Forschungsmonitor Schule
Rezensionen für die Praxis

Nils Voelzke

Rezension zu

Lademann, J., Henze, J. & Becker-Genschow, S. (2025). Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables. *Physical Review Physics Education Research, 21*, Article 010147.

Kommentierter Kurzbefund

Die zunehmende Verfügbarkeit generativer KI-Systeme wie ChatGPT wirft die Frage auf, wie KI-generierte Materialien sinnvoll zur Unterstützung von Lernprozessen eingesetzt werden können. In der Studie wird untersucht, ob durch einen KI-basierten Chatbot generierte Erklärungen als ergänzende Lernmaterialien das Lernen von Schülerinnen und Schülern unterstützen können.

In einer randomisierten kontrollierten Studie mit 214 Schülerinnen und Schülern der 6. Jahrgangsstufe wurde untersucht, wie sich KI-generierte Erklärungen als ergänzende Lernmaterialien zu proportionalen Zusammenhängen im mathematisch-physikalischen Kontext auswirken. Eine Experimentalgruppe erhielt hierbei zusätzlich zum Schulbuch KI-generierte Erklärungen, während eine Kontrollgruppe ausschließlich mit traditionellen Materialien arbeitete.

Zwischen den Gruppen konnte kein signifikanter Unterschied in der Lernleistung festgestellt werden. Allerdings berichten Schülerinnen und Schüler, die mit den KI-generierten Materialien arbeiteten, höhere Selbstwirksamkeit, stärkeres situatives Interesse und positivere Emotionen. Gleichzeitig war ihre wahrgenommene kognitive Belastung geringer als bei Lernenden, die nur mit Schulbuchmaterial arbeiteten.

Die Untersuchung spricht dafür, dass KI-generierte ergänzende Lernmaterialien das affektive Erleben des Lernens verbessern und die empfundene kognitive Belastung reduzieren können, auch wenn kurzfristig kein zusätzlicher Leistungszuwachs messbar ist.

Hintergrund

Der Einsatz generativer Systeme künstlicher Intelligenz wird in den letzten Jahren zunehmend auch im Bildungsbereich diskutiert. Große Sprachmodelle wie ChatGPT können Texte generieren, Fragen beantworten oder fachliche Inhalte erläutern und werden daher als mögliche Unterstützung von Lernprozessen betrachtet. Lademann, Henze und Becker-Genschow weisen darauf hin, dass solche Systeme zur Erstellung ergänzender Lernmaterialien genutzt werden könnten. Zugleich verweisen sie unter Bezug auf aktuelle Forschung darauf, dass der Einsatz von KI-Chatbots im Bildungskontext auch kritisch diskutiert werde. Thematisiert würden unter anderem mögliche Einschränkungen des tiefen konzeptionellen Verständnisses, die Qualität und Zuverlässigkeit KI-generierter Antworten sowie mögliche Auswirkungen auf kognitive Verarbeitungsprozesse. In ersten Studien sei zugleich berichtet worden, dass KI-generierte Lernhilfen motivationale Unterstützung böten und subjektive Hemmschwellen beim Lernen reduzieren könnten (vgl. Adiguzel, Kaya & Cansu, 2023; Kotsis, 2024).

CC BY NC ND

Bislang sei jedoch nur begrenzt empirisch untersucht, welche konkreten Auswirkungen KI-generierte Lernmaterialien auf Lernprozesse haben. Insbesondere fehle es an Studien, die systematisch den Einfluss solcher Materialien auf Lernleistung, kognitive Belastung und affektive Variablen des Lernens analysieren.

Die Autorin und die Autoren verorten ihre Untersuchung in der Forschung zu KI-generierten Lernmaterialien und digital unterstütztem Lernen. Als theoretischen Bezugspunkt führen sie insbesondere die Cognitive Load Theory nach Sweller (1988) an. In dieser wird davon ausgegangen, dass Lernprozesse durch die begrenzte Kapazität des Arbeitsgedächtnisses beeinflusst werden. Lernmaterialien gelten demnach als lernförderlich, wenn sie Informationen verständlich strukturieren und dadurch die kognitive Belastung der Lernenden reduzieren.

