LITERATUR – MEDIEN – DIDAKTIK

DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz

Hans-Georg Müller / Maurice Fürstenberg (Hg.)

Frank & Timme

Hans-Georg Müller / Maurice Fürstenberg (Hg.)
DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz

Sebastian Bernhardt (Hg.)
Literatur – Medien – Didaktik
Band 16

Hans-Georg Müller / Maurice Fürstenberg (Hg.)

DeutschGPT – Deutschunterricht im Dialog mit Künstlicher Intelligenz

Frank & Timme
Verlag für wissenschaftliche Literatur

Umschlagabbildung: Maurice Fürstenberg unter Verwendung von ChatGPT-4o

Diese Veröffentlichung wurde aus Mitteln des Publikationsfonds für Open-Access-Monografien des Landes Brandenburg gefördert. /
This publication was supported by funds from the Publication Fund for Open Access Monographs of the Federal State of Brandenburg, Germany.

Die Beiträge haben eine intensive Qualitätsprüfung und mehrere inhaltliche wie formale Überarbeitungsschleifen durch die Herausgeber des Bandes und den Reihenherausgeber erfahren.

ISBN 978-3-7329-1120-2
ISBN Open Access 978-3-7329-8796-2
DOI 10.26530/20.500.12657/104475
ISSN 2749-5620

Frank & Timme GmbH Verlag für wissenschaftliche Literatur
Berlin 2025.

Herstellung durch Frank & Timme GmbH
Wittelsbacherstraße 27a, 10707 Berlin
info@frank-timme.de
Gedruckt auf säurefreiem, alterungsbeständigem Papier.

www.frank-timme.de

Inhaltsverzeichnis

Maurice Fürstenberg

Zur Qualität KI-generierten Feedbacks

Ein explorativer Vergleich menschlicher und künstlicher Intelligenzen

Abstract

Ziel des vorliegenden Beitrags ist, die Qualität der Rückmeldungen eines didaktisch systemgeprompteten, generativen Sprachmodells zu Schüler:innentexten zu untersuchen. Hierzu schrieben in einer explorativen Studie 19 Schüler:innen eine erste Textversion, erhielten KI-generiertes Feedback und überarbeiteten daraufhin noch einmal ihre Texte, die wiederum ein maschinelles Feedback bekamen. Die kriteriengeleiteten Rückmeldungen des KI-basierten Systems wurden im Anschluss mit menschlichen Expertenurteilen qualitativ und quantitativ verglichen. Darüber hinaus wurden die Überzeugungen der Schüler:innen untersucht. Die Arbeit liefert vier zentrale Ergebnisse: Erstens verbesserten die Schüler:innen die Qualität ihrer Texte durch die Überarbeitung – ob diese Verbesserung *dank* oder *trotz* des KI-generierten Feedbacks erwirkt werden konnte, muss ob der fehlenden Kontrollgruppe unbeantwortet bleiben. Zweitens weist der Vergleich der analytischen Urteile zwischen einem menschlichen Experten und dem KI-System nur auf schwache Übereinstimmungen (ICC = 0,279, p = 0,001) hin. Drittens deuten die qualitativen Analysen an, dass lernförderliches Feedback zu Schüler:innentexten durch das KI-System zwar möglich ist, es treten aber auch eindeutige Probleme in Bezug auf die Konsistenz und die inhaltliche Richtigkeit des Feedbacks zutage. Viertens haben diese Fehler zur Folge, dass einige Schüler:innen dem Feedback kein Vertrauen schenken, was eine der zentralen Herausforderungen von KI-generiertem Feedback ist und bleiben wird.

Schlagwörter: Künstliche Intelligenz, Sprachmodelle, Feedback, Argumentieren, Feedback

1 Hinführung

Im Bildungssektor löste der ChatGPT-Hype – analog zu anderen Entwicklungen der Digitalisierung – eine vorschnelle Absage an etablierte Teile des Systems aus: Hausaufgaben (Emmerich 2023) und Prüfungsformate wie Seminararbeiten (Batzlen 2023) werde es so nicht mehr geben. In Lehrerzimmern rund um den Globus fragt(-e) man sich, wie damit umzugehen sei, dass plötzlich alle Zugriff auf eine Technologie haben, die bessere Texte als Schüler:innen (Herbold et al. 2023) produziert, die nicht einmal mehr von erfahrenen Lehrkräften als maschinell erstellte Texte entdeckt werden können (Fleckenstein et al. 2024). Nach den anfänglichen Sorgen wurden auch Potenziale aufgedeckt, wie diese stochastischen Papageien (Bender et al. 2021) für das Bildungssystem genutzt werden könnten.

Für das Unterrichtsfach Deutsch, das im Zentrum dieses Beitrags steht, aber auch andere korrekturintensive Fächer wurde schon oft genug betont, dass Aufsatzkorrekturen Lehrkräfte enorm beanspruchen (Mußmann et al. 2017, Mußmann et al. 2020). Dies wird durch eine einfache Rechnung offenbar: Geht man von einer durchschnittlichen Klassengröße von 25 Schüler:innen pro Klasse und vier großen schriftlichen Leistungsnachweisen inklusive eines jeweils vorbereitenden Übungsaufsatzes aus, summiert sich die Anzahl der zu korrigierenden Text allein für das Fach Deutsch und bei ‚nur' zwei Deutschklassen bereits auf 400 Texte pro Schuljahr. Deutschlehrkräfte müssen folglich mehr als einen Text pro Tag im Jahr korrigieren. Es kommt erschwerend hinzu, dass dieser hohe Aufwand nicht unbedingt didaktisch zielführend ist, weil Kriterien lernförderlichen Feedbacks nicht eingehalten werden (Müller/Utesch/Busse 2023) und das Feedback teils mit starker zeitlicher Verzögerung erfolgt.

