Vergleich von Bewertungen standardisierter Prüfungen durch Menschen und Maschine

Amir Mahdi Meshkin Mehr, Teheran

ISSN 1470 – 9570

Vergleich von Bewertungen standardisierter Prüfungen durch Menschen und Maschine

Amir Mahdi Meshkin Mehr, Teheran

Die Bewertung schriftlicher Prüfungsleistungen ist für Lehrkräfte zeitaufwendig und mit hohen Kosten für Institutionen sowie langen Wartezeiten für Lernende verbunden. Generative Sprachmodelle wie ChatGPT eröffnen neue Perspektiven für die Automatisierung dieser Prozesse. In der Studie wurden Texte von 20 Lernenden auf der Niveaustufe B2-1, die an einer standardisierten B1-Prüfung des Goethe-Instituts teilnahmen, von zwei zertifizierten Prüfenden sowie von ChatGPT korrigiert und nach den offiziellen Kriterien des Goethe-Instituts bewertet. Mithilfe von ICC-Analysen und Standardabweichungen wurde die Übereinstimmung zwischen menschlichen und KI-basierten Bewertungen untersucht. Die Ergebnisse zeigen, dass menschliche Bewertungen größere Streuungen aufweisen, während ChatGPT konstantere Urteile liefert. Eine besonders hohe Reliabilität ergibt sich durch die Aggregation beider Bewertungsarten. Die Befunde verdeutlichen Potenziale und Herausforderungen beim Einsatz von ChatGPT in standardisierten Prüfungen.

1 Einleitung

Die produktive Fertigkeit Schreiben ist eine der vier Grundfertigkeiten im Fremdsprachenunterricht und ein fester Bestandteil standardisierter Prüfungen. Die Korrektur und Bewertung schriftlicher Texte nehmen dabei eine zentrale Rolle in der Fremdsprachendidaktik ein. Es ist jedoch entscheidend, zwischen beiden Prozessen zu unterscheiden: Während Korrektur auf die Identifikation und Berichtigung sprachlicher, grammatikalischer sowie inhaltlicher Fehler abzielt, beschreibt Bewertung einen kognitiven Prozess der qualitativen Leistungseinschätzung. Diese erfolgt stets auf Grundlage von Wertmaßstäben, die in Form von Bewertungskriterien – bewusst oder unbewusst – konkretisiert werden. „Die Qualität einer schriftlichen Leistung wird als ‚gelungen' oder ‚nicht gelungen' bewertet" (Jost & Böttcher 2012: 123).

In standardisierten Prüfungen werden diese Kriterien systematisch operationalisiert: Mithilfe von Bewertungsskalen werden die Leistungen der Lernenden in Punktzahlen überführt, die schließlich in eine Gesamtnote verdichtet werden. Diese Form der Leistungsbeurteilung wird als *rater-mediated assessment* bezeichnet und verdeutlicht die zentrale Rolle der Bewertenden im Bewertungsprozess. Anders als maschinelle Bewertungssysteme transformieren menschliche Bewertende die beobachtete Leistung

nicht mechanisch in eine Punktzahl, sondern konstruieren aktiv ein individuelles Urteil. Eigenschaften wie Strenge oder Milde, das Verständnis der Kriterien sowie der Umgang mit der Skala beeinflussen dabei direkt das Bewertungsergebnis. Solche subjektiven Einflüsse werden als Bewertereffekte (*rater effects*) bezeichnet und führen dazu, dass die vergebenen Punktzahlen nicht ausschließlich die Leistung der Testteilnehmenden widerspiegeln (Eckes 2020: 153). Erwartungseffekte, Projektionsfehler, der Halo-Effekt sowie Strenge-, Milde- und Mittelwerttendenzen sind typische Ausprägungen dieser Effekte (Schott 2018). Die daraus resultierenden Schwankungen werden als Bewertervariabilität zusammengefasst (Eckes 2020: 153).

Um Bewertereffekte zu minimieren, setzen Bildungseinrichtungen auf Maßnahmen wie gezielte Prüferqualifizierung, unabhängige Mehrfachbewertungen und die Überprüfung der Bewertungsübereinstimmung (ebd.: 154). Das Goethe-Institut folgt diesem Ansatz, indem Lehrkräfte, die Prüfungen abnehmen möchten, eine spezifische Schulung absolvieren müssen, in der sie intensiv mit den Bewertungskriterien arbeiten und diese anhand von Prüfungsbeispielen anwenden. Die abschließende Qualifikationsprüfung überprüft die Bewertungskompetenz, indem die Ergebnisse mit offiziellen Referenzbewertungen abgeglichen werden. Nur bei hinreichender Übereinstimmung wird das Prüfzertifikat vergeben.

Trotz dieser Standardverfahren wird in der Forschung die begrenzte Wirksamkeit von Prüfertrainings und die alleinige Fokussierung auf Bewertungsübereinstimmung als Qualitätsmaß kritisch hinterfragt (Wind & Peterson 2018). In einer der ersten Studien im deutschsprachigen Raum untersuchte Eckes (2005) im Rahmen einer Studie, inwieweit Bewertereffekte das Bewertungsergebnis beeinflussen. Eine Many-Facet Rasch-Analyse von über 1.300 Prüfungsleistungen in der TestDaF-Prüfung zeigte, dass sich Bewertende in ihrer Strenge oder Nachsicht deutlich unterschieden, auch wenn sie innerhalb ihrer Bewertungen konsistent blieben. Besonders bei den Bewertungskriterien im Schreiben (37 %) und den Aufgaben im Sprechen (16 %) traten signifikante Inkonstanzen auf.

Neben den Bewertereffekten erschweren auch strukturelle Faktoren den Bewertungsprozess. Der hohe Zeitaufwand für manuelle Korrekturen, finanzielle Belastungen für Bildungseinrichtungen sowie verspätetes oder ausbleibendes Feedback beeinträchtigen Effizienz und Qualität. Besonders in Institutionen mit großem Prüfungsvolumen führen diese Herausforderungen zu langen Wartezeiten für Lernende und einer hohen

Belastung der Bewertenden. Die genannten Herausforderungen haben das Interesse an der Entwicklung automatisierter Bewertungssysteme verstärkt. In diesem Zusammenhang entstanden computergestützte Systeme, die schriftliche Texte automatisch bewerten und benoten. Solche Systeme, bekannt als Automated Essay Scoring (AES) oder Computer-Automated Scoring (CAS) (Yun 2023: 106), nutzen Algorithmen zur Bewertung von Aufgaben mit offenen Antworten (Ramineni & Williamson, 2013).

Obwohl AES-Systeme bislang nur vereinzelt in standardisierten Prüfungen Anwendung finden, eröffnen KI-gestützte generative Sprachmodelle wie ChatGPT neue Perspektiven für die automatische Bewertung schriftlicher Texte im Fremdsprachenunterricht. Im Rahmen einer explorativen Vergleichsstudie am Goethe-Institut Teheran sollen nun Hinweise auf die Beantwortung folgender Fragen gefunden werden:

1. Welche Übereinstimmungen und Abweichungen zeigen sich zwischen den Bewertungen zertifizierter Prüfender und den Bewertungen von ChatGPT bei der Beurteilung schriftlicher Texte?
2. Welche Muster in der Bewertungsstreuung zeigen sich bei ChatGPT im Vergleich zu menschlichen Prüfenden und welche Hinweise lassen sich daraus für den Einsatz generativer KI-Modelle in standardisierten Bewertungskontexten ableiten?