Als fachlichen Hintergrund beschreiben die Autorin und die Autoren zudem Schwierigkeiten von Schülerinnen und Schülern beim Umgang mit unterschiedlichen mathematischen und physikalischen Darstellungsformen wie Graphen, Tabellen oder Formeln. Lernprobleme entstünden insbesondere dann, wenn Beziehungen zwischen diesen Repräsentationen nicht verstanden oder übertragen werden könnten (vgl. Ainsworth, 2006). KI-gestützte Erklärungen könnten hier unterstützend wirken, indem sie Zusammenhänge zwischen verschiedenen Darstellungen adressatengerecht erläutern. Vor diesem Hintergrund wird untersucht, inwiefern KI-gestützte Erklärungen zur Gestaltung fachlich geeigneter Unterstützungsmaterialien beitragen können.

Darüber hinaus werden Forschungserkenntnisse zu affektiven Variablen des Lernens einbezogen, um zu prüfen, ob diese auch hier einen Einfluss haben. Dazu zählen beispielsweise Interesse, Emotionen oder Selbstwirksamkeitserwartungen, die den Lernprozess beeinflussen können.

Vor diesem Hintergrund untersuchen Lademann et al., welche Auswirkungen KI-generierte Erklärungen als ergänzendes Lernmaterial auf

1. die Lernleistung,
2. affektive Variablen wie situatives Interesse, Emotionen und Selbstwirksamkeit sowie
3. die wahrgenommene kognitive Belastung

von Schülerinnen und Schülern haben.

Design

Stichprobe

An der Studie nahmen 214 Schülerinnen und Schüler der 6. Jahrgangsstufe weiterführender Schulen in Deutschland teil. Die Teilnehmenden wurden zufällig einer Experimentalgruppe oder einer Kontrollgruppe zugeordnet.

Untersuchungsanlage

Durch die Aufteilung in eine Experimental- und eine Kontrollgruppe wurde untersucht, welchen Einfluss die KI-generierten Erklärungen auf Lernleistung, kognitive Belastung und affektive Variablen des Lernens haben. Die Lernenden bearbeiteten Unterrichtsmaterial zu „proportionalen Zusammenhängen“ in mathematischen und physikalischen Kontexten. Während die Kontrollgruppe ausschließlich mit traditionellen Schulbuchmaterialien arbeitete, erhielt die Experimentalgruppe zusätzlich KI-generierte

Erklärungen, die mithilfe eines speziell konfigurierten Chatbots erstellt und nach mehrfachem menschlichem Feedback durch diesen jeweils überarbeitet worden waren.

Erhebungsinstrumente

Um die Vergleichbarkeit der zwei Gruppen zu prüfen, wurden vor Bearbeitung der Lernmaterialien die letzte Mathematiknote sowie das fachbezogene Interesse an Mathematik erhoben. Nach der Bearbeitung der Lernmaterialien erfassten die Autorin und die Autoren zunächst subjektive Einschätzungen der Lernenden mithilfe standardisierter, zum Teil adaptierter Fragebogenskalen, die aufgrund des Alters der Probanden zum Teil jedoch leicht verkürzt wurden und an den Kontext angepasste Itemformulierungen enthielten. Erfasst wurden insbesondere

- intrinsische und extrinsische kognitive Belastungen (je 3 Items; basierend auf einer ursprünglichen 10-Item-Skala),
- situatives Interesse (4 Items),
- Emotionen während des Lernens (2 Items Freude, Zufriedenheit; 3 Items zu Langeweile, Frustration und Unsicherheit) sowie
- Selbstwirksamkeitserwartungen (5 Items).

Die Antworten wurden über vierstufige Antwortskalen (Likert-Skalen) erhoben, auf denen die Lernenden ihre Einschätzungen zu verschiedenen Aussagen und Empfindungen abgaben.

Anschließend bearbeiteten die Schülerinnen und Schüler einen fachlichen Leistungstest zu proportionalen Zusammenhängen, mit dem die Lernleistung erfasst wurde.

Datenauswertung

Die Analysen wurden mit dem Statistikprogramm R (Version 4.4.0) durchgeführt. Zunächst wurden die internen Konsistenzen der adaptierten und teilweise gekürzten Fragebogenskalen überprüft (Cronbachs Alpha). Die internen Konsistenzen der adaptierten und gekürzten Fragebogenskalen liegen im akzeptablen bis exzellent Bereich ($\alpha > 0.7$ bis $\alpha > 1.0$). Zusätzliche Modifikationen der Skalen erwiesen sich daher nicht als erforderlich.

