Diyalog 2025/ Sonderausgabe: Germanistik im 21. Jahrhundert- Band I: 501-525 (Research Article)

Smart Graders? Untersuchung des Potenzials von Sprachmodellen in der Fremdsprachenevaluation

Bora Başaran, Eskişehir – Yaşar Ali Sarkiler, Eskişehir

https://doi.org/10.37583/diyalog.1824385

Abstract (Deutsch)

Bewertungen sind ein integraler Bestandteil des Bildungssystems und erfordern ihrer Natur nach häufig einen hohen Zeitaufwand, da Genauigkeit und Konsistenz erwartet werden. Diese Studie untersucht, inwieweit große Sprachmodelle (LLMs) die Leistungsbewertung im Bereich des Fremdsprachenunterrichts unterstützen können. Grundlage sind mehrere Deutsch-Prüfungen, die sowohl von Lehrkräften als auch von LLMs bewertet wurden. Ziel ist es, KI-gestützte Bewertungen mit traditionellen Bewertungen qualitativ zu vergleichen.

Die Analyse konzentriert sich auf Aspekte wie Genauigkeit, Effizienz und Konsistenz und berücksichtigt zudem die Komplexität der Aufgaben sowie die Art der Antworten. Darüber hinaus bietet die Studie eine differenzierte Betrachtung darüber, in welchen Bereichen KI-Leistungen die Arbeitsbelastung von Lehrkräften verringern kann, ohne die pädagogische Qualität der Bewertung zu beeinträchtigen. Abschließend werden praxisnahe Empfehlungen gegeben, wie KI sinnvoll und nachhaltig in den Unterricht integriert werden kann.

Durch den Vergleich von KI-durchgeführten Bewertungen mit Menschlichen, identifiziert die Studie zentrale Bereiche, in denen große Sprachmodelle (LLMs) entweder erfolgreich sind oder nicht. Die technischen und ethischen Grenzen des Einsatzes von KI als eigenständiges Bewertungssystem werden auch thematisiert. Durch die vielsichtige Darstellung sowohl des revolutionären Potenzials von KI als auch der damit verbundenen Risiken leistet diese Studie einen Beitrag zur zunehmend kontrovers geführten Debatte über die Integration von LLMs in die pädagogische Praxis.

Schlüsselwörter: *Benotungsautomatisierung, Bewertung, Deutsch als Fremdsprache, KI in der Lehre, Sprachmodelle.*

Abstract (English)

Smart Graders? Exploring the Potential of Language Models in Foreign Language Evaluation

Assessments function as part of the fabric of education, and by their very nature, are often time-intensive because of the expectation of accuracy and consistency. This study aims to explore how large language models (LLMs) can mediate assessment in the space of a foreign language based on several German exam papers that were graded and assessed by both LLMs and teachers, while ultimately comparing AI assessments to traditional assessments using a qualitative approach.

The analyses focused on aspects of accuracy, efficiency and consistency, while also noting the

Einsendedatum: 11.08.2025 Freigabe zur Veröffentlichung: 30.11.2025

'complexity' of the tasks and response types. In addition, the study provides a detailed overview of how AI could help reduce teacher workload without compromising the pedagogical quality of assessment and offers practical suggestions for the meaningful and sustainable integration of AI into the classroom.

By comparing AI-output to human judgment, the research determines principal areas of LLM failure or success. The technical and moral boundaries of using AI as a standalone assessor are also covered, especially where subtle or linguistically advanced judgments are required. By adding a balanced viewpoint that emphasizes both the potentially revolutionary ability of AI and the wariness in its application, this study adds to the increasingly heated debate regarding the incorporation of LLMs into pedagogic practice.

Keywords: *Assessment automation, evaluation, German as a Foreign Language, AI in teaching, language models.*

EXTENDED ABSTRACT

Assessment is a deeply valuable and essential activity for teachers, but also among the most time-intensive and intellectually taxing. The fast pace of advancements in artificial intelligence (AI), especially through the development of Large Language Models (LLMs), allows to re-think assessment processes. This study explores whether, and to what extent, LLMs like ChatGPT-4o can support, fast-track, or even better, assessment processes in the specific context of teaching German as a Foreign Language (DaF).

The research is guided by three key questions: (1) How closely do AI assessments correspond to those completed by humans? (2) What are the variances in terms of precision, reliability, and efficiency when using LLMs? (3) Under what conditions can AI be integrated in a meaningful way into the assessment process?

Our data corpus includes 29 authentic exam papers from beginner level DaF courses (CEFR levels A1–A2). The key assessments comprised both open-ended and structure items, which were rated both by teachers with experience and by the LLM ChatGPT-4o. For all human assessments by the same two experienced teachers, two forms of assessment were completed: a traditional correction process of using reference keys, and a brand new 'overlay method' which visually allows teachers to assess whether answers are correct, which makes the grading/correction process clear and significantly reduces the time needed to grade.

As for the AI-based assessment, we originally planned to use two formats for input: annotated answer sheets, and a reference list of correct answers. Interestingly, the LLM independently extracted and processed the solution key via Optical Character Recognition (OCR) producing its own version of the answer key. This exemplifies the model's ability to self-interpret and act upon unstructured data - a useful skill in real-world educational applications.

The AI followed a structured prompting process; first, it was given an overview of the assessment task and a series of sample answers. Then the AI evaluated new and unseen exam papers. This method mimics how a language model can operate without a defined or hard- coded evaluation framework, instead relying on contextual learning to create generalizations.

From a quantitative perspective, the model's agreement with human scores were somewhat high given a numeric match of 87.9%, however, looking qualitatively there is a marked difference in reasoning: the Cohen's Kappa value between human raters and AI was only .22 indicating low agreement at the content level. By contrast human raters achieved an inter-rater reliability of .85. Although both AI and human evaluators achieved similar outcomes they often did so using fundamentally different interpretative frameworks.

In relation to efficiency, GPT-4o was peerless. The average time response per item was a lit1le over 1.5 seconds per item, which provided a significant temporal advantage over human assessment. However, the transparency and interpretability in the model's judgement fell short of that which is expected in an educational context, whereas human raters may be slower, yet their reasoning and interpretations are generally more tractable or pedagogy justified.

Another intriguing finding was the stability of AI timing. The GPT-4o response is stable in terms of speed regardless of item difficulty and degree of accuracy. In contrast, humans adjust their speed of work based on difficulty and often generate more errors when working too quickly. This finding also suggests speed alone is not a reliable metric of assessment quality; true for both AI and human raters.

This study has illustrated that meaningful engagement of LLMs in educational assessment also requires explicit conditions: fine-tuning of models against assessment aims, transparency of evaluative criteria, perhaps most importantly, that independent human oversight is engaged. LLMs can be supportive allies to teachers and help them reduce some of the mundane decision making, but should not replace live professional pedagogical judgement (yet).

In summary, LLMs like GPT-4o are valuable and useful tools for educational assessment; they are speedy and often accurate when scoring assessments, but still lack the depth of interpretation and

contextualisation available from human educators. Thus, whilst the teacher role will remain central to the assessment process, a hybrid assessment approach that integrates human expertise and professional judgement with AI assistive efficiency, may provide the equilibrium.

Further development of OCR technologies may provide AI assessments improved accuracy and research into prompt engineering and fine tuning needs to continue to facilitate responsible use within education.

Ethical considerations need to also be unambiguously stated and reflectively considered. Data privacy, algorithmic bias, and transparency in automated decision making are only a few of the ethical and professional considerations that need to be acknowledged. Above all prompt-based AI should be seen as a supportive system, that may help create a more sustainable and quality teaching and learning ambience, not as a replacement.

1 Einleitung

Der Bildungssektor erlebt derzeit einen tiefgreifenden Wandel, der auf den zunehmenden Einfluss Künstlicher Intelligenz (KI) zurückzuführen ist. Diese Entwicklung zeigt deutlich, dass traditionelle Lehrmethoden in verschiedenen Fachbereichen einem stetigen Wandel unterliegen werden (Adiguzel et al. 2023; Ahmad et al. 2021; Başaran 2025; Chen et al. 2020; Kafadar 2022; Soliman et al. 2024; Wang et al. 2024; Yamtinah et al. 2025; Zafari et al. 2022). KI ist ein Teilbereich der Informatik, der sich mit der Entwicklung von Systemen beschäftigt, die Aufgaben übernehmen können, für die normalerweise menschliche Intelligenz erforderlich ist, wie Lernen, Entscheidungsfindung oder Problemlösung (Sarker 2022). In diesem Zusammenhang erfährt insbesondere der Bereich der Sprachbildung nahezu eine Revolution. Neue Technologien verändern kontinuierlich die Dynamiken des Sprachunterrichts, und der Einfluss von KI auf den Fremdsprachenerwerb wird insbesondere in beruflichen Kontexten, in denen Sprache situativ und praxisnah eingesetzt wird, dauerhaft spürbar sein. Zudem hat das gestiegene Kommunikationsaufkommen zwischen Ländern und Gesellschaften infolge des technologischen Fortschritts das Interesse und den Bedarf an Fremdsprachen deutlich erhöht (Tanrıkulu/Üstün 2020).

Der Einsatz von KI-Werkzeugen in Unterricht, Lernprozessen, Leistungsbewertung und Verwaltung hat nachweislich positive Reaktionen bei den Beteiligten hervorgerufen (Aldosari 2020). So konnten durch KI-Instrumente die pädagogischen Kompetenzen verbessert werden, indem sie Lehrpersonen zur Reflexion anregen und neue Impulse bieten (Jaiswal/Arun 2021), ebenso wie die Lernergebnisse von Schüler*innen gesteigert werden konnten (Aldeman et al. 2021). Vor diesem Hintergrund gewinnt auch die Bewertung von Schülerleistungen zunehmend an Bedeutung, da sie nach wie vor eine zentrale, jedoch äußerst belastende Aufgabe für Lehrkräfte im Sprachunterricht darstellt, insbesondere im Fremdsprachenunterricht, wo sprachliche Fähigkeiten auf unterschiedlichen Niveaus beurteilt werden müssen und der Bewertungsprozess ein hohes Maß an Genauigkeit, Fairness und Konsistenz erfordert (Bachman 1990). Gleichzeitig beansprucht die Korrektur von Schülerarbeiten sehr viel Zeit und schränkt die für andere pädagogische Aufgaben verfügbare Zeit erheblich ein.

