Konrad Zerr[1] & Margarita Bidler[2]

Menschliche vs. KI-basierte Bewertung von Abschlussarbeiten als Goldstandard

Zusammenfassung

Die Bewertung wissenschaftlicher Abschlussarbeiten durch menschliche Gutachter:innen gilt traditionell als Goldstandard. Der vorliegende Beitrag hinterfragt diese Annahme, diskutiert typische Schwächen und Stärken menschlicher Bewertungen und stellt sie sechs KI-Modellen gegenüber. Auf Basis von 52 Bachelorarbeiten werden statistische Maßzahlen der Übereinstimmung sowie qualitative Tiefe der Gutachten verglichen. Die Ergebnisse zeigen zumindest für die hier betrachtete Gutachter:innen-Stichprobe, dass Menschen die Notenskala breiter nutzen und oft milder bewerten, während KI-Modelle zu Nivellierung neigen. In der qualitativen Leistungsbewertung finden KI-Modelle systematisch ergänzende Ansatzpunkte; ein starkes Argument für hybride Mensch-KI-Ansätze.

Schlüsselwörter

Künstliche Intelligenz; Automated Essay Scoring; Abschlussarbeiten; Notengebung; Hybrid Assessment

[1] Corresponding Author; Hochschule Pforzheim; konrad.zerr@hs-pforzheim.de

[2] Hochschule Pforzheim; margarita.bidler@hs-pforzheim.de; ORCiD 0000-0002-2881-4140

https://doi.org/10.21240/zfhe/21-3/09

Human vs. AI-based: Evaluating the evaluation standards for academic theses

Abstract

The assessment of academic theses by human examiners is traditionally considered the gold standard. This article challenges this assumption, discusses typical strengths and weaknesses of human evaluations, and contrasts them with assessments of six AI models. Based on 52 bachelor's theses, statistical measures of similarity as well as the qualitative depth of the reports are compared. The results show – at least for the sample of assessors considered here – that people tend to use the grading scale more broadly and often award more lenient marks, whilst AI models tend to level out marks. In qualitative performance assessment, AI models systematically identify complementary aspects—a strong argument for hybrid human-AI approaches.

Keywords

artificial intelligence; automated essay scoring; theses; grading;
hybrid assessment

1 Einleitung

Die Bewertung studentischer Abschlussarbeiten erfolgt traditionell durch Menschen – z. B. betreuende Dozierende. Diese Form der Leistungsbeurteilung durch menschliche Gutachter:innen dient in der Hochschulpraxis weithin als Goldstandard, an dem alternative Verfahren gemessen werden. Dieser Praxis liegt die implizite Annahme zugrunde, dass das menschliche Urteil objektiv, fachlich fundiert und zuverlässig ist. Doch ist der Mensch tatsächlich ein verlässlicher Maßstab, gerade im Vergleich zu modernen KI-Systemen?

Vorangegangene Arbeiten zur Notenvergabe zeigen, dass selbst erfahrene Beurteiler:innen bei identischen Texten zu divergierenden Urteilen gelangen können (Bloxham et al., 2016; Starch & Elliott, 1912). Die Literatur identifiziert dabei eine Vielzahl systematischer Einflussfaktoren, darunter Subjektivität, inkonsistente Kriteriengewichtung, situative Effekte sowie bewusste und unbewusste Biases (Erturk et al., 2022; Malouff & Thorsteinsson, 2016; Link & Koltovskaia, 2023).

Parallel zur Erforschung dieser menschlichen Schwächen hat sich in den letzten Jahren die Forschung zu KI-basierten Bewertungssystemen stark entwickelt. Diese Systeme versprechen eine höhere Konsistenz, Reproduzierbarkeit und Skalierbarkeit, werden jedoch ebenfalls kritisch diskutiert (Dikli, 2006; Link & Koltovskaia, 2023). Damit stehen sich zwei Bewertungsansätze gegenüber, die es in ihren spezifischen Stärken und Schwächen zu verstehen gilt. Vor diesem Hintergrund stellt der vorliegende Beitrag diese zwei Verfahren – Mensch vs. KI – gegenüber. Ziel ist es, Fehlerquellen menschlicher und künstlicher Bewertungen systematisch aufzuzeigen, empirisch zu prüfen und tatsächliche Unterschiede in der Bewertung sichtbar zu machen.

Im Folgenden werden zunächst typische Schwächen menschlicher Bewertung beleuchtet. Dem gegenübergestellt werden bekannte Fehlerquellen von KI. Anschließend werden die identifizierten Schwächen einer empirischen Überprüfung unterzogen und Urteile beider Systeme verglichen. Die Ergebnisse werden diskutiert und hybride Verfahren als potenzielle Handlungsempfehlung vorgestellt.

2 Theoretischer Rahmen

2.1 Schwächen menschlicher Bewertungen

Menschliche Beurteilungen gelten zwar als Goldstandard, sind aber keineswegs frei von Fehlerquellen. Die Literatur identifiziert Schwächen und Verzerrungen in menschlichen Bewertungsprozessen – insbesondere bei komplexen schriftlichen Leistungen wie Abschlussarbeiten.

