MedienPädagogik
Zeitschrift für Theorie und Praxis der Medienbildung
www.medienpaed.com **ISSN 1424-3636**

Themenheft 65: Tagungsband des 1. AK Mediendidaktik der DGfE-Sektion Medienpädagogik (MEDIDA24). Herausgegeben von Maria Klar, Josef Buchner, Barbara Getto, Marco Kalz und Michael Kerres

Bewertung durch eine künstliche Intelligenz?

Auswertungs- und Interpretationsobjektivität von ChatGPT-4o bei der Bewertung von Lerntagebucheinträgen

Lhea Reinhold[1] und Marion Händel[2]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg

[2] Hochschule Ansbach

Zusammenfassung
Künstliche Intelligenz (KI) kann im Prozess der Leistungsbewertung assistieren und diesen transformieren. Besonders lohnend scheint eine KI-Assistenz bei der Bewertung von komplexem, geschriebenem Text. Jedoch ist der Einsatz von KI im Bewertungsprozess «hochriskant» (EU 2024) und bedarf umfangreicher Analysen. Die vorliegende Studie untersucht, inwiefern ChatGPT-4o die Auswertung und Interpretation von Lerntagebucheinträgen objektiv vornehmen kann. Dafür werden 757 Lerntagebucheinträge aus der geförderten Weiterbildung in Deutschland von Mensch und Maschine bewertet. Sowohl Mensch als auch Maschine erhalten hierzu Kriterien, nach denen die Bewertung vorzunehmen ist; ChatGPT-4o wird diesbezüglich mit einem Prompt unterstützt. Die Übereinstimmung der Bewertungen wird anhand der Masse Sensitivität und Spezifität gemessen. Die Ergebnisse zeigen, dass die Bewertungsvorschläge von ChatGPT-4o eine moderate bis hohe Übereinstimmung mit den menschlichen Bewertungen aufweisen; gleichzeitig neigt ChatGPT-4o jedoch zu einer optimistischen Bewertung der Lerntagebucheinträge. Die Ergebnisse weisen darauf hin, dass eine hybride Intelligenz, also eine Kombination der Stärken von Mensch und Maschine, gewinnbringend für Bewertungsprozesse sein kann. Künftig denkbar sind halbautomatisierte Bewertungsprozesse von Lerntagebucheinträgen, in denen die KI die Bewertung der Lerntagebucheinträge übernimmt und Lehrkräfte bei kritischen Fällen regulierend eingreifen. So könnte die Korrektureffizienz ohne bedeutende Qualitätsverluste gesteigert werden.

Reinhold, Lhea, und Marion Händel. 2025. «Bewertung durch eine künstliche Intelligenz? Auswertungs- und Interpretationsobjektivität von ChatGPT-4o bei der Bewertung von Lerntagebucheinträgen». *MedienPädagogik* 65 (MEDIDA24): 227–250. https://doi.org/10.21240/mpaed/65/2025.08.03.X.

Assessment Objectivity of Artificial Intelligence? Analyzing the Evaluation and Interpretation Objectivity of ChatGPT-4o in Learning Journals

Abstract

Artificial Intelligence (AI) can assist in and transform the performance evaluation process. AI assistance seems particularly rewarding when it comes to evaluating complex, written text. However, the use of AI in the evaluation process is considered «high-risk» (EU 2024) and requires extensive analysis. This study examines the extent to which ChatGPT-4o can objectively assess and interpret learning journals. For this purpose, 757 learning journals from government-funded training programs in Germany are assessed by both humans and machine. Both humans and the machine are provided with criteria by which the evaluation is to be carried out; ChatGPT-4o is supported by a prompt. The agreement between the evaluations is measured using sensitivity and specificity metrics. The results show that ChatGPT-4o's evaluation proposals exhibit moderate to high agreement with human evaluations. However, ChatGPT-4o tends to lean towards more optimistic assessments of the learning journals. The findings suggest that hybrid intelligence – a combination of human and machine strengths – can be beneficially applied to evaluation processes. In the future, semi-automated evaluation processes for learning journals could be conceivable, where AI handles the evaluations and educators intervene in critical cases. This could increase correction efficiency without significant loss of quality.

1. Einleitung

Künstliche Intelligenz (KI) schafft neue berufliche Einsatzmöglichkeiten, z. B. Machine Learning Engineer oder KI-Prompter, und verändert Tätigkeitsprofile von Berufen (Grienberger et al. 2024). Die transformatorische Wirkungsweise von KI wird auch im Bildungsbereich diskutiert; besonders die Lehrperson wird durch die schrittweise Implementierung von KI in den Lehr-Lern-Prozess mit veränderten Aufgaben konfrontiert (Celik 2023; Mishra et al. 2023; Molenaar 2021; Ning et al. 2024).

Die Lehrkraft ist eine wichtige Einflussgrösse im Lernprozess des Lernenden (Hattie et al. 2015), und so ist deren Substitution durch KI nicht zu erwarten (Gentile et al. 2023; Lameras und Arnab 2022). Vielmehr kann durch eine Zusammenarbeit von Lehrkraft und KI die Lehrperson in unterschiedlichen Handlungssituationen unterstützt werden. Beispielsweise können KI-Technologien im Prozess der Leistungsbeurteilung assistieren (Ninaus und Sailer 2022).

Inwiefern KI eine Lehrkraft bei der Beurteilung von Leistungen beraten oder unterstützen kann, ist jedoch abhängig von der Qualität der KI-Ausgaben, d. h. in diesem Fall von der Qualität der vorgenommenen Bewertungen der KI. Somit adressiert die vorliegende quantitative Studie die Fragestellung, inwiefern KI eine

objektive Leistungsbewertung abbilden kann. Sie zielt darauf ab, die Diskrepanz zwischen theoretisch möglichen und tatsächlichen KI-Tätigkeiten in Bildungskontexten zu minimieren (Zhang und Aslan 2021).

Die Fragestellung wird im Kontext der geförderten Weiterbildung in Deutschland untersucht, indem die von den Lernenden erstellten prüfungsrelevanten Lerntagebucheinträge von einer Lehrkraft und zeitgleich von ChatGPT-4o bewertet werden. Die Ergebnisse ermöglichen eine vergleichende Analyse der Bewertungen durch Mensch und Maschine. Dem Vorschlag von Alers et al. (2024) folgend, werden die Möglichkeiten von KI in verschiedenen Prüfungssituationen ausgelotet und untersucht.

2. Theoretische Konzepte

2.1 Subjektivierte Bewertungsobjektivität

Eine Leistung zu bewerten bzw. zu benoten, meint, eine durch Dritte durchgeführte kriterienbasierte Entscheidung über die Kompetenz des Lernenden zu treffen (Jönsson et al. 2021). Zum jetzigen Zeitpunkt impliziert «Dritte», dass dem Bewertungsprozess ein menschliches Urteil zugrunde liegt.

Mit dem Ziel, ein reliables und valides Urteil vorzunehmen, muss der Bewertungsprozess einer Leistung objektiviert werden, d. h. die Durchführung, Auswertung und Interpretation einer Leistung muss reproduzierbar sein bzw. standardisiert verlaufen: (1) Die Durchführungsobjektivität ist dabei hoch, wenn für alle Lernenden möglichst identische Testbedingungen vorherrschen. (2) Eine hohe Auswertungsobjektivität wird realisiert, indem die Bewertungskriterien für die Lernenden gleichermassen angewendet werden. (3) Eine hohe Interpretationsobjektivität ist gegeben, wenn dieselben Ergebnisse in derselben Bewertung bzw. Entscheidung münden (Elsäßer 2017; Lerche 2022).

Die Messung der Auswertungs- und Interpretationsobjektivität kann mittels intrapersonaler und interpersoneller Übereinstimmung erfolgen. Bei der intrapersonalen Überstimmung kontrolliert die Lehrkraft sich selbst, indem sie die zu beurteilende Leistung nochmals kritisch prüft. Im Gegensatz dazu meint die interpersonelle Übereinstimmung, dass mehrere Lehrkräfte unabhängig voneinander dieselbe Leistungserbringung bewerten. Aus Zeitgründen wird davon im Lehr- und Unterrichtsalltag in aller Regel abgesehen und dieses aufwendige Verfahren lediglich dann praktiziert, wenn durch die Bewertung der schulische oder berufliche Werdegang beeinflusst wird wie bspw. bei Abschlussprüfungen (Kircher et al. 2020). Gleichzeitig zeigen prominente Studien, dass die Vergabe von Noten nicht standardisiert verläuft und «Subjektivismen» (Benischek et al. 2012) die Auswertungs- und

Interpretationsobjektivität beeinflussen (Ingenkamp und Lissmann 2008). Dies ist besonders bei divergenten Aufgaben zu beobachten, welche durch mehrere Lösungswege charakterisiert sind (Ehlers et al. 2013; Sonnenschein 2015). In einer Studie von Birkel und Birkel (2002) wurden bspw. dieselben Texte von unterschiedlichen Lehrkräften bewertet; die vergebenen Noten umfassten nahezu die gesamte Notenskala. Das Erfahrungswissen einer Lehrkraft trägt allerdings nicht per se zu einer «objektiveren» Bewertung bei: In einer Studie, die sich mit der Bewertung von online geschriebenen Texten von Neuntklässler:innen befasste (Möller et al. 2022), zeigte sich, dass die erfahrenen Lehrkräfte ungenauer und heterogener bewerteten als Lehramtsstudierende. Ein Erklärungsansatz für solch unterschiedliche Bewertungen ist der Einbezug von personenbezogenen Faktoren, die sich nicht an definierten Bewertungskriterien orientieren. So berücksichtigt eine Lehrkraft möglicherweise die Anstrengung und das Verhalten der zu bewertenden Person, anstatt sich ausschliesslich auf inhaltliche Kriterien zu fokussieren (Brookhart et al. 2016). Neben dem Einbezug der Entwicklung des Lernenden beeinflussen weitere Faktoren, z. B. die Lehrkraft selbst, die Urteilsart und Testcharakteristika, die Bewertung der Lehrperson. In einer Metaanalyse korrelativer Studien zeigte sich, dass die Beurteilungsgenauigkeit von Lehrkräften noch deutliches Verbesserungspotenzial aufweist und höher ausfällt, wenn Lehrkräfte informierte Urteile abgeben, d. h. den Vergleichsstandard kennen (Südkamp et al. 2012).