KI bietet Lehrkräften zudem neue Wege zur beruflichen Weiterentwicklung, etwa durch Vorschläge zur Unterrichtsevaluation oder zur Optimierung von Lehrpraktiken (Gunawan et al. 2021; Hu 2021). Darüber hinaus kann KI auf Grundlage der individuellen Lernleistungen Bewertungen generieren (Durall/Kapros 2020). KI-gestützte Chatbots könnten zur Entwicklung automatisierter und intelligenter Systeme beitragen, die es Lehrkräften ermöglichen, das Lernverhalten und den Lernfortschritt ihrer Schüler*innen systematisch zu analysieren und zu bewerten (Durall/Kapros 2020). Im Kontext der rasanten Entwicklungen im Bereich der Künstlichen Intelligenz, insbesondere durch Fortschritte bei großen Sprachmodellen (Large Language Models, LLMs), eröffnen sich neue Möglichkeiten, um traditionelle Bewertungsverfahren zu ergänzen und zu unterstützen.

Forschungsergebnisse zeigen, dass einige KI-basierte Systeme Aufgaben unter

unterschiedlichsten Bedingungen zuverlässig bewerten können und dabei eine hohe Konsistenz aufweisen (Igaki et al. 2023; Kankanamge et al. 2025). Dabei wurde auch gezeigt, dass KI-gestützte Formate wie videobasierte Hörverstehensaufgaben mit digitalen Plattformen die Leistung und das Angstniveau von Fremdsprachenlernenden positiv beeinflussen können (Tanır 2023). Mit Hilfe Künstlicher Intelligenz können Lehrkräfte große Mengen an Lerndaten analysieren, Muster und Trends erkennen und daraus individuelle Lernbedarfe ableiten. Dies ermöglicht eine personalisierte Unterrichtsplanung, gezielte Interventionen und eine differenzierte Unterstützung entlang des individuellen Bildungswegs der Lernenden (Maghsudi et al. 2021). Darüber hinaus lässt sich die Integration solcher Systeme gewinnbringend mit didaktischen Kompetenzen verbinden, da insbesondere die Unterrichtsplanung als zentrales Element eines erfolgreichen Lehr-Lern-Prozesses gilt (Üstün 2025).

Allerdings bleibt offen, inwieweit diese Technologien auch komplexere, kontextabhängige Aufgaben, wie etwa Aufsätze oder offene Sprachproduktionen, bewerten können, ohne dabei die bestehenden pädagogischen Qualitätsstandards zu gefährden (Hao et al. 2024). Weitere Studien belegen, dass Lehrkräfte zwar die Bedeutung von Leistungsbewertung anerkennen, jedoch weiterhin Unterstützung und Fortbildungsbedarf im Umgang mit KI-gestützten Bewertungssystemen haben (Mede/Atay 2017; Tsagari 2011).

Vor diesem Hintergrund untersucht die vorliegende Studie die Einsatzfähigkeit großer Sprachmodelle (LLMs) als Bewertungsinstrument im Fremdsprachenunterricht und vergleicht deren Leistungen mit jenen traditionellen, lehrkraftbasierten Bewertungen. Ziel ist es herauszufinden, ob KI-gestützte Bewertungssysteme eine sinnvolle Unterstützung für Lehrkräfte darstellen können und welche Rahmenbedingungen dafür erforderlich sind.

Im Zentrum dieser Studie stehen folgende Forschungsfragen:

- **RQ1:** Inwieweit stimmen die Bewertungsergebnisse von LLMs mit denen menschlicher Beurteilung überein?
- **RQ2:** Wie unterscheiden sich die Genauigkeit, Konsistenz und Effizienz zwischen KI-gestützter und traditioneller Bewertung?
- **RQ3:** Welche Voraussetzungen müssen erfüllt sein, damit KI-Systeme als sinnvolle Ergänzung im Bewertungsprozess dienen können?

Das Hauptziel der Arbeit ist es, praxisorientierte Erkenntnisse über die Integration von KI in den Bewertungsprozess zu gewinnen und Empfehlungen für einen pädagogisch verantwortungsvollen Einsatz zu formulieren.

2 Theoretischer Rahmen

2.1 Bewertung im Bildungskontext: Grundlegende Konzepte

Bewertung ist ein zentraler Bestandteil schulischer Bildungsprozesse (William 2011) und

dient der systematischen Erstellung, Erhebung und Rückmeldung von Lernerfolgen. Sie kann einerseits als Diagnose des Lernstandes fungieren, andererseits kontinuierliche Unterstützung für den Lernfortschritt bieten und als Grundlage für Entscheidungen über den weiteren Bildungsweg der Lernenden dienen. Der Bewertungsprozess unterstützt das Lernen durch Rückmeldungen und Ermutigung, mit dem Ziel, dass Lernende ihren eigenen Fortschritt reflektieren (Boud 2000).

Im Bildungskontext wird typischerweise zwischen formativer und summativer Bewertung unterschieden. Formative Bewertung findet während des Lernprozesses statt und liefert kontinuierlich Feedback an die Lernenden, während summative Bewertung eine abschließende Leistungsbeurteilung darstellt – also eine abschließende Note oder Bewertung. Eine qualitativ hochwertige Bewertung zeichnet sich durch Zuverlässigkeit (Reliabilität), Gültigkeit (Validität) und Fairness aus (Dunbar et al. 1991; Efremova et al. 2019; Norcini et al. 2011; Sullivan 2011).

Bewertung stellt hohe Anforderungen an Lehrkräfte: Sie müssen komplexe Schülerleistungen möglichst objektiv und konsistent einschätzen, individuelle Lernfortschritte berücksichtigen und dabei pädagogisch verantwortungsvoll handeln. Die Komplexität dieses Bereichs verdeutlicht die Notwendigkeit, bestehende Bewertungspraktiken kritisch zu hinterfragen und zugleich mögliche Synergien mit neuen Technologien in den Blick zu nehmen.

Die vom türkischen Bildungsministerium am 11. Oktober 2023 eingeführte neue Prüfungsverordnung stellt eine plötzliche Veränderung dar, die die Arbeitsbelastung der Lehrkräfte erheblich erhöht hat (2023). Obwohl die Verordnung das Ziel verfolgt, die vier grundlegenden Sprachfertigkeiten stärker zu betonen, bringt sie tiefgreifende Änderungen in den bestehenden Prüfungsformaten mit sich. Diese Veränderung wird inzwischen auch kritisch diskutiert. Göçer berichtet, dass die befragten Lehrkräfte die Einbeziehung von Hör- und Sprechfertigkeiten – bisher lange vernachlässigt – als einen wichtigen und notwendigen Schritt in Richtung kompetenzorientierten Unterrichts betrachten (2024). Zugleich wird jedoch kritisch angemerkt, dass diese Umstellung zu abrupt erfolgt sei und ohne eine angemessene konzeptionelle Vorbereitung umgesetzt wurde. Gleichzeitig wird in aktuellen Diskussionen betont, dass insbesondere bei der Leistungsbewertung neue Technologien das Potenzial haben, sowohl diagnostische als auch förderorientierte Funktionen zu übernehmen, die traditionell als zentrale Elemente von Lernprozessen gelten (Üstün et al. 2025).

2.2 Bewertung im Fremdsprachenunterricht: Besondere Anforderungen

Die Bewertung im Fremdsprachenunterricht unterscheidet sich in vielerlei Hinsicht von anderen Schulfächern, da sie sowohl sprachliche Kompetenz als auch kommunikative Leistungsfähigkeit umfasst. Im Sprachunterricht werden verschiedene Teilkompetenzen bewertet, darunter Hörverstehen, Leseverstehen, Schreiben, Rechtschreibung und Grammatikkenntnisse, oft auch im Zusammenhang mit kulturellem Wissen.

Ein zentrales Problem stellt die subjektive Natur der Bewertung dar, insbesondere bei offenen Aufgabenformaten wie Aufsätzen oder bei der Einschätzung von rezeptiven Leistungen. Lehrkräfte stehen vor der Herausforderung, reichhaltige sprachliche Leistungen zuverlässig und vergleichbar zu bewerten. Häufig kommen daher Kriterienraster oder Punktesysteme zum Einsatz, um Transparenz und Vergleichbarkeit zu gewährleisten.

Darüber hinaus spielt die Motivation der Lernenden für den Spracherwerb eine entscheidende Rolle. Unsicherheit oder das Empfinden von Ungerechtigkeit in der Bewertung können sich negativ auf die Motivation und das Selbstvertrauen der Lernenden auswirken (Shao et al. 2024). Daher ist es besonders wichtig, dass Bewertungsverfahren – insbesondere solche mit formativem Charakter, klare, lernförderliche Strukturen aufweisen. Diese spezifischen Anforderungen verdeutlichen die Notwendigkeit, unterstützende technologische Mittel einzusetzen, um Objektivität, Effizienz und Fairness bei der Leistungsbewertung im Fremdsprachenunterricht zu fördern – etwa durch KI-gestützte Bewertungssysteme.

2.3 Künstliche Intelligenz und Large Language Models (LLMs)

Künstliche Intelligenz (KI) bezeichnet die Fähigkeit von Maschinen, Aufgaben auszuführen, die üblicherweise mit menschlicher Intelligenz verbunden sind, wie etwa Problemlösen, Lernen, Sprachverarbeitung oder Entscheidungsfindung (Sarker 2022; Winston 1992; Yin et al. 2023). In jüngster Zeit hat insbesondere ein Bereich der KI bemerkenswerte Fortschritte gemacht: sogenannte Large Language Models (LLMs) (Makridakis 2023).

LLMs wie GPT (Generative Pre-trained Transformer) lernen aus riesigen Mengen an Textdaten und sind in der Lage, Sprache zu verstehen, zu analysieren und zu generieren. Sie können Fragen beantworten, Texte zusammenfassen, Dialoge führen und komplexe sprachliche Aufgaben bewältigen (Felder/Kückelhaus 2025). Chatbots wie ChatGPT und Bard generieren laut Douglas (2023) Sätze auf der Grundlage statistischer Muster, die sie aus umfangreichen Textkorpora gelernt haben.