Subjektivität:
Menschen nutzen oft implizite, persönliche Maßstäbe bei der Notenfindung. Zwei Gutachter:innen können dasselbe Werk unterschiedlich einschätzen, weil sie verschiedene Aspekte subjektiv höher gewichten. Studien zur Bewertung von Hochschulleistungen zeigen, dass menschliche Bewerter:innen ohne strikte Kriterien häufig zu uneinheitlichen Urteilen kommen und nur moderate Interrater-Reliabilitäten erreichen (Brimi, 2011; Bloxham et al., 2016). Auch bleiben menschliche Urteile anfällig für Drift und Ermüdung. So zeigt etwa eine Studie von Erturk et al. (2022), dass monotone, sich wiederholende Korrekturphasen zu Langeweile führen, die wiederum die Genauigkeit und Strenge von Bewertungen beeinflusst.

Biases:
Menschen lassen sich – oft unbewusst – von irrelevanten Informationen beeinflussen. Eine Meta-Analyse von Malouff und Thorsteinsson (2016) zeigt, dass Bias auftritt, sobald zusätzliche Details über Studierende bekannt sind. So könnte eine Dozentin, die den Studierenden als sozial engagiert kennt, zu tendenziell milderen Urteilen neigen. Die Studie identifiziert eine Reihe solcher Einflussfaktoren: Unterschiede im ethnischen Hintergrund, im sozialen Status oder in der physischen Attraktivität. Besonders problematisch ist, dass diese Verzerrungen häufig implizit wirken – die Gutachter:innen halten sich selbst weiterhin für objektiv (Malouff, 2008).

Halo-Effekte und Verlust von Detaildifferenzen:
Der Halo-Effekt ist ein klassischer Urteilsfehler, bei dem ein dominanter Gesamteindruck alle Einzelbewertungen überstrahlt. Erscheint eine Arbeit insgesamt sehr hochwertig, so werden auch einzelne Kriterien milde bewertet – selbst wenn in einigen Bereichen Schwächen vorliegen. Empirisch schlägt sich dies in der sehr hohen inneren Konsistenz der Teilnoten nieder (Schmidt et al., 2023). Dies führt neben der offensichtlichen Verzerrung zu einem Verlust der Detaildifferenzen in der Bewertung.

Inkonsistente Kriteriengewichtung und Kalibrierung:
Ohne klare Raster läuft die menschliche Bewertung Gefahr verschieden kalibriert zu sein. Eine Prüferin kann etwa bei einer Serie schwacher Arbeiten die eigenen Maßstäbe anpassen (Framingeffekt). Auch der Vergleich der Kriterien ist mitunter intransparent: Manche Gutachter:innen legen viel Wert auf Formalia und weniger auf Kreativität, andere umgekehrt. Die Forschung empfiehlt daher ausdrücklich den Einsatz standardisierter Rubrics, moderierter Absprachen und systematischer Kalibrierung, um die Bewertung zu stabilisieren (z. B. Chakrabarty et al., 2021; Villa et al., 2020).

Grenzen der Expertise:
Ein oft übersehener Faktor ist die Begrenzung des Fachwissens einzelner Gutachter:innen. Fehlendes Detailwissen kann dazu führen, dass wichtige Fehler nicht erkannt werden. Auch besteht die Gefahr, dass Prüfer:innen sich außerhalb ihres Kerngebietes stärker an formalen Aspekten orientieren, anstatt inhaltliche Tiefe angemessen zu beurteilen. Studien zeigen dabei, dass Menschen unterschiedliche inhaltliche Schwerpunkte setzen, was sich direkt in variierender Benotung niederschlägt (Stolpe et al., 2021).

Zusammenfassend weisen menschliche Bewertungen eine Reihe von Schwächen auf, was einen berechtigten Zweifel daran aufkommen lässt, ob die menschliche Benotung als „Goldstandard“ gelten kann.

2.2 Fehlerquellen von KI-Bewertungen

KI-basierte Bewertungssysteme, insbesondere solche auf Basis großer Sprachmodelle (LLMs) oder automatisierter Essay-Scoring-Algorithmen (AES), versprechen objektive und konsistente Beurteilungen. Dennoch sind auch KI-Bewertungen keineswegs frei von Fehlern, wie nachfolgend dargestellt wird.

Grauzonenproblem der KI – Tendenz zur „Lehrbuchkonformität" und Schwierigkeiten mit Kreativität:
KI-Systeme orientieren sich an den Daten, mit denen sie trainiert wurden. Entsprechend bevorzugen sie Arbeiten, die den „dominanten Mustern" der Trainingsdaten entsprechen. Unkonventionelle Thesen werden oft schlechter erkannt oder gar als fehlerhaft bewertet. Studien zeigen, dass die Validität der Bewertungen bei originellen oder atypischen Lösungen sinkt (z. B. Li & Ng, 2024; Ifenthaler, 2022). Eine KI sucht nach bestimmten Strukturelementen, nach Schlüsseltermini und einer linearen Argumentation. Weicht ein exzellenter Text bewusst von diesen Schemata ab – etwa durch einen innovativen Aufbau oder sehr originelle Beispiele – kann die KI dies als negativ interpretieren. Erste Studien zeigen, dass KI-Modelle Originalität im Vergleich zu Expertenurteilen nur eingeschränkt zuverlässig erfassen und eher konservativ bewerten (Chakrabarty et al., 2024; Zhu et al., 2025).