In der Berufspraxis wird der Bewertungsprozess so nachweislich durch den Faktor Mensch subjektiviert. Daher könnten computergestützte, automatisierte Bewertungsverfahren Abhilfe schaffen. Automatisierte Bewertungssysteme für textbasierte Aufsätze (AES) werden in standardisierten Tests wie dem TOEFL eingesetzt, nutzen zur Bewertung Trainingsdaten, d. h. vom Menschen bereits bewertete Aufsätze, und wenden diese mittels regelbasierter, statistischer Modelle oder neuronaler Netzwerke an. So untersuchen sie z. B. Wortzahl, Satzlänge, Grammatik und Satzstruktur (Hussein et al. 2019). Maschinelles Lernen und Deep Learning beziehen zudem inhaltliche und semantische Merkmale des Textes ein (Ramesh und Sanampudi 2022). Mit der Bereitstellung von ChatGPT und der Integration von Large Language Models (LLM) in AES konnte die Bewertungsgenauigkeit von AES verbessert werden: Mizumoto und Eguchi 2023 zeigten, dass die höchste Bewertungsgenauigkeit durch die Kombination aus regelbasierten linguistischen Analysen des AES und ChatGPT-3 erzielt wurde. Während die inhaltliche Qualität vom LLM analysiert wurde, bewertete das regelbasierte System die Wort-, Satz- und Textebene, wodurch die Bewertungsqualität des AES stieg (Li und Ng 2024). Folgerichtig liegt die Vermutung nahe, dass der Einsatz von KI zu einer Objektivierung der Bewertungspraxis beitragen könnte.

2.2 Der Einsatz von ChatGPT als Korrektor

ChatGPT ist ein Sprachmodell, bereitgestellt von OpenAI, und ermöglicht mittels Natural Language Processing-Algorithmen die Verarbeitung und Generierung von Sprache (Pinto et al. 2023). Der bewusste Einsatz von Sprachmodellen im Bildungsbereich, speziell in Prüfungskontexten, ist aufgrund des AI Acts der Europäischen Kommission gesetzlich vorgeschrieben (EU 2024). Beeinflusst ein KI-System den beruflichen Werdegang einer Person, z. B. durch die Bewertung einer Leistung, ist das KI-System als «hochriskant» (ebd.) einzustufen.

Die im Folgenden diskutierten empirischen Studien sind in diesen rechtlichen Rahmen eingebettet und zeigen exemplarisch die Möglichkeiten und Grenzen von KI bei der Bewertung von Leistungen auf (González-Calatayud et al. 2021): So wurden bereits offene Fragen im Bereich der Informations- und Kommunikationstechnologie (Alers et al. 2024), Programmieraufgaben (Jukiewicz 2024), Aufsätze in der Politikwissenschaft (Lundgren 2024) und kurze schriftliche Aufgaben zur Globalisierung und Privatisierung öffentlicher Güter (Kooli und Yusuf 2024) mithilfe verschiedener ChatGPT-Modelle bewertet. Untersuchungsgegenstand war studienübergreifend die fachliche Leistung von Studierenden, wobei alle Studien die Bewertungen von Mensch und Maschine getrennt voneinander durchführten.

Die Ergebnisse der Untersuchungen zeigten moderate bis starke Korrelationen zwischen den Bewertungen durch Mensch und Maschine. Beispielsweise erfasste Jukiewicz (2024) eine hohe positive Korrelation ($r = .81$) bei der Bewertung von Programmieraufgaben mit dem Einsatz von ChatGPT-3.5-turbo. Eine moderate Übereinstimmung ($r = .46$) wurde in der Studie von Kooli und Yusuf (2024) bei der Analyse mit ChatGPT-3.5 festgestellt. In der Studie von Alers et al. (2024) vergab ChatGPT-4 in fast 30% der Fälle dieselbe Note wie die Lehrkraft; in ca. 82% der Fälle über das Bestehen und Nichtbestehen einer Prüfungsleistung entschied sich ChatGPT-4 wie die Lehrkraft.

Teils wurden jedoch entweder grosse oder kleine Standardabweichungen berichtet: Kooli und Yusuf (2024) stellten eine grössere Streuung in den KI-generierten Bewertungen fest und interpretierten diese unter Einbezug der moderaten positiven Korrelation als Indiz für einen «objektiveren» Bewertungsprozess der KI. In der Studie von Lundgren (2024) wurden geringe Standardabweichungen bei den KI-generierten Punktzahlen ermittelt, welche auf eine Tendenz zur Mitte bzw. zu einem «höflichen» oder «optimistischen» Bewertungsstil der KI hindeuteten.

Unabhängig vom rechtlichen Rahmen zeigen die ausgewählten Studien, dass zum jetzigen Zeitpunkt von einem vollautomatisierten Bewertungsprozess durch KI abgesehen werden muss. Eine Möglichkeit, die durch die empirischen Studien identifizierten Potenziale zu erschliessen, ist ein halbautomatisierter Bewertungsprozess, welcher eine hybride Intelligenz einbindet. Dabei meint hybride Intelligenz, dass die sich ergänzenden Stärken von menschlicher und künstlicher Intelligenz

genutzt werden, sodass diese gemeinsam eine bessere Leistung erbringen können als separat voneinander (Dellermann et al. 2019). Das Verhältnis der Zusammenarbeit kann variieren und ist abhängig von der Aufgabenverteilung zwischen Mensch und Maschine.

Molenaar (2021) zufolge gibt es sechs verschiedene Kooperationsmöglichkeiten zwischen Mensch und Technologie, welche prozessual aufeinander aufbauen (vgl. Abb. 1). Die sechs Phasen werden nachfolgend anhand des Automatisierungsgrads des Bewertungsprozesses exemplarisch dargestellt:

- In der Phase «Teacher only» hat die Lehrkraft die volle Kontrolle über den Lehr-Lern-Prozess. Der:Die Korrektor:in führt die Bewertung einer Lernaufgabe ohne KI durch.
- In der Phase «Teacher assistance» unterstützt bzw. assistiert die Technologie erstmalig. Die Technologie berät den:die Korrektor:in bei der Bewertung, indem KI einen Bewertungsvorschlag für die Lernaufgabe gibt. Der:Die Korrektor:in kann diesen Bewertungsvorschlag annehmen oder anpassen.
- Die dritte Phase, «Partial automation», wird umgesetzt, indem die Technologie spezifische Aufgaben ausführt. Die Lehrkraft überwacht die Technologie gänzlich. KI kann z. B. einen Teil der Lernaufgabe eigenverantwortlich bewerten.
- In der vierten Phase «Conditional automation» übernimmt die Technologie einen Grossteil aller Aufgaben, jedoch unter fortlaufender, stichprobenartiger Kontrolle durch die Lehrkraft, welche jederzeit regulierend eingreifen kann. Zudem kann die Technologie die Unterstützung der Lehrkraft im Rahmen vordefinierter Regeln gezielt anfordern. Würde bspw. eine Lernaufgabe als «nicht bestanden» charakterisiert werden, signalisiert KI dem:der Korrektor:in, dass es einer menschlichen Überprüfung bedarf.
- Die Phase «High automation» ist definiert über eine nahezu autonom handelnde Technologie; die Lehrkraft überwacht die Technologie nicht. So übernimmt KI die Bewertung der Lernaufgabe vollständig. Lediglich in Ausnahmefällen signalisiert die KI dem:der Korrektor:in, dass eine menschliche Überprüfung erforderlich ist, z. B. wenn eine Lernaufgabe mit 0 Punkten bewertet würde.
- In der sechsten Phase «Full automation» übernimmt die Technologie alle Aufgaben autonom; das Eingreifen einer Lehrkraft ist nicht erforderlich. Alle Bewertungen werden von der KI automatisch und ohne die Überprüfung eines:einer Korrektor:in an den Lernenden weitergegeben.