2 Automatische Textkorrektur und Textbewertung

In den 1960er Jahren begann der amerikanische Bildungspsychologe Ellis Batten Page mit der Entwicklung des Project Essay Grade (PEG), das automatische Textkorrekturen ermöglichen sollte (Page 1968). Pages Forschung unterschied zwischen intrinsischen (Trins) und approximativen (Prox) Bewertungskriterien, die in der automatischen Textkorrektur Anwendung fanden. Während Trins die menschlichen Bewertungskriterien wie Satzstruktur und Wortwahl umfassen, sind Proxen messbare Textmerkmale, die zur Qualitätsbewertung herangezogen werden (Bai et al. 2022). Die technischen Beschränkungen der damaligen Zeit verhinderten jedoch eine sofortige Implementierung von PEG, bis die Computerentwicklung in den 1990er Jahren fortgeschritten war und zur Einführung von AES-Systemen führte (vgl. Milewski 2022). Mit dem technologischen Fortschritt ab den 1990er Jahren wurden verschiedene AES-Systeme entwickelt und kommerzialisiert. Shermis (2014) bezeichnet e-rater (Educational Testing Service), Intelligent Essay Assessor (IEA, Pearson Knowledge Technologies), IntelliMetric

(Vantage Learning) sowie eine Weiterentwicklung des Project Essay Grade (PEG, Measurement Incorporated) als die bekanntesten AES-Systeme. Moderne AES-Systeme analysieren nicht mehr nur oberflächliche Textmerkmale wie Grammatik, Rechtschreibung und Textlänge, sondern beziehen zunehmend inhaltsbezogene Merkmale wie Kohärenz, Organisation, lexikalische Vielfalt und stilistische Aspekte in die Bewertung ein (Yun 2023). Dabei kommen unterschiedliche Technologien zum Einsatz: Während IEA auf Latent Semantic Analysis (LSA) basiert, verwenden e-rater und IntelliMetric Natural-Language-Processing-Verfahren (NLP), um syntaktische, semantische und diskursive Merkmale zu extrahieren (Koul, Clariana & Salehi 2005). Die Modelle dieser Systeme werden anhand umfangreicher Korpora von Texten trainiert, die zuvor von menschlichen Prüfenden bewertet wurden. Durch statistische Skalierungsmethoden werden die maschinell generierten Scores an die Verteilung menschlicher Bewertungen angepasst. Auf diese Weise können AES-Systeme menschliche Bewertungstendenzen emulieren und bewertendebedingte Variabilitäten abbilden (Yun 2023).

Trotz der Fortschritte in der AES-Technologie und der nachgewiesenen hohen Übereinstimmungsrate zwischen menschlicher und maschineller Bewertung (vgl. Kukich 2000; Attali & Burstein 2006; Ben-Simon & Bennett 2007), wird die Anwendung von AES in der Fremdsprachendidaktik kontrovers diskutiert. Während einige Systeme, wie e-rater, erfolgreich in standardisierten Tests wie dem Test of English as a Foreign Language (TOEFL) und dem Graduate Record Examination (GRE) eingesetzt werden (Uto et al. 2020), werden in anderen Studien Unterschiede zur menschlichen Bewertung und die Bedeutung der Lehrkraft in der Sprachbewertung betont (vgl. Zribi & Smaoui 2021; Toranj & Ansari 2012). In einer Meta-Analyse untersuchte Yun (2023) die Übereinstimmung und Abweichungen zwischen menschlicher und automatisierter Bewertung von englischen Aufsätzen, wobei die vier genannten AES-Systeme im Mittelpunkt standen. Ziel der Studie war es, die Zuverlässigkeit dieser AES-Systeme zu bewerten, indem sie die Effektstärken von Korrelationen und standardisierten Mittelwertsdifferenzen analysierte. In der Studie wurden Daten aus 15 Studien aus dem Zeitraum 1999–2014 gesammelt und analysiert, wobei E-Rater am häufigsten vertreten war. Die Ergebnisse zeigten eine hohe Korrelation zwischen menschlicher und maschineller Bewertung sowie minimale Abweichungen, was auf eine starke Übereinstimmung hindeutet. Allerdings variierten die Korrelationen

zwischen den Studien, während die Abweichungen konsistent waren. Die Forschungslage zum Einsatz von AES-Systemen im DaF/DaZ-Unterricht sowie in standardisierten Prüfungen im Bereich DaF ist derzeit noch lückenhaft.

3 Neue Möglichkeiten durch künstliche Intelligenz

Mit künstlicher Intelligenz (KI) und generativen Sprachmodellen können Ideen, deren Umsetzung vor einigen Jahren unrealistisch aussah, realisiert werden. Ein Beispiel dafür ist das generative Sprachmodell ChatGPT von der Firma OpenAI. Dieses Modell versteht, verarbeitet und reagiert auf menschliche Sprache, wodurch es die Interaktion und das Lernen im Fremdsprachenunterricht fördern kann. Es kann die Sprachkenntnisse der Lernenden analysieren und konstruktives Feedback geben, was zuvor hauptsächlich in den Aufgabenbereich menschlicher Lehrkräfte fiel. Dadurch können individuelle Lernprozesse unterstützt werden (Hein et al. 2024). Die Einsatzmöglichkeiten von ChatGPT im Fremdsprachenunterricht sind vielfältig. Tekin (2023) unterscheidet zwischen den folgenden Einsatzmöglichkeiten:

- ChatGPT als Generator: Texterfindung, Texterstellung, Erstellung von Unterrichtsplänen, Erstellung von Übungen und Prüfungen.
- ChatGPT als Transformator: Zusammenfassung, Vereinfachung, Übersetzung, Veränderung des Schreibstils und des Formats
- ChatGPT als Evaluator: Korrektur, Fehleranalyse, Bewertung, Abwägung
- ChatGPT als Kommunikator: Dialog, Sekretär, Hilfe bei Fragen.

Nach Horn (2023) eignet sich ChatGPT für die Korrektur, Erstellung sowie Anpassung von Texten im Fremdsprachenunterricht. Baskara (2023) hebt hervor, dass ChatGPT zahlreiche Vorteile bietet, darunter die Aufrechterhaltung des Interesses und der Motivation sowie Unterstützung bei der Entwicklung von Sprachkenntnissen. Da Lernende ihre Texte mit ChatGPT selbstständig auf Fehler überprüfen und zugleich erklärendes Feedback erhalten können, bietet das System Potenzial zur Förderung autonomen Lernens.

Mit der Funktion MyGPT bietet die Firma OpenAI die Möglichkeit, einen personalisierten Chatbot nach eigenen spezifischen Bedürfnissen und Vorlieben zu generieren. Den Chatbot kann man trainieren und anpassen, indem man ihm persönliche Daten oder spezifische Anweisungen vorgibt, um ihn auf ein spezielles Thema und

bestimmte Aufgaben zu konfigurieren. Ein Chatbot zur automatischen Bewertung von schriftlichen Texten auf B1-Niveau könnte wie folgt gestaltet sein:

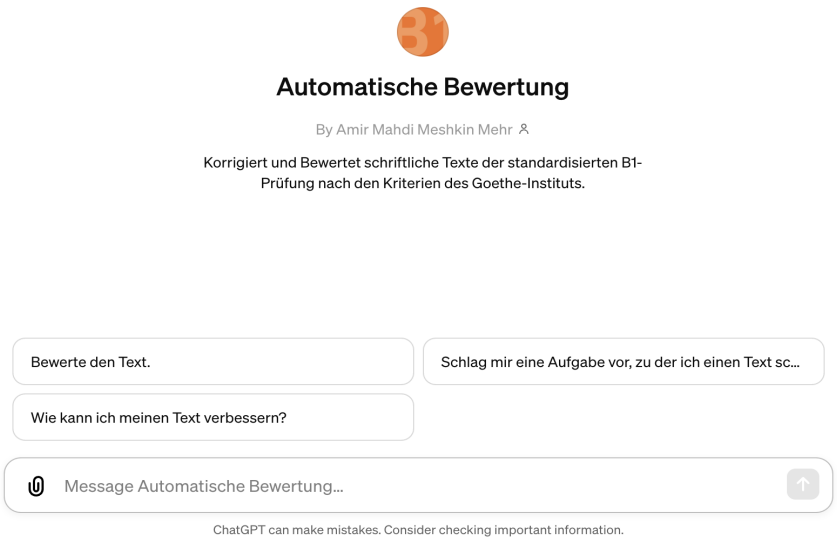

Abbildung 1: Ein Chatbot zur automatischen Bewertung schriftlicher Texte.