LLMs können eine Vielzahl unterschiedlicher Aufgaben übernehmen, bringen jedoch auch Risiken mit sich – etwa ungenaue Ausgaben, sicherheitsrelevante Instabilitäten oder ein gewisses Manipulationspotenzial. Aus diesem Grund vertreten einige Forschende die Auffassung, dass LLMs – sofern sie angemessen eingesetzt werden – lediglich als unterstützende Werkzeuge in Bewertungskontexten mit hoher Tragweite dienen sollten, während die endgültigen Entscheidungen stets von Menschen getroffen werden müssen (Endres/Ibisch 2025).

Domänenspezifisch trainierte LLMs können in verschiedensten Fachgebieten leistungsstarke Werkzeuge darstellen (Minaee et al. 2024). Werden sie mit fachspezifischen Datensätzen aus Bereichen wie Medizin, Recht, Bildung oder Ingenieurwesen trainiert, sind sie in der Lage, kontextgerechte, präzise und inhaltlich fundierte Antworten zu generieren. Über die reine Wissensvermittlung hinaus können

diese spezialisierten Modelle komplexe Inhalte analysieren, Entscheidungsprozesse unterstützen und fachspezifische Texte generieren. Ein Blick in die aktuelle Fachliteratur zeigt, dass LLMs zunehmend in Bereichen wie Gesundheitswesen, Bildung und Finanzwesen eingesetzt werden (Mohan et al. 2024; Thirunavukarasu et al. 2023). Soliman et al. konnten zeigen, dass ein LLM-basierter Chatbot, der mit gezielten Materialien trainiert wurde, Lehramtsstudierende deutlich besser unterstützte als ein Chatbot mit allgemeinen Trainingsdaten (2019). Diese Ergebnisse unterstreichen den Mehrwert einer fokussierten und fachspezifischen Schulung von Chatbots, um die Wirksamkeit der Lernunterstützung für Studierende zu erhöhen (Chen et al. 2023).

Im Bildungsbereich eröffnen LLMs eine Vielzahl von Anwendungsmöglichkeiten, darunter die automatisierte Textanalyse, die Erstellung von Lernmaterialien und die Unterstützung bei Bewertungsprozessen. Besonders im Fremdsprachenunterricht gelten sie aufgrund ihrer Fähigkeit, sprachliche Strukturen zu erkennen und inhaltliche Kohärenz zu beurteilen, als vielversprechend. Gleichzeitig werfen sie neue Fragen im Hinblick auf Bewertung und Lernprozesse auf: Wie objektiv, transparent und lernförderlich ist eine KI-gestützte Bewertung? In welchem Ausmaß können LLM-Systeme Lehrkräfte ergänzen, oder sogar ersetzen? Diese Fragen stehen im Zentrum aktueller bildungswissenschaftlicher und ethischer Debatten.

2.4 Bildungstechnologische Anwendungen von KI: Aktuelle Forschungsperspektiven

In den letzten Jahren ist der Einsatz Künstlicher Intelligenz (KI) im Bildungsbereich erheblich gestiegen (Chen 2020; Popenici et al. 2017). Heutzutage werden KI-Systeme in verschiedenen Bereichen eingesetzt, von adaptiven Lerntechnologien und Spracherkennungssystemen bis hin zur automatisierten Bewertung von Schülerleistungen. Insbesondere im Fremdsprachenunterricht bieten KI-Technologien großes Potenzial für Personalisierung und Effizienzsteigerung.

Aktuelle Forschungsergebnisse zeigen, dass KI in der Lage ist, Lernverläufe nachzuverfolgen und durch die Identifikation von Fehlerquellen gezieltes und präzises Feedback zu geben (Luckin et al. 2016). Systeme wie automatische Essay-Bewerter oder Chatbots im Sprachunterricht verdeutlichen, wie KI-Technologien interaktive Lernprozesse von Lernenden unterstützen können (Evenddy 2024; IU Internationale Hochschule 2024; Zawacki-Richter et al. 2019). Dabei spielen die Transparenz der Bewertung und deren Akzeptanz durch Lehrkräfte und Lernende eine zentrale Rolle.

Zunehmend wird auch untersucht, wie KI-Systeme im Bildungsbereich als unterstützende Instrumente und nicht als Ersatz für pädagogische Fachkräfte eingesetzt werden können. Insbesondere im Bereich der Leistungsbewertung wird weiterhin diskutiert, ob KI-basierte Bewertungen objektiver, schneller oder verlässlicher sein können – wobei gleichzeitig pädagogische Sensibilität berücksichtigt werden muss (Holstein et al. 2020).

Trotz der jüngsten Entwicklungen befindet sich dieses Forschungsfeld noch in einem frühen Stadium. Es besteht weiterhin ein erheblicher Bedarf an empirischen Studien, die die Vorteile, Grenzen und pädagogischen Implikationen des KI-Einsatzes im Klassenzimmer systematisch untersuchen.

3 Methode

3.1 Forschungsdesign

Die Untersuchung verwendete ein vergleichendes experimentelles Design, um Bewertungsmethoden miteinander zu vergleichen. Die durchgeführte Studie fand an der Fakultät für Erziehungswissenschaften der Anadolu-Universität statt. Zwei unterschiedliche Bewertungsverfahren wurden gegenübergestellt: (1) eine automatische Bewertung der Antworten mithilfe eines Large Language Models (LLM), bei der sowohl schriftliche als auch visuelle Lösungsschlüssel zum Einsatz kamen, und (2) eine manuelle Bewertung durch menschliche Bewerter unter Verwendung eines traditionellen Referenzbogens sowie einer innovativen Overlay-Methode zur Antwortabstimmung.

Der Datensatz umfasste 29 authentische, anonymisierte Zwischenprüfungsklausuren von Studierenden. Die Klausur bestand aus Multiple-Choice-Fragen, die gemäß dem standardisierten Bewertungsformat des betreffenden Kurses erstellt wurden. Sämtliche personenbezogenen Daten wurden gemäß den ethischen Genehmigungsverfahren und den Datenschutzrichtlinien der Institution aus den Prüfungsunterlagen entfernt. Im Rahmen der Studie wurden KI-gestützte Methoden eingesetzt, um strukturierte Eingabeformate auf standardisierte Weise anhand von Lösungsschlüsseln zu analysieren.

3.2 Bewertungsverfahren

3.2.1 KI-gestützte Bewertung

Der erste Bewertungsansatz nutzte das kommerziell verfügbare, allgemein einsetzbare große Sprachmodell ChatGPT 4o. Das Modell wurde ausgewählt, da es mehrere Sprachen unterstützt, für akademische Zwecke geeignet ist und über Webschnittstellen oder APIs zugänglich ist. Das Modell sollte ursprünglich folgende standardisierte Eingaben zur Bewertung erhalten:

- Einen digitalisierten, farbcodierten Antwortbogen
- Eine Referenzliste mit den korrekten Antworten im Textformat
- In der praktischen Durchführung:
- Verwendete das Modell den Antwortbogen nicht direkt, sondern erstellte mithilfe von OCR-Technologie selbstständig eine textuelle Antwortliste
- Die Prüfungsbögen der Studierenden wurden dem Modell ohne Noten übermittelt, entweder in Textform oder als Bilddateien

Im Rahmen der Prompting-Methode wurde der Lösungsschlüssel in das System geladen, und das Modell wurde anschließend angewiesen, die eingereichten Antwortbögen sowie die zugehörigen Bewertungsbögen auf Basis dieses Schlüssels zu bewerten. Auf diese Weise konnte das Modell seine Mustererkennungsfähigkeiten und kontextbasierte Bewertung nutzen, ohne dass ein Bewertungsschema explizit entwickelt werden musste.

Die Bewertungsmethode diente dazu, das Potenzial allgemein einsetzbarer LLMs für Bildungsbewertungsprozesse zu untersuchen, indem die Fähigkeit der Modelle zur Mustererkennung und kontextbezogenen Argumentation, ohne explizite Definition eines Bewertungsschemas, analysiert wurde.

3.2.2 Menschliche Bewertung

Die Studie verwendete zwei Methoden der menschlichen Bewertung, die sowohl eine konventionelle referenzbasierte Bewertung als auch eine eigens entwickelte Overlay-Bewertung für denselben Satz von Prüfungsarbeiten umfassten.

Konventionelle Methode: Zwei Dozierende mit Erfahrung im Bereich „Deutsch als Fremdsprache“ bewerteten die Prüfungsarbeiten unabhängig voneinander anhand einer standardisierten, referenzbasierten Bewertungsmethode. Den Bewertenden wurden Antwortschlüssel sowie strukturierte Bewertungstabellen zur Verfügung gestellt, um die Punktevergabe für jede Schülerantwort zu bestimmen. Während des Bewertungsprozesses wurde gleichzeitig die zur Durchführung benötigte Zeit erfasst.

Overlay-Methode: Die Forschenden entwickelten die Overlay-Methode als Lösung zur Verbesserung sowohl der Bewertungsgeschwindigkeit als auch der Übersichtlichkeit bei Multiple-Choice-Prüfungen. Dabei kamen transparente Overlays mit den richtigen Antworten zum Einsatz, die direkt auf die Prüfungsbögen der Studierenden gelegt wurden, um richtige und falsche Antworten schnell identifizieren zu können. Die gleichen beiden Dozierenden führten auch die Bewertungen dieses Prüfungssatzes mit dieser Methode durch.

3.3 Cohen’s Kappa

Die von Cohen entwickelte Kappa-Statistik ist ein statistisches Verfahren, das darauf abzielt, zu quantifizieren, inwieweit zwei unabhängige Bewerterinnen bei der Kategorisierung von Daten übereinstimmen (Cohen 1960). Die Kappa-Statistik wurde als präzisere Methode zur Erfassung der Interrater-Reliabilität entwickelt, da sie auch zufällige Übereinstimmungen berücksichtigt (McHugh 2012). Der Einsatz der Kappa-Statistik ist besonders dann sinnvoll, wenn einfache Übereinstimmungsraten die Möglichkeit einer zufälligen Übereinstimmung der Bewerterinnen bei der Wahl derselben Kategorien ignorieren, was die Aussagekraft solcher Maße einschränkt.