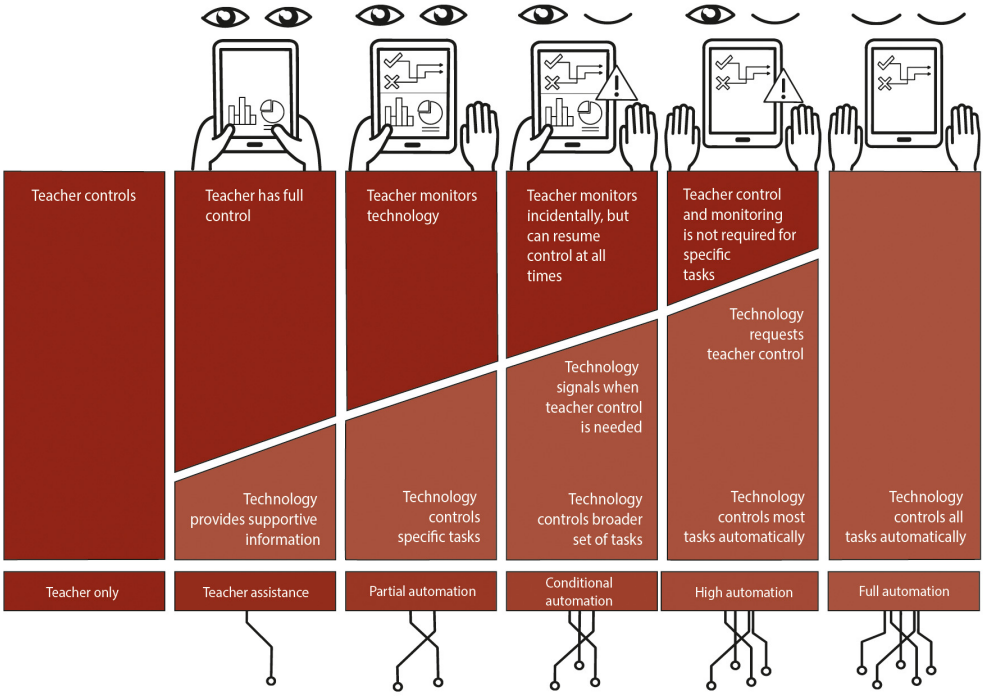

Abb. 1: Sechs Stufen des Automatisierungsmodells für personalisiertes Lernen (Molenaar 2022, 636).

Die Phasen «Teacher assistance» bis «High automation» können dem Konzept der hybriden Intelligenz zugeordnet werden. Arbeiten Mensch und Maschine im Korrekturprozess zusammen, gilt, dass die Gründe der KI für die Bewertung transparent aufgezeigt werden müssen. Ohne die dazugehörige Erklärung würde Swiecki et al. (2022) zufolge die Lehrkraft dazu tendieren, einen KI-gestützten Vorschlag ohne dazugehörige Erklärung anzunehmen. Ein Aushöhlungsprozess – also das unkritische Akzeptieren des KI-generierten Vorschlags – könne so verhindert werden. Des Weiteren bedarf es einer benutzerfreundlichen Implementierung von KI, damit die Bereitschaft erhöht wird, die Technologie zu nutzen (Brüggemann und Wiepcke 2023).

Der Einsatz von KI sollte für alle am Bewertungsprozess Beteiligten vorteilig sein: Lernende könnten zeitnah individuelles Feedback erhalten, daraufhin ihren Lernprozess anpassen und den Lernoutput z. B. durch das frühzeitige Aufdecken von Fehlkonzepten verbessern. Lehrende könnten die eingesparte Zeit für das Initiieren von personalisierten Lehr-Lern-Prozessen nutzen. Besonders formativ angelegte Prüfungsformate, die einen erhöhten Korrekturaufwand für Lehrpersonen bedeuten, könnten dadurch häufiger eingesetzt werden. Lerntagebücher sind ein solch formatives Prüfungsformat, welches zugleich durch mehrere Lösungsmöglichkeiten gekennzeichnet ist. Inwiefern KI auf Lerntagebücher und deren Ergebnisoffenheit reagiert bzw. diese kriterienbasiert bewertet, ist bislang noch nicht untersucht worden.

2.3 Lerntagebücher

«Lerntagebücher dienen der kontinuierlichen Erfassung des Lernprozesses und werden von den Lernenden selbst geführt» (Spörer und Brunstein 2006, 154). So setzen sich Lernende mittels Lerntagebüchern regelmässig und gegebenenfalls frühzeitig mit dem Lernstoff auseinander, reflektieren und optimieren ihr eigenes Lernverhalten. Im Hochschulkontext wurden Lerntagebücher als formatives Prüfungsformat bereits eingesetzt und untersucht (Bürgermeister et al. 2021; Wilkens 2020). Der Aufbau dieser Lerntagebücher war meist offen und vorstrukturiert: Offen meint, dass Lernende einen Fliesstext als Antwort formulieren sollten (Naujoks und Händel 2020). Vorstrukturiert bedeutet, dass z. B. Leitfragen die Lernenden in der Erstellung des Lerntagebuchs unterstützen. Dies ist lernförderlich, da Lernende meist Lernstrategien nicht spontan anwenden können (Nückles et al. 2010); sie dienen als «strategische Aktivatoren» (Nückles et al. 2020, 1102).

Die dabei anzuwendenden Lernstrategien können in kognitive und metakognitive eingeteilt werden: *Kognitive Lernstrategien*, also Organisations- oder Elaborationsstrategien, unterstützen die unmittelbare «Aufnahme und Verarbeitung von Informationen» (Naujoks und Händel 2020, 223). Dabei ermöglichen Organisationsstrategien die Transformation der Lerninhalte in ein für die Lernenden nachvollziehbares Lernformat und Elaborationsstrategien die Anbindung der Lerninhalte an bereits vorhandene Wissensstrukturen. Mit anderen Worten: Werden die Inhalte durch die Lernenden *strukturiert*, z. B. durch die Identifizierung und Beschreibung von Hauptthemen, werden Organisationsstrategien angewendet. *Verknüpfen* die Lernenden die Lerninhalte mit Beispielen und *wenden* diese *an*, indem sie bspw. die berufliche Relevanz herausarbeiten, werden Elaborationsstrategien genutzt (Wilkens 2020). Metakognitive Lernstrategien werden zur Überwachung und Regulation des Verstehensprozesses eingesetzt. So können durch die aktive Identifikation von Wissenslücken und Verständnisschwierigkeiten diese geschlossen werden, indem die Lernenden idealerweise kognitive Lernstrategien nutzen, um die offen gebliebenen Fragen zu beantworten (Nückles et al. 2020). *Reflektieren* die Lernenden die Inhalte, z. B. indem sie offene oder weiterführende Fragen formulieren und Lösungsvorschläge skizzieren, werden metakognitive Lernstrategien angewendet.

Folgerichtig sind Lerntagebücher eine Möglichkeit, regelmässig kognitive und metakognitive Lernstrategien in den Lernprozess zu integrieren, indem die Lernenden die Lerninhalte strukturieren, verknüpfen und anwenden sowie ihr Lernverhalten reflektieren. Beantworten Lernende die Leitfragen des Lerntagebuchs, das bspw. Wilkens (2020) skizziert, kann eine hohe Lernwirksamkeit erzielt (Nückles et al. 2010), die Entlastung des Arbeitsgedächtnisses gefördert (Nückles et al. 2020) und der Lernoutput verbessert werden (Glogger et al. 2012).

3. Methodisches Vorgehen

3.1 Zielstellung

Das Ziel der vorliegenden Studie ist, die Bewertungsobjektivität von ChatGPT-4o am Beispiel prüfungsrelevanter Lerntagebucheinträge der Erwachsenenbildung zu quantifizieren. Zur Messung der Bewertungsobjektivität werden die Auswertungs- und Interpretationsobjektivität von ChatGPT-4o untersucht, indem die Bewertungen – vorgenommen von Mensch und KI – gegenübergestellt werden. Indem die Bewertung des Menschen als Referenzgrösse angenommen wird, können folgende Fragen zur Auswertungs- (F1 und F2) und Interpretationsobjektivität (F3 und F4) adressiert werden. Dabei adressieren F1 und F3 jeweils die Sensitivität, F2 und F4 die Spezifität. Diese Masse beschreiben die Wahrscheinlichkeit, mit der Mensch und KI in der Vergabe bzw. Nicht-Vergabe von Bewertungen übereinstimmen:

F1: Wie hoch ist die Wahrscheinlichkeit, dass ChatGPT-4o eine vom Menschen vorgenommene Punktzahl für die Leistung ebenfalls vergibt?

F2: Wie hoch ist die Wahrscheinlichkeit, dass ChatGPT-4o eine vom Menschen nicht vorgenommene Punktzahl für die Leistung ebenfalls nicht vergibt?

F3: Wie hoch ist die Wahrscheinlichkeit, dass ChatGPT-4o einen vom Menschen als «bestanden» bewerteten Lerntagebucheintrag ebenfalls als «bestanden» bewertet?

F4: Wie hoch ist die Wahrscheinlichkeit, dass ChatGPT-4o einen vom Menschen als «nicht bestanden» bewerteten Lerntagebucheintrag ebenfalls als «nicht bestanden» bewertet?