Bei der Erstellung des Chatbots müssen ein Name, eine Beschreibung sowie die Anweisung mit den Prompts (ähnlich wie in dieser Studie) vorgegeben werden. Der erstellte Bot lässt sich über einen Link mit anderen teilen. Um die Bewertung zu erhalten, genügt es, die Aufgabenstellung und den verfassten Text einzugeben. Von dieser Funktion können sowohl Bildungseinrichtungen als auch Lernende profitieren. Lernende, die sich auf standardisierte Prüfungen vorbereiten, haben oft keine klare Vorstellung davon, wie ihre schriftlichen Texte bewertet werden. Auch wenn Lehrkräfte über die notwendigen Prüferzertifikate und Erfahrung in der Bewertung solcher Texte verfügen sollten, fehlt es ihnen häufig an Zeit, zusätzliche Schülerarbeiten zu korrigieren und zu bewerten. Mit dem genannten Tool könnten Lernende ihre Texte eigenständig bewerten lassen, wodurch sie auch ein besseres Verständnis für ihre Schwachstellen erhalten können.

Polakova & Lenz (2024) untersuchten in einer quasi-experimentellen Studie die Wirkung KI-gestützten Feedbacks auf die Schreibkompetenz von Englischlernenden. Der Lernfortschritt wurde mittels Pre- und Post-Tests erfasst, wobei insbesondere Aspekte wie Präzision, Grammatik, Informationsdichte und Passivgebrauch analysiert wurden. Die Ergebnisse zeigten signifikante Verbesserungen in der Textqualität sowie eine hohe Akzeptanz des KI-basierten Feedbacks. Obwohl generative Sprachmodelle ein großes Potenzial für schnelles und individualisiertes Feedback bieten, sollte dieser

Vorteil mit Vorsicht betrachtet werden, da sich diese Technologien rasant weiterentwickeln und die empirische Evidenz zur langfristigen Wirksamkeit und Qualität des KI-gestützten Feedbacks bislang noch begrenzt ist.

Der Einsatz von generativen Sprachmodellen wie ChatGPT als AES-Systeme hat in der letzten Zeit deutlich an Aufmerksamkeit gewonnen. Erste empirische Untersuchungen zeigen sowohl Potenziale als auch Grenzen dieser neuen Technologien. Yamashita (2024) untersuchte in einer weiteren Studie die Leistungsfähigkeit von ChatGPT-4.0 als automatisiertes Bewertungssystem für argumentative Schreibaufgaben von Englisch-lernenden aus asiatischen Ländern. Die Bewertung von insgesamt 136 Aufsätzen erfolgte parallel durch menschliche Prüfende und ChatGPT-4.0. Mit Hilfe eines umfassenden Analysetools, das auch Verzerrungen in der Bewertung identifizierte, zeigte sich, dass ChatGPT-4.0 eine ähnliche Strenge wie menschliche Bewertende aufwies und konsistente Ergebnisse lieferte. Dennoch variierte die Übereinstimmung je nach Textqualität. Mizumoto & Eguchi (2023) gingen einen Schritt weiter, indem sie GPT-3 mit linguistischen Analyseverfahren kombinierten. In ihrer groß angelegten Studie mit über 12.000 TOEFL-Aufsätzen integrierten sie lexikalische, syntaktische und kohäsive Merkmale in den Bewertungsprozess. Diese Kombination ermöglichte eine deutlich präzisere Vorhersage der Bewertungsniveaus als GPT-3 allein. Besonders die lexikalischen Merkmale, wie etwa der Wortschatzumfang, verbesserten die Bewertungs-ergebnisse erheblich. Die Übereinstimmung mit menschlichen Referenzbewertungen erreichte dabei ein beachtliches Niveau. Allerdings gibt es auch Studien, die Schwächen in der Bewertung durch generative Sprachmodelle aufzeigen. Bui & Barrot (2024) analysierten 200 Essays, die sowohl von ChatGPT als auch von erfahrenen mensch-lichen Bewertenden beurteilt wurden. Im Gegensatz zu anderen Studien zeigte sich hier, dass ChatGPT durchgängig strengere Noten vergab und die Korrelation zu den menschlichen Bewertungen vergleichsweise schwach ausfiel. Zudem traten Inkonsisten-zen bei wiederholten Bewertungen desselben Textes auf, was auf Stabilitätsprobleme in der Bewertung hinweist. Die Autoren führen diese Ergebnisse auf systemische Ein-schränkungen im Trainingsdatensatz und auf Modifikationen durch Modell-Updates zurück, die das Bewertungsverhalten der KI beeinflussen können.

Bei der Bewertung schriftlicher Texte durch generative Sprachmodelle spielen sowohl die Gestaltung der Prompts (Eingabeaufforderungen) als auch die Feinabstimmung (Fine-Tuning) der Modelle eine entscheidende Rolle. Studien wie jene von Wang und

Gayed (2024) zeigen, dass eine gezielte Feinabstimmung die Bewertungsgenauigkeit erheblich steigern kann. Liu et al. (2025) untersuchten, inwiefern ein feinjustiertes ChatGPT-Modell in der Lage ist, argumentative Texte von Lernenden mit unterschiedlichen Erstsprachen (Deutsch, Französisch, Spanisch, Chinesisch) zuverlässig zu bewerten. Das Modell erzielte insgesamt hohe Übereinstimmungen mit menschlichen Bewertungen, zeigte jedoch Schwächen bei der Bewertung leistungsschwächerer Texte sowie bei Texten von L1-deutschen Schreibenden. Zusätzlich prüften die Autoren, ob sich die Kombination aus GPT-Bewertungen und linguistischen Komplexitätsmaßen positiv auf die Bewertungsergebnisse auswirkt. Diese Integration führte zwar nur zu geringen Verbesserungen im Gesamtergebnis, erwies sich jedoch bei schwächeren Texten als vorteilhaft. Die Studie macht deutlich, dass Fein-Tuning das Potenzial von Sprachmodellen im Bereich der automatisierten Schreibbewertung weiter steigern kann. Gleichzeitig zeigt sie, dass Faktoren wie L1-bedingte Verzerrungen und ungleich verteilte Leistungsniveaus in den Trainingsdaten gezielt adressiert werden müssen.

Ähnlich bei AES-Systemen gibt es bislang nur wenige empirische Studien, die den Einsatz generativer Sprachmodelle wie ChatGPT zur Bewertung schriftlicher Leistungen im DaF/DaZ-Unterricht und in standardisierten Prüfungen untersuchen, insbesondere im Vergleich zu menschlichen Bewertungen. Ziel des vorliegenden Artikels ist es, im Rahmen einer explorativen Vergleichsstudie das Potenzial des generativen Sprachmodells ChatGPT im Hinblick auf die Bewertung schriftlicher Texte in standardisierten Prüfungen für DaF zu analysieren und dessen Leistungsfähigkeit im Vergleich zu menschlichen Bewertungen kritisch zu reflektieren.