An eben dieser Herausforderung setzt die vorliegende Untersuchung an: Sprachmodelle produzieren im Handumdrehen sprachlich einwandfreies Textfeedback und könnten damit einerseits Lehrkräfte entlasten und andererseits Schüler:innen in kürzester Zeit lernförderliches Feedback geben. Der inhaltlichen Qualität dieses Feedbacks geht der vorliegende Text nach.

2 Forschungseinblick

Der Gedanke, manuelle Beurteilung von Texten mithilfe von Computern zu automatisieren, ist nicht neu (vgl. auch den Überblick in Wendt (2023)): Page (1966) beschäftigte sich schon vor fast 60 Jahren mit automatisierter Bewertung von Schüler:innentexten. Die meisten Entwicklungen in diesem Bereich, die insbesondere im angloamerikanischen Raum stattfanden, sind unter *Automatic Short Answer Grading* oder *Automated Essay Scoring* zu finden. Dabei stützten sich dafür eingesetzte Systeme bis 2022 bei der Bewertung von (kurzen) Texten vor allem auf sprachliche Oberflächenphänomene (Ramesh/Sanampudi 2022). Relativ neu ist der Einsatz von Sprachmodellen, die durch Verwendung hochdimensionaler Vektorräume (Simbeck i. d. B.) einen qualitativen Sprung versprechen.

So setzen beispielsweise Sawatzki et al. (2022) ein auf Wikipedia und Open Legal Data trainiertes und mit 233 Fragen und Antworten aus der Betriebswirtschaftslehre (Moodle-Prüfungen) spezialisiertes BERT-Modell ein. Das Modell sollte Punktbewertungen (0–6/8/10) zu Kurzantworten ($M_{Wörter}$=87,6) vorhersagen. Die Modellgenauigkeit wurde statt mit dem für ordinale Daten geeigneten Spearman- mit dem Pearson-Korrelationskoeffizient geschätzt und liegt bei nicht gesehen Testsets bei r = 0,78, was die Autoren so interpretieren, dass BERT-Modelle effektiv zur automatischen Bewertung von Kurzantworten genutzt werden können, ohne jedoch einen p-Wert anzugeben.

Padò et al. (2023) prüfen die Inter-Rater-Reliabilität von speziell trainierten SBERT-Modellen und menschlichen Ratings beim Short Answer Grading (richtige vs. falsche Antwort). Das beste Modell erreichte eine Genauigkeit von *M*=71,4 % [*MIN*=64,7, *MAX*=86,3].

Zahlreiche Studien haben den Einsatz von nicht speziell trainierten Modellen für automatisierte Rückmeldung zu englischen Texten untersucht (u. a. Mizumoto/Eguchi (2023), Naismith et al. (2023)). Beispielsweise verglichen Chiang und Lee (2023) die Bewertungen von drei Lehrkräften mit denen von GPT-3. In ihrer Studie wurden 400 englische Textfragmente ($M_{Wörter}$=150) analysiert. Die Beurteilung der Texte durch Mensch und Modell erfolgte anhand von vier Kategorien (*Grammaticality*, *Cohesion*, *Likability* und *Relevance*) auf einer fünfstufigen Likert-Skala. Die Ergebnisse zeigten lediglich für die Kate-

gorie *Relevance* (Aufgabenpassung des Textes) einen moderaten Zusammenhang zwischen den menschlichen und den GPT-3-Bewertungen, mit einem Kendall's Tau von $\tau = 0{,}38$.

Eine darauf aufbauende Untersuchung von Stahl et al. (2024) vergleicht verschiedene Prompting-Strategien, um zu überprüfen, wie gut Mistral Rückmeldungen zu englischen Texten geben kann. Dabei wurden die generierten Rückmeldungen mit Bewertungen von 12 Laien sowie mit den Rückmeldungen von LLaMA-2 und Mistral selbst verglichen. Die Autoren begründeten ihre Entscheidung, das Urteil über die Qualität des Textfeedbacks eines Sprachmodell durch ein anderes Sprachmodell bewerten zu lassen, mit dem vermeintlichen Ergebnis von Chiang und Lee (2023): *„Using LLMs to assess the quality of generated texts has been shown to be consistent with human expert annotations for some free-text generation tasks.*" (Stahl et al. 2024: 7)

Die Studie von Seßler et al. (2025) untersucht die Fähigkeit großer Sprachmodelle, Schüler:innenaufsätze anhand von zehn Kriterien zu beurteilen. Dabei wurden fünf Modelle (GPT-3.5, GPT-4, o1, LLaMA 3-70B und Mixtral 8x7B) mit den Bewertungen von 37 Lehrkräften verglichen. Die Modelle und die Lehrkräfte bewerteten 20 Schüler:innenaufsätze auf einer sechs-stufigen Likert-Skala. Die Modelle wurden mit einem Zero-Shot-Prompt instruiert und es wurden zu jedem Text jeweils zehn Bewertungsdurchläufe durchgeführt, um die interne Konsistenz der Modellbewertungen zu testen. Die Untersuchung zeigt, dass die Urteile der geschlossenen Modelle (GPT-3.5, GPT-4, o1), insbesondere aber o1, eine hohe Korrelation mit denen der Lehrkräfte aufweisen. Besonders hoch ist die Übereinstimmung bei oberflächennahen Kriterien wie *Ausdruck*, *Satzstruktur* und *Rechtschreibung*. Das Modell o1 erreichte eine Übereinstimmung mit menschlichen Bewertungen von $r = 0{,}74$ und eine interne Konsistenz von $ICC = 0{,}80$. Offen zugängliche Modelle (LLaMA 3-70B, Mixtral 8x7B) zeigten hingegen eine geringe Reliabilität und keine bis geringe Korrelation mit den Lehrerbewertungen. Zudem neigten alle Sprachmodelle dazu, Schüler:innenaufsätze im Vergleich zu Lehrkräften systematisch milder zu bewerten.

