen, inwiefern tutorielle Systeme langfristig zur Förderung vertieften literarischen Lernens in der Hochschullehre beitragen können.

LITERATURVERZEICHNIS

- **Bach, Friedrich / Bernhardt, Sebastian / Reuvekamp, Silvia / Schmiedgen, Ninja (2025)**: SHIFT happens. Lernen mit und von textgenerierender KI. In: Fürstenberg, Maurice / Müller, Hans-Georg (Hg.): *DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz*. Berlin: Frank & Timme, 87–105. <https://doi.org/10.26530/20.500.12657/104475>. — **Bajohr, Hannes (2022)**: *Schreibenlassen. Texte zur Literatur im Digitalen*. Berlin: Matthes & Seitz. — **Bernhardt, Sebastian / Koppenhöfer, Ann-Kathrin / Schmiedgen, Ninja (angenommen)**: Tutorielle LLMs als dialogischer Support? – Nutzungsszenarien Studierender während eines Semesters in quantitativer und qualitativer Perspektive. In: Jagdschian, Larissa Carolin / Hesse, Florian (Hg.): *Literatur \\\ Künstliche Intelligenz // Didaktik*. Würzburg: Königshausen&Neumann. — **Bond, Melissa et al. (2024)**: A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, 4. <https://doi.org/10.1186/s41239-023-00436-z>. — **Corvacho del Toro, Irene / Fuhlrott, Mareike / Steinhoff, Torsten (2025)**: Didaktische Agenten. KI als Lehr-/Lernpartnerin im Deutschunterricht im Forschungsprojekt KIMADU. In: Fürstenberg, Maurice / Müller, Hans-Georg (Hg.): *DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz*. Berlin: Frank & Timme, 65–86. <https://doi.org/10.26530/20.500.12657/104475>. — **Gibbs, Graham / Simpson, Claire (2005)**: Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education* (1), 3–31. — **Kosmyna, Nataliya / Hauptmann, Eugene / Yuan, Ye Tong et al. (2025)**: Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. *arXiv Preprint*, 10. Juni 2025. <https://doi.org/10.48550/arXiv.2506.08872>. — **Landis, J. Richard / Koch, Gary G. (1977)**: The Measurement of Observer Agreement for Categorical Data. In: *Biometrics*, 33 (1), 159–174. <https://doi.org/10.2307/2529310>. — **Lewis, Patrick, Perez, Ethan, Piktus, Aleksandra, Petroni, Fabio, Karpukhin, Vladimir, Goyal, Naman, Küttler, Heinrich, Lewis, Mike, Yih, Wen-tau, Rocktäschel, Tim, Riedel, Sebastian, Kiela, Douwe (2020)**: Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, 9459–9474. — **Müller, Hans-Georg / Fürstenberg, Maurice (2023)**: Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. In: *Mitteilungen des Deutschen Germanistenverbandes* 70 (4), S. 327–345. — **Simbeck, Katharina (2025)**: Von Wörtern zu Wundern. Die Technologie großer Sprachmodelle und ihre Grenzen. In: Müller, Hans-Georg / Fürstenberg, Maurice (Hg.): *DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz*. Berlin: Frank & Timme, 13–37. <https://doi.org/10.26530/20.500.12657/104475>. — **Steinhoff, Torsten / Lehnen, Katrin (2025)**: Schreiben mit Künstlicher Intelligenz: Das GPT-Modell (Ghost, Partner, Tutor). In: *Leseräume*, 11 (12), 1–14. — **Walter, Yoshija (2024)**: Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1), 15. <https://doi.org/10.1186/s41239-024-00448-3>. — **Zawacki-Richter, Olaf et al. (2019)**: Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16, 39. <https://doi.org/10.1186/s41239-019-0171-0>.

ÜBER DIE AUTOR:INNEN

Prof. Dr. Sebastian Bernhardt ist Professor für Literatur und Mediendidaktik am Germanistischen Institut der Universität Münster. Zu seinen Forschungsinteressen gehören Literatúrausstellungen und ihre Potenziale für das literarische Lernen, diversitätssensibler Literaturunterricht, Hörspiellidaktik und transmediale Literaturdidaktik sowie AI in der Hochschullehre und im Literaturunterricht.

Friedrich Bach, M.Sc., ist wissenschaftlicher Mitarbeiter am REACH EUREGIO Start-up Center an der Universität Münster und Co-Leiter des Projekts Incub.AI. Seine Forschungsinteressen liegen in den Bereichen KI, Big Data und KI in der Hochschullehre.

Ann-Kathrin Koppenhöfer ist als studentische Hilfskraft bei Prof. Dr. Sebastian Bernhardt beschäftigt und studiert Germanistik und Geographie. Ihr Forschungsinteresse liegt im diversitätssensiblen Literaturunterricht.

Ninja Schmiedgen, M.A., ist wissenschaftliche Mitarbeiterin an der Forschungsstelle Innovation und Transfer digitaler Lehre (ITdL) im Fachbereich 09 der Universität Münster. Ihre Forschungsinteressen sind Digitalität in der Hochschullehre sowie KI in der Hochschullehre.