4 Vergleichsstudie

In der vorliegenden Studie wurden die Bewertungen schriftlicher Texte in standardisierten Prüfungen, die von ChatGPT erstellt wurden, mit den Bewertungen menschlicher Prüfender verglichen. Im Anschluss wurde die Streuung der Bewertungen innerhalb beider Gruppen analysiert. Dafür absolvierten 20 Lernende zu Beginn der Niveaustufe B2-1 am Goethe-Institut Teheran den schriftlichen Teil der standardisierten B1-Prüfung. Die Lernenden auf diesem Niveau hatten etwa 450 Unterrichtseinheiten absolviert, wobei für die Teilnahme an der B1-Prüfung üblicherweise ein Lernumfang von 300 bis 600 Unterrichtseinheiten empfohlen wird. Aus diesem Grund wurde diese Lerngruppe für die Untersuchung ausgewählt. Von den 20 Teilnehmenden waren 12 Frauen und 8

Männer im Alter zwischen 20 und 45 Jahren. Die geringe Stichprobengröße schränkt die Aussagekraft und Repräsentativität der Studie jedoch deutlich ein. Idealerweise hätte die Stichprobe mindestens 30 Personen auf dem unteren, 30 auf dem soliden und 30 auf dem höheren B1-Niveau umfassen sollen. Ursprünglich war vorgesehen, weitere Studien mit vergleichbarer Teilnehmerzahl durchzuführen; diese konnten jedoch aufgrund der Schließung des Goethe-Instituts in Teheran nicht realisiert werden.

Das Schreibmodul der B1-Prüfung am Goethe-Institut gliedert sich in drei Segmente, in denen die Teilnehmenden jeweils Texte unterschiedlicher Art verfassen müssen: eine persönliche Nachricht, eine Meinungsäußerung und eine formelle E-Mail.

Aufgabe 1:

Sie mussten wegen einer Sportverletzung zwei Wochen im Bett bleiben und haben eine besorgte Nachricht von Ihrem Freund/ Ihrer Freundin erhalten, weil er/sie so lange nichts von Ihnen gehört hat.

- Beschreiben Sie: Wie haben Sie sich verletzt?
- Begründen Sie: Warum hatten Sie keinen Kontakt zu Ihrem Freund/ Ihrer Freundin?
- Machen Sie einen Vorschlag für ein Treffen.

Schreiben Sie eine persönliche Nachricht (circa 80 Wörter). Schreiben Sie etwas zu allen drei Punkten. Achten Sie auf den Textaufbau (Anrede, Einleitung, Reihenfolge der Inhaltspunkte, Schluss).

Aufgabe 2:

Sie haben im Radio eine Diskussion Sendung zum Thema „Wie kann man den richtigen Beruf wählen?" gehört. Im Online-Gästebuch der Sendung finden Sie folgende Meinung:

Alexander: Ich finde, die Berufswahl ist eine sehr wichtige Entscheidung und deshalb braucht man dabei Hilfe und Beratung. Ich habe mit meinen Eltern darüber diskutiert. Das war für mich eine große Hilfe, denn es genügt nicht, dass der Beruf einem Spaß macht. Ganz wichtig ist, dass man später auch eine Arbeit finden kann.

Schreiben Sie nun Ihre Meinung (circa 80 Wörter).

Aufgabe 3:

Sie sind normalerweise täglich bis 16 Uhr im Büro. Diesen Montag möchten Sie aber eine Stunde früher gehen.
Schreiben Sie an Ihren Chef, Herrn Bäuerle. Beschreiben Sie Ihr Problem und bitten Sie höflich um Erlaubnis.
- Schreiben Sie eine E-Mail (circa 40 Wörter).
- Vergessen Sie nicht die Anrede und den Gruß am Schluss.

Abbildung 2: Beispielaufgabe zum Modul Schreiben in einer B1-Prüfung

Die Bewertung orientiert sich an den vom Goethe-Institut (2015: 42) definierten Kriterien Erfüllung, Kohärenz, Wortschatz und Struktur. Für jedes Kriterium werden fünf Leistungsstufen (A–E) unterschieden, wobei A die vollständige und angemessene Realisierung der Anforderungen und E eine durchgängig unangemessene Leistung bzw. deutliche Themenverfehlung kennzeichnet. Die maximale Punktzahl beträgt 100 (40 Punkte für die Teile 1 und 2, 20 Punkte für Teil 3).

5 Auswertung der Bewertungen mit ChatGPT

Für die Validierung der ChatGPT-Bewertungen wurden die Texte sowohl von zwei Prüfenden des Instituts als auch von ChatGPT 4 bewertet. Die Prüfenden verfügten über

das Prüferzertifikats des Goethe-Instituts für die B1-Prüfung und waren seit mehreren Jahren bei den Prüfungen gesetzt und hatten demnach jahrelange Erfahrung bei der Bewertung dieser Texte. Um die Validität der mit ChatGPT durchgeführten Bewertungen zu überprüfen, wurden die Texte dreimal von ChatGPT analysiert, um die Datenstreuung bei wiederholten Bewertungen zu berechnen. Während die Prüfenden mit den Bewertungskriterien vertraut sind, wurden diese ChatGPT explizit vorgegeben. Im Folgenden sind die Prompts sowie Beispiele dieser Prompts aufgeführt, wie sie in dieser Studie verwendet wurden.

- **Erläuterung der Anforderung (Was soll ChatGPT machen?)**: „Ich möchte, dass du die schriftlichen Texte von einer standardisierten Prüfung bewertest. Die Texte befinden sich am Ende dieses Abschnitts. Jeder Text wird nach den Kriterien Erfüllung, Kohärenz, Wortschatz sowie Strukturen korrigiert. Für jedes Kriterium können die Punkte A, B, C, D und E vergeben werden, wobei A die beste Note ist und E die Schlechteste. Die Kriterien werden im Folgenden detailliert aufgelistet."
- **Ausgangslage der Lernenden**: „Meine Lernenden lernen bei mir Deutsch und sind auf der Niveaustufe B2-1. Die folgende Prüfung ist eine standardisierte B1-Prüfung. Achte darauf, dass Fehler normal sind. Es geht darum zu wissen, was sie können und nicht was sie nicht können. Achte auf die Ausgangslage."
- **Kurze Beschreibung der B1-Prüfung des Goethe-Instituts**: „Die B1-Prüfung des Goethe-Instituts ist eine Deutschprüfung für Jugendliche und Erwachsene. Sie bestätigt eine selbstständige Verwendung der deutschen Sprache und entspricht der dritten Stufe (B1) auf der sechsstufigen Kompetenzskala des Gemeinsamen europäischen Referenzrahmens für Sprachen (GER)."
- **Kann-Beschreibung des Moduls Schreiben auf Niveau B1**: „Fremdsprachenlernende auf der Niveaustufe B1 können beim Modul Schreiben über Themen, die ihnen vertraut sind oder sie persönlich interessieren, einfache zusammenhängende Texte schreiben."
- **Detaillierte Beschreibung der Kriterien**:
- **Die Aufgabe, zu der die Lernenden die Texte schreiben müssen:** (vgl. Abbildung 2).
- **Geschriebene Texte der Lernenden**.