Die besprochenen Studien, die stellvertretend für einen Teil der aktuellen Forschung zur automatisierten Beurteilung von Texten stehen, nutzen verschiedene quantitative Bewertungsmaße, mit denen sie ein überaus komplexes Phänomen (Textqualität) zu quantifizieren suchen, untersuchen fast nie

deutsche, kaum Texte von Lernenden und verwenden nur selten authentische Expertenurteile (Lehrkräfte, Sprachdidaktiker:innen oder Linguist:innen), um die Qualität der automatisierten Beurteilungen zu prüfen. Trotz vergleichsweise einfacher Bewertungsaufgaben (richtig oder falsch, Punktebewertung zu definierten Kategorien) liefern die Modelle dennoch oft keine zufriedenstellenden Ergebnisse, was den Eindruck vermitteln kann, dass der Weg zu lernförderlichem Feedback in Textform, das didaktischer Zielpunkt all dieser Bemühungen ist, noch überaus weit ist, wie auch Seßler et al (2025: 471) betonen.

Die Qualität der maschinellen Beurteilungen wird in allen vorgestellten Studien daran festgemacht, wie hoch sie mit einer menschlichen Beurteilung übereinstimmen. Dadurch entstehen mindestens zwei Probleme: Erstens sind sich Menschen bei der Beurteilung von Textqualität nicht unbedingt einig, auch wenn sie vergleichbar qualifizierte Expert:innen (z. B. Lehrkräfte) sind (Birkel/Birkel 2002, Schröter et al. 2023) – ihre Urteile über Textqualität sind also nicht unbedingt ein zuverlässiger „Goldstandard". Zweitens wird in den Studien die qualitative Analyse der einzelnen Texte und Rückmeldungen durch Expert:innen ausgespart. Dadurch gerät das aus didaktischer Sicht relevanteste Qualitätskriterium aus dem Blick: der potenzielle Beitrag des Feedbacks zum Lernerfolg der Lernenden. Es ist daher dringend geboten, den quantitativen Vergleichen zwischen Mensch und Maschine auch qualitative Analysen von Textfeedback aus fachdidaktischer Perspektive an die Seite zu stellen und auch die Perspektive der Lernenden zu integrieren.

Obwohl verlässliche Qualitätsanalysen maschineller Rückmeldungen zu Schüler:innentexten also noch ausstehen, finden schon vielfach Systeme Einsatz – bisher allerdings meist[1] ohne fachdidaktische Begleitung oder Überprüfung –, die ein eben solches Feedback zur Verfügung stellen (u. a. *PEER*, *FelloFish*, *fobizz*, *cornelsen-AI*). Der Beitrag geht daher der Frage nach der Qualität solcher Rückmeldungen nach. Dazu werden im folgenden Kapitel die für diese Untersuchung notwendigen forschungsmethodischen Entscheidungen dargestellt, bevor zentrale Ergebnisse und deren Auswertung vorgestellt und diskutiert werden.

1 Der Autor gehört seit dem Frühjahr 2024 dem wissenschaftlichen Beirat der FelloFish GmbH an.

3 Methodik

3.1 Stichprobe

Da Sprachmodelle maschinell lesbare Texte[2] benötigen und junge Schüler:innen teils noch größere Probleme mit der Schreibflüssigkeit[3] bei digitalen Texten haben, wurde die Studie mit einer neunten Klasse (N=19; n_w=7, n_m=12) durchgeführt. Die teilnehmende Klasse gehörte einem bayerischen Gymnasium einer Kleinstadt im ländlichen Raum an. Die Stichprobe ist selbstredend weder randomisiert noch repräsentativ für die Grundgesamtheit, weshalb die vorzustellenden Ergebnisse zwar nicht verallgemeinerbar sind, wohl aber relevante methodische und inhaltliche Ergebnisse für zukünftige Studien liefert.

3.2 Erhebungsinstrumente

3.2.1 Textsorte

Zur Untersuchung der Qualität von KI-generierten Rückmeldungen wurde die Textsorte *Materialgestütztes Argumentieren* (Feilke/Tophinke 2017) gewählt. Das (schriftliche) Argumentieren und die Vermittlung entsprechender Kompetenzen führt die Kultusministerkonferenz gleich zu Beginn ihrer Bildungsstandards für das Fach Deutsch für die Allgemeine Hochschulreife (KMK 2014: 13) unter allgemeinen Zielen des Faches auf und verleiht dieser Kompetenz auch in der Formulierung eine hohe Relevanz: „Besonderes Gewicht erhält die Entwicklung der Argumentations- und Reflexionsfähigkeit". Diese nimmt spätestens ab der siebten Jahrgangsstufe eine zentrale Rolle im Schreibprogramm des Gymnasiums ein, die es bis zum Abitur nicht einbüßt.[4] KI-Systeme bieten insbesondere in diesem Bereich didaktische Anknüpfungspunkte weit über das

2 Zwar machen es OCR-Techniken auch möglich, die Handschrift von Schüler:innen erkennen zu lassen, allerdings produzieren sie nach wie vor keine verlässlichen Ergebnisse.