Fokus auf oberflächliche Merkmale:
KI-Bewertungssysteme neigen dazu, leicht messbare, oberflächliche Textmerkmale stark zu gewichten. Rechtschreibung, lexikalische Muster und Textlänge werden zuverlässig erkannt. Übersichtsarbeiten zu AES betonen seit Jahren, dass viele Modelle vor allem solche „oberflächlichen" Features nutzen (Ifenthaler, 2022; Fleckenstein et al., 2020; Li & Ng, 2024). Neuere Studien mit LLMs bestätigen dies: Seßler et al. (2025) zeigen, dass GPT-4 bei der Bewertung von Aufsätzen Kriterien wie „Spelling & Punctuation" deutlich stärker gewichtet als Menschen. Dadurch kann eine inhaltlich eher durchschnittliche Arbeit mit makelloser Form überbewertet werden. Zudem finden sich Hinweise auf Längen- und Komplexitäts-Bias: Längere Texte mit komplexeren Satzstrukturen schneiden besser ab, unabhängig von der tatsächlichen Qualität (Faseeh et al., 2024).

Schwierigkeiten bei der Erkennung inhaltlicher Fehler:
Während menschliche Fachexpert:innen inhaltliche Fehler oder Plausibilitätslücken erkennen können, sind KI-Modelle auf Trainingsdaten angewiesen. Wenn eine Arbeit einen neuartigen Denkfehler enthält, kann eine KI dies übersehen, sofern er sprachlich plausibel ist. Studien zur Nutzung von AES und ChatGPT berichten Fälle, in denen faktisch falsche Inhalte hohe Bewertungen erhalten, wenn sie formal sauber formuliert waren (Fleckenstein et al., 2020; Perelman, 2014; Quah et al., 2024).

Black-Box-Problem:
Ein weiterer Kritikpunkt ist die begrenzte Nachvollziehbarkeit von KI-Bewertungen. Viele Systeme liefern zwar neben der Note ein kurzes Feedback, doch bleibt oft unklar, wie genau das Urteil zustande gekommen ist. Während menschliche Gutachten fachlich erläutern können, wie sie zu ihrem Urteil gelangen, erlaubt die KI nur unzureichende Einblicke in die algorithmischen Prozesse hinter der Bewertung. So stellen ausgegebene Begründungen häufig lediglich nachträglich generierte und standardisierte Rationalisierungen dar (Gurrapu et al., 2023). Sowohl Übersichtsarbeiten zu AES (Ifenthaler, 2022) als auch neuere Arbeiten zur Nutzung von KI im Assessment (Johnson & Zhang, 2024; Gaudeau, 2025) bemängeln diese unzureichende Erklärbarkeit aus bildungsethischer Sicht.

Systematische Verzerrungen (Bias):
Auch wenn KI keine eigenen Vorurteile hat, kann sie Verzerrungen aus Trainingsdaten reproduzieren. Fairness-Analysen zu AES-Systemen zeigen, dass bestimmte Gruppen systematisch andere Scores erhalten (Doewes et al., 2022). Eine aktuelle Studie von Johnson und Zhang (2024) berichtet etwa, dass GPT-4o im Mittel niedrigere Scores vergab als menschliche Rater:innen – und dass der Abstand für asiatisch-amerikanische Schüler:innen besonders ausgeprägt war. Andere Studien weisen auf mögliche Nachteile für bestimmte Sprachstile hin (Li & Liu, 2024).

Nivellierungseffekt und eingeschränkte Differenzierung:
KI-Bewertungen neigen dazu, die Notenskala enger zu nutzen und Extrembewertungen zu vermeiden. So zeigen Studien, dass GPT-4o im Vergleich zu menschlichen Rater:innen weniger Höchstnoten vergibt bei gleichzeitig deutlich kleiner Standardabweichung der Notenverteilung (Johnson & Zhang, 2024; Manning et al., 2025). Aus pädagogischer Sicht ist diese Nivellierung ambivalent: Einerseits reduziert sie Zufallsschwankungen; andererseits demotiviert sie sehr gute Studierende, wenn ihre besondere Leistung nicht angemessen honoriert wird.

Zusammengefasst haben KI-Bewertungssysteme zwar Stärken in der Konsistenz, kämpfen aber mit einigen Herausforderungen. Diese Fehlerquellen werden beim nachfolgenden Vergleich von KI- und menschlichen Bewertungen untersucht.

3 Empirische Analyse: Vergleich von Mensch und KI

3.1 Methode

Zur Beantwortung der Frage, ob menschliche Bewertungen von Abschlussarbeiten als „Goldstandard“ gelten sollten wurden KI und menschliche Bewertungen von 52 Bachelorarbeiten auf die bekannten Fehlerquellen und Übereinstimmungen hin untersucht. Die Analyse basiert dabei auf Bachelorarbeiten aus marketingorientierten Studiengängen, bewertet von jeweils einer menschlichen Gutachter:in. Alle Arbeiten entstanden im Rahmen regulärer Prüfungen und wurden für die Analyse pseudonymisiert. Der quantitativen Analyse lag ein 14-Kriterien-Rubric zugrunde, das u. a. nach theoretischer Fundierung, Methodik, Argumentation, Ergebnisdarstellung, Transferleistung sowie sprachlich-formaler Gestaltung differenziert. Die menschliche Bewertung vergab zu jedem Kriterium Teilnoten, aus denen eine gewichtete Gesamtnote berechnet wurde.