Cohens Kappa zeigt auf, wie viel größer die beobachtete Übereinstimmung zwischen zwei Bewerter*innen ist als die erwartete Übereinstimmung durch Zufall. Der Kappa-Wert kann zwischen 0 und 1 liegen, wobei 1 eine perfekte Übereinstimmung und

0 eine rein zufällige Übereinstimmung bedeutet (Warrens 2015; Warrens 2025). Die Kappa-Statistik standardisiert die Differenz zwischen der beobachteten und der erwarteten Übereinstimmung, indem sie diese Differenz durch (1 – erwartete Übereinstimmung) teilt.

Obwohl Kappa-Werte theoretisch zwischen –1 und +1 liegen können, werden in der Praxis typischerweise nur positive Werte zwischen 0 und 1 als relevant angesehen. In der Fachliteratur gelten Werte ab 0,60 üblicherweise als Zeichen für eine „gute Übereinstimmung“ und Werte über 0,80 als „sehr gute Übereinstimmung“. Dennoch sollten Kappa-Werte stets im Kontext der jeweiligen Studie interpretiert werden.

Cohens Kappa wird in Disziplinen wie Psychologie, Medizin, Bildungswissenschaften und maschinellem Lernen verwendet, um die Klassifikationsübereinstimmung zwischen zwei Bewerterinnen objektiv zu messen (Perez et al., 2020). Allerdings ist Cohens Kappa ausschließlich für Designs mit zwei Bewerterinnen geeignet; bei mehreren Bewerter*innen kommen generalisierte Verfahren wie Fleiss’ Kappa zum Einsatz (Marasini et al. 2016).

In der vorliegenden Studie wurde der Kappa-Koeffizient berechnet, um die Übereinstimmung zwischen den Codierungen zweier Bewerter*innen zu bestimmen und so die Objektivität und Verlässlichkeit des Bewertungsprozesses sicherzustellen.

4 Ergebnisse

Dieser Abschnitt präsentiert die Ergebnisse der Studie zur Beantwortung der Forschungsfragen (RQ1–RQ3), die sich auf die Genauigkeit, Effektivität und Benutzerfreundlichkeit von KI-basierten und menschlichen Bewertungsmethoden im Kontext der Bewertung von Deutschprüfungen konzentrieren.

- **RQ1:** Inwieweit stimmen die Bewertungsergebnisse von LLMs mit denen menschlicher Bewertung überein?

Die Ergebnisse der Bewertung durch das Large Language Model (ChatGPT-4o) zeigen, dass eine Gesamtübereinstimmungsrate von 87,9 % mit den menschlichen Bewertern (Mensch 1 und Mensch 2) durchaus beachtlich ist. Dieser Wert legt nahe, dass das Modell in der Regel zu ähnlichen Bewertungen gelangt wie menschliche Korrektor*innen, zumindest in Bezug auf die Gesamtpunktzahl.

Bei genauerer Betrachtung der inhaltlichen Bewertungskonsistenz wird jedoch eine kritische Einschränkung deutlich: Die Cohen’s-Kappa-Werte, die neben der reinen Übereinstimmung auch den Zufallsfaktor berücksichtigen, fallen mit jeweils .22 für GPT vs. Mensch 1 sowie GPT vs. Mensch 2 sehr niedrig aus. Im Vergleich dazu erreichen die beiden menschlichen Bewerter*innen einen sehr hohen Kappa-Wert von .85. Dies deutet darauf hin, dass GPT zwar häufig zur gleichen Gesamtpunktzahl wie ein Mensch gelangt, dabei jedoch möglicherweise auf einer abweichenden Bewertungslogik basiert. Dem Modell fehlt das differenzierte interpretative und kontextuelle Verständnis, das für eine Bewertung auf vergleichbarem Niveau mit menschlichen Prüfer*innen erforderlich ist.

Diese Ergebnisse machen deutlich, dass numerische Übereinstimmungen allein nicht ausreichen, um die Qualität eines KI-gestützten Bewertungssystems zu beurteilen. Erst durch die Einbeziehung konsistenzbasierter Kennzahlen wie etwa Cohen's Kappa lässt sich eine fundierte Aussage über die Vergleichbarkeit mit menschlicher Leistung treffen. Die nachfolgende Abbildung veranschaulicht diesen Widerspruch zwischen scheinbarer Übereinstimmung und tatsächlicher Bewertungsgüte.

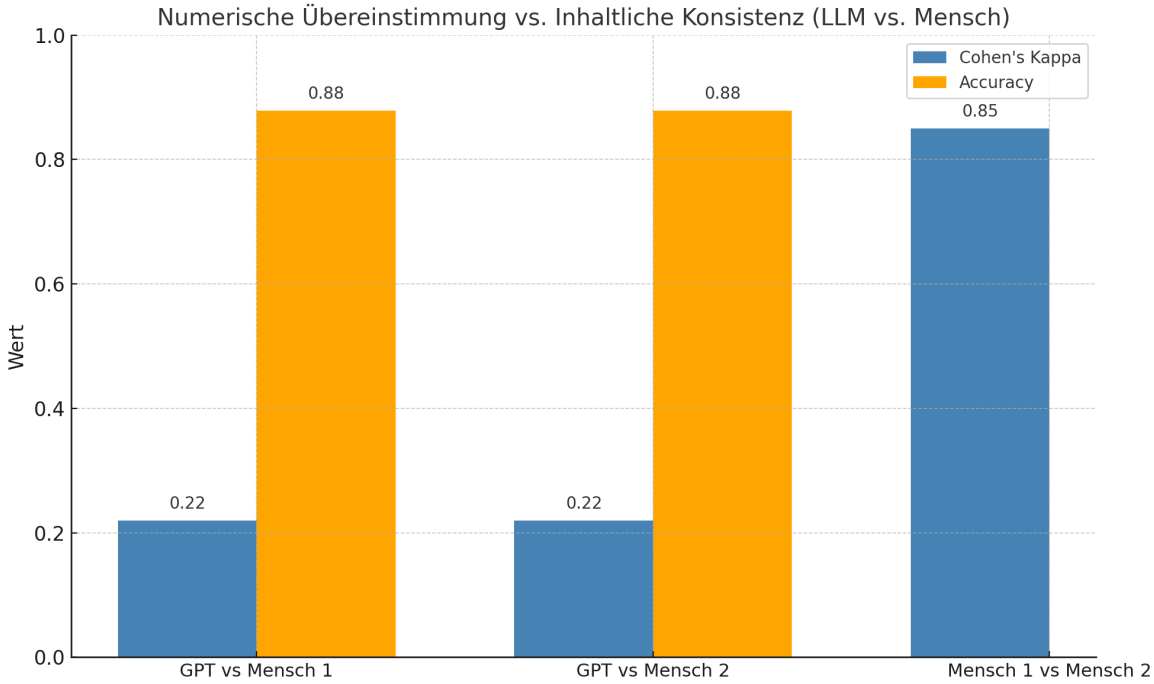

Abb. 1. Numerische Übereinstimmung vs. Inhaltliche Konsistenz (LLM vs. Mensch)[1]

[1] Erstellt von den Verfassern.

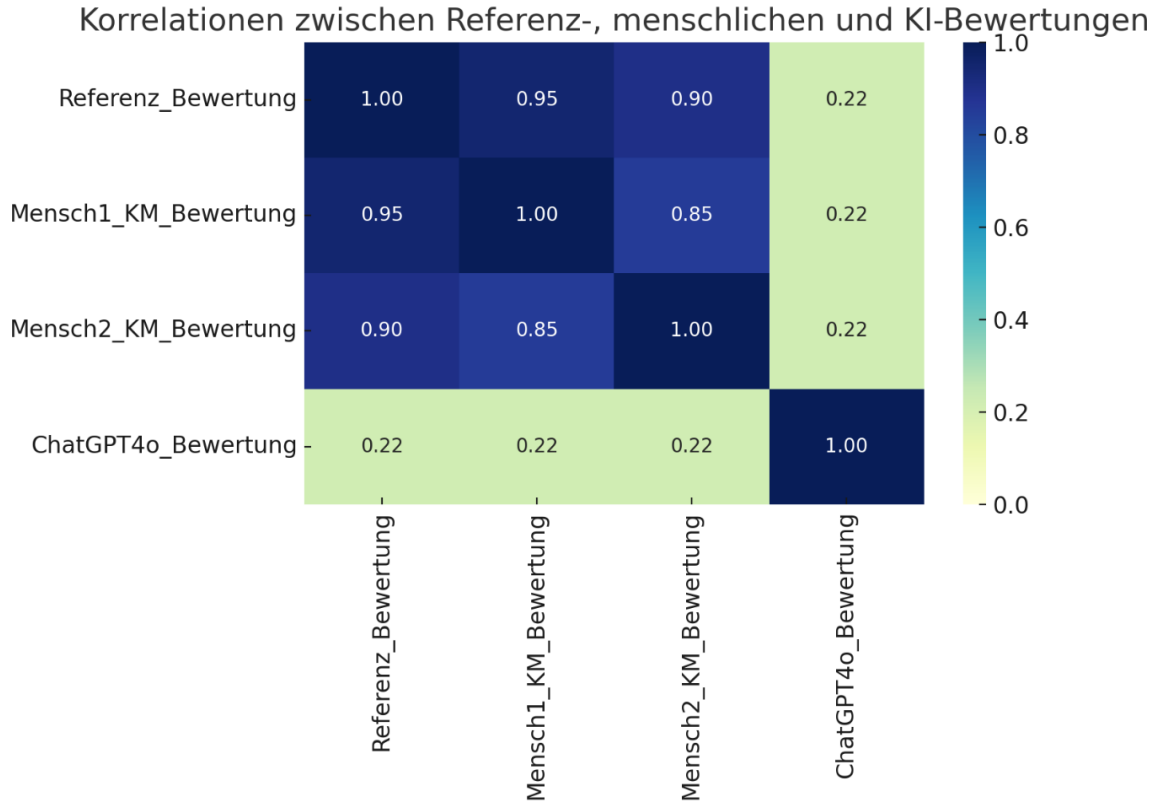

Abb. 2. Korrelationen zwischen Referenz-, menschlichen und KI-Bewertungen[2]

Die Korrelationen zwischen den Bewertungen geben Aufschluss darüber, wie stark die Punktwerte systematisch miteinander zusammenhängen. Die zugehörige Heatmap zeigt dabei ein klares Bild:

- Sehr hohe Korrelation zwischen den beiden menschlichen Bewertenden (Mensch 1 und Mensch 2),
- Hohe Korrelation zwischen den menschlichen Bewertungen und der Referenzbewertung,
- Eine etwas schwächere, aber dennoch deutliche Korrelation der KI-Bewertung (GPT-4o) mit allen anderen.