Jeder Textabschnitt wurde separat eingegeben und bewertet, um eine detaillierte Analyse der Bewertungskonsistenz zwischen den menschlichen Prüfenden und ChatGPT zu ermöglichen. Ein Beispiel für die Bewertung mit ChatGPT 4 sieht wie folgt aus:

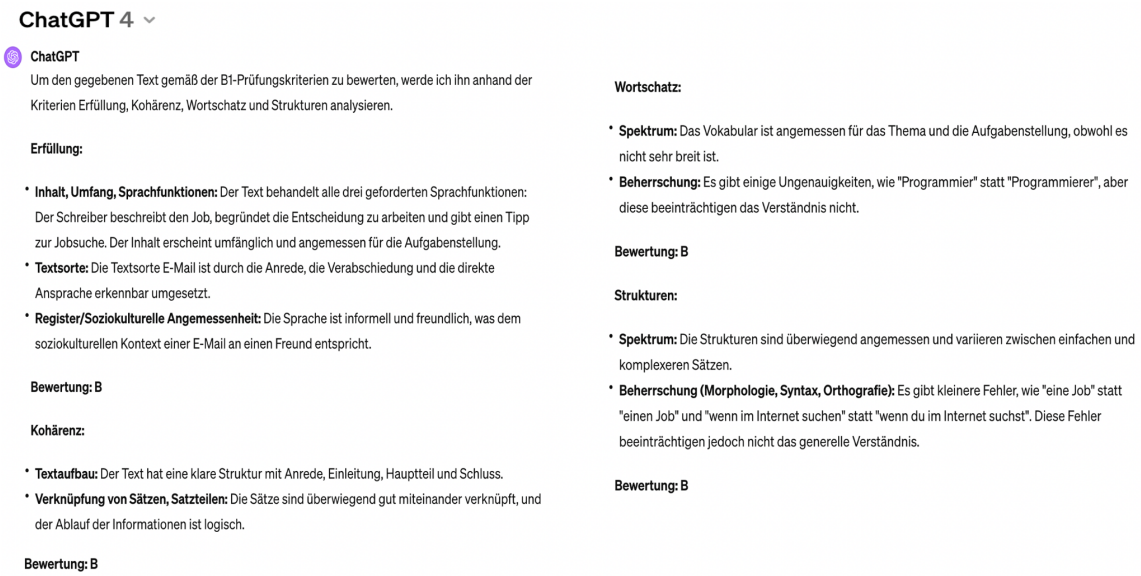

Abbildung 3: Bewertung eines Textes mit ChatGPT 4.

Im Folgenden finden sich die Bewertungen der Texte durch die Prüfenden sowie die drei Bewertungen mit ChatGPT:

Tabelle 1: Bewertungen der Texte von Prüfenden sowie ChatGPT 4.

	Prüfende 1	Prüfende 2	1. Bewertung ChatGPT	2. Bewertung ChatGPT	3. Bewertung ChatGPT	Durchschnitt Prüfende	Durchschnitt ChatGPT
TN 1	65	64,50	59,50	56,50	66	64,75	60,66
TN 2	30	30	47,50	40	40	30	42,50
TN 3	80	86	90	81	87	83	86
TN 4	45,5	53	59,50	78,50	64,50	49,25	67,50
TN 5	77,50	87,50	88,50	87	80	82,50	85,16
TN 6	54,50	73	52	70	47	63,75	56,33
TN 7	40	52,50	43,50	32,50	31	46,25	35,66
TN 8	49,50	57	60,50	59,50	53,50	53,25	57,83

	Prüfende 1	Prüfende 2	1.Bewertung ChatGPT	2. Bewertung ChatGPT	3. Bewertung ChatGPT	Durchschnitt Prüfende	Durchschnitt ChatGPT
TN 9	47,50	52,50	82	**80**	77	50	79,66
TN 10	68,50	73,50	76	79,50	78,50	71	78
TN 11	37,50	44,50	43,50	42	51	41	45,50
TN 12	70,50	78	73,50	87	72,50	74,25	77,66
TN 13	63,50	77	68,50	82	71	70,25	73,83
TN 14	76	86	81	78,50	78,50	81	79,33
TN 15	63,50	58,50	78,50	61	73,50	61	71
TN 16	63,50	65	80	77,50	65	64,25	74,16
TN 17	72,50	77,5	83,50	64,50	71	75	73
TN 18	67,50	77,50	86	61	77	72,50	74,66
TN 19	63,50	62,50	61	71	75	63	69
TN 20	89	86,50	82	83,50	80	87,75	81,83

Die Datenanalyse erfolgte mit der statistischen Software R. Nachfolgend erfolgt die Auswertung der beiden Fragestellungen.

5.1 Vergleichbarkeit der Bewertungen der Prüfende mit den Bewertungen von ChatGPT

Da bei den standardisierten Prüfungen des Goethe-Instituts die Endnote aus dem Mittelwert zweier Bewertungen ermittelt wird, wird im Folgenden zunächst auch der Vergleich der Durchschnittsbewertungen der Prüfenden und von ChatGPT betrachtet:

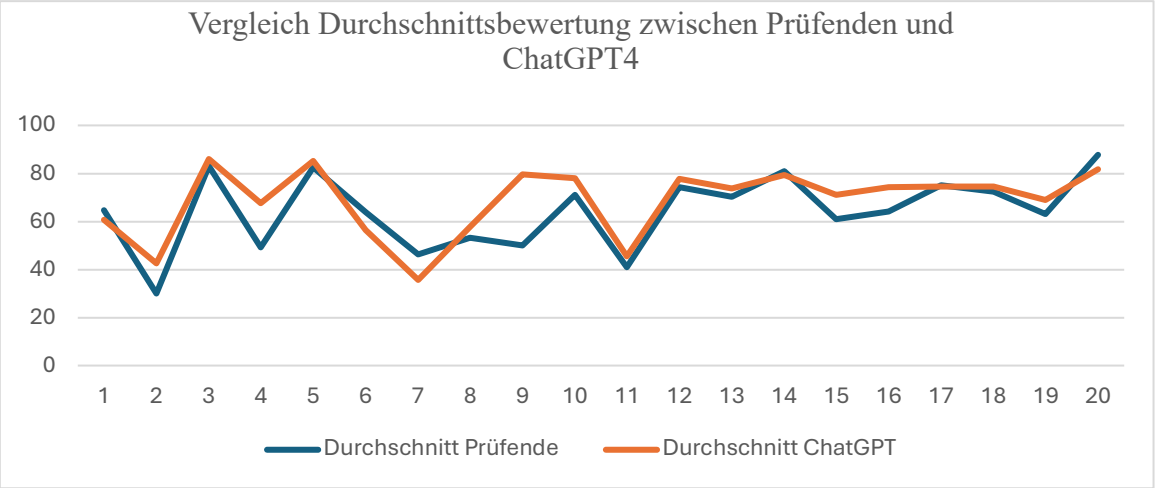

Diagramm 1: Durchschnittbewertungen Prüfende und ChatGPT.