Für den KI Vergleich wurden die gleichen Arbeiten von mehreren großen Sprachmodellen (gpt4_1_mini, mod3o, mod4_1, mod4o, gpt4o_b, gpt4o_mini, GPT-5.1 nur qual. Analyse) bewertet. Der KI-Bewertungsprozess wurde vollständig mittels Python-Skript automatisiert: PDF-Import und Textextraktion, kontextgerechte Segmentierung, Einsatz eines standardisierten System-Prompts mit Rubric und Regeln zur fairen und reproduzierbaren Bewertung, Übergabe der Texte per User-Prompt, API-Aufrufe mit einheitlichen Parametern (u. a. temperature = 0,2, fester seed, API-Standardeinstellungen für reasoning-Modelle) sowie Zusammenführung der Ergebnisse zu konsistenten Gutachten. Alle Modelle erhielten identische Bewertungsgrundlagen, und die Gesamtnoten wurden im selben Notensystem wie bei den menschlichen Gutachter:innen berechnet.

3.2 Ergebnisse quantitativer Noten

Im ersten Teil der empirischen Untersuchung wurde die Bewertung von Mensch und KI auf Gesamtnotenebene verglichen. 12 der 52 zugrundeliegenden Arbeiten verwendeten ein leicht unterschiedliches Bewertungsraster. Es wurden daher nur 40 Arbeiten herangezogen, welche im Rahmen der regulären (menschlichen) Benotung mittels identischem Raster bewertet wurden. Es wurden Kennwerte zu Notenniveau und Strengegrad (Mittelwerte und mittlere Differenzen zwischen Mensch und KI), zur Streuung und Notenkompression (Varianzen, Spannweiten und Varianzratios), zur Rangtreue (Rangkorrelationen und paarweise Trefferquoten) sowie zur Messübereinstimmung (Intraklassenkorrelationen und Bland-Altman) analysiert. Zusätzlich wurden Konsistenzmaße innerhalb der Rubrics berechnet (Cronbachs α über die Teilnoten), um mögliche Halo-Effekte bei Mensch und KI abzuschätzen.

Die Ergebnisse zeigen *erstens* eine deutlich höhere Streuung menschlicher Noten: Die Varianz der menschlichen Gesamtbewertungen war vielfach höher als jene der KI-Noten (Varianzratio je nach Modell zwischen 2,1 und 9,9). Die menschlichen Noten reichten von 1,0 bis 3,3 (M = 1,67; SD = 0,55). Die sechs getesteten KI-Modelle erzielten dagegen Mittelwerte zwischen 1,61 und 2,93, bei deutlich geringeren Standardabweichungen (vgl. Tabelle 1). Damit nutzen KI-Modelle die Notenskala

deutlich komprimierter. Vergleichbare Muster reduzierter Varianz und seltener Extremwerte werden auch in Studien zur automatisierten Essaybewertung berichtet (Johnson & Zhang, 2024; Seßler et al., 2025).

Model	M	SD	Range	
Mensch	1.67	0.55	1	3.3
mod4_1	1.68	0.22	1.4	2.3
gpt4_1_mini	1.65	0.17	1.3	2.1
mod4o	2.30	0.33	1.5	3.2
mod3o	1.61	0.20	1.3	2.2
gpt4o_b	2.36	0.27	1.7	3
gpt4o_mini	2.93	0.38	2.3	3.9

Tabelle 1: Mittelwerte, Standardabweichungen und Notenspanne

Zweitens zeigten die menschlichen Gutachten eine gnädigere Benotung. Über alle sechs KI-Modelle hinweg lagen die vom Menschen vergebenen Noten im Mittel um etwa 0,41 Notenpunkte besser (d. h. niedriger) als die KI-Urteile. Dabei ist der Strengegrad der Modelle heterogen: Während gpt4_1_mini (Δ = -0,02), mod3o (Δ = -0,06) und mod4_1 (Δ = +0,01) im Mittel nur minimal anders als Menschen bewerten, liegen mod4o (Δ = +0,63), gpt4o_b (Δ = +0,69) und insbesondere gpt4o_mini (Δ = +1,25) im Schnitt deutlich über dem menschlichen Niveau. Dieser systematische Lenienzeffekt belegt, dass menschliche Bewertungen tendenziell wohlwollender ausfallen. Ähnliche Befunde strenger KI-Bewertungen zeigt auch die Studie von Johnson und Zhang (2024): GPT-4o vergab auf einer sechsstufigen Skala knapp einen Punkt niedriger als menschliche Bewertungen.