Die Heatmap zeigt, dass GPT-4o in der Regel ähnliche Bewertungstendenzen wie menschliche Prüfer erkennen kann, die Korrelationen liegen durchweg im oberen Bereich. Das bedeutet vor allem, dass ein vergleichbares Verhalten bei der Punktevergabe wahrscheinlich ist.

Dennoch darf diese numerische Übereinstimmung nicht mit inhaltlicher Bewertungsübereinstimmung gleichgesetzt werden. So fällt beispielsweise die inhaltliche Konsistenz, gemessen anhand von Cohen's Kappa, deutlich geringer aus. Das weist darauf hin, dass GPT-4o zwar zu ähnlichen Ergebnissen gelangen kann, dabei jedoch nicht zwingend denselben Bewertungslogiken oder pädagogischen Überlegungen folgt wie menschliche Prüfer*innen.

[2] Erstellt von den Verfassern.

Die Heatmap verdeutlicht somit einen wichtigen methodischen Unterschied: Sie macht den Unterschied zwischen statistischer Korrelation (einer oberflächlichen Übereinstimmung) und inhaltlicher Bewertungskonsistenz (einer substanziellen Übereinstimmung) sichtbar – ein zentraler Aspekt zur Sicherstellung fairer und nachvollziehbarer Bewertungsentscheidungen.

- **RQ2:** Wie unterscheiden sich die Genauigkeit, Konsistenz und Effizienz zwischen KI-gestützter und traditioneller Bewertung?

Die Analyse der Bewertungsergebnisse zeigt ein vielschichtiges Bild hinsichtlich der Leistungsfähigkeit von Large Language Models wie ChatGPT-4o im Vergleich zu menschlichen Bewertungsverfahren. In drei zentralen Bereichen, Genauigkeit, Konsistenz und Effizienz, lassen sich deutliche Unterschiede feststellen, die sowohl Potenziale als auch Grenzen der KI-gestützten Bewertung verdeutlichen.

1. Genauigkeit – Oberflächliche Übereinstimmung

Mit einer Übereinstimmungsrate von 87,9 % mit den Bewertungen von Mensch 1 und Mensch 2 erreicht GPT-4o einen hohen Wert. Dieser scheinbar positive Befund legt nahe, dass das Modell in der Lage ist, in den meisten Fällen dieselbe Bewertung wie menschliche Korrektoren zu vergeben. Die numerische Genauigkeit auf dieser Ebene ist daher durchaus als leistungsstark zu bewerten – zumindest oberflächlich betrachtet.

2. Konsistenz – Unterschiedliche Entscheidungswege

Ein kritischer Blick auf die inhaltliche Konsistenz offenbart jedoch Schwächen. Der Cohen's Kappa-Wert, ein statistisches Maß, das nicht nur Übereinstimmungen zählt, sondern auch den Zufallsfaktor einbezieht, liegt bei der Übereinstimmung zwischen GPT und den menschlichen Bewertern bei lediglich 0.22. Demgegenüber erreichen die beiden menschlichen Bewerter untereinander einen sehr hohen Kappa-Wert von 0.85. Diese Diskrepanz zeigt deutlich, dass GPT-4o inhaltlich andere Bewertungslogiken verfolgt. Es kommt zwar oft zum gleichen Ergebnis, aber aus Gründen, die sich von denen menschlicher Prüfer unterscheiden. Dadurch fehlt dem Modell eine nachvollziehbare und konsistente Bewertungsgrundlage.

3. Effizienz – Geschwindigkeit als Stärke der KI

Im Bereich der Effizienz zeigt sich der größte Vorteil von GPT-4o: Das Modell benötigt im Schnitt 1536 Millisekunden pro Bewertung und ist damit schneller als beide menschlichen Verfahren:

Konventionelle Methode: ca. 2500–2700 ms

Overlay-Methode: ca. 1690–1876 ms

Besonders bemerkenswert ist die Geschwindigkeit von GPT-4o bei der Bewertung bzw. Punktevergabe vor allem dann, wenn große Datenmengen verarbeitet oder standardisierte Aufgaben bewertet werden. Selbst im Vergleich zur bereits sehr schnellen menschlichen Bewertung mittels Overlay-Methode arbeitet das Modell noch etwas schneller. Fazit: Abwägung zwischen Geschwindigkeit und Bewertungsqualität.

Die Ergebnisse zeigen, dass GPT-4o in Bezug auf Effizienz und numerische Genauigkeit eindeutig überlegen ist. Hinsichtlich konzeptueller Konsistenz und Interpretierbarkeit kann das KI-Modell jedoch nicht mit menschlichen Bewertenden mithalten. Die scheinbar hohe Übereinstimmung täuscht, da sie auf völlig unterschiedlichen Denk- und Bewertungsprozessen beruht.

Für den Einsatz in realen Bewertungsszenarien bedeutet dies, dass KI-Modelle wie GPT-4o derzeit eher als unterstützende Werkzeuge und nicht als vollwertiger Ersatz zu betrachten sind. Sie können die Bewertung beschleunigen und Empfehlungen generieren, sollten jedoch durch menschliche Aufsicht ergänzt werden – insbesondere in Kontexten, die transparente und nachvollziehbare Bewertungsentscheidungen erfordern.

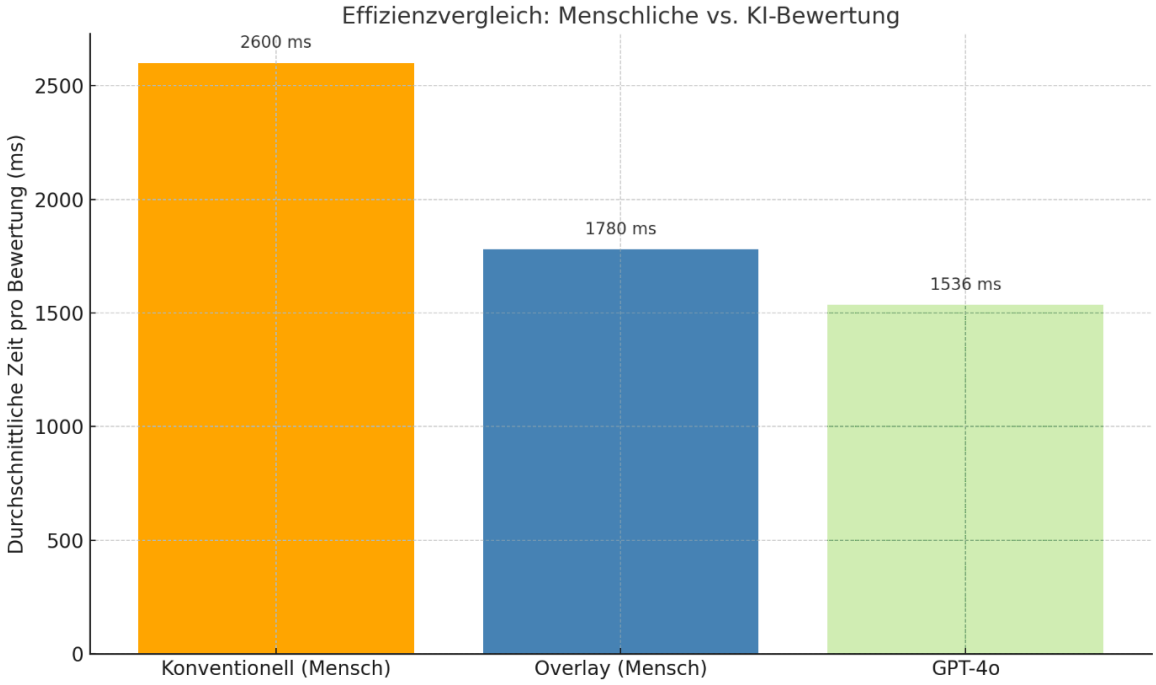

Abb. 3. Effizienzvergleich: Menschliche vs. KI-Bewertung[3]

[3] Erstellt von den Verfassern.

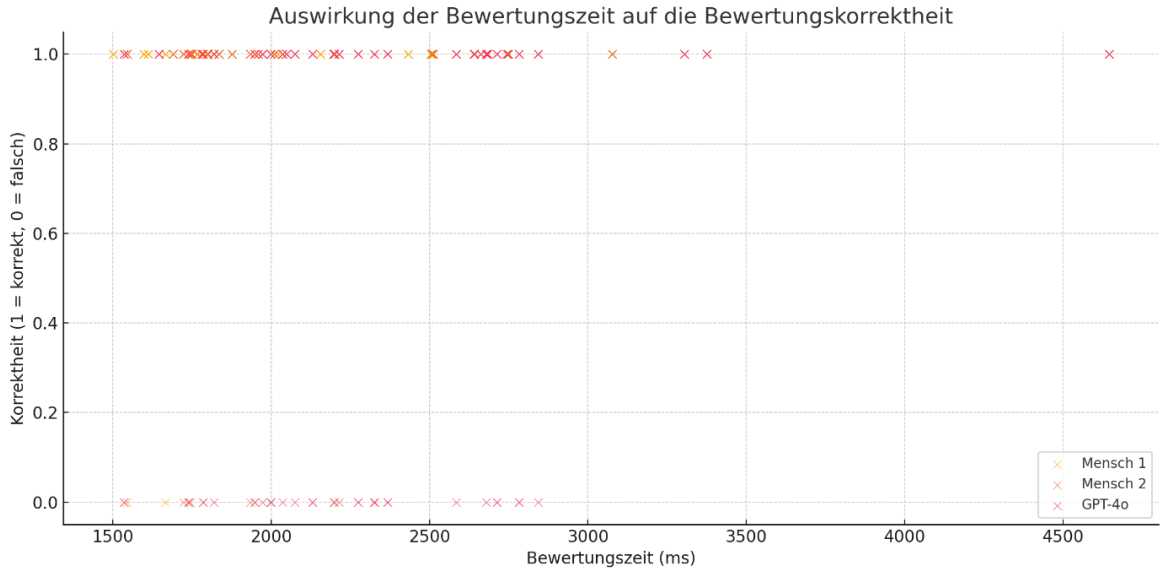

Abb. 4. Auswirkung der Bewertungszeit auf die Bewertungskorrektheit[4]

Die Bewertungsgeschwindigkeit von GPT-4o ist bemerkenswert konstant. Alle Zeitstempel der Bewertungen liegen bei etwa 1536 Millisekunden, was darauf hindeutet, dass das Modell über eine eingebaute Fähigkeit zur konsistenten Bewertung verfügt. Obwohl die überwiegende Mehrheit der Ergebnisse korrekt ist, gibt es auch einzelne Bewertungen, die deutlich von der Referenz abweichen – und dennoch in derselben kurzen Zeitspanne erfolgten. Dies zeigt, dass Geschwindigkeit bei KI nicht gleichbedeutend mit inhaltlicher Korrektheit ist.