Das Diagramm zeigt, dass bei vielen Texten eine hohe Übereinstimmung zwischen den Durchschnittsbewertungen der Prüfenden und den Bewertungen von ChatGPT besteht. Die Durchschnittsbewertung stellt jedoch kein alleiniges Kriterium dar, weshalb die Bewertungen einer vertieften Analyse bedurften. Zur Untersuchung der Bewertungsübereinstimmung zwischen den menschlichen Prüfenden und den ChatGPT-Bewertungen wurde hierfür der Intraclass Correlation Coefficient (ICC) herangezogen. Der ICC ist ein statistisches Maß, das die Übereinstimmung numerischer Bewertungen innerhalb einer Gruppe im Verhältnis zur Variation zwischen den bewerteten Objekten quantifiziert und damit sowohl systematische als auch zufällige Bewertungsunterschiede berücksichtigt. Da sowohl die Bewertungen der menschlichen Prüfenden als auch jene von ChatGPT als zufällige Stichproben aus größeren Grundgesamtheiten betrachtet werden, kam ein Two-Way-Random-Effects-Modell (ICC2) mit der Definition *absolute agreement* zur Anwendung, bei die exakte Übereinstimmung der Bewertungswerte berücksichtigt wird (Shrout & Fleiss, 1979). Zur Bestimmung der Reliabilität einzelner Bewertungen wurde der ICC(2,1) berechnet, zur Bestimmung der Reliabilität von Mittelwerten der menschlichen sowie maschinellen Bewertungen ICC(2,k). Die Berechnungen basierten auf den Rohdaten, um den Einfluss der Mittelwertbildung auf die Reliabilität sichtbar zu machen. Alle ICC-Werte werden mit 95%-Konfidenzintervallen, F-Statistik und p-Werten berichtet. Nach Koo & Li (2016) gelten ICC-Werte unter 0,50 als Hinweis auf geringe Reliabilität, Werte zwischen 0,50 und 0,75 auf mäßige, zwischen 0,75 und 0,90 auf gute und Werte über 0,90 auf ausgezeichnete Reliabilität. In der Tabelle 2 finden sich die berechneten ICC-Werte:

Tabelle 2: Berechnete ICC-Werte bei den Gesamtbewertungen.

Bewertungsgruppe	ICC-Wert
Alle fünf Bewertungen (ICC2,1)	0,75
Nur menschliche Bewertungen (ICC2,1)	0,87
Nur maschinellen Bewertungen (ICC2,1)	0,79
Mittelwerten der menschlichen und maschinellen Bewertungen (ICC2k)	0,94

Die Analyse der Bewertungsübereinstimmung zeigt für alle fünf Bewertungen (zwei menschliche und drei maschinelle Bewertungen) eine mäßige bis gute Reliabilität, was auf eine insgesamt solide, jedoch nicht perfekte Übereinstimmung zwischen allen Bewertungsgruppen hinweist. Die Übereinstimmung innerhalb der beiden menschlichen Prüfenden ist hoch und entspricht einer guten Reliabilität. Die drei maschinellen Bewertungen (ChatGPT) weisen ebenfalls eine gute Reliabilität auf, jedoch leicht unterhalb der Werte der menschlichen Prüfenden. Betrachtet man den Mittelwert der menschlichen Bewertungen im Vergleich zum Mittelwert der maschinellen Bewertungen, so zeigt sich eine ausgezeichnete Übereinstimmung. Dies verdeutlicht, dass sich Unterschiede zwischen einzelnen Bewertungen durch Mittelwertbildung deutlich reduzieren lassen und die Gesamtabstimmung zwischen Mensch und Maschine in aggregierter Form sehr hoch ist.

5.2 Übereinstimmung und Streuung der Bewertung innerhalb der Kriterien

Eine weitere Fragestellung betrifft die Unterschiede in den Bewertungen nach den einzelnen Kriterien. Zu diesem Zweck wurden zunächst die Bewertungen jedes Textabschnitts in eine Tabelle übertragen, deren Aufbau dem unten dargestellten Beispiel entspricht. Die vollständigen Daten sind dem Anhang zu entnehmen. Da sich die Bewertungsskalen der Kriterien in den Segmenten 1 und 2 der Prüfung vom dritten Teil unterscheiden, wurden alle Werte auf eine einheitliche Skala von 4 bis 0 normiert. Der Wert 4 entspricht der besten Bewertung (Note A), der Wert 0 der schlechtesten Bewertung (Note E). Anschließend wurde der ICC für jede Bewertungsgruppe separat berechnet.

Tabelle 3: Bewertungen der Texte von TN 1 nach den Kriterien.

TN/ Text	Bewertungsgruppe	Erfüllung	Kohärenz	Wortschatz	Struktur
TN 1 Text 1	Prüfende 1/2	3/2	3/4	2/3	2/2
	ChatGPT 1/2/3	2/3/3	2/2/2	2/2/2	2/2/2
TN 1 Text 2	Prüfende 1/2	3/3	3/3	2/2	2/2
	ChatGPT 1/2/3	3/3/3	2/2/3	2/2/3	2/1/3
TN 1 Text 3	Prüfende 1/2	3/4	3/2	3/2	3/2
	ChatGPT 1/2/3	4/2/3	2/4/3	2/2/2	2/2/2

Bei der Berechnung der ICC-Werte wurde analog zur vorherigen Vorgehensweise verfahren. Die berechneten ICC-Werte sind in der Tabelle 4 dargestellt:

Tabelle 4: Berechnete ICC-Werte nach den Kriterien.

Bewertungsgruppe	Erfüllung	Kohärenz	Wortschatz	Struktur
Alle fünf Bewertungen (ICC2,1)	0,54	0,62	0,61	0,62
Prüfende (ICC2,1)	0,67	0,66	0,70	0,78
ChatGPT (ICC2,1)	0,71	0,73	0,60	0,66
Mittelwerten der Prüfenden und ChatGPT (ICC2k)	0,85	0,89	0,88	0,89

Die ICC-Werte zeigen ein differenziertes Bild der Übereinstimmung zwischen menschlichen und maschinellen Bewertungen. Während die Einzelbewertungen alle Bewertungsgruppen (zwei menschliche und drei maschinelle) für die Kriterien Kohärenz (0,62), Wortschatz (0,61) und Struktur (0,62) eine moderate Reliabilität aufweisen, fällt die Übereinstimmung für das Kriterium Erfüllung mit 0,54 deutlich niedriger aus. Dies deutet darauf hin, dass die Bewertung der Erfüllung besonders stark von subjektiven Interpretationen abhängt und möglicherweise einer klareren Operationalisierung bedarf. Die unterschiedliche Auslegung des Kriteriums „Aufgabenerfüllung“ zeigt sich exemplarisch bei TN 9: Während die menschlichen Prüfenden den dritten Text als thematisch verfehlt einstuften und gemäß den Bewertungsregeln sämtliche übrigen Kriterien automatisch mit 0 Punkten bewerteten, vergab ChatGPT in allen drei Durchläufen für die Aufgabenerfüllung gute Noten.

Betrachtet man ausschließlich die menschlichen Prüfenden, steigt die Reliabilität für alle Kriterien an, insbesondere für Struktur (0,78) und Wortschatz (0,70). Dies unterstreicht die Fähigkeit menschlicher Prüfender, strukturelle und lexikalische Aspekte konsistent zu bewerten. Interessanterweise zeigt die maschinelle Bewertung für Kohärenz (0,73) und Erfüllung (0,71) sogar höhere Übereinstimmung als die menschlichen Prüfenden, während sie bei Wortschatz (0,60) schwächer abschneidet. Diese Diskrepanz könnte darauf hindeuten, dass die verwendeten Algorithmen semantische Kohärenz gut erfassen, jedoch Schwierigkeiten mit der Bewertung lexikalischer Vielfalt haben. Die Aggregation aller Bewertungen zu Mittelwerten führt durchweg zu exzellenter Reliabilität (0,85–0,89). Dies bestätigt, dass die Kombination menschlicher und maschineller Bewertungen die Zuverlässigkeit der Ergebnisse signifikant erhöht. Besonders hervorzuheben ist, dass dieser Effekt für alle vier Kriterien gleichermaßen gilt, auch für das anspruchsvolle Kriterium Erfüllung.