Drittens ist die ordinale Übereinstimmung – also die Frage, ob Gutachter:in und KI dieselbe Rangfolge der Arbeiten erzeugen – nur begrenzt (vgl. Tabelle 2). Spearman-

Rangkorrelationen zeigen für die besten Modelle mod4o und gpt4_1_mini moderate Zusammenhänge ($\rho \approx 0{,}30$–$0{,}39$, Kendall $\tau \approx 0{,}26$), während mod4_1, gpt4o_b und mod3o nur noch schwache Korrelationen erreichen ($\rho \approx 0{,}11$–$0{,}26$). GPT4o_mini weist sogar eine leicht negative Rangkorrelation ($\rho \approx -0{,}06$) auf. Entsprechend liegt der Anteil, die Gutachter:in und Modell in derselben Reihenfolge einstufen, je nach Modell zwischen rund 47 % und 68 %.

Modell	Spearman ρ	Kendall τ	Paar-Übereinstimmung*	Top-10-Overlap	Bottom-10-Overlap
mod4o**	0,39	0,30	68 %	50%	40 %
gpt4_1_mini	0,30	0,26	65,1 %	30 %	50 %
mod4_1	0,26	0,20	61,9 %	30 %	40 %
gpt4o_b**	0,25	0,20	62,1 %	30 %	50 %
mod3o	0,11	0,09	56,0 %	20 %	40 %
gpt4o_mini	−0,06	−0,05	47,2 %	20 %	40 %

Tabelle 2: Ordinaleffekte in den Bewertungen

* Paar-Übereinstimmung: Anteil der Paare, die Mensch und Modell in gleicher Reihenfolge einstufen (ohne Gleichstände).
** Datenbasis: 33 (mod4o) bzw. 34 (gpt4o_b) von 40 Arbeiten

Viertens zeigen sich leistungsabhängige Differenzen zwischen KI- und Menschennoten. Eine Bland-Altman-Analyse (Abb. 1) weist eine deutliche positive Steigung (≈ +1,36) der Differenz über dem Notenmittelwert auf: Sehr gute Arbeiten werden von menschlichen Gutachter:innen meist besser bewertet als von der KI, während

schwächere Arbeiten tendenziell schlechtere menschliche Noten als KI-Noten erhalten.

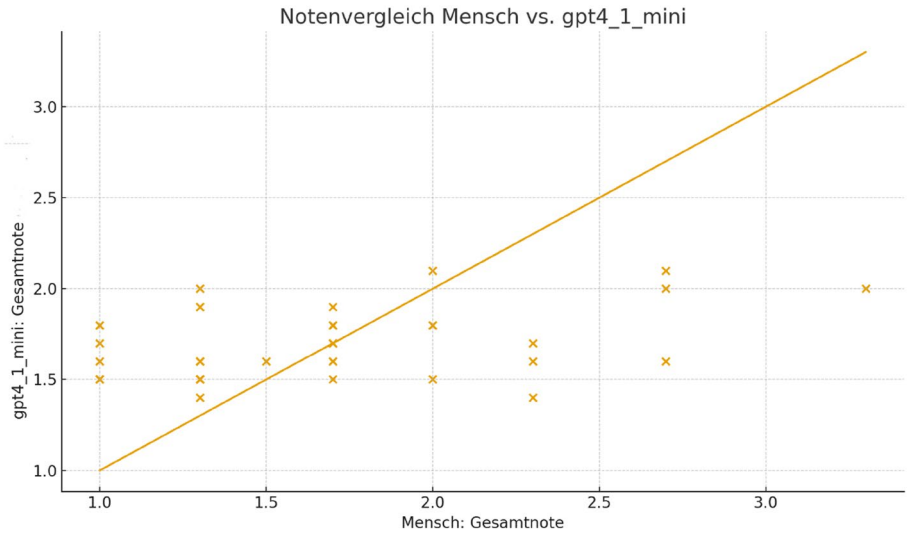

Abbildung 1. Mensch-KI-Noten im Vergleich für gpt4_1_mini

Bemerkung: Die Darstellung ist streng genommen formal keine Blend-Altman-Darstellung, zeigt aber die Ergebnisse der Analyse in einer einfacheren und besser nachvollziehbaren Form

Fünftens wurden Unterschiede in der Konsistenz über Bewertungskriterien hinweg beobachtet. Die menschlichen Gutachten wiesen eine sehr hohe interne Konsistenz in den einzelnen Teilbewertungen auf (Cronbachs $\alpha \approx 0{,}96$), was ein Hinweis auf einen Halo-Effekt sein kann.

Neuere Untersuchungen deuten darauf hin, dass KI-Modelle wie GPT-4 ebenfalls zu stark korrelierten Kriterienscores neigen (Seßler et al., 2025). In der vorliegenden

Studie ist die interne Konsistenz der KI-Bewertungen allerdings geringer als die des menschlichen Gutachtens.

Insgesamt lässt sich festhalten, dass es substanzielle Unterschiede in Streuung, Niveau, Rangtreue, Konsistenz und Übereinstimmung der Noten zwischen menschlicher und KI-Bewertung gibt. Menschliche Gutachter:innen nutzen die Skala breiter und sind im Mittel milder, weisen aber eine starke Tendenz zu globalen Gesamteindrücken auf; KI-Modelle sind konsistenter, komprimieren die Noten jedoch und reproduzieren die Rangordnung menschlicher Gutachter:innen nur begrenzt. Im nächsten Schritt sollten die qualitativen Urteile von KI und Mensch verglichen werden.