Im Vergleich dazu weisen menschliche Bewerter eine deutlich größere zeitliche Streuung auf. Manche Bewertungen erfolgen sehr schnell, andere nehmen wesentlich mehr Zeit in Anspruch. Bemerkenswert ist, dass sowohl schnelle als auch langsame Bewertungen korrekt sein können. Es scheint also eine nichtlineare Beziehung zwischen Bearbeitungsdauer und Bewertungskorrektheit zu geben – nach dem Motto „je länger, desto besser“ lässt sich hier kein einfaches Muster ableiten.

Dennoch zeigt sich ein gewisses Muster: Viele der fehlerhaften Bewertungen treten bei sehr kurzen Bearbeitungszeiten auf – besonders auffällig bei Mensch 2. Dies legt nahe, dass eine längere Bearbeitungszeit die Fehlerwahrscheinlichkeit verringern kann, auch wenn sie keine Garantie für Richtigkeit bietet. Der Zeitfaktor kann somit als potenzieller Einflussfaktor bei menschlicher Bewertung betrachtet werden – insbesondere bei komplexeren Aufgabenstellungen.

- **RQ3:** Welche Voraussetzungen müssen erfüllt sein, damit KI-Systeme als sinnvolle Ergänzung im Bewertungsprozess dienen können?

Zusammenfassend zeigen die Ergebnisse dieser Studie deutlich, dass große Sprachmodelle (LLMs) wie ChatGPT-4o zwar ein gewisses Potenzial für den Einsatz in der Bildungspraxis bieten, jedoch derzeit nicht als eigenständige Bewertungsinstanz geeignet sind. Die schnelle Bewertung und oberflächliche Übereinstimmung mit

[4] Erstellt von den Verfassern.

menschlichen Prüfer*innen können die inhaltlichen Inkonsistenzen und die mangelnde Transparenz in der Argumentation nicht ausgleichen. Damit LLMs als sinnvoller Bestandteil des Bewertungsprozesses bzw. der Bewertungskontinuität anerkannt werden können, müssen verschiedene Voraussetzungen erfüllt sein – sowohl technischer als auch didaktischer Natur.

1. Feinabstimmung auf Bewertungskriterien (Fine-Tuning und Prompt-Engineering)

Eine wesentliche Einschränkung besteht in der Unfähigkeit des Modells, auf bestimmte Bewertungskriterien adäquat zu reagieren. Häufig liefert das Modell zwar die korrekte Antwort, stützt sich dabei jedoch auf andere Argumentationsprinzipien als die beteiligten menschlichen Bewerter*innen. Um dieser Problematik entgegenzuwirken, ist entweder eine gezielte Feinabstimmung anhand sektorspezifischer Bewertungsstandards erforderlich oder der Einsatz präziser Prompt-Anweisungen, die das Modell auf die im jeweiligen Bewertungskontext benötigten Kriterien ausrichten.

2. Einsatz als unterstützendes System, nicht als Ersatz

Auf dem aktuellen Stand ihrer Entwicklung sollten KI-Systeme nicht als Ersatz für menschliche Bewertung betrachtet werden. Sie eignen sich besser als Feedbackgeber, Filter oder Vorschlagsinstanz, um Lehrkräfte bei der Bewältigung umfangreicher Bewertungsaufgaben zu unterstützen, ohne die pädagogische Verantwortung aus der Hand zu geben. In der praktischen Umsetzung könnte dies bedeuten, dass die KI eine Bewertung vornimmt, die anschließend von der Lehrkraft überprüft und bei Bedarf angepasst wird.

3. Transparenz und Vergleichbarkeit der Bewertungskriterien

Eine grundlegende Voraussetzung für die Akzeptanz und Transparenz KI-gestützter Bewertung ist die klare Festlegung der Kriterien, auf deren Grundlage das Modell bewertet. Die Bewertungslogik des Systems sollte – soweit möglich – in einzelne Bestandteile unterteilt und öffentlich zugänglich gemacht werden, sodass sie von Nutzer*innen eingesehen und nachvollzogen werden kann. Idealerweise sollten diese Kriterien mit langfristig etablierten menschlichen Bewertungsstandards übereinstimmen. Nur auf diese Weise ist es wahrscheinlich, dass Studierende, Lehrende und Prüfungsverantwortliche den algorithmischen Entscheidungsprozess verstehen und akzeptieren.

4. Menschliche Validierung und Kontrollmechanismen

Selbst bei einem hohen Automatisierungsgrad darf die menschliche Aufsicht nicht entfallen. KI ist nicht in der Lage, alle kontextuellen, nuancierten oder impliziten Bedeutungen zu erfassen, die für komplexe, offene oder kreative Aufgaben erforderlich sind. Daher erfordert jeder Einsatz von KI im Bewertungsprozess eine Validierung durch Lehrpersonen. Dieser Prozess gewährleistet pädagogische Verlässlichkeit und stärkt das Vertrauen in das Bewertungssystem.

Fazit: Potenzial mit klaren Rahmenbedingungen

Selbst bei einem hohen Automatisierungsgrad sollte die menschliche Aufsicht nicht abgeschafft werden. KI ist nicht in der Lage, alle kontextuellen, nuancierten oder impliziten Bedeutungen zu erfassen, die für komplexe, offene oder kreative Aufgaben erforderlich sind. Daher erfordert jeder Einsatz von KI im Bewertungsprozess eine Validierung durch Lehrkräfte. Dieser Prozess gewährleistet pädagogische Verlässlichkeit und stärkt das Vertrauen in den Bewertungsablauf.

5 Diskussion

5.1 Vorteile und Grenzen KI-gestützter Bewertung

Die Forschung legt nahe, dass GPT-basierte KI-Systeme zwar Bewertungen in sehr kurzer Zeit nachbilden können, doch „schnell" bedeutet nicht automatisch „präzise". Bei Bewertungen, die auf feinskaliger Genauigkeit und Konsistenz basieren, erreicht das KI-Modell noch nicht das Niveau menschlicher Prüfer. Zudem bestehen weiterhin mehrere Hürden – wie etwa Datenschutz und Informationssicherheit – die nach wie vor relevant sind und den Einsatz solcher Systeme in komplexen Prüfungs- und Bewertungssituationen problematisch machen können.

5.2 Pädagogische Qualität der Bewertung

Pädagogisch sinnvolle Bewertung umfasst nicht nur die Rückmeldung zu Fehlern, sondern auch die Berücksichtigung von Lernfortschritt, individueller Ausdrucksweise und der Motivation der Lernenden. Menschliche Bewertende bringen in diesen Prozess Erfahrung, Empathie und pädagogisches Bewusstsein ein. KI kann diese Dimensionen nur teilweise erfassen und sollte daher zum jetzigen Zeitpunkt als ergänzender Akteur und nicht als Ersatz betrachtet werden. Da dem Modell keine ergänzenden Informationen außer den Prompts und den schriftlichen Prüfungsantworten zur Verfügung standen – wie etwa die individuelle Lernentwicklung, persönliche Fortschritte oder emotionale Aspekte – ist eine umfassende pädagogische Bewertung auf dieser Grundlage nur in begrenztem Maße möglich.

5.3 Die Rolle der Lehrkräfte: Unterstützende und überwachende Funktionen

Die Ergebnisse deuten darauf hin, dass KI-Systeme Lehrkräfte bei der Bewertung unterstützen können, insbesondere bei standardisierten Aspekten des Bewertungsprozesses. Letztlich behalten Lehrkräfte jedoch eine zentrale Rolle als pädagogische Instanz, um getroffene Entscheidungen zu reflektieren, den Lernstand der Schüler*innen einzuschätzen und individuelles Feedback zu geben. In diesem Sinne könnte sich die Rolle der Lehrkraft von der primären Bewertungsinstanz hin zur reflektierenden Moderation im Bewertungsprozess weiterentwickeln.

6 Fazit und Ausblick

Die Studie untersuchte das Potenzial eines allgemeinen Large Language Models (LLM) im Hinblick auf seine mögliche Anwendung bei der Bewertung im Rahmen groß angelegter Sprachprüfungen und bezog dabei Bewertungen menschlicher Beurteilender unter Verwendung etablierter und innovativer Bewertungsmodelle mit ein. Die Ergebnisse zeigten, dass das LLM zwar Effizienzvorteile in Form schneller Bewertungen bot, jedoch bei der genauen Bewertung von Schülerantworten teilweise Defizite aufwies.

Die Bewertungsgüte des allgemeinen LLM zeigte sich variabel und erwies sich letztlich als unzureichend für den Einsatz in prüfungsrelevanten Kontexten mit hohen Anforderungen. Im Gegensatz dazu erzielte die menschliche Bewertung – insbesondere unter Verwendung der Overlay-Methode – die höchste Genauigkeit und war zugleich effizienter als die Bewertung durch das LLM.