Neben der Berechnung der ICC-Werte wurde für jedes Kriterium auch die Standardabweichung der Bewertungen ermittelt. Die Standardabweichung liefert ergänzende Informationen zur Streuung der vergebenen Bewertungen und ermöglicht so eine differenziertere Interpretation der Reliabilitätsergebnisse. Hohe Standardabweichungen weisen auf größere Unterschiede in den Bewertungen und damit potenziell auf ein höheres Maß an Interpretationsspielraum oder Unsicherheit bei der Anwendung des Kriteriums hin, während niedrige Standardabweichungen auf eine einheitlichere Bewertungspraxis schließen lassen. Durch den Vergleich der Standardabweichungen zwischen menschlichen Prüfenden und ChatGPT lassen sich zudem Unterschiede in der Tendenz zur differenzierten oder gleichmäßigen Bewertung identifizieren. In Tabelle 5 finden sich die berechneten Standardabweichungen.

Tabelle 5: Standardabweichung der Bewertungen nach den Kriterien.

Kriterien	Berechnete Standardabweichung für alle Bewertungsgruppen				
	Alle Bewertungen	Prüfende	Nur ChatGPT	Durchschnitt Prüfende	Durchschnitt ChatGPT
Erfüllung	0,9548	1,0365	0.8946	0,9436	0,8022
Kohärenz	0,9893	1,0500	0.9266	0,9566	0,8279
Wortschatz	0,9100	0,9803	0,8500	0,9058	0,7527
Struktur	0.8977	0,9151	0,8786	0,8635	0,7668

Die berechneten Werte lassen sich zudem aus zwei Perspektiven analysieren:

I. Vergleich der Standardabweichung eines Kriteriums zwischen allen Bewertungen:

Erfüllung:
Prüfende > Alle Bewertungen > D Prüfende > Nur ChatGPT > D ChatGPT

Kohärenz:
Prüfende> Alle Bewertungen > D Prüfende > Nur ChatGPT > D ChatGPT

Wortschatz:
Prüfende > Alle Bewertungen > D Menschen > Nur ChatGPT > D ChatGPT

Struktur:
Prüfende > Alle Bewertungen > Nur ChatGPT > D Menschen > D ChatGPT

Die Ergebnisse zeigen, dass die Prüfenden über alle Bewertungskriterien hinweg die größte Streuung in ihren Urteilen aufweisen, während die durchschnittlichen Bewertungen von ChatGPT die geringste Streuung zeigen. Die zweitgrößte Streuung findet sich in den aggregierten Bewertungen der gesamten Bewertungsgruppe.

II. Vergleich der Standardabweichung einer Bewertung über alle Kriterien hinweg.

Alle Bewertungen: Kohärenz > Erfüllung > Wortschatz > Struktur
Prüfende: Kohärenz > Erfüllung > Wortschatz > Struktur
Nur ChatGPT: Kohärenz > Erfüllung > Struktur > Wortschatz
D Prüfende: Kohärenz > Erfüllung > Wortschatz > Struktur
D ChatGPT: Kohärenz > Erfüllung > Struktur > Wortschatz

Das Kriterium Kohärenz weist insgesamt die höchste Bewertungsstreuung auf, gefolgt von Erfüllung. Bei den Kriterien Wortschatz und Struktur ist die Bewertungstendenz insgesamt klarer als bei den anderen Kategorien. Rechtschreibfehler und grammatikalische Mängel führen hier häufig zu einer hohen Übereinstimmung unter den Prüfenden. Für automatische Bewertungssysteme lassen sich diese Kriterien zudem vergleichsweise gut abbilden, da sie auf definierten Regeln zu grammatikalischen Kompetenzen und zum Wortschatzumfang der Lernenden basieren können. Demgegenüber erfordern die Kriterien Kohärenz und Erfüllung eine stärker inhaltlich-analytische Beurteilung. Während Struktur und Wortschatz überwiegend formale und klar abgrenzbare Aspekte betreffen, sind Kohärenz und Erfüllung stärker durch interpretative Einschätzungen geprägt. Die Bewertung inhaltlicher Zusammenhänge und der Aufgabenerfüllung ist daher mit einer größeren Streuung zwischen den Prüfenden verbunden und weist eine höhere Komplexität auf.

6 Fazit

Die Ergebnisse verdeutlichen, dass sowohl menschliche als auch maschinelle Bewertungen eine gute Reliabilität aufweisen, wobei die Übereinstimmung zwischen den menschlichen Prüfenden etwas höher ausfällt. Die Aggregation der Bewertungen zu Mittelwerten führt in allen Fällen zu einer ausgezeichneten Übereinstimmung und unterstreicht das Potenzial einer kombinierten Nutzung von Mensch und Maschine, um die Bewertungsgenauigkeit in standardisierten Prüfungen zu erhöhen. Die Ergebnisse zeigen zudem, dass die Reliabilität je nach Kriterium und Bewertungsgruppe variiert: Menschliche Prüfende waren bei formalen Kriterien wie Struktur und Wortschatz insgesamt konsistenter, während ChatGPT bei Kohärenz und Aufgabenerfüllung teils höhere Übereinstimmung erreichte. Am geringsten war die Übereinstimmung jedoch genau bei Kohärenz und Aufgabenerfüllung. Diese inhaltlich geprägten Kriterien lassen mehr Interpretationsspielraum und führen dadurch zu größerer Streuung. Die Kombination beider Bewertungsarten führte in allen Kriterien zu exzellenter Reliabilität, auch wenn bei Kohärenz und Aufgabenerfüllung die Streuung am größten blieb. Es ist jedoch zu beachten, dass diese Ergebnisse aufgrund des begrenzten Umfangs der Stichprobe mit einer gewissen Zurückhaltung interpretiert werden sollten. Nichtsdestotrotz legen die Befunde nahe, dass generative Sprachmodelle wie ChatGPT potenziell in der Lage sein können, die Bewertung schriftlicher Prüfungsleistungen zu automatisieren, was eine bedeutende Innovation im Bereich der Bewertung und Korrektur von standardisierten Prüfungen darstellen könnte. Sie könnten wie bei der TOEFL-Prüfung die Zweitbewertung übernehmen oder die Prüfungsbewertung voll automatisieren, was beträchtliche Vorteile für Bildungseinrichtungen und Lernende mit sich bringen würde: Eliminierung langer Wartezeiten auf Prüfungsergebnisse, Entlastung der Lehrkräfte, Steigerung der Bewertungseffizienz sowie ausformuliertes Feedback. Ob der Einsatz generativer Sprachmodelle tatsächlich zu einer Reduktion von Aufwand und Kosten führen kann, lässt sich derzeit nur schwer vorhersagen. Die Kosten standardisierter Prüfungen umfassen unter anderem das Honorar für die Prüfenden. Einerseits liegt die Annahme nahe, dass sich durch den Einsatz automatisierter Bewertungssysteme diese Kosten senken lassen. Andererseits ist zu berücksichtigen, dass die Entwicklung eines generativen Sprachmodells für die Bewertung schriftlicher Texte in standardisierten Prüfungen mit erheblichen Investitionen verbunden ist. Hinzu kommen laufende Ausgaben für die

technische Infrastruktur, den IT-Support sowie gegebenenfalls für regelmäßige Anpassungen und Qualitätssicherungsmaßnahmen.

Dennoch sollten weiterführende Untersuchungen klären, inwieweit generative Sprachmodelle wie ChatGPT auch bei Texten höherer Niveaustufen zuverlässige Ergebnisse liefern können, da hier sowohl die Textlänge als auch die Komplexität der Sprache und somit die Anforderungen an die Korrektur und Bewertung zunehmen. Im Gegensatz dazu könnten Texte auf Anfängerniveau, wie A1, aufgrund ihrer Einfachheit möglicherweise leichter von solchen Systemen bewertet werden. Diese Aspekte gilt es in zukünftigen Studien zu erforschen, um das volle Potential der maschinellen Bewertung zu erschließen und deren Grenzen zu definieren. Ein weiterer vielversprechender Ansatzpunkt für zukünftige Forschung stellt die maschinelle Bewertung von mündlichen Prüfungsteilen dar. Es sei zu untersuchen, ob mündliche Texte unter anderem nach Bewertungskriterien wie Aussprache oder Interaktion automatisch und maschinell bewertet werden können.