3.3 Ergebnisse qualitativer Evaluation Mensch – KI im Vergleich

Für 12 Arbeiten lagen digitalisiert qualitative Gutachtenkommentare vor. Für diese Arbeiten wurden entsprechende Kommentare von zwei GPT-4-basierten Modellen und einer GPT-5.1-Variante generiert. Die Kommentare sowohl von Mensch als auch KI-Systemen wurden zunächst deskriptiv ausgewertet, anschließend wurden regelbasierte Qualitätsindikatoren berechnet (Spezifität, Begründungstiefe und Prüfbarkeit, thematische Kategorie). Zur Identifikation von „KI-only"-Befunden wurde die Textähnlichkeit zwischen menschlichen und KI Kommentaren berechnet.

Ergebnisse. Es zeigen sich *erstens* deutliche Qualitätsunterschiede zwischen den KI-Modellen und *zweitens* systematische Differenzen zwischen KI und Mensch.

Auf der Ebene der Ausführlichkeit und Begründungstiefe erzeugte GPT-5.1 die umfangreichsten Bewertungen (die meisten Stärken- und Schwächepunkte und die umfangreichste Textlänge). o4mini-r lieferte hingegen kürzere, stärker „prüfbare" Kommentare durch einen hohen Anteil an Fundstellenverweisen.

Inhaltlich verdeutlichen die Themenprofile, dass die Modelle unterschiedliche Schwerpunkte setzen. Die menschlichen Kommentare betonen vergleichsweise stark Sprache, Theorie, Struktur sowie Literatur, während die KI-Modelle stärker Methodik, Ergebnisse sowie Analyse aufgreifen. Die thematische Nähe zur menschlichen

Bewertung (Jaccard-Ähnlichkeit) fiel bei GPT-5.1 am höchsten aus, während GPT-4o mini am stärksten abwich.

Eine Differenzanalyse zeigt ergänzend, dass KI-Modelle zusätzliche Punkte identifizieren, die in menschlichen Gutachten nicht explizit genannt werden; je nach Modell betrifft dies mindestens 40 % der Schwächen. Diese „KI-only"-Aspekte konzentrieren sich vor allem auf Methodik und Ergebnisse, bei GPT-5.1 zusätzlich auf Theorie, Statistik und Zielpräzision. Fälle von KI-Konsens (mindestens zwei Modelle nennen ein zusätzliches Themenfeld) finden sich mehrmals.

Insgesamt liefern die KI-Modelle unterschiedliche qualitative Stärken, wobei GPT-5.1 die höchste Qualität zeigt. Die Befunde legen nahe, dass KI ergänzende Perspektiven einbringt und Präzision und Vollständigkeit menschlicher Bewertungen erhöhen könnte.

4 Diskussion: Ist der Mensch ein gerechtfertigter Goldstandard?

Traditionell wurden automatische Bewertungssysteme darauf trainiert, menschliche Urteile nachzuahmen – in der impliziten Annahme, dass die menschliche Note die „Wahrheit" darstellt (Doewes & Pechenizkiy, 2021). Die vorliegenden Befunde zeichnen jedoch ein komplexeres Bild.

Unbestreitbar verfügen menschliche Gutachter:innen über Fähigkeiten, die KI bislang weitgehend fehlen: Kontextsensitivität, interpretative Tiefe sowie die Würdigung kreativer und origineller Leistungen. Sie bringen bildungsethische Aspekte ein und begründen ihre Urteile transparent.

Gleichzeitig zeigen die vorliegenden Ergebnisse übereinstimmend mit vorangegangener Literatur (Malouff & Thorsteinsson, 2016; Schmidt et al., 2023) Grenzen menschlicher Bewertungen, etwa eine ausgeprägte Übermilde und Halo-Effekte.

Vor diesem Hintergrund plädieren mehrere Autor:innen dafür, den traditionellen, monolithischen Goldstandard aufzubrechen. Gaudeau (2025) argumentiert, dass die Aggregation von Bewertungsdisparitäten zu einem einzigen „Goldstandard"-Label wichtige Gutachtervariabilität verdeckt. Aydın et al. (2025) fordern mehrdimensionale Bewertungsmetriken mit mehreren Gutachter:innen. Damit rückt ein Kollektivmaßstab in den Fokus.

Zunehmend wird außerdem hinterfragt, ob qualitative Ansätze nicht nachhaltiger sind als die dominante quantitative Leistungsbeurteilung. Die vorliegende Studie zeigt, dass KI-Modelle trotz unterschiedlicher Leistungsfähigkeit ergänzende Schwerpunkte setzen („KI-only"-Befunde). Insgesamt sprechen diese Ergebnisse dafür, KI Systeme als komplementäre Instanz in die qualitative Leistungsbewertung einzubetten.

5 Limitationen und weiterführende Forschungsfragen

Es sei angemerkt, dass die Aussagekraft der vorliegenden Studie unter Berücksichtigung methodischer Limitationen zu bewerten ist. So dienten in dieser Studie als Referenzpunkt jeweils nur die Gutachten eines Menschen. Dies wirft die Frage auf, in welchem Maße die beobachteten Muster individuelle Charakteristika widerspiegeln oder ob sie als stabilere Merkmale menschlicher Bewertungsprozesse interpretiert werden können. Da Gutachter:innen im Hinblick auf Strenge, Bewertungsschwerpunkte und Erfahrungsstand teils erheblich variieren, bleibt offen, wie robust die Befunde gegenüber einer veränderten Referenzbasis wären. Auch erlaubte die Datenbasis keine empirische Analyse von Mensch-Mensch- und Mensch-KI-Korrelationen. Künftige Forschung hierzu könnte weitere interessante Erkenntnisse bezüglich der Interrater-Reliabilität liefern. Insbesondere ein Vergleich aggregierter menschlicher Gutachten mit KI-Bewertungen sollte künftig untersucht werden.