3.2 Untersuchungsgegenstand

In der Studie wurden digitale Lerntagebucheinträge untersucht. Diese waren offen, d. h. die Lernenden dokumentierten und reflektierten ihr Lernverhalten mit ihren eigenen Worten. Die Erstellung eines Lerntagebucheintrags wurde durch vier Kategorien vorstrukturiert, welche bei jeder digitalen Abgabe von den Lernenden bearbeitet werden mussten (Nückles et al. 2020; Wilkens 2020):

1. In «Einordnen und Strukturieren» erklären die Lernenden die zentralen Lerninhalte mit eigenen Worten.
2. In «Verstehen und Verknüpfen» stellen sie die ausgewählten Lerninhalte mit eigenen Beispielen dar.
3. In «Anwenden und Bewerten» erörtern sie die Bedeutsamkeit der Lerninhalte für die eigene berufliche Zukunft.
4. In «Reflektieren und Hinterfragen» formulieren sie offene oder weiterführende Fragen.

Pro Lerntagebuch-Kategorie kann die Punktzahl 0, 1 und 2 erreicht werden; diese kleinschrittige Korrektur ermöglicht die Umsetzung eines «analytischen Bewertungsmodells» (Jönsson et al. 2021, 216). Die Punktzahl orientiert sich an den folgenden Bewertungskriterien:

- 2 Punkte werden erreicht, wenn die Lerntagebuch-Kategorie in hoher Qualität umgesetzt wird. Es findet eine begründete und differenzierte Auseinandersetzung mit den Lerninhalten statt.
- 1 Punkt wird erreicht, wenn die Lerntagebuch-Kategorie in ausreichender Qualität umgesetzt wird. Es findet eine oberflächliche Auseinandersetzung mit den Lerninhalten statt.
- 0 Punkte werden erreicht, wenn die Lerntagebuch-Kategorie in unzureichender Qualität umgesetzt wird. Es findet keine Auseinandersetzung mit den Lerninhalten statt.

Lernende können in allen vier Lerntagebuch-Kategorien maximal acht Punkte erreichen. Zusätzlich erhalten sie maximal zwei Punkte, wenn die Bewertungskriterien – Einhaltung der deutschen Rechtschreibung und fristgerechte Abgabe – erfüllt werden.

3.3 Studiendesign

Die Studie wurde bei velpTEC GmbH, einem Bildungsinstitut der Erwachsenenbildung, durchgeführt. Die digitalen Lerntagebucheinträge wurden als formatives Prüfungsformat eingesetzt; die Lernenden sollten in Abhängigkeit von der Länge ihrer geförderten Weiterbildungsmassnahme mindestens zwei und maximal elf Lerntagebucheinträge schreiben. Die Studie wurde durch die Datenschutzbeauftragten des Weiterbildungsinstituts sowie der Friedrich-Alexander-Universität Erlangen-Nürnberg geprüft und freigegeben.

An der Studie nahmen sieben Korrektor:innen teil. Sie lernten im April 2024 den Lerntagebucheintrag als Prüfungsformat und die zugrunde liegenden Bewertungskriterien kennen (Wilkens 2020). Gleichzeitig wurden Musterlösungen zur Verfügung gestellt (sog. informiertes Urteil; Südkamp et al. 2012) und diese in wöchentlichen Workshops erörtert. In zusätzlichen Einzelsessions besprachen ein Experte und jede:r Korrektor:in die zu Übungszwecken zur Verfügung gestellten Lerntagebucheinträge; individuelle Rückfragen zum Prüfungsformat und den Bewertungskriterien konnten an den Experten gestellt werden. Zwei wichtige Grundsätze zur

Durchführung der Bewertung wurden eingeübt: Jede Lerntagebuch-Kategorie wird trennscharf von den anderen bewertet. Es können pro Lerntagebuch-Kategorie 0, 1 oder 2 Punkte vergeben werden. Diese Grundsätze galten fortan für jede Korrekturphase.[1]

Im Mai 2024 durchlief der Bewertungsprozess die Phase «Teacher only» (Molenaar 2021). Das bedeutet, dass die Korrigierenden ohne KI die Bewertung von Lerntagebucheinträgen durchführten. Von Juni bis August 2024 wurde die Phase «Teacher assistance» (Molenaar 2021) umgesetzt. In dieser Phase las ein:e Korrektor:in den abgegebenen Lerntagebucheintrag.[2] Über einen Button «KI-generierter Bewertungsvorschlag» erhielt der:die Korrektor:in die *vorgeschlagene* Bewertung von ChatGPT-4o inkl. Angaben zur Punktzahl pro Lerntagebuch-Kategorie und einer kurzen Erklärung (Swiecki et al. 2022). Diese KI-Ausgaben wurden automatisiert in einen Bewertungsbogen eingetragen. Darin trug der:die Korrektor:in die *tatsächliche*, von ihm:ihr vorgenommene Bewertung pro Lerntagebuch-Kategorie ein. Die Bewertung der Korrigierenden galt als Referenzgrösse und wurde an den:die Lernende:n weitergegeben (vgl. Abb. 2).

Das ChatGPT-4o[3] Modell wurde durch einen vom Bildungsinstitut erstellten Prompt[4] unterstützt (Jacobsen und Weber 2023). Dabei wurden nur die Inhalte des Lerntagebucheintrags und keine weiteren personenbezogenen Daten an die KI mitgegeben. Mit dem Ziel einer praxisnahen Implementierung von KI in den Bewertungsprozess sowie aus Gründen der Datensparsamkeit wird der erste Bewertungsvorschlag von ChatGPT-4o für die Studie verwendet.

1 Die umfassende Vorbereitung der Lehrkräfte wird begründet, da der Lerntagebucheintrag für Bewertende ein hochkomplexes Prüfungsformat darstellt, für welches keine Musterlösung, wie z. B. für Abituraufgaben, ausgegeben werden kann. Zudem ist das Prüfungsformat in der geförderten Weiterbildungsbranche unbekannt. Auch kannten die sieben Korrigierenden es nicht, sodass eine Vorbereitung notwendig war.

2 Anzumerken ist hierzu, dass die korrigierenden Personen jeweils unterschiedliche Lerntagebucheinträge lasen und bewerteten, sodass keine Information zur Übereinstimmung zwischen Lehrkräften vorliegt.

3 Zu Beginn der Studie war ChatGPT-4o das leistungsstärkste (kostenpflichtige) Sprachmodell von OpenAI.

4 Der Prompt ist Eigentum der velpTEC GmbH. Er ist in einem zweimonatigen, iterativen Prozess zwischen einem Experten für Lerntagebücher und einem Prompt Engineer entstanden und nicht zur Veröffentlichung freigegeben.

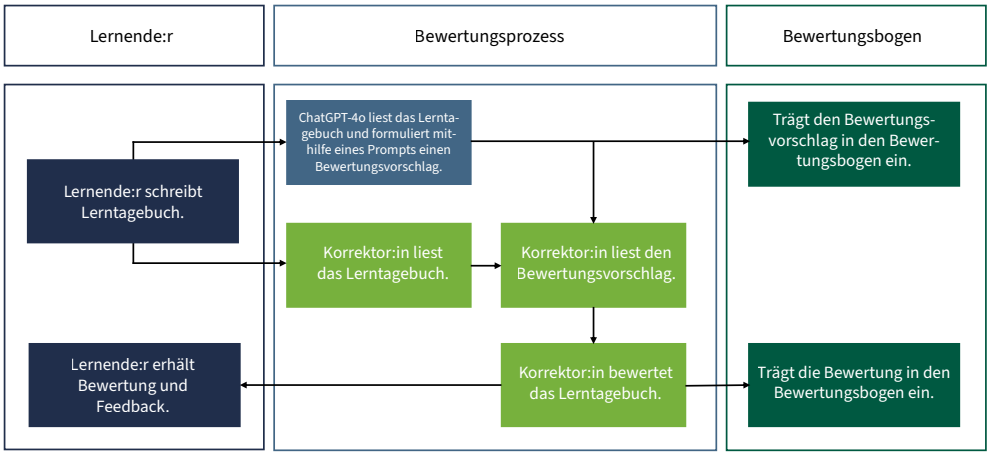

Abb. 2: Der Bewertungsprozess in der Phase «Teacher assistance» bei velpTEC GmbH.

3.4 Stichprobe

Insgesamt wurden im Zeitraum von Juni bis August 757 Lerntagebucheinträge aus einer in Deutschland geförderten Weiterbildungsmassnahme von sieben Lehrkräften sowie von ChatGPT-4o korrigiert. Dabei kam es zu insgesamt 3028 vorgenommenen Bewertungen von Mensch und ChatGPT-4o. Dieser Wert ist das Produkt aus den Faktoren «Anzahl der Lerntagebucheinträge» und «vier Lerntagebuch-Kategorien».