Diese Erkenntnisse bestätigen erneut, dass menschliche Beurteilung der Ausgangspunkt jeder Sprachbewertung sein sollte, da interpretative Urteile unverzichtbar sind, wenn es um komplexe sprachliche Strukturen geht. KI-gestützte Werkzeuge können zwar zu bestimmten Bewertungszwecken beitragen, zeigen jedoch derzeit nicht das erforderliche Maß an Differenzierung und Tiefe für eine vollständig autonome Bewertung im Bereich der Sprachbildung.

Die zukünftige Weiterentwicklung KI-gestützter Sprachbewertungssysteme erfordert drei grundlegende Verbesserungen:

1. **Anpassungsfähigkeit an Inhalte** – Lehrkräfte sollten in der Lage sein, KI-Werkzeuge über benutzerfreundliche, nicht-technische Schnittstellen an spezifische Lehrpläne und Bewertungsformate anzupassen.

2. **Transparenz in Entscheidungsprozessen** – Um die Nachvollziehbarkeit von Bewertungsentscheidungen zu erhöhen, sollten erklärbare KI-Modelle eingesetzt werden, die darlegen, wie und warum bestimmte Bewertungen vorgenommen wurden.

3. **Mensch-KI-Kollaboration** – Es sollten strukturierte Bewertungsabläufe entwickelt werden, in denen KI-Systeme menschliche Bewertende bei ihren Entscheidungen unterstützen, jedoch die endgültige Entscheidungskompetenz weiterhin beim Menschen verbleibt.

Das bedeutet, dass KI-gestützte Bewertungssysteme in der schulischen Praxis als unterstützende Ergänzungen betrachtet werden sollten. Lehrkräfte können durch den gezielten Einsatz von KI zeitlich entlastet werden und sich verstärkt komplexen pädagogischen Aufgaben widmen. Voraussetzung dafür ist jedoch die Entwicklung transparenter, didaktisch fundierter Bewertungsmodelle sowie eine fundierte Aus- und Weiterbildung des pädagogischen Personals im Umgang mit KI-Technologien.

6.1 Empfehlungen für zukünftige Forschung

Die Ergebnisse dieser Studie haben weitreichende Implikationen für Forschende und Bildungseinrichtungen, die planen, Künstliche Intelligenz (KI) in ihren Bewertungsstrategien einzusetzen. Die Befunde legen nahe, dass allgemeine Sprachmodelle (LLMs) mit Vorsicht verwendet werden sollten – insbesondere im Kontext summativer und hochrelevanter Leistungsbewertungen. Zwar lassen sich bestimmte Einsatzmöglichkeiten im formativen Bereich erkennen (z. B. für Feedback oder die Bewertung von Entwürfen), doch treten Genauigkeitsprobleme hinreichend häufig auf, um zu bestätigen, dass LLMs nicht mit menschlichen Bewertenden gleichgesetzt werden können und keine Ersatzfunktion bei abschließenden Bewertungen erfüllen sollten. Lehrkräfte, die derartige Systeme einsetzen möchten, müssen gegenüber den Lernenden klare Erwartungen an die Rolle der KI formulieren sowie Verfahren etablieren, durch die KI-Ausgaben durch menschliche Prüfende validiert werden können.

Ein verantwortungsvoller Einsatz von KI in der Leistungsbewertung, der die damit verbundenen Potenziale erschließt, erfordert mehr als nur technisches Verständnis. Für eine sinnvolle Implementierung müssen verschiedene Rahmenbedingungen geschaffen werden: Schulungsprogramme für Lehrkräfte im Umgang mit KI, transparente Kriterien für deren Nutzung sowie hybride Bewertungsmodelle, in denen sich menschliche und maschinelle Bewertungen gegenseitig ergänzen. Langfristig müssen auch normative Fragen – etwa Fairness, Datenschutz und Rechenschaftspflicht – stärker ins Zentrum der Implementierungsstrategien rücken.

Ein weiterer zentraler Aspekt für den verantwortungsvollen Einsatz von KI in der Leistungsbewertung ist die Transparenz der zugrunde liegenden Prozesse. KI-Systeme müssen in der Lage sein, ihre Bewertungsentscheidungen klar und nachvollziehbar zu begründen. Das bedeutet, dass Modelle Erklärungen für ihre Bewertungsergebnisse liefern können müssen. Ebenso müssen die eingesetzten Texterkennungstechnologien (z. B. OCR) kontinuierlich weiterentwickelt werden, um handschriftliche Antworten oder gescannte Seiten zuverlässig und präzise verarbeiten zu können. Nur durch transparente Abläufe und eine präzise Datenerfassung kann Vertrauen in KI-gestützte Bewertungssysteme aufgebaut und ihre Akzeptanz im Bildungskontext nachhaltig gefördert werden.

Vor diesem Hintergrund ergibt sich ein erweitertes Forschungsdesiderat, das über die technischen Anforderungen hinausgeht. Künftige Forschungsarbeiten sollten sich auf die Eignung KI-gestützter Bewertungssysteme unter verschiedenen didaktischen Bedingungen und in unterschiedlichen sprachlichen Kontexten konzentrieren. Dabei sind insbesondere die Perspektiven von Lernenden und Lehrenden in Bezug auf die Akzeptanz und das Vertrauen in KI-Systeme von Bedeutung. Ethische Fragestellungen – wie der sensible Umgang mit personenbezogenen Daten, potenzielle Diskriminierungstendenzen (bewusst oder unbewusst) sowie die Orientierung an lernförderlichen Prinzipien – sollten integraler Bestandteil aller zukünftigen Untersuchungen sein.

Neben ethischen und didaktischen Aspekten rückt auch die technologische Weiterentwicklung zunehmend in den Fokus. KI-Plattformen müssen künftig eine

größere Bandbreite an Anpassungsmöglichkeiten im Design bieten. Es liegt deutliche Evidenz vor, dass generische Sprachmodelle den komplexen Anforderungen der Sprachbewertung nicht gerecht werden können – insbesondere bei morphologisch komplexen Sprachen wie dem Deutschen. Daher ist es besonders wichtig, diese Modelle weiterzuentwickeln und spezifisch an solche Sprachsysteme anzupassen. Nur so kann gewährleistet werden, dass KI-Systeme sprachliche Feinheiten und kontextuelle Bedeutungen adäquat erkennen.

Darüber hinaus erfordert der Einsatz dieser Systeme im schulischen Kontext eine flexible technische Gestaltung, die es Lehrkräften ermöglicht, Bewertungskriterien curricular passgenau zu definieren – nur auf diese Weise kann eine faire und verlässliche Bewertung sichergestellt werden. Die Weiterentwicklung von KI sollte darauf abzielen, Lehrkräften anpassungsfähige Bewertungsprotokolle zur Verfügung zu stellen, die es erlauben, Bewertungssysteme flexibel an ihre curriculare Praxis anzupassen – und dies ohne die Notwendigkeit umfangreicher Programmierkenntnisse. Bildungstechnologische Systeme, die solche lehrkraftgesteuerten Anpassungen ermöglichen, dürften nachhaltiger und erfolgreicher in den Sprachunterricht integriert werden können und gleichzeitig zur Einhaltung ethischer Standards beitragen.

Literaturverzeichnis

Adiguzel, Tufan / Kaya, Mehmet Haldun / Cansu, Fatih Kürşat (2023): Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. In: *Contemporary Educational Technology*, 15(3), ep429. https://doi.org/10.30935/cedtech/13152.

Ahmad, Sayed Fayaz / Rahmat, Mohd. Khairil / Mubarik, Muhammad Shujaat (2021): Artificial intelligence and its role in education. In: *Sustainability*, *13*(22), 12902. https://doi.org/10.3390/su132212902.

Aldosari, Share Aiyed M. (2020): The future of higher education in the light of artificial intelligence transformations. In: *International Journal of Higher Education*, 9(3), 145-151. https://doi.org/10.5430/ijhe.v9n3p145.

Bachman, Lyle F. (1990): *Fundamental considerations in language testing*. Oxford University Press.

Başaran, Bora (2025): The cultural dance of words: The transforming value of language teaching in the age of AI. In: Çevikkilıç, Deniz Beste (Hg.): International studies in educational sciences (Chapter 1). Serüven Yayınevi.

Boud, David (2000): Sustainable assessment: Rethinking assessment for the learning society. In: *Studies in Continuing Education*, *22*(2), 151–167. https://doi.org/10.1080/713695728.

Cerf, Vinton G. (2023): Large Language Models. *Communications of the ACM*, 66, 7 - 7. https://doi.org/10.1145/3606337.

Chen, Lijia / Chen, Pingping / Lin, Zhijian (2020): Artificial intelligence in education: A review. In: *IEEE Access*, *8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510.

Chen, Yong / Chen, Hongpeng / Su, Songzhi (2023): Fine-tuning large language models in education. In: *2023 13th International Conference on Information Technology in Medicine and Education (ITME)*. IEEE, 718–723. https://doi.org/10.1109/itme60234.2023.00148.

Cohen, Jacob (1960): A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104.

Dunbar, Stephen B. / Koretz, Daniel M. / Hoover, H.D. (1991): Quality control in the development and use of performance assessments. In: *Applied Measurement in Education*, *4*(4), 289–303. https://doi.org/10.1207/S15324818AME0404_3.

Durall, Eva / Kapros, Evangelos (2020): Co-design for a competency self-assessment chatbot and survey in science education. In: *International conference on human-computer interaction*. Springer, 13-24. https://doi.org/10.1007/978-3-030-50506-6_2.

Efremova, Nadezhda / Shvedova Svetlana / Huseynova, Anastasia (2019): The influence of assessment on learning motivation. *SHS Web of Conferences, 70*, 04003. https://doi.org/10.1051/shsconf/20197004003.

Endres, Christoph / Ibisch, Andrea (2025) Why one size doesn't fit all – Differenzierte Absicherung von LLMs. In: *Datenschutz Datensich* 49, 220–224. https://doi.org/10.1007/s11623-025-2075-6.

Evenddy, Sutrisno Sadji (2024): Investigating AI's automated feedback in English language learning. In: *FLIP: Foreign Language Instruction Probe, 3*(1), 76–87. https://doi.org/10.54213/flip.v3i1.401.