Zur Vorbereitung der weiteren statistischen Analysen überprüften die Autorin und die Autoren mithilfe eines Tests auf Normalverteilung (Shapiro-Wilk-Test), ob die Daten die Voraussetzungen zur Analyse durch parametrische Verfahren erfüllen.

Da für einzelne Variablen keine Normalverteilung festgestellt wurde, erfolgten die Gruppenvergleiche zwischen Experimental- und Kontrollgruppe mithilfe eines nichtparametrischen Verfahrens für unabhängige Stichproben (Mann-Whitney-U-Test).

Analysiert wurden Unterschiede hinsichtlich der Lernleistung, der wahrgenommenen kognitiven Belastung sowie verschiedener affektiver Variablen wie situatives Interesse, Emotionen und Selbstwirksamkeitserwartungen.

Zusätzlich wurden die jeweiligen Effektstärken (Cohen's d) angegeben, um einordnen zu können, wie deutlich die Unterschiede zwischen den Gruppen ausfielen.

Ergebnisse

Auswirkungen KI-generierter Erklärungen auf die Lernleistung

Zur Untersuchung der Lernleistung wurden die Testergebnisse der Experimental- und der Kontrollgruppe mithilfe statistischer Gruppenvergleiche (Mann-Whitney-U-Tests) verglichen. Dabei zeigt sich kein signifikanter Unterschied im Leistungstest zwischen den beiden Gruppen. Für den Gesamtwert des Tests ergibt sich kein statistisch signifikanter Effekt des Einsatzes KI-generierter Erklärungen ($p = 0.750$).

Auswirkungen KI-generierter Erklärungen auf die kognitive Belastung

Die kognitive Belastung wurde getrennt für intrinsische kognitive Belastung und extrinsische kognitive Belastung untersucht. Für beide Bereiche zeigen sich signifikante Unterschiede zugunsten der Experimentalgruppe. Lernende der Experimentalgruppe nehmen sowohl die intrinsische kognitive Belastung mit kleinem Effekt ($p = 0.00060$, $d = 0.47$) als auch die extrinsische kognitive Belastung mit mittlerem Effekt ($p = 0.0001$, $d = 0.59$) geringer wahr als Lernende der Kontrollgruppe.

Auswirkungen KI-generierter Erklärungen auf affektive Variablen

Darüber hinaus untersuchten die Autorin und die Autoren mehrere affektive Variablen.

Situatives Interesse

Für das situative Interesse zeigt sich ein signifikanter Unterschied mit kleinem Effekt zugunsten der Experimentalgruppe ($p = 0.00223$; $d = 0.45$). Lernende der Experimentalgruppe berichteten ein höheres situatives Interesse als Lernende der Kontrollgruppe.

Emotionen

Bei den positiv aktivierenden Emotionen (z. B. Freude oder Zufriedenheit) zeigt sich ebenfalls ein signifikanter Unterschied mit kleinem Effekt zugunsten der Experimentalgruppe ($p = 0.00093$; $d = 0.48$).

Für negativ-deaktivierende Emotionen (z. B. Langeweile oder Frustration) wird hingegen kein signifikanter Unterschied zwischen den Gruppen festgestellt ($p = 0.391$).

Selbstwirksamkeitserwartungen

Die Analyse der Selbstwirksamkeitserwartungen ergibt einen signifikanten Unterschied mit mittlerem Effekt zugunsten der Experimentalgruppe ($p = 0.00001$; $d = 0.63$). Lernende der Experimentalgruppe berichteten höhere Erwartungen, Aufgaben zum behandelten Thema erfolgreich bearbeiten zu können, als Lernende der Kontrollgruppe.

Während sich für mehrere affektive Variablen signifikante Unterschiede zugunsten der Experimentalgruppe zeigen, unterscheiden sich die Testergebnisse beider Gruppen kaum.

Diskussion und Einschätzung

Zum Hintergrund

Die Studie greift ein aktuelles Forschungsdesiderat auf: den pädagogisch sinnvollen Einsatz generativer KI im Unterricht. Während viele Diskussionen über KI im Bildungsbereich spekulativ bleiben, liefert die Untersuchung empirische Evidenz zum Einsatz KI-generierter Erklärungen als ergänzende Lernmaterialien. Die theoretische Einbettung in Ansätze der Cognitive Load Theory sowie

motivationspsychologische Perspektiven erscheint plausibel.