3.5 Auswertungsobjektivität

Zur Quantifizierung der Auswertungsobjektivität werden die Punktzahlen trennscharf und unabhängig von der bewerteten Lerntagebuch-Kategorie analysiert. Folgerichtig entstehen drei Auswertungen: Pro Punktzahl entsteht eine 3x3-Konfusionsmatrix (a, b oder c), welche in Tabelle 1 dargestellt wird (Bernius et al. 2022; Schraw et al. 2013; Tharwat 2018). Die Konfusionsmatrix a repräsentiert die 2-Punkte Bewertung (b: 1-Punkte Bewertung, c: 0-Punkte Bewertung). Mittels der Konfusionsmatrix 3x3 wird der Bewertungsvorschlag von ChatGPT-4o der Bewertung der Korrigierenden gegenübergestellt. Die Bewertung der Korrigierenden gilt als Massstab. Die in den Tabellen genutzten Abkürzungen repräsentieren folgende Werte:

- TP steht für *True Positive*. Die zu untersuchende Punktzahl wird sowohl vom Korrigierenden als auch von ChatGPT-4o vergeben.
- TN steht für *True Negative*. Die zu untersuchende Punktzahl wird sowohl vom Korrigierenden als auch von ChatGPT-4o nicht vergeben.
- FP steht für *False Positive*. Die zu untersuchende Punktzahl wird von ChatGPT-4o vergeben; der Korrigierende vergibt aber eine andere Punktzahl.
- FN steht für *False Negative*. Die zu untersuchende Punktzahl wird vom Korrigierenden vergeben; ChatGPT-4o vergibt diese nicht.

n=x		Korrektor:in								
		2 Punkte			1 Punkt			0 Punkte		
Konfusionsmatrix		**a**	**b**	**c**	**a**	**b**	**c**	**a**	**b**	**c**
ChatGPT-4o	**2 Punkte**	TP	TN	TN	FP	FN	TN	FP	TN	FN
	1 Punkt	FN	FP	TN	TN	TP	TN	TN	FP	FN
	0 Punkte	FN	TN	FP	TN	FN	FP	TN	TN	TP

Tab. 1: Konfusionsmatrix 3x3 (in Anlehnung an Bernius et al. 2022; Schraw et al. 2013; Tharwat 2018).

Die Analyse der Auswertungsobjektivität pro Punktzahl wird mithilfe der Masse Sensitivität und Spezifität vorgenommen (Bernius et al. 2022; Schraw et al. 2013; Tharwat 2018). Die Indizes wurden ausgewählt, da diese bei unausgewogenen Datensätzen aussagekräftig sind (Tharwat 2018) und die Klassifikationsleistung von KI eindeutig messbar machen. Im Gegensatz zu Korrelationsanalysen, welche die lineare Beziehung zweier Variablen angeben, quantifizieren Sensitivität und Spezifität die Bewertungsgenauigkeit in Bezug auf die definierten Punktzahlen.

Die Formeln für Sensitivität und Spezifität nach Schraw et al. (2013) lauten:

$$\text{Sensitivität} = \frac{TP}{(TP + FN)}$$

$$\text{Spezifität} = \frac{TN}{(TN + FP)}$$

Die Werte können jeweils zwischen 0 und 100% variieren; dabei impliziert 0% eine geringe und 100% eine hohe Sensitivität oder Spezifität.

3.6 Interpretationsobjektivität

Die Interpretationsobjektivität wird anhand der Entscheidung «bestanden» oder «nicht bestanden» untersucht. Dabei gilt ein Lerntagebucheintrag als «bestanden», wenn er mit mindestens 5 von 10 möglichen Punkten bewertet worden ist. Folgerichtig ist ein Lerntagebucheintrag «nicht bestanden», wenn er mit weniger als 5 Punkten bewertet worden ist. Die Quantifizierung der Interpretationsobjektivität ermöglicht eine 2x2-Konfusionsmatrix; auch hier gilt die Interpretation des:der Korrigierenden als Massstab.

- TP meint, dass der Lerntagebucheintrag sowohl vom Korrigierenden als auch von ChatGPT-4o als «bestanden» interpretiert wird.
- TN meint, dass der Lerntagebucheintrag weder vom Korrigierenden noch von ChatGPT-4o als «bestanden» interpretiert wird.

- FP meint, dass der Lerntagebucheintrag vom Korrigierenden als «nicht bestanden» interpretiert wird; jedoch interpretiert ChatGPT-4o den Lerntagebucheintrag als «bestanden».
- FN meint, dass der Lerntagebucheintrag vom Korrigierenden als «bestanden» interpretiert wird; jedoch interpretiert ChatGPT-4o den Lerntagebucheintrag als «nicht bestanden».

4. Ergebnisse

4.1 Auswertungsobjektivität

Ziel ist es, die Auswertungsobjektivität von ChatGPT-4o in Abhängigkeit von der 0-, 1- oder 2-Punkte Bewertung zu quantifizieren (siehe F1 und F2). Durchschnittlich hat der Mensch 1.69 Punkte vergeben ($SD = 0.52$), ChatGPT-4o 1.65 Punkte ($SD = 0.54$). Die detaillierten Ergebnisse sind in Tabelle 2 dargestellt; der Chi-Quadrat-Test ergibt einen Wert von $\chi^2 = 2537.38$ ($p < .001$).

n=3028		**Korrektor:in**			
		2 Punkte	**1 Punkt**	**0 Punkte**	**Summe**
ChatGPT-4o	**2 Punkte**	1935	145	4	2084
	1 Punkt	219	598	22	839
	0 Punkte	11	34	60	105
	Summe	2165	777	86	3028

Tab. 2: Konfusionsmatrix 3x3 mit Daten.

Am Beispiel der 2-Punkte-Bewertung (vgl. Tabelle 1, Konfusionsmatrix a) sind die Daten wie folgt auszuwerten: Insgesamt haben Mensch und Maschine 3028 Bewertungen vorgenommen. In 1935 Fällen haben Korrektor:innen und ChatGPT-4o beide die 2 Punkte vergeben (TP). In 714 (598+22+34+60) Fällen hat weder der Mensch noch ChatGPT-4o 2 Punkte vergeben (TN). In 145 Fällen vergab der:die Korrektor:in eine 1-Punkt-Bewertung und in 4 Fällen 0 Punkte, während ChatGPT-4o in diesen 149 (145+4) Fällen die 2 Punkte vergab (FP). In 230 (219+11) Fällen hat der Mensch 2 Punkte vergeben; in diesen Fällen hat ChatGPT-4o 219-mal einen Punkt und 11-mal 0 Punkte gegeben (FN).

False Positive ist immer die Summe aus zwei Werten, z. B. 145+4 bei der 2-Punkte Bewertung (vgl. Tabelle 1, Konfusionsmatrix a) und 11+34 bei der 0-Punkte Bewertung (vgl. Tabelle 1, Konfusionsmatrix c). Sowohl bei der 2- als auch bei der 0-Punkte Bewertung resultiert der grössere Summand aus dem Zusammenhang, dass

ChatGPT-4o 2 oder 0 Punkte vergibt, obwohl der Mensch 1 Punkt vergab. Ähnliches gilt für False Negative. Der grössere Summand, z. B. 219 bei FN der 2-Punkte Bewertung (vgl. Tabelle 1, Konfusionsmatrix a: 219+11), entsteht, indem ChatGPT-4o 1 Punkt vergibt, obwohl der:die Korrektor:in die Leistung mit 2 Punkten bewertete. Dasselbe gilt für FN der 0-Punkte Bewertung (vgl. Tabelle 1, Konfusionsmatrix c): In 22 Fällen vergab ChatGPT-4o 1 Punkt und in vier Interkationen 2 Punkte; der Mensch nahm eine 0-Punkte Bewertung vor.

Die Werte der Sensitivität (F1) und Spezifität (F2) können Tabelle 3 entnommen werden. Bei der 2-Punkte Bewertung ist der Sensitivitäts-Wert grösser als der Spezifitäts-Wert. Im Gegensatz dazu sind die Spezifitäts-Werte der 1- und 0-Punkte Bewertungen grösser als die Sensitivitäts-Werte.

n=3028	Sensitivität	Spezifität
2 Punkte	89.4%	82.9%
1 Punkt	77.0%	89.3%
0 Punkte	69.8%	98.5%

Tab. 3: Sensitivität (F1) und Spezifität (F2) in Abhängigkeit von der Punktzahl.

4.2 Interpretationsobjektivität

Zur Untersuchung von F3 und F4 wurde die Interpretationsobjektivität der KI-generierten Bewertung in Abhängigkeit von der Grenze «bestanden» vs. «nicht bestanden» erneut anhand der Masse Sensitivität und Spezifität quantifiziert (vgl. Tabelle 4).

n=757		Korrektor:in		
		bestanden	**nicht bestanden**	**Summe**
ChatGPT-4o	**bestanden**	633 (TP)	20 (FP)	653
	nicht bestanden	15 (FN)	89 (TN)	104
	Summe	648	109	757

Tab. 4: Konfusionsmatrix 2x2 mit Daten.

Bei einer Anzahl von 757 korrigierten Lerntagebucheinträgen haben der Mensch und ChatGPT-4o 633 Lerntagebucheinträge als «bestanden» (TP) und 89 Lerntagebucheinträge als «nicht bestanden» bewertet (TN). In 15 Fällen hat der Mensch einen Lerntagebucheintrag als «bestanden» bewertet, ChatGPT-4o bewertete diese Leistung der Lernenden als «nicht bestanden» (FN). In 20 Fällen hat die KI die Leistung der Lernenden als «bestanden» bewertet; der:die Korrektor:in bewertete

diese Lerntagebucheinträge als «nicht bestanden» (FP). Die Sensitivität liegt somit bei 97.7% (F3) und die Spezifität bei 81.7% (F4). Der Chi-Quadrat-Test beträgt $\chi^2 = 495.5616$ ($p < .001$).