3 Schreibflüssigkeit gilt als Prädiktor für Textqualität (Sturm/Schneider 2021: 40 f.).

4 Hier zeichnet sich in der Grundschule ein Wandel ab: Waren argumentierende Kompetenzen in den Bildungsstandards für die Primarstufe in der Fassung von 2004 (KMK 2005) noch vor allem auf den Kompetenzbereich *Sprechen und Zuhören* fokussiert, weist die neue Fassung von 2022 auch schriftliche Argumentationskompetenzen aus (KMK 2022: 15).

Feedback hinaus, welche die Unnatürlichkeit des Argumentierens per Aufsatz auflösen können (Fürstenberg/Matz 2025).

3.2.2 KI-System

Doch auch über die Auflösung der zerdehnten Kommunikationssituation hinaus können sich KI-Systeme als lernförderlich erweisen: Insbesondere die Unmittelbarkeit des Feedbacks, aber auch Rückmeldungen zu eher formal orientierten Kompetenzen wie dem Einsatz von Informationen oder Zitaten Materialien, aber auch die Einhaltung der sprachlichen Normen sollten durch KI-Systeme schnell und effektiv überprüft werden können (Neff 2023). Gleichzeitig gelten Sprachmodelle im Vergleich zu Lehrkräften als objektiver, was Vor- und Nachteil zugleich ist: In Bezug auf die Fairness der Beurteilung wäre das zu begrüßen. Andererseits haben KI-Systeme (noch) keinen Zugriff auf den „Kontext des Feedbacks" (Philipp 2023: 13), wissen also nicht, wie der bisherige Kompetenzerwerb der jeweiligen Person abgelaufen ist, und können daher das Produkt nicht vor dem Hintergrund der Lernbiografie beurteilen. Zudem muss kritisch auf den Einsatz von Sprachmodellen als Feedbackmaschinen geblickt werden, denn „[e]s liegt auf der Hand, dass […] Feedback hohes metakognitives Wissen zum Schreiben (darunter auch: schreibdidaktisches Wissen) impliziert, um sinnvolle Handlungsoptionen vorzuschlagen." (ebd.: 12) Mit Blick auf das Training von großen Sprachmodellen (Simbeck i. d. B.; Müller/Fürstenberg 2023) ist es sehr wahrscheinlich, dass dabei „deutschdidaktisch relevante Inhalte allenfalls einen Bruchteil ausgemacht haben" (Fürstenberg/Müller 2024: 2f.). Allgemeine Sprachmodelle wie ChatGPT können also nicht auf solches Wissen zugreifen. Daher wurde ein didaktisch systemgeprompttes KI-System für die Studie genutzt. Von den bei der Erhebungsplanung zur Verfügung stehenden Werkzeugen war *fiete.ai* das ausgereifteste, da es nicht nur Bewertungskriterien, die Aufgabenstellung, Jahrgangsstufe und Material in Bezug auf das Feedback miteinbezieht, sondern zusätzlich mithilfe von System-Prompts für das Feedbackgeben spezialisiert wurde (Haverkamp/Hecht/Schindler 2024). Das KI-System gibt den Schüler:innen neben qualitativem Feedback auch eine quantitative Einschätzung der Leistung in Form von Fortschrittsbalken (s. Abb. 1):

Abb. 1: KI-Feedback zu einem Schüler:innentext

Der Ausschnitt zeigt einen Teil der KI-generierten Rückmeldungen zu einem der abgegebenen Texte. Es wird zunächst ein allgemeiner Eindruck wiedergegeben, der zur Überarbeitung des Textes motivieren soll (Guter Anfang, aber Verbesserungspotential) und stellt zwei Kriterien (Materialnutzung und Aufbau) in den Mittelpunkt der Kritik. Im Anschluss wird jedes Kriterium erst wiederholt (kursiv), dann wird qualitatives und im Anschluss quantitatives Feedback gegeben. Durch die farbigen Fortschrittsbalken können sich die Schüler:innen schnell orientieren, wo noch am meisten Handlungsbedarf besteht. Das qualitative Feedback nennt stets zuerst Positives im Sinne eines *Feed backs* (Philipp 2023: 11f.) und formuliert im Anschluss Kritik in Form eines *Feed forwards* (ebd.) – beides basiert auf den zugrundeliegenden Beurteilungskriterien.

3.3 Beurteilungskriterien

U. a. Philipp (2015: 34) zeigt, dass der Aufbau argumentativer Textkompetenz von einer kriteriengeleiteten Überarbeitung des eigenen Textes profitieren kann. Solche Kriterien erarbeitete die Klasse in einer Unterrichtssequenz, die hier nur skizziert wird: Die Teile des Zieltextes wurden dazu von den Schüler:innen produziert, im Anschluss mithilfe einer Mischung[5] aus Schreibkonferenz (Spitta 1992) und Textlupe (Böttcher/Wagner 1993) miteinander sowie mit KI-generierten Lösungen verglichen. Daraus wurden folgende Kriterien abgeleitet:

ID	Kriterium (aufgeteilt in Unterkriterien)
1	a) Die Einleitung steigt mit einer sprachlichen Auffälligkeit (Frage, These, Zitat) ein, b) skizziert die Thematik und c) stellt kurz die Kontroverse dar. d) Sie führt dann knapp zur Themafrage hin, e) mit der in den Hauptteil übergeleitet wird.
2	a) Der Hauptteil ist in Gegenposition, Eigenposition und Synthese gegliedert, b) wobei die Argumente der Gegenseite vom stärksten zum schwächsten und c) bei der Eigenposition vom schwächsten zum stärksten Argument sortiert sind.
3	a) Die Argumente sind vollständig b) (enthalten jeweils Behauptung, c) Begründung und d) ein konkretes Beispiel), e) sind inhaltlich richtig, f) passen zur Erörterungsfrage, g) nutzen eine sachliche und unterstützende Sprache und h) sind inhaltlich kohärent.
4	a) Die Synthese wägt vor dem Schluss nochmal die stärksten Argumente gegeneinander ab und b) begründet die eigene Position abschließend.
5	a) Der Schluss schließt an die Einleitung an und b) enthält einen Appell oder einen Ausblick.
6	a) Es werden alle notwendigen Kommas gesetzt und b) die großzuschreibenden Wörter werden großgeschrieben.
7	a) Aus dem Material wird mit Anführungszeichen und b) unter Angabe der Quelle zitiert.

Tab. 1: Bewertungskriterien materialgestütztes schriftliches Argumentieren

5 Die engere Führung durch die Textlupe kann für Schüler:innen hilfreich sein (Reichardt/Kruse/Lipowsky 2014: 81).

Die ersten fünf Kriterien sind weitgehend textchronologisch angeordnet. Kriterien 6 und 7 zielen auf eher formalsprachliche Aspekte ab. Die Einzelkriterien werden hier nicht detailliert besprochen. Sie wurden von der Lehrkraft im Unterricht vermittelt und enthalten sowohl gelungene wie auch durchaus strittige Inhalte: So kann mithilfe der überleitenden Themafrage lokale Textkohärenz (Averintseva-Klisch 2018: 15) hergestellt werden und die Qualitätskriterien der Argumente sind an Schwarze (2016) angelehnt. Das in der Schule beliebte Sanduhr-Prinzip hingegen darf mit Dax (2023: 254 f.) als umstritten[6] gelten und die Kriterien beziehen sich weder auf prozessuale (z. B. metakognitive) noch auf sonstige personale Fähigkeiten. Darüber hinaus sind es zwar auf erster Ebene nur sieben Kriterien, die jedoch bei näherer Betrachtung in 24 Einzelkriterien zerfallen. Es sollte allerdings nicht unberücksichtigt bleiben, dass viele der Anforderungen den Schüler:innen aus früheren Jahrgangsstufen bereits bekannt gewesen sein sollten.

3.4 Untersuchungsablauf

Die Schüler:innen schrieben den Übungsaufsatz auf Basis einer ansatzweise profilierten Schreibaufgabe (Bachmann/Becker-Mrotzek 2010):

Die Schülerzeitung am Gymnasium XY ist noch auf der Suche nach einem abwägenden Beitrag zur Frage, ob Influencer ein Traumberuf ist. Schreibe diesen Beitrag, indem du die gegebenen Materialien (M1–7) und eigenes Wissen nutzt. Wähle eine geeignete Überschrift und beziehe gegen Ende deines Textes Position.

Abb. 2: Schreibaufgabe

Die Aufsätze wurden auf den schuleigenen iPads mit dem Textverarbeitungsprogramm *Microsoft Word* verfasst. Direkt nach der Abgabe analysierte das KI-System die Texte und gab in weniger als einer Minute ein entsprechendes Feedback (s. Abb. 1). Die Klasse hatte am Ende der ersten Stunde und in der

6 Eine zentrale Herausforderung bildet die Gewichtung der Argumente nach Güte, denn hierbei sind Wissen und Kompetenzen gefragt, die der Deutschunterricht eher sekundär ausbildet. Ganz konkret mussten die Schüler:innen entscheiden (s. Abb. 2), ob bspw. die hohe Verdienstmöglichkeit von Influencern oder die große Freiheit in der Ausübung des Berufes ein wichtigeres Argument darstellt.

Folgestunde Zeit, um dieses Feedback zu analysieren, die Aufsätze zu überarbeiten und erneut einzureichen. Auch nach dieser zweiten Abgabe gab das KI-System Feedback auf die überarbeitete Version. Beide Textversionen wurden anonymisiert an den Untersuchungsleiter weitergeleitet, der im Anschluss alle Abgaben der Klasse (N=34[7]) randomisiert anordnete und die Qualität jedes Aufsatzes sechs Wochen später mit einer Punktzahl (0–15) holistisch beurteilte. In derselben Reihenfolge benotete auch ein weiterer schreibdidaktischer Experte die Arbeiten. Einen Monat später beurteilte nur der Untersuchungsleiter auf derselben Skala (1–11) wie das KI-System im Unterricht für alle 34 Texte anhand von 7 Kriterien (s. Tab. 1) analytisch die Qualität jedes Aufsatzes. Daraus ergibt sich folgender Überblick:

Rater	**Beurteilungsform**	**Likert-Skala**
KI-System	analytisch (7 Kriterien)	1–11
Untersuchungsleiter	analytisch (7 Kriterien) und holistisch	1–11 und 0–15
Schreibdidaktik-Experte	holistisch	0–15

Tab. 2: Überblick über die Rater

Somit wird ein Vergleich der analytischen Urteile von KI-System und Untersuchungsleiter sowie der holistischen Urteile von Untersuchungsleiter und Schreibdidaktik-Experte möglich.

Die folgenden Ergebnisse wurden mit *R* berechnet. Die Ergebnisse des Fragebogens wurden direkt in *Google-Forms* ausgewertet und in *Excel* teilweise aufbereitet.

7 Zwei Schüler:innen waren in der Überarbeitungsstunde nicht anwesend und holten die Überarbeitung zu Hause nach. Diese wurden für die Untersuchung ausgeschlossen.