Zum Design

Die randomisierte Zuordnung der Lernenden zu Experimental- und Kontrollgruppe stärkt die interne Validität der Untersuchung, jedoch fehlen zentrale Informationen zur Stichprobenzusammensetzung und zur konkreten Randomisierung. In der Studie wird zwar von einer zufälligen Zuweisung der Schülerinnen und Schüler zu Experimental- und Kontrollgruppe berichtet, es fehlen jedoch Angaben dazu, aus wie vielen Schulen oder Lerngruppen die Teilnehmenden stammen und ob ganze Klassen einer Bedingung zugeordnet wurden. Dadurch bleibt unklar, inwieweit mögliche Klassen- oder Schuleffekte kontrolliert werden konnten.

Andererseits überprüften die Autorin und die Autoren die Vergleichbarkeit der Gruppen hinsichtlich der letzten Mathematiknote sowie des fachspezifischen Interesses an Mathematik vor Beginn der Intervention. Für beide Merkmale zeigten sich keine signifikanten Unterschiede zwischen Experimental- und Kontrollgruppe, was zumindest im Hinblick auf diese Aspekte für eine vergleichbare Ausgangslage der Gruppen spricht.

Da die KI-generierten Materialien iterativ überarbeitet und durch menschliches Feedback angepasst wurden, lassen sich die beobachteten Effekte nur eingeschränkt auf die Qualität des verwendeten Sprachmodells selbst zurückführen. Vielmehr dürften die Ergebnisse auch durch didaktische Entscheidungen bei der Auswahl und Gestaltung der ergänzenden Lernmaterialien beeinflusst worden sein. In der Studie werden hierbei insbesondere strukturierende Erklärungen, adressatengerechte Formulierungen sowie die Unterstützung beim Umgang mit unterschiedlichen Darstellungsformen hervorgehoben.

Die interne Konsistenz der eingesetzten Skalen wurde mithilfe eines statistischen Maßes zur Überprüfung der Zuverlässigkeit von Fragebogenskalen (Cronbachs Alpha) getestet. Dies ist insbesondere vor dem Hintergrund relevant, dass mehrere Instrumente übersetzt, gekürzt und an den Untersuchungskontext angepasst wurden. Die berichteten Reliabilitätswerte sprechen insgesamt für eine hohe Zuverlässigkeit der Skalen.

Für eine graphische Teilaufgabe erfolgte eine unabhängige Zweitbewertung. Die hohe Übereinstimmung der beiden Bewertungen (gewichtetes Cohen's Kappa: $\kappa = 0.8469$) spricht für eine insgesamt konsistente Kodierung der Leistungen.

Auffällig ist (zudem) der unterschiedliche Umgang der Autorin und der Autoren mit nicht-signifikanten Befunden. Während bei den negativ-deaktivierenden Emotionen trotz fehlender Signifikanz auf eine deskriptive Tendenz zugunsten der Experimentalgruppe hingewiesen wird, werden mögliche deskriptive Unterschiede bei der Lernleistung deutlich zurückhaltender eingeordnet. Die in einer Abbildung (Figur 6) dargestellten Mittelwerte deuten eher auf leicht höhere Testwerte der Kontrollgruppe hin. Die dargestellten Werte zeigen jedoch auch, dass daraus kein belastbarer Gruppenunterschied abgeleitet werden kann; in der Studie wird dieser Befund jedoch nur begrenzt diskutiert. Zudem wird keine Effektstärke für die Lernleistung berichtet. Dies führt an diesen Stellen zu einer ungleichen Interpretation nicht-signifikanter Befunde.

Zu den Ergebnissen

Die Ergebnisse sind besonders interessant, weil sie keinen direkten Leistungszuwachs, aber positive Effekte auf Motivation, Selbstwirksamkeitserwartungen und wahrgenommene kognitive Belastung

zeigen. Sie lassen vermuten, dass KI-generierten Erklärungen von Lernenden als kognitiv entlastend wahrgenommen werden. Als mögliche Erklärung wird für die Studie diskutiert, dass die Materialien der Experimentalgruppe verständlicher strukturiert und stärker an das Wissensniveau der Lernenden angepasst gewesen seien.

Damit stehen die Ergebnisse zwar im Einklang mit Befunden der Lernforschung, nach denen motivationale Variablen und wahrgenommene Verständlichkeit wichtige Bedingungen von Lernprozessen darstellen, lassen jedoch offen, warum kein zusätzlicher Leistungszuwachs beobachtet werden konnte.