6 Praktische Implikation: Hybridansätze als Zukunftsperspektive

Die vorliegenden Ergebnisse deuten darauf hin, dass sowohl Mensch als auch KI Schwächen und Stärken in der Leistungsbeurteilung akademischer Abschlussarbeiten haben. Während KI-basierte Systeme konsistente, regelgeleitete und skalierbare Bewertungen ermöglichen, bleibt die menschliche Urteilskraft zentral für die kontextuelle Einbettung und die inhaltlich interpretative Tiefe (Agrawal et al., 2025; Bahadure et al., 2025). Hybride Bewertungsansätze stellen vor diesem Hintergrund eine vielversprechende Perspektive dar. Voraussetzung hierfür sind transparente Bewertungsprozesse sowie eine klare Rollenverteilung zwischen Mensch und KI. Denkbar sind z. B. Modelle mit KI-basierter Erstbewertung und anschließender menschlicher Einordnung aber auch umgekehrt. Insbesondere bei uneindeutigen oder grenzwertigen Bewertungsfällen kann die komplementäre Perspektive zu fundierteren Entscheidungen beitragen, indem die KI als Instrument der kritischen Selbstreflexion genutzt wird. Ein kompakter Reflexions-Prompt könnte beispielsweise wie folgt aussehen (Abb. 2):

Lies die beigefügte Abschlussarbeit und überprüfe das dazu vorliegende Gutachten in Bezug auf folgende Aspekte:
*- **Konsistenz:** Stimmen Gesamtnote und Einzelbewertungen widerspruchsfrei mit den zugrunde liegenden Kriterien überein?*
*- **Fairness:** Ist die Kritik objektiv und verhältnismäßig formuliert (keine unpassenden persönlichen Präferenzen erkennbar)?*
*- **Methodische Kohärenz:** Entsprechen die festgelegten Bewertungsmaßstäbe dem fachlichen Standard, und ist die Evidenzbasis schlüssig?*
*- **Qualitätsdiagnose:** Wurden alle wesentlichen Stärken und Schwächen erfasst, oder existieren blinde Flecken in der Analyse?*
*- **Maßnahmen:** Welche Anpassungen oder Verbesserungen ergeben sich daraus (z. B. Kalibrierung der Bewertungsskala, Ergänzung von Kommentaren)?*

Abbildung 2. Reflexions-Prompt

Die Diskussion legt nahe, dass die Zukunft der akademischen Leistungsbewertung weniger im Entweder-oder, sondern vielmehr im Sowohl-als-auch von Mensch und KI liegt. Auch wenn menschliche Bewertungen historisch als Standard gelten, erscheint eine kritisch-reflektierte Haltung angebracht, die technologische Hilfsmittel nicht als Bedrohung, sondern als Chance zu mehr Fairness, Transparenz und methodischer Belastbarkeit sieht. Ziel sollte es sein, die Stärken beider Seiten systematisch zu integrieren, um die Qualität und Nachvollziehbarkeit akademischer Bewertungen nachhaltig zu erhöhen.

Literaturverzeichnis

Agrawal, A., Gans, J. S., & Goldfarb, A. (2025). Machine Intelligence and Human Judgment. *Finance & Development*, *62*(2), 54-57. https://doi.org/10.5089/9798400297359.022.a018

Aydın, B., Kışla, T., Tan Elmas, N., & Bulut, O. (2025). Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System*, *133*, 103784. https://doi.org/10.1016/j.system.2025.103784

Bahadure, N. B., Khomane, R., Shafaq, Pathan, S., Sharma, A., Satokar, P. (2025). Importance of Human Judgment in Artificial Intelligence: A Panoramic Overview. In S. Rama Sree & S. Kumar (Eds), *Algorithms and Computational Theory for Engineering Applications*. Springer, Cham. https://doi.org/10.1007/978-3-031-72747-4_34

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466-481. https://doi.org/10.1080/02602938.2015.1024607

Brimi, H. M. (2011). Reliability of Grading High School Work in English. *Practical Assessment, Research, and Evaluation*, *16*(1). https://doi.org/10.7275/j531-fz38

Chakrabarty, S., Dann, C., Mandal, A., Dann, B., Paul, M., & Hafeez-Baig, A. (2021). Effects of rubric quality on marker variation in higher education. Studies in Educational Evaluation, 70, 100997. https://doi.org/10.1016/j.stueduc.2021.100997

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C.-S. (2024). Art or Artifice? Large Language Models and the False Promise of Creativity. Proceedings of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.48550/arXiv.2309.14556

Dikli, S. (2006). An overview of automated essay scoring. *Journal of Technology, Learning, and Assessment, 5*(1), 1–35.