4 Ergebnisse und Diskussion

4.1 Qualitätsentwicklung der Texte

Ein kurzfristiges Ziel von Feedback zu Texten ist, dass die Schüler:innen diese überhaupt überarbeiten: Jansen et al. (2025) stellen in Übereinstimmung mit dem Forschungsstand fest, dass lediglich 48 % der Schüler:innen ihre Texte nach einem KI-generierten Feedback verändern. Sollten die Schüler:innen überarbeiten, sind grundsätzlich drei Folgen dieser Tätigkeit denkbar: Die Texte können verschlechtert werden, auf einem ähnlichen Niveau bleiben oder verbessert werden.

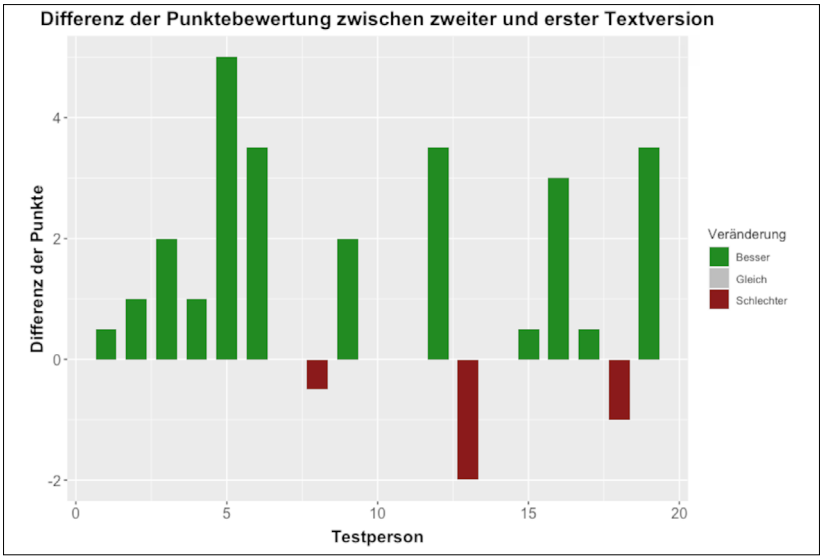

Abb. 3: Säulendiagramm zum Vergleich der beiden Textversionen

Um ein möglichst genaues Bild zeichnen zu können, wurde für Abb. 3, die eben diese Entwicklung für 17 Schüler:innen in einem Säulendiagramm zeigt, ein Mittelwert aus den Punkten der beiden menschlichen Rater[8] für 34 Texte

8 Die holistischen Gesamturteile der beiden menschlichen Rater liegen bei einem ICC von 0,754 (p = 0,001), was auf eine hohe Übereinstimmung hindeutet.

gebildet. Im Anschluss wurde der mittlere Punktewert der zweiten Abgabe vom mittleren Punktewert der ersten Abgabe subtrahiert, wodurch die Veränderung von der ersten zur zweiten Abgabe sichtbar wird.

Von den 17 Schüler:innen verschlechterten drei ihren Text, zwei Texte blieben auf demselben Niveau und zwölf Texte wurden bei der zweiten Abgabe besser. Zwei Ausprägungen der kategorialen Variable *Testperson* auf der x-Achse bleiben leer, weil diese keine zweite Version einreichten. Damit verbesserten 70 % der Proband:innen durch die Überarbeitung ihren Text. Fünf Testpersonen waren sogar in der Lage, mit der überarbeiteten Textversion drei bis fünf Punkte mehr zu erreichen, was im Mittel einer Leistungssteigerung von einer ganzen Notenstufe entspricht. Selbstredend kann die Verbesserung nicht direkt auf die Qualität des Feedbacks durch das KI-System rückgeführt werden. Dazu wären eine Kontrollgruppe und mehr Testpersonen nötig. Nichtsdestotrotz kann festgehalten, dass das KI-generierte, kriteriengeleitete Feedback bei den meisten Schüler:innen nicht zu einer Leistungsverschlechterung geführt hat. Womit die Verbesserung der meisten Texte zusammenhängt, muss hingegen ungeklärt bleiben. Die explorative Untersuchung zeigt, dass es durchaus lohnenswert sein könnte, das Experiment mit mehr Testpersonen zu replizieren. Es bleibt das ermutigende Ergebnis, dass die meisten Schüler:innen kompetent genug waren, ihre Texte entweder mithilfe oder vielleicht sogar entgegen dem Feedback durch das KI-System zu verbessern.

4.2 Quantitatives Feedback

Um die Qualität des KI-generierten Feedbacks zu bestimmen, wird in vielen Fällen auf den Vergleich zwischen menschlichem und maschinellem Feedback zurückgegriffen. Wenngleich diese Qualitätsprüfung problematisch ist, stellt es für diese frühe Phase der Analyse von KI-Feedback dennoch eine pragmatische Zwischenlösung dar.