5. Diskussion

5.1 Auswertungsobjektivität

Die Betrachtung der Mittelwerte und Standardabweichungen des Menschen und von ChatGPT-4o sind ähnlich; die Sensitivitäts- und Spezifitäts-Werte der zu vergebenden Punktzahlen sind moderat bis hoch. Folgerichtig implizieren die Werte, dass ChatGPT-4o mithilfe des dafür entwickelten Prompts die für die Lerntagebucheinträge definierten Bewertungskriterien grundsätzlich versteht und richtig anwendet. ChatGPT-4o kann eine moderate bis hohe Auswertungsobjektivität im punktebezogenen Korrekturprozess von Lerntagebucheinträgen der geförderten Weiterbildung abbilden.

Infolgedessen könnte im Hinblick auf das sechsstufige Automatisierungsmodell nach Molenaar (2021) (s. Abb. 1) die dritte Phase, «Partial automation», initiiert werden. Eine Analyse und Interpretation der Sensitivitäts- und Spezifitäts-Werte pro Lerntagebuch-Kategorie könnte Aufschluss über die unterschiedlichen Auswertungsobjektivitäten pro Kategorie geben. Die Kategorie mit der höchsten Auswertungsobjektivität könnte anschliessend von ChatGPT-4o bewertet und alle anderen Lerntagebuch-Kategorien von einem Menschen korrigiert werden. Bei der Umsetzung der Zusammenarbeit von Lehrkräften mit der KI gilt, dass der Mensch vorbereitet werden muss und kompetent im Umgang mit der Anwendung sein sollte, da sonst keine Entlastung der Lehrkraft bzw. ein Qualitätsverlust in der Bewertung zu erwarten ist (Goh et al. 2024). Würden die Phasen «Conditional automation» oder «High automation» umgesetzt, müsste beachtet werden, dass KI aktuell statistisch signifikant optimistischer in der Bewertung ist: Die Sensitivität der 2-Punkte-Bewertung ist grösser als die Spezifität, sodass ChatGPT-4o häufiger 2 Punkte vergibt, als dass diese Punktzahl gerechtfertigt wäre. Die Sensitivität der 1- und 0-Punkte-Bewertung ist kleiner als die Spezifität, sodass die tatsächliche Vergabe von 1 oder 0 Punkten nicht immer erfolgt. Der Zusammenhang, also die häufig fälschliche Vergabe der höchsten Punktzahl und die Zurückhaltung bei der Vergabe niedrigerer Punktzahlen, impliziert eine den Lernenden zugewandte Bewertung. Im Fall des formativen Prüfungsformats und unter Einbezug der konsequenten Vergabe höherer Punktzahlen kann dies bedeuten, dass gute Leistungen als sehr gut bewertet werden und so die Dokumentation einer positiven Lern*entwicklung* von vornherein

ausbleibt. Es besteht die Möglichkeit, dass aus bildungspolitischer Perspektive das durchschnittliche Lernniveau steigt, das tatsächliche Niveau aber gleich bleibt bzw. aufgrund nicht-diagnostizierter Lerndefizite sinkt.

Mit dem Ziel, das optimistische Anwenden der Bewertungskriterien und die vom Menschen signifikant unterschiedliche Bewertung zu adressieren, könnte eine trennschärfere Abgrenzung von unmittelbar aneinander grenzenden Punktzahlen vorgenommen werden. Die vom Menschen abweichende KI-Vergabe der Punktzahlen tritt nämlich häufiger zwischen 2 und 1 Punkt(en) bzw. 1 und 0 Punkt(en) auf als zwischen 2 und 0 Punkten.

Im Vergleich zu anderen Studien zum Einsatz von KI in Prüfungsformaten sind zwei Interpretationen möglich: (1) Ein Vergleich der vorliegenden Mittelwerte und Standardabweichungen mit den Werten von Lundgren (2024) sowie Kooli und Yusuf (2024) impliziert, dass möglicherweise die im Studiendesign verankerte hybride Intelligenz zu einem «objektiveren» Bewertungsprozess beiträgt. So besteht im Sinne der Definition von hybrider Intelligenz die Möglichkeit, dass die beim Menschen vorliegenden «Subjektivismen» (Benischek et al. 2012) minimiert wurden. (2) Die vorliegenden Ergebnisse können als Fortführung der bereits vorliegenden empirischen Studien gelten. Indem ChatGPT-4o, das leistungsstärkste (kostenpflichtige) Sprachmodell zur Ausführung von komplexen Aufgaben von OpenAI, genutzt wurde und die bisherigen Studien mit ChatGPT-3.5-turbo oder ChatGPT-4 durchgeführt wurden, wird die Lernfähigkeit von Sprachmodellen deutlich. Der gleichberechtigte Zugang zu KI und die kontinuierliche pädagogische Analyse und Interpretation von Sprachmodellen wird in Zukunft erforderlich sein.

5.2 Interpretationsobjektivität

Die Analyse der Interpretationsobjektivität zeigt, dass Mensch und Maschine 98% der 757 untersuchten Lerntagebucheinträgen gleichermassen als «bestanden» interpretieren. Interpretiert der Mensch einen Lerntagbucheintrag als «nicht bestanden», nimmt ChatGPT-4o diese Interpretation in 82% der Fälle vor. Folgerichtig ist die objektive Interpretationsleistung von ChatGPT-4o in Bezug auf bestandene Lerntagebucheinträge zuverlässiger als gegenüber nicht bestandenen.

Aufgrund dieser Werte scheint künftig ein halbautomatisierter Bewertungsprozess möglich, z. B. ähnlich den Phasen «Conditional automation» oder «High automation» (Molenaar 2021), wenn Lernende nicht über die konkrete Punktzahl informiert werden, sondern nur den Status des Lerntagebucheintrags – also «bestanden» oder «nicht bestanden» – einsehen können. Mit dem Ziel, den hohen Sensitivitäts-Wert zu berücksichtigen und gleichzeitig die Anzahl an FN zu minimieren, könnte folgende Regel für den halbautomatisierten Prozess definiert und angewendet werden: «Wird ein Lerntagebucheintrag als ‹nicht bestanden› bewertet, benachrichtige eine

Lehrkraft». So kann die Lehrkraft ihre Aufmerksamkeit darauf konzentrieren, die Lerntagebucheinträge zu prüfen, die von der KI als «nicht bestanden» kategorisiert werden. Würde diese Regel auf die 757 untersuchten Lerntagebucheinträge übertragen, müsste die Lehrkraft hiervon lediglich 104 (ca. 14%) lesen und bewerten. Bei einer durchschnittlichen Bearbeitungszeit von 15 Minuten pro Lerntagebucheintrag, ergibt sich eine Zeitersparnis von ca. 163 Stunden. Um die Anzahl an FP zu reduzieren, könnten zufällig etwa 3% der durch die KI als bestanden kategorisierten Lerntagebucheinträge zur Kontrolle an eine Lehrkraft weitergeleitet werden.

Würde ein vollautomatisierter Bewertungsprozess des formativen Prüfungsformats angestrebt, wäre im besonderen Masse eine Untersuchung der 15 Fälle notwendig, in welchen der Mensch die Leistung als «bestanden» und die KI als «nicht bestanden» charakterisierte (FN). In 13 von 15 Lerntagebucheinträgen traf die abweichende KI-genierte Kategorisierung jeweils einen anderen Lernenden. Für das in dieser Studie untersuchte Setting würde der Abschluss der beruflichen Weiterbildungsmassnahme fast gar nicht negativ durch die KI-gestützte Bewertung beeinflusst, da mehrere Lerntagebucheinträge abzugeben sind. Jedoch wurden zwei von insgesamt vier Lerntagebucheinträgen desselben Lernenden von ChatGPT-4o als «nicht bestanden» charakterisiert, obwohl diese vom Menschen als «bestanden» kategorisiert wurden. Die Untersuchung der 20 Fälle, in welchen die KI eine Leistung als «bestanden» charakterisiert hatte und der:die Korrektor:in den Lerntagebucheintrag als «nicht bestanden» kategorisierte (FP), ergibt 16 Einzelfälle. Das bedeutet, dass die KI-generierte Kategorisierung der Lerntagebucheinträge als «bestanden» einmal pro Lernenden auftrat und die Auswirkungen auf den erfolgreichen Abschluss der Weiterbildung sehr gering wären. Bei einem Lernenden wurden zwei von insgesamt drei Prüfungsleistungen und bei einem anderen Lernenden zwei von fünf Lerntagebucheinträgen von der KI als «bestanden», vom Menschen als «nicht bestanden» charakterisiert. Dass die Lernenden hierfür negative berufsbiografische Konsequenzen zu erwarten haben, ist nicht anzunehmen und somit sind alle 20 FP Fälle im Hinblick auf den AI Act vernachlässigbar. Nichtsdestotrotz gilt, dass für das Individuum selbst wichtige Feedbackprozesse und demzufolge Lernmomente ausbleiben würden; gleichermassen würden die Selektions- und Allokationsfunktion von Prüfungsformaten eingeschränkt.