Für den schulischen Kontext könnte dies bedeuten, dass KI-generierte ergänzende Lernmaterialien vor allem als unterstützende Lernhilfen sinnvoll sind, etwa um komplexe Inhalte verständlicher zu machen oder Lernende stärker zu motivieren. Gleichzeitig verdeutlichen die Ergebnisse, dass technologische Innovation allein nicht automatisch zu signifikant besseren Leistungen führt. Entscheidend bleibt die didaktische Einbettung in den Unterricht.

Die Aussagekraft der Ergebnisse bleibt insofern eingeschränkt, als es sich um eine kurzfristige Intervention in einem spezifischen Themenbereich mit nur einer Altersgruppe handelt. Zudem wurden KI-Materialien ausschließlich in Form ergänzender Erklärungen eingesetzt. Langfristige Effekte sowie Auswirkungen interaktiver KI-Lernumgebungen bleiben daher offen.

Weitere Forschung sollte insbesondere langfristige Effekte KI-gestützter Lernmaterialien, unterschiedliche Formen der Interaktion mit Chatbots sowie Auswirkungen auf verschiedene Altersgruppen und Fachkontexte untersuchen.

Weiterführende Hinweise im Kontext

Adiguzel, T., Kaya, M. H. & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Kotsis, K. T. (2024). Correcting students' misconceptions in physics using experiments designed by ChatGPT. *European Journal of Contemporary Education and E-Learning, 2*(2), 83–100. https://doi.org/10.59324/ejceel.2024.2(2).07

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Reflexionsfragen für die Praxis

Nachfolgende Reflexionsfragen sind ein Angebot, die Befunde der rezensierten Studie auf das eigene Handeln als Lehrkraft oder Schulleitungsmitglied zu beziehen und zu überlegen, inwiefern sich

Anregungen für die eigene Handlungspraxis ergeben. Die Befunde der rezensierten Studien sind nicht immer generalisierbar, was z. B. in einer begrenzten Stichprobe begründet ist. Aber auch in diesen Fällen können die Ergebnisse interessante Hinweise liefern, um über die eigene pädagogische und schulentwicklerische Praxis zu reflektieren.

Reflexionsfragen für Lehrkräfte

- In welchen Unterrichtssituationen könnte eine gezielte kognitive Entlastung durch zusätzliche KI-generierte Erklärungen oder strukturierende Lernhilfen sinnvoll sein?
- Wie könnten KI-generierte Lernhilfen zur individuellen Förderung und Binnendifferenzierung genutzt werden, etwa durch unterschiedlich aufbereitete Unterstützungsmaterialien?
- Wie könnten Lernmaterialien so gestaltet werden, dass sie Lernende kognitiv entlasten und zugleich Motivation sowie Selbstwirksamkeit fördern?

Reflexionsfragen für Schulleitungen

- Welche didaktischen Potenziale und Grenzen KI-generierter Lernmaterialien sollten Lehrkräfte erkennen und reflektieren können?
- Wie kann Schule Lehrkräfte dabei unterstützen, KI-gestützte Lernhilfen lernförderlich und fachlich reflektiert einzusetzen?
- Welche organisatorischen und pädagogischen Voraussetzungen braucht unsere Schule, um KI-Angebote sinnvoll in bestehende Unterrichtskonzepte einzubinden?

Rezensent/-in

Dr. Nils Voelzke, Lehrbeauftragter am Institut für Politikwissenschaft der Universität Münster und Oberstudienrat für die Fächer Sozialwissenschaften und Mathematik an der Europaschule Gymnasium Wolbeck in Münster

Zitiervorschlag

Voelzke, N. (2026). Rezension zu Lademann, J., Henze, J. & Becker-Genschow, S. (2025). Augmenting learning environments using AI custom chatbots: Effects on learning performance, cognitive load, and affective variables. Physical Review Physics Education Research, 21, Article 010147. *Forschungsmonitor Schule, 207*. Abgerufen von https://www.forschungsmonitor-schule.de/print.php?id=207

Urheberrecht

Dieser Text steht unter der CC BY-NC-ND 4.0 Lizenz. Der Name des Urhebers / der Urheberin soll bei einer Weiterverwendung wie folgt genannt werden: Nils Voelzke (2026) für den Forschungsmonitor Schule.