Dazu wurden die sieben Bewertungskriterien (s. Tabelle 1) vom Untersuchungsleiter und dem KI-System jeweils auf einer Skala (1–11 Punkte) bewertet. Rein deskriptiv lässt sich für jede Kategorie zählen, wie häufig die beiden Rater (KI und Untersuchungsleiter) in Bezug auf die Punkte wie weit auseinanderliegen:

Rater-differenz	0	+/–1	+/–2	+/–3	+/–4	+/–5	+/–6	+/–7	+/–8	+/–9	+/–10
Anzahl	19	53	52	35	30	22	12	11	2	1	1
Prozent	8	22,3	21,8	14,7	12,6	9,2	5	4,6	0,8	0,4	0,4

Tab. 3: Raterdifferenz über alle Beurteilungskategorien

Maschine und Mensch stimmen bei 30,3 % ($n = 72$) der Beurteilungen über alle Kategorien hinweg entweder perfekt (Differenz = 0) oder annähernd perfekt (Differenz = 1/–1) überein. Bei 159, also etwa in etwa zwei Drittel aller 238 Ratings, stimmen die beiden Rater mit einer Differenz von 3 oder weniger Punkte überein. Das wirkt wie ein sehr guter Wert. Diese intuitive Einschätzung wird aber durch die Angabe eines statistisch verlässlicheren Wertes infrage gestellt: Da es sich um zwei feste Rater und sieben Kategorien handelt, die mit in ihren Abständen sinnvoll interpretierbaren Punkten bewertet wurden, wurde die Interrater-Reliabilität mit der Intraklassen-Korrelation (ICC3) berechnet (Hedderich/Sachs 2020: 434 ff.). Die ICC liegt hier bei 0,279 ($p = 0{,}0001$), was eine geringe bis mäßige Übereinstimmung zwischen dem menschlichen und maschinellen Rater bei der kategorialen Bewertung der Texte andeutet.

Betrachtet man die Übereinstimmung pro Kategorie, zeigt sich die überaus komplexe Kategorie 3 (Argumentgüte) mit 53 % als diejenige, mit dem höchsten Zusammenhang zwischen Mensch und Maschine. Es ist also möglich, dass innerhalb der globalen Einschätzung der Übereinstimmung von Mensch und Maschine einzelne, besonders schlecht oder gut übereinstimmende Kategorien enthalten sind, die durch das Gesamtergebnis verschleiert werden.

Weiter oben wurde bereits darauf hingewiesen, dass Sprachmodelle ob ihrer Funktionsweise gerade die Kategorien, welche sich auf die sprachliche Oberfläche (z. B. Orthografie oder Zitate) beziehen, besonders sicher bewerten sollten. Daher beschreiben die Boxplots in Abb. 4 für jede der sieben Kategorien, wie hoch die Differenz zwischen menschlichem und maschinellem Rater ist. Dazu wurde für alle 34 Texte in jeder der sieben Kategorien jeweils die Differenz zwischen maschinellem und menschlichem Rater berechnet. Der Plot fasst die wichtigsten deskriptiven Daten der Abweichungen in Punkten (y-Achse) für jede Kategorie (x-Achse) zusammen.

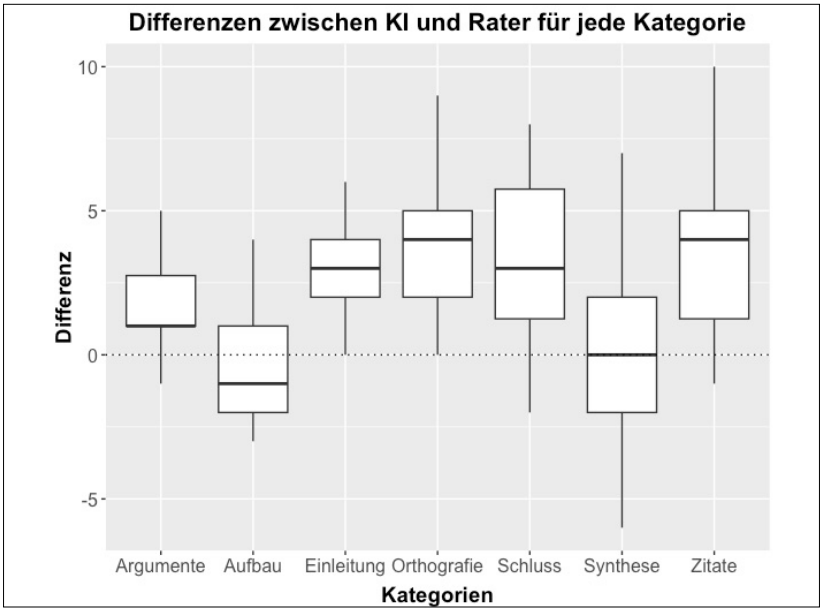

Abb. 4: Boxplots zur Abweichung von menschlichem und maschinellem Rater nach Bewertungskategorie

Die maschinelle Bewertung ist bei fünf der sieben Kategorien höher als die menschliche, weshalb die Boxen über der gestrichelten Nulllinie liegen. Die KI hat also die Texte der Schüler:innen grundsätzlich eher milder beurteilt als der Untersuchungsleiter, was zu den Ergebnissen von Seßler et al. (2025: 468) passt. Lediglich bei der Kategorie *Aufbau* war der menschliche Rater strenger und bei der Kategorie *Synthese* liegt der Median exakt auf der Null-Linie – hier stimmen also die menschlichen und maschinellen Beurteilungen im Mittel überein. Das ist insofern weniger erstaunlich, als die Kategorie *Synthese* mit der Abwägung der stärksten Argumente gegeneinander und dem Begründen der eigenen Position durchaus sprachliche Handlungen (abwägen, begründen) fordert, die konkrete Oberflächenmerkmale wie adversative Konjunktionaladverbien (*allerdings*, *einerseits … andererseits* etc.) oder kausale Subjunktionen (*weil, da* etc.) mit sich bringen, die von Sprachmodellen besonders gut erkannt werden. Auch die stärksten Argumente sind bei den meisten Texten sprachlich (z. B. *am wichtigsten aber ist*) und/oder formal (durch Absätze)

gekennzeichnet. Jedoch ist die Box der Kategorie *Synthese* an sich relativ groß und auch die Whisker-Linien (vertikale Linien, enthalten den Großteil der Daten außerhalb der Box) zeigen eine breite Streuung der Daten an. Wie bereits die ICC kategorienübergreifend gezeigt haben, kann von Einigkeit also nicht die Rede sein. Eine Reliabilitätsanalyse der einzelnen Ergebnisse kam zu keinen signifikanten Ergebnissen, was vor allem auf die geringe Anzahl an Daten pro Kategorie zurückzuführen sein dürfte.