Im Hinblick auf den AI Act (EU 2024) könnte folgendes gelten: Würden Lerntagebucheinträge als formatives Prüfungsformat in der geförderten Weiterbildung in Deutschland eingesetzt und deren Interpretation, «bestanden» oder «nicht bestanden», an die Lernenden ausgegeben, könnte ein halbautomatisierter Bewertungsprozess umgesetzt werden. Um negative Auswirkungen auf den beruflichen Werdegang der Personen auszuschliessen, müsste der Mensch alle von der KI als «nicht bestanden» bewerteten Leistungen kritisch nachprüfen. Von einem vollständig automatisierten Bewertungsprozess müsste aufgrund der gesetzlichen Regelung

derzeit abgesehen werden, obwohl die Analysen dieser Studie zeigen, dass die Art des Prüfungsformats die vom Menschen abweichenden KI-generierten Bewertungen und deren Einfluss auf den Abschluss der Weiterbildungsmassnahme nahezu vernachlässigbar macht. So geht möglicherweise mit der Implementierung von KI im Bewertungsprozess der Bedarf und somit die erhöhte Nutzung von formativen Assessments einher.

5.3 Limitationen der Studie

Abschliessend sind Limitationen der Studie zu beachten: Die Ergebnisse sind nur auf den spezifischen Fall des Lerntagebucheintrags in der Erwachsenenbildung anwendbar, da der zugrundeliegende Prompt spezifisch für diesen Anwendungsfall entwickelt wurde. Gleichzeitig stellt die Studie den Versuch dar, die Bewertungsobjektivität von ChatGPT-4o anhand einer menschenzentrierten Perspektive zu erschliessen, obwohl das menschliche Urteil nicht gänzlich objektiv ist. Zwar wurden die Bewertungskriterien mithilfe eines Experten und Musterlösungen umfassend mit den korrigierenden Lehrkräften besprochen und eingeübt (Brookhart 2018) und transparent kommuniziert (Reddy und Andrade 2010). Die Beurteilungsgenauigkeit der Lehrkräfte selbst (Südkamp et al. 2012; Möller et al. 2022) wurde jedoch nicht geprüft und es ist unklar, wie stark die Lehrkräfte untereinander in ihren Bewertungen übereinstimmen, da die bewertenden Lehrkräfte jeweils unterschiedliche Lerntagebucheinträge korrigierten. Schliesslich bedeutet die Phase «Teacher only» und die Vorbereitung der Lehrkräfte auf den Bewertungsprozess, dass die Ergebnisse in einer kontrollierten Untersuchungsumgebung entstanden sind. Unklar ist daher, inwiefern die Ergebnisse auf die allgemeine Bewertungspraxis übertragbar sind, in der Lehrkräfte nicht explizit geschult werden.

6. Fazit und Ausblick

Die Studie untersuchte die hybride Intelligenz von Mensch und KI am Beispiel des Bewertungsprozesses von Lerntagebucheinträgen in einer geförderten Weiterbildungsmassnahme in Deutschland. Die Forschungsfrage, inwiefern ChatGPT-4o eine objektive Bewertung durchführen kann, erschloss sich mittels der Analyse von Auswertungs- und Interpretationsobjektivität. Hierfür wurde der KI-generierte Bewertungsvorschlag mit der menschlichen Bewertung verglichen.

Von ChatGPT-4o wurden alle Punktzahlen vergeben. Die Bewertungsvorschläge der KI zeigten eine moderate bis hohe Ähnlichkeit zur tatsächlichen Bewertung durch den:die Korrektor:in auf, was darauf hindeutet, dass ChatGPT-4o die Lehrkraft im Bewertungsprozess künftig unterstützen könnte. Basierend auf den Ergebnissen

zur Interpretationsobjektivität könnte ein halbautomatisierter Bewertungsprozess implementiert werden unter der Bedingung, dass eine Lehrkraft im Fall eines «nicht bestandenen» Lerntagebucheintrages informiert wird und die Bewertung überprüft.

Die Studienergebnisse führen zu vielfältigen Anschlussperspektiven: (1) Die Ergebnisse sollten anhand anderer KI-Sprachmodelle, aber auch mittels mehrerer Lehrkräfte repliziert werden, die dieselben Lerntagebucheinträge bewerten. Somit kann die auf ein Sprachmodell zentrierte Forschungsperspektive erweitert und mit der Urteilsgenauigkeit der Lehrkräfte in Zusammenhang gebracht werden. (2) Im Hinblick auf die Lerntagebuchforschung bietet sich eine differenzierende Analyse der Einträge pro Lerntagebuch-Kategorie an, da die aktuelle Studie lediglich den Gesamtwert untersucht hat: Ziel der vier Lerntagebuch-Kategorien war, unterschiedliche Lernstrategien zu aktivieren. Folglich stellt sich die Frage, inwiefern ChatGPT-4o sowohl kognitive als auch metakognitive Lernstrategien differenzieren und bewerten kann. (3) Die im Forschungsdesign implizierte Interaktion von Korrektor:in und KI könnte zukünftig näher erforscht werden. Mit dem Ziel, die Interaktion und deren Wirkungen zu verstehen, bedarf es der Untersuchung, wie Korrigierende die Zusammenarbeit mit ChatGPT-4o wahrnehmen, wie hoch sie den Nutzen von KI im Bewertungsprozess einschätzen oder auch, welche Veränderungen des Arbeitsalltags mit der Implementierung von KI in den Korrekturprozess einhergehen.

Insgesamt beabsichtigte die Studie, einen transformatorischen, keinen substituierenden Ansatz von KI zu untersuchen, mit dem Ziel, Lehrkräfte bei Korrekturtätigkeiten zu unterstützen.

Literatur

Alers, Hani, Aleksandra Malinowska, Gregory Meghoe, und Enso Apfel. 2024. «Using ChatGPT-4 to Grade Open Question Exams». In *Advances in Information and Communication Proceedings of the 2024 Future of Information and Communication Conference (FICC), Volume 1*, 919: 1–9. Lecture Notes in Networks and Systems. Berlin: Springer Nature. https://doi.org/10.1007/978-3-031-53960-2_1.

Benischek, Isabella, Angela Forstner-Ebhart, Hubert Schaupp, und Herbert Schwetz, Hrsg. 2012. «Empirische Forschung zu schulischen Handlungsfeldern. Ergebnisse der ARGE Bildungsforschung an pädagogischen Hochschulen in Österreich. 2.» Austria: Forschung und Wissenschaft. Erziehungswissenschaft. 15. Berlin u. a.: Lit.

Bernius, Jan Philip, Stephan Krusche, und Bernd Bruegge. 2022. «Machine learning based feedback on textual student answers in large courses». *Computers and Education: Artificial Intelligence* 3 (Januar): 100081. https://doi.org/10.1016/j.caeai.2022.100081.

Birkel, Peter, und Claudia Birkel. 2002. «Wie einig sind sich Lehrer bei der Aufsatz beurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss.» *Psychologie in Erziehung und Unterricht* 49 (3): 219–24.

Brookhart, Susan M. 2018. «Appropriate Criteria: Key to Effective Rubrics». *Frontiers in Education* 3. https://doi.org/10.3389/feduc.2018.00022.

Brookhart, Susan M., Thomas R. Guskey, Alex J. Bowers, James H. McMillan, Jeffrey K. Smith, Lisa F. Smith, Michael T. Stevens, und Megan E. Welsh. 2016. «A Century of Grading Research: Meaning and Value in the Most Common Educational Measure». *Review of Educational Research* 86 (4): 803–48. https://doi.org/10.3102/0034654316672069.

Brüggemann, Tim, und Claudia Wiepcke. 2023. «Der EdTech-Index (ETX)», herausgegeben von Claudia Wiepcke. https://nbn-resolving.org/urn:nbn:de:bsz:751-opus4-4570.

Bürgermeister, Anika, Inga Glogger-Frey, und Henrik Saalbach. 2021. «Supporting Peer Feedback on Learning Strategies: Effects on Self-Efficacy and Feedback Quality». *Psychology Learning & Teaching* 20 (3): 383–404. https://doi.org/10.1177/14757257211016604.

Celik, Ismail. 2023. «Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education». *Computers in Human Behavior* 138 (Januar):107468. https://doi.org/10.1016/j.chb.2022.107468.

Dellermann, Dominik, Philipp Ebel, Matthias Söllner, und Jan Marco Leimeister. 2019. «Hybrid Intelligence». *Business & Information Systems Engineering,*: Vol. 61, No. 5, S.637–43. http://dx.doi.org/10.1007/s12599-019-00595-2.

Ehlers, Jan P., Christian Guetl, Susan Höntzsch, Claus A. Usener, und Susanne Gruttmann. 2013. «Prüfen mit Computer und Internet. Didaktik, Methodik und Organisation von E-Assessment». In *Lehrbuch für Lernen und Lehren mit Technologien*, herausgegeben von Martin Ebner und Sandra Schön, 2. Auflage. Berlin: epubli GmbH: 227–237.

Elsäßer, Sibylle Dorothee. 2017. «Komponenten von schulischen Leistungen: Eine Analyse zu Einflussfaktoren auf die Notengebung in der Grundschule». München: Ludwig-Maximilians-Universität, Elektronische Hochschulschriften. http://nbn-resolving.de/urn:nbn:de:bvb:19-226439.

EU. 2024. *Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 16. Juli 2024*. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32024R1689.

Gentile, Manuel, Giuseppe Città, Salvatore Perna, und Mario Allegra. 2023. «Do we still need teachers? Navigating the paradigm shift of the teacher's role in the AI era». *Frontiers in Education* 8. https://doi.org/10.3389/feduc.2023.1161777.