Doewes, A., & Pechenizkiy, M. (2021). On the limitations of human-computer agreement in automated essay scoring. *Proceedings of the 14th International Conference on Educational Data Mining*. 475–480

Doewes, A., Saxena, A., Pei, Y., & Pechenizkiy, M. (2022). Individual fairness evaluation for automated essay scoring system. *Proceedings of the 15th International Conference on Educational Data Mining*.

Erturk, S., van Tilburg, W. A., & Igou, E. R. (2022). Off the mark: Repetitive marking undermines essay evaluations due to boredom. *Motivation and Emotion, 46*(2), 264-275. https://doi.org/10.1007/s11031-022-09929-2

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y. I. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics, 12*(21), 3416. https://doi.org/10.3390/math12213416

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a Long Essay Always a Good Essay? The Effect of Text Length on Writing Assessment. *Frontiers in Psychology, 11*, 562462. https://doi.org/10.3389/fpsyg.2020.562462

Gaudeau, G. (2025). Beyond the Gold Standard in Analytic Automated Essay Scoring. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics.* 18–39. https://doi.org/10.18653/v1/2025.acl-srw.2

Gurrapu, S., Kulkarni, A., Huang, L., Lourentzou, I., & Batarseh, F. A. (2023). Rationalization for explainable NLP: a survey. *Frontiers in Artificial Intelligence, 6*, 1-19. https://doi.org/10.3389/frai.2023.1225093

Ifenthaler, D. (2022). Automated Essay Scoring Systems. In O. Zawacki-Richter & I. Jung (Eds.), *Handbook of open, distance and digital education* (pp. 1–15). Springer. https://doi.org/10.1007/978-981-19-0351-9_59-1

Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports, 14*, 30064. https://doi.org/10.1038/s41598-024-79208-2

Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications, 11*(1), 1-15. https://doi.org/10.1057/s41599-024-03209-9

Li, S., & Ng, V. (2024). Automated Essay Scoring: Recent Successes and Future Directions. *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*. 8114-8122. https://doi.org/10.24963/ijcai.2024/897

Link, S., & Koltovskaia, S. (2023). *Automated Scoring of Writing*. In O. Kruse et al. (Eds.), *Digital Writing Technologies in Higher Education*. Springer. https://doi.org/10.1007/978-3-031-36033-6_21

Malouff, J. M. (2008). Bias in Grading. *College Teaching, 56*(3), 191-192. https://doi.org/10.3200/CTCH.56.3.191-192

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60*(3), 243– 256. https://doi.org/10.1177/0004944116664618

Manning, J., Baldwin, J., & Powell, N. (2025). Human versus machine: The effectiveness of ChatGPT in automated essay scoring. *Innovations in Education and Teaching International, 62*(5), 1500-1513. https://doi.org/10.1080/14703297.2025.2469089

Perelman, L. (2014). When "the state of the art" is counting words: The problem of automated essay scoring. *Assessing Writing, 21*(2), 104-111. https://doi.org/10.1016/j.asw.2014.05.001

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education, 24*(1), 962. https://doi.org/10.1186/s12909-024-05881-6

Schmidt, F. T. C., Kaiser, A., & Retelsdorf, J. (2023). Halo effects in grading: An experimental approach. *Educational Psychology, 43*(2–3), 246–262. https://doi.org/10.1080/01443410.2023.2194593

Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *Proceedings of the 15th International Learning Analytics & Knowledge Conference*. https://doi.org/10.1145/3706468.3706527

Starch, D., & Elliott, E. C. (1912). Reliability of the Grading of High-School Work in English. *The School Review*, *20*(7), 442–457. https://doi.org/10.1086/435971

Stolpe, K., Björklund, L., Lundström, M. et al. (2021). Different profiles for the assessment of student theses in teacher education. *Higher Education, 82*, 959–976. https://doi.org/10.1007/s10734-021-00692-w

Villa, K. R., Sprunger, T. L., Walton, A. M., Costello, T. J., & Isaacs, A. N. (2020). Inter-rater Reliability of a Clinical Documentation Rubric Within Pharmacotherapy Problem-Based Learning Courses. *American Journal of Pharmaceutical Education, 84*(7). https://doi.org/10.5688/ajpe7648

Zhu, T., Weissburg, I., Zhang, K., & Wang, W. Y. (2025). Examining Human Judgment Against Text Labeled as AI Generated. *Findings of the Association for Computational Linguistics*, 25907-25914. https://doi.org/10.18653/v1/2025.findings-acl.1329

Danksagungen

Die Forschungsarbeit wurde gefördert mit einem Stipendium aus besonderen Mitteln, die der Freistaat Bayern zur Realisierung der Chancengleichheit für Frauen in Forschung und Lehre im Staatshaushalt bereitstellt, eingeworben von Rosario Lopez-Gavira.

This publication is part of the grant RYC2024-048361-I awarded to Inmaculada Orozco, funded by the Spanish Ministry of Science, Innovation and Universities, the State Research Agency /10.13039/501100011033, and the European Social Fund Plus (ESF+).

This study was carried out in collaboration with the team of the research project funded by the Ministry of Science and Innovation of Spain, State Research Agency and FEDER funds European Union [grant numbers EDU2020-112761RB-100/ Feder Funds].