Felder, Ekkehard / Kückelhaus, Marcel (2025): Das definierende Sprachmodell (LLM): Anthropomorphisierung in der Mensch-Maschine-Interaktion. In: *Zeitschrift für Literaturwissenschaft und Linguistik,* 431-448. https://doi.org/10.1007/s41244-025-00380-7

Göçer, Ali (2024): Öğrencilerin dinleme ve konuşma becerilerinin uygulamalı sınavlarla ölçülüp değerlendirilmesine yönelik Türkçe öğretmenlerinin görüşleri. In: *International Journal of Language Academy*, *12*(3), 120–144.

Gunawan, Kadek Dwi Hendratma / Liliasari, Liliasari / Kaniawati, Ida / Setiawan, Wawan (2021): Implementation of competency enhancement program for science teachers assisted by artificial intelligence in designing HOTS-based integrated science learning. In: *Journal Penelitian dan Pembelajaran IPA*, 7(1), 55-65. https://doi.org/10.30870/jppi.v7i1.8655.

Hao, Jiangang / Davier, Alina A. / Yaneva, Victoria / Lottridge, Susan / Davier, Matthias / Harris, Deborah J. (2024): Transforming assessment: The impacts and implications of large language models and generative AI. In: *Educational Measurement: Issues and Practice*, 16-29. https://doi.org/10.1111/emip.12602.

Hu, Jingjing (2021): Teaching evaluation system by use of machine learning and artificial intelligence Methods. In: *International Journal of Emerging Technologies in Learning*, 16(5), 87-101. https://doi.org/10.3991/ijet.v16i05.20299.

Igaki, Takahiro / Kitaguchi, Daichi / Matsuzaki, Hiroki / Nakajima, Kei / Kojima, Shigehiro / Hasegawa, Hiro / Takeshita, Nobuyoshi / Kinugasa, Yusuke / Ito, Masaaki (2023): Automatic surgical skill assessment system based on concordance of standardized surgical field development using artificial intelligence. *JAMA Surgery*, e231131. https://doi.org/10.1001/jamasurg.2023.1131.

IU Internationale Hochschule (2024): *Lernreport 2024: Was treibt Menschen in Deutschland zum Lernen an?* https://www.iu.de/forschung/studien/lernreport-2024/ (Zugriff am 05.05.2025).

Jaiswal, Akanksha / Arun, C. Joe (2021): Potential of artificial intelligence for transformation of the education system in India. In: *International Journal of Education and Development Using Information and Communication Technology*, 17(1), 142-158.

Kafadar, Tuğba (2022): Oyunlaştırmanın eğitimdeki yeri. In: Kafadar, Tuğba / Can, Asena Ayvaz (Hg.), *Eğitimde oyunlaştırma*. Nobel Akademik Yayıncılık, 1–16.

Kankanamge, Dinesha / Wijiweera, C. / Ong, Z. / Preda, T. / Carney, T. / Wilson, M. / Preda, V. (2025): Artificial intelligence based assessment of minimally invasive surgical skills using standardised objective metrics – A narrative review. In: *The American Journal of Surgery*, *241*, 116074. https://doi.org/10.1016/j.amjsurg.2024.116074.

Lucena Sangreman Aldeman, Nayze / Sá Urtiga Aita, Keylla Maria de / Ponte Machado, Vinícius / Demes da Mata Sousa, Luiz Claudio / Gilberto Borges Coelho, Antonio / Socorro da Silva, Adalberto / Silva Mendes, Ana Paula da / Oliveira Neres, Francisco Jair de / Jamil Hadad do Monte, Semíramis (2021): Smartpath (k): A platform for teaching glomerulopathies using machine learning. In: *BMC Medical Education*, 21(1), 248. https://doi.org/10.1186/s12909-021-02680-1.

Luckin, Rose / Holmes, Wayne (2016): *Intelligence unleashed: An argument for AI in education*. Pearson.

Maghsudi, Setareh / Lan, Andrew / Xu, Jie / Schaar, Michaela (2021): Personalized education in the artificial intelligence era: What to expect next. In: *IEEE Signal Processing Magazine*, *38*(2), 37–50. https://doi.org/10.1109/MSP.2021.3055032.

Makridakis, Spyros / Petropoulos, Fotios / Kang, Yanfei (2023): Large Language Models: Their Success and Impact. In: *Forecasting*, 5(3), 536-549 https://doi.org/10.3390/forecast5030030.

Mede, Enisa / Atay, Derin (2017): English language teachers' assessment literacy: The Turkish context. *Dil Dergisi*, *168*(1), 43–60.

Millî Eğitim Bakanlığı (2023): Yazılı ve Uygulamalı Sınavlar Yönergesi. https://odsgm.meb.gov.tr/meb_iys_dosyalar/2023_10/12115933_MEB_yazili_ve_uygulamali_sinavlar_yonergesi.pdf (Zugriff am 05.05.2025).

Minaee, Shervin / Mikolov, Tomas / Nikzad, Narjes / Chenaghlu, Meysam / Socher, Richard / Amatriain, Xavier / Gao, Jianfeng (2024): Large language models: A survey. https://doi.org/10.48550/arXiv.2402.06196.

Mohan, G. Bharathi / Kumar, R. Prasanna / Krishh, P. Vishal / Keerthinathan, A. / Lavanya, G. / Meghana, Meka Kavya Uma / Sulthana, Sheba / Doss, Srinath (2024): An analysis of large language models: their impact and potential applications. In: *Knowl. Inf. Syst.*, 66, 5047-5070. https://doi.org/10.1007/s10115-024-02120-8.

Norcini, John / Anderson, Brownell / Bollela, Valdes / Burch, Vanessa / Costa, Manuel João / Duvivier, Robbert / Galbraith, Robert / Hays, Richard / Kent, Athol / Perrott, Vanessa / Roberts, Trudie (2011): Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. In: *Medical Teacher*, *33*, 206–214. https://doi.org/10.3109/0142159X.2011.551559.

Popenici, Stefan A. D. / Kerr, Sharon (2017): Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning, 12(1): 22*. Epub 2017 Nov 23. PMID: 30595727; PMCID: PMC6294271. https://doi.org/10.1186/s41039-017-0062-8.

Sarker, Ikbal H. (2022): AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, *3*(2), 158. https://doi.org/10.1007/s42979-022-01043-x.

Shao, Yueyang / Liu, Qimeng / Dong, Yaoyao / Liu, Jian (2024): Perceived formative assessment practices in homework and creativity competence: The mediating effects of self-confidence in learning and intrinsic motivation. *Studies in Educational Evaluation, 80*, 101376. https://doi.org/10.1016/j.stueduc.2024.101376.

Soliman, Hassan / Kravcik, Milos / Neumann, Alexander Tobias / Yin, Yue / Pengel, Norbert / Haag, Maike / Wollersheim, Heinz-Werner (2024): *Generative KI zur Lernenbegleitung in den Bildungswissenschaften: Implementierung eines LLM-basierten Chatbots im Lehramtsstudium*. *Proceedings of DELFI 2024*. Gesellschaft für Informatik e.V. 171-177 https://doi.org/10.18420/delfi2024_15.

Sullivan, Gail M. (2011): A primer on the validity of assessment instruments. *Journal of Graduate Medical Education*, *3*(2), 119–120. https://doi.org/10.4300/JGME-D-11-00075.1.

Tanır, Ahmet (2023): YouTube-assisted listening instruction (YALI): A study of listening comprehension and listening anxiety of university students of german as a foreign language. In: *Research on Education and Psychology (REP)*, 7(Special Issue 2), 270-299.

Tanrıkulu, Lokman / Üstün, Bilal (2020): Almanca öğretmenliği yüksek lisans öğrencilerinin lisansüstü eğitim yapma nedenlerine ilişkin nitel bir çalışma. In: *International Journal of Language Academy*, 8(5), 104–114. https://doi.org/10.29228/ijla.47061.

Thirunavukarasu, Arun James u.a. (2023): Large language models in medicine. In: *Nature Medicine, 29*(9), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8.

Tsagari, Dina (2011): Investigating the 'assessment literacy' of EFL state school teachers in Greece. In: Tsagari, Dina / Csépes, Ildikó (Hg.), *Classroom-based language assessment*. Peter Lang, 169–190.

Üstün, Ebru (2025): Kursplanung, Materialentwicklung und Kompetenzaufbau bei angehenden Fremdsprachenlehrkräften in der Türkei: Eine qualitative Fallstudie. In: *Diyalog Interkulturelle Zeitschrift Für Germanistik,* 13(1), 193-215. https://doi.org/10.37583/diyalog.1714784.

Üstün, Ebru / Üstün, Bilal / Karataş, Fatih (2024): K.I.-Literacy von Studierenden im Grundstudium. In: *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, (42), 404-415. DOI: https://doi.org/10.5281/zenodo.13980839.

Wang, Shan / Wang, Fang / Zhu, Zhen / Wang, Jingxuan / Tran, Tam / Du, Zhao (2024): Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, *252*, 124167. https://doi.org/10.1016/j.eswa.2024.124167.

Wiliam, Dylan (2011): What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001.

Winston, Patrick Henry (1992): *Artificial intelligence* (3rd ed.). Addison-Wesley Longman Publishing Co., Inc.

Yamtinah, Sri / Wiyarsi, Antuni / Widarti, Hayuni Retno / Shidiq, Ari Syahidul / Ramadhani, Dimas Gilang (2025): Fine-tuning AI models for enhanced consistency and precision in chemistry educational assessments. *Computers and Education: Artificial Intelligence*, *8*, 100399. https://doi.org/10.1016/j.caeai.2025.100399.

Yin, Shukang / Fu, Chaoyou / Zhao, Sirui / Li, Ke / Sun, Xing / Xu, Tong / Chen, Enhong (2023): A survey on multimodal large language models. *National Science Review, 11*. https://doi.org/10.1093/nsr/nwae403.

Zafari, Mostafa / Safari Bazargani, Jalal / Sadeghi-Niaraki, Abolghasem / Choi, Soo-Mi (2022): Artificial intelligence applications in K-12 education: A systematic literature review. *IEEE Access*, *PP*, 1–1. https://doi.org/10.1109/ACCESS.2022.3179356.