Glogger, Inga, Rolf Schwonke, Lars Holzäpfel, Matthias Nückles, und Alexander Renkl. 2012. «Learning Strategies Assessed by Journal Writing: Prediction of Learning Outcomes by Quantity, Quality, and Combinations of Learning Strategies». *Journal of Educational Psychology* 104 (Januar):452–68. https://doi.org/10.1037/a0026683.

Goh, Ethan, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, et al. 2024. «Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial». *JAMA Network Open* 7 (10): e2440969–e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969.

González-Calatayud, Víctor, Paz Prendes-Espinosa, und Rosabel Roig-Vila. 2021. «Artificial Intelligence for Student Assessment: A Systematic Review». *Applied Sciences* 11 (12). https://doi.org/10.3390/app11125467.

Grienberger, Katharina, Britta Matthes, und Wiebke Paulus. 2024. «Folgen des technologischen Wandels für den Arbeitsmarkt: Vor allem Hochqualifizierte bekommen die Digitalisierung verstärkt zu spüren». IAB-Kurzbericht 2024, 5. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung. https://doi.org/10.48720/IAB.KB.2405.

Hattie, John, Deb Masters, und Kate Birch. 2015. *Visible Learning into Action International Case Studies of Impact*. London: Routledge. https://doi.org/10.4324/9781315722603.

Hussein, Mohamed Abdellatif, Hesham Hassan, und Mohammad Nassef. 2019. «Automated language essay scoring systems: a literature review». *PeerJ Computer Science 5:e208.* https://doi.org/10.7717/peerj-cs.208.

Ingenkamp, Karlheinz, und Urban Lissmann. 2008. *Lehrbuch der pädagogischen Diagnostik*. 6., neu ausgestattete Aufl. Beltz Pädagogik. Weinheim u. a.: Beltz.

Jacobsen, Lucas, und Kira Weber. 2023. «The Promises and Pitfalls of ChatGPT as a Feedback Provider in Higher Education: An Exploratory Study of Prompt Engineering and the Quality of AI-Driven Feedback.» https://doi.org/10.31219/osf.io/cr257.

Jönsson, Anders, Andreia Balan, und Eva Hartell. 2021. «Analytic or holistic? A study about how to increase the agreement in teachers' grading». *Assessment in Education: Principles, Policy & Practice* 28 (3): 212–27. https://doi.org/10.1080/0969594X.2021.1884041.

Jukiewicz, Marcin. 2024. «The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process». *Thinking Skills and Creativity* 52:101522. https://doi.org/10.1016/j.tsc.2024.101522.

Kircher, Ernst, Raimund Girwidz, und Hans E. Fischer. 2020. *Physikdidaktik: Grundlagen*. 4. Auflage. Berlin: Springer. https://doi.org/10.1007/978-3-662-59490-2.

Kooli, Chokri, und Nadia Yusuf. 2024. «Transforming Educational Assessment: Insights Into the Use of ChatGPT and Large Language Models in Grading». *International Journal of Human–Computer Interaction* 0 (0): 1–12. https://doi.org/10.1080/10447318.2024.2338330.

Lameras, Petros, und Sylvester Arnab. 2022. «Power to the Teachers: An Exploratory Review on Artificial Intelligence in Education». *Information* 13 (1). https://doi.org/10.3390/info13010014.

Lerche, Thomas. 2022. *Leistungsbeurteilung an Schulen*. München: Lehrstuhl für Schulpädagogik, Ludwig-Maximilians-Universität München. https://epub.ub.uni-muenchen.de/91872/.

Li, Shengjie, und Vincent Ng. 2024. «Automated Essay Scoring: Recent Successes and Future Directions». In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, herausgegeben von Kate Larson, 8114–22. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2024/897.

Lundgren, Magnus. 2024. «Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders.» https://doi.org/10.13140/RG.2.2.27630.42561.

Mishra, Punya, Melissa Warr, und Rezwana Islam. 2023. «TPACK in the age of ChatGPT and Generative AI». *Journal of Digital Learning in Teacher Education* 39 (4): 235–51. https://doi.org/10.1080/21532974.2023.2247480.

Mizumoto, Atsushi, und Masaki Eguchi. 2023. «Exploring the potential of using an AI language model for automated essay scoring». *Research Methods in Applied Linguistics* 2 (2): 100050. https://doi.org/10.1016/j.rmal.2023.100050.

Molenaar, Inge. 2021. «Personalisation of learning: Towards hybrid human-AI learning technologies». In *OECD digital education outlook 2021: Pushing the frontiers with AI, blockchain, and robots*, herausgegeben von S. Vincent-Lancrin, 57–77. OECD. https://doi.org/10.1787/589b283f-en.

Molenaar, Inge. 2022. «Towards Hybrid human-AI Learning Technologies». *European Journal of Education* 57 (4): 632–45. https://doi.org/10.1111/ejed.12527.

Möller, Jens, Thorben Jansen, Johanna Fleckenstein, Nils Machts, Jennifer Meyer, und Raja Reble. 2022. «Judgment accuracy of German student texts: Do teacher experience and content knowledge matter?» *Teaching and Teacher Education* 119 (November):103879. https://doi.org/10.1016/j.tate.2022.103879.

Naujoks, Nick, und Marion Händel. 2020. «Nur vertiefen oder auch wiederholen? Differenzielle Verläufe kognitiver Lernstrategien im Semester». *Unterrichtswissenschaft* 48 (2): 221–41. https://doi.org/10.1007/s42010-019-00062-7.

Ninaus, Manuel, und Michael Sailer. 2022. «Closing the loop – The human role in artificial intelligence for education». *Frontiers in Psychology* 13. https://doi.org/10.3389/fpsyg.2022.956798.

Ning, Yimin, Cheng Zhang, Binyan Xu, Ying Zhou, und Tommy Tanu Wijaya. 2024. «Teachers' AI-TPACK: Exploring the Relationship between Knowledge Elements». *Sustainability* 16 (3). https://doi.org/10.3390/su16030978.

Nückles, Matthias, Sandra Hübner, Sandra Dümer, und Alexander Renkl. 2010. «Expertise reversal effects in writing-to-learn». *Instructional Science* 38 (Mai):237–58. https://doi.org/10.1007/s11251-009-9106-9.

Nückles, Matthias, Julian Roelle, Inga Glogger-Frey, Julia Waldeyer, und Alexander Renkl. 2020. «The Self-Regulation-View in Writing-to-Learn: Using Journal Writing to Optimize Cognitive Load in Self-Regulated Learning». *Educational Psychology Review* 32 (4): 1089–1126. https://doi.org/10.1007/s10648-020-09541-1.

Pinto, Gustavo, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, und Kiev Gama. 2023. «Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT». In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*, 293–302. SBES '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3613372.3614197.

Ramesh, Dadi, und Suresh Kumar Sanampudi. 2022. «An automated essay scoring systems: a systematic literature review». *Artificial Intelligence Review* 55 (3): 2495–2527. https://doi.org/10.1007/s10462-021-10068-2.

Reddy, Y., und Heidi Andrade. 2010. «A review of rubric use in higher education». *Assessment & Evaluation in Higher Education – ASSESS EVAL HIGH EDUC* 35 (Juli): 435–48. https://doi.org/10.1080/02602930902862859.

Schraw, Gregory, Fred Kuch, und Antonio P. Gutierrez. 2013. «Measure for measure: Calibrating ten commonly used calibration scores». *Calibrating Calibration: Creating Conceptual Clarity to Guide Measurement and Calculation* 24 (April):48–57. https://doi.org/10.1016/j.learninstruc.2012.08.007.

Sonnenschein, Katharina. 2015. *Die Objektivität der Leistungsbewertung: Inwieweit sind Leistungsbeurteilungen von Lehrkräften einheitlich und somit vergleichbar?* Hamburg: Diplomica.

Spörer, Nadine, und Joachim C. Brunstein. 2006. «Erfassung selbstregulierten Lernens mit Selbstberichtsverfahren». *Zeitschrift für Pädagogische Psychologie* 20 (3): 147–60. https://doi.org/10.1024/1010-0652.20.3.147.

Südkamp, Anna, Johanna Kaiser, und Jens Möller. 2012. «Accuracy of Teachers' Judgments of Students' Academic Achievement: A Meta-Analysis». *Journal of Educational Psychology* 104 (März):743–62. https://doi.org/10.1037/a0027627.

Swiecki, Zachari, Hassan Khosravi, Guanliang Chen, Roberto Martinez-Maldonado, Jason M. Lodge, Sandra Milligan, Neil Selwyn, und Dragan Gašević. 2022. «Assessment in the age of artificial intelligence». *Computers and Education: Artificial Intelligence* 3 (Januar):100075. https://doi.org/10.1016/j.caeai.2022.100075.

Tharwat, Alaa. 2018. «Classification Assessment Methods: a detailed tutorial», August. https://doi.org/10.1016/j.aci.2018.08.003.Wilkens, Robert. 2020. «Bewerten ohne Klausur: Kompetenzorientierte, semesterbegleitende Leistungsmessung Studierender.» *die hochschullehre* 6: 499–503. https://doi.org/10.3278/HSL2038W.

Zhang, Ke, und Ayse Begum Aslan. 2021. «AI technologies for education: Recent research & future directions». *Computers and Education: Artificial Intelligence* 2 (Januar):100025. https://doi.org/10.1016/j.caeai.2021.100025.