Bibliographie

Attali, Yigal; Burstein, Jill (2006) Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment* 4: 3, 13–18. https://ejournals.bc.edu/index.php/jtla/article/view/1650

Bai, John Y. H.; Zawacki-Richter, Olaf; Bozkurt, Aras; Lee, Kyungmee; Fanguy, Mik; Cefa Sari, Bin; Marín, Vikcotira I. (2022) Automated essay scoring (AES) systems: Opportunities and challenges for open and distance education. *Proceedings of the Tenth Pan-Commonwealth Forum on Open Learning (PCF10), 14.–16. September.* https://doi.org/10.56059/pcf10.8339

Baskara, F. X. Risang (2023) Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning* 5: 1, 44–55. https://doi.org/10.31763/ijele.v5i1.858

Ben-Simon, Anat; Bennett, Randy Elliot (2007) Toward theoretically meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment* 6: 1. https://files.eric.ed.gov/fulltext/EJ838611.pdf

Bui, Ngoc My; Barrot, Jessie S. (2025) ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies* 30, 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Eckes, Thomas (2005) Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly* 2: 3, 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, Thomas (2020) Many-facet Rasch measurement: Implications for rater-mediated language assessment. In: Aryadoust, Vahid; Raquel, Michelle (Hrsg.) *Quantitative Data Analysis for Language Assessment. Volume II: Advanced Methods.* Routledge, 153–176.

Goethe-Institut (2015) *Zertifikat B1: Deutschprüfung für Jugendliche und Erwachsene – Modellsatz Erwachsene.* https://www.goethe.de/pro/relaunch/prf/materialien/B1/b1_modellsatz_erwachsene.pdf

Goethe-Institut (2023) *Goethe-Zertifikat B1: Durchführungsbestimmungen.* https://www.goethe.de/pro/relaunch/prf/de/Durchfuehrungsbestimmungen_B1.pdf

Hein, Laura; Högemann, Malte; Illgen, Katharina-Maria; et al. (2024) ChatGPT als Unterstützung von Lehrkräften – Einordnung, Analyse und Anwendungsbeispiele. *HMD* 61, 449–470. https://doi.org/10.1365/s40702-024-01052-9

Horn, Christian (2023) Im Dialog mit ChatGPT – Vier potenzielle Einsatzbereiche des Chatbots im Sprachunterricht und beim Fremdsprachenlernen im Test. *DaF-Szene Korea* 56, 109–125. https://lvk-info.org/wp-content/uploads/2023/06/DaF-Szene-Korea-56-WEB.pdf

Jost, Jörg; Böttcher, Ingrid (2012) Leistungen messen, bewerten und beurteilen. In: Becker-Mrotzek, Michael; Böttcher, Ingrid (Hrsg.) *Schreibkompetenz entwickeln und beurteilen.* Cornelsen, 113–144.

Koo, Terry K.; Li, Mae Y. Li BPS (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15: 2, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koul, Ravinder; Clariana, Roy B.; Salehi, Roya (2005) Comparing several human and computer-based methods for scoring concept maps and essays. *Journal of Educational Computing Research* 32: 3, 227–239. https://doi.org/10.2190/5X9Y-0ETN-213U-8FV7

Kukich, Karen (2000) Beyond automated essay scoring. *IEEE Intelligent Systems* 15: 5, 22–27. https://people.ischool.berkeley.edu/~hearst/papers/ieee-essay-grading.pdf

Liu, Yingying; Qi, Huilei; Lu, Xiaofei (2025) Enhancing GPT-based automated essay scoring: The impact of fine-tuning and linguistic complexity measures. *Computer Assisted Language Learning* 1–20. https://doi.org/10.1080/09588221.2025.2518430

Milewski, Glenn (2022) Ahead of the curve: How PEG™ has led automated scoring for years. https://www.erblearn.org/blog/ahead-of-the-curve-how-peg-has-led-automated-scoring-for-years/

Mizumoto, Atsushi; Eguchi, Masaki (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

Page, Ellis B. (1968) The use of the computer in analyzing student essays. *International Review of Education* 14: 2, 210–225. https://doi.org/10.1007/BF01419938

Polakova, Polakova; Ivenz, Petra (2024) The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education* 11: 1. https://doi.org/10.1080/2331186X.2024.2410101

Ramineni, Chaitanya; Williamson, David M. (2013) Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing* 18: 1, 25–39. https://doi.org/10.1016/j.asw.2012.10.004

Schott, Reinhard (2018) Beurteilen von Prüfungen. *Infopool besser lehren.* https://infopool.univie.ac.at/fileadmin/user_upload/p_infopool/PDFs/Pruefen_u_Beurteilen/04_Beurteilen_von_Pru__fungen.pdf

Shermis, Mark D. (2014) State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing* 20, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shrout, Patrick; Fleiss, Joseph L. (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86: 2, 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Tekin, Özlem (2023) ChatGPT im Unterricht Deutsch als Fremdsprache. *Alman Dili Ve Kültürü Araştırmaları Dergisi* 5: 2, 135–163. https://doi.org/10.55143/alkad.1390420

Toranj, Somaye; Ansari, Dariush Nejad (2012) Automated versus human essay scoring: A comparative study. *Theory and Practice in Language Studies* 2, 719–725. https://doi.org/10.4304/tpls.2.4.719-725

Uto, Masaki; Xie, Yikuan; Ueno, Maomi (2020) Neural automated essay scoring incorporating handcrafted features. In: *Proceedings of the 28th International Conference on Computational Linguistics,* 6077–6088. https://aclanthology.org/2020.coling-main.535.pdf

Wang, Qiao; Gayed, John Maurice (2024) Effectiveness of large language models in automated evaluation of argumentative essays: Finetuning vs. zero-shot prompting. *Computer Assisted Language Learning* 1–29. https://doi.org/10.1080/09588221.2024.2371395

Wind, Stefanie A.; Peterson, Meghan E. (2018) A systematic review of methods for evaluating rating quality in language assessment. *Language Testing* 35, 161–192. https://doi.org/10.1177/0265532216686999

Yamashita, Taichi (2024) An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics* 3: 3, 100133. https://doi.org/10.1016/j.rmal.2024.100133

Yun, Jiyeo (2023) Meta-analysis of inter-rater agreement and discrepancy between human and automated English essay scoring. *English Teaching* 78: 3, 105–124. https://doi.org/10.15858/engtea.78.3.202309.105

Zribi, Rania; Smaoui, Chokri (2021) Automated versus human essay scoring: A comparative study. *International Journal of Information Technology and Language Studies* 5: 1, 62–71. https://journals.sfu.ca/ijitls/index.php/ijitls/article/view/199

Kurzbiografie

Amir Mahdi Meshkin Mehr ist Gastdozent an der Universität Teheran an der Fakultät für deutsche Sprache und Literatur. Zuvor unterrichtete er über acht Jahre am Goethe-Institut in Teheran. Er promovierte an der Pädagogischen Hochschule Schwäbisch Gmünd. Seine Forschungsschwerpunkte liegen auf computergestütztem Fremdsprachenunterricht und multimedialem Lernen. E-Mail: Meshkinmehr@davilo.org

Schlagwörter

Automatisierte Aufsatzbewertung, Generative Sprachmodelle, ChatGPT, Deutsch als Fremdsprache/Zweitsprache

Keywords

Automated essay scoring, large language models, ChatGPT, German as a foreign/second language