Das mit Abstand komplexeste Beurteilungskriterium ist die Qualität der *Argumente*, da es acht – vor allem inhaltlich geprägte – Unterkriterien enthält. Es wurde gemutmaßt, dass das KI-System mit der Beurteilung dieser komplexen Kategorie Probleme haben sollte. Die prozentualen Daten zur Rater-Übereinstimmung ließen schon vermuten, dass diese Kategorie aber die höchste Übereinstimmung aufweist, was die Boxplots entsprechend spiegeln, da die Box am kleinsten ist und die Whisker-Linien am kürzesten sind. Da die Box nur nahe der Null-Linie liegt, haben Mensch und KI hier zwar tendenziell ähnlich, die KI aber mit durchschnittlich mehr Punkten beurteilt. Ähnliches gilt noch für die zweitkomplexeste Kategorie (*Einleitung*), bei der sich allerdings die Box noch weiter von der Null-Linie entfernt, die KI also noch einmal im Schnitt mit einer höheren Punktzahl bewertet hat, wenn auch die Tendenz ähnlich ist wie beim menschlichen Rater. Die Kategorien *Schluss*, *Orthografie* und der Einsatz von *Zitaten* sind allesamt überaus wenig übereinstimmend, was insbesondere mit Blick auf die beiden sprachformalen Kategorien, *Orthografie* und *Zitate*, überrascht.

Insgesamt lässt sich keine zufriedenstellende Übereinstimmung zwischen der kategorialen Bewertung durch den Untersuchungsleiter und dem Sprachmodell festhalten. Zwar liegen einzelne und teilweise auch inhaltlich komplexe Kriterien näher an einer Übereinstimmung. Die ICC-Analyse legt aber nicht nahe, dass das eingesetzte KI-System eine der menschlichen Bewertung ähnliche Beurteilung vornehmen würde. Die Punkte des Sprachmodells lagen im Mittel fast durchgehend höher als die Bepunktung durch den Untersuchungsleiter. Eine durchaus denkbare Begründung für diesen Umstand ist, dass potenzielle Attribuierungen im System-Prompt wie *Du bist ein* freundlicher *Lehrer, der Texte beurteilen soll.* dazu führen, dass *freundlich* in den Vektorräumen des Sprachmodells auch mit besserer Benotung assoziiert ist. Ebenso plausibel ist,

dass Sprachmodelle grundsätzlich einen Höflichkeits-Bias aufweisen, der sich hier und bei Seßler et al. 2025 in (zu) hohen Bewertungen von Schüler:innentexten zeigt.

4.3 Qualitatives Feedback

Um den quantitativen Abweichungen zwischen menschlichen und maschinellen Ratings qualitativ auf den Grund zu gehen, werden kursorisch einige der Textteile (orthografisch unverändert) mitsamt dem Feedback dargestellt, um Herausforderungen KI-generierten Feedbacks zu entdecken.[9] Es ist bei diesem Vorgehen wichtig, sich vor Augen zu führen, dass diese Beispiele bewusst ausgewählt wurden, um Herausforderungen herauszuarbeiten, die sich in den Daten zeigen. Es könnten ebenso viele Beispiele gezeigt werden, die potentiell hilfreiches Feedback aufweisen.

In der Gesamtschau über alle qualitativen Textrückmeldungen zeigt sich das KI-System als genau der Phrasenspeicher (Müller/Fürstenberg 2023: 336) bzw. Papagei, als der große Sprachmodelle gelten, und zwar offenbar insbesondere dann, wenn sie durch den Systemprompt einen sehr spezifischen Auftrag erhalten. So werden stets dieselben syntaktischen Strukturen (Matrixsatz, häufig mit einer Aufzählung, gefolgt von einem Satzgefüge mit einer finalen Infinitivgruppe; s. Abb. 5) mit den entsprechenden Wörtern aus der jeweiligen Beurteilungskategorie gefüllt, was insgesamt eher an Produkte eines deterministischen Systems erinnert und wohl auf eine sehr niedrige *temperature* (Fürstenberg/Müller 2024: 85) zurückzuführen ist. In der Einzelbetrachtung zeigen sich dann aber durchaus Unterschiede zwischen den jeweiligen Feedbacktexten, deren inhaltliche Qualität in der Folge beleuchtet wird.

Überraschende Einigkeit zeigte in der quantitativen Untersuchung die komplexe Beurteilungskategorie *Argumente*. Bei der qualitativen Beurteilung moniert das KI-System in jedem zweiten Feedbacktext, die Sprache solle sachlicher und/oder unterstützender sein. Auch die inhaltliche Kohärenz

9 Qualitative Analysen zu den Kategorien *Einleitung* und *Synthese* liegen in Fürstenberg (2025) bereit.

der Argumente wird relativ häufig (n = 12) kritisiert. Der menschliche Rater hingegen kritisiert am häufigsten fehlende oder zu wenig konkrete Beispiele in den Argumenten und damit die Vollständigkeit der Argumente und weniger sprachliche Faktoren. Es zeigen sich qualitativ also deutliche Unterschiede, wenngleich die quantitative Übereinstimmung in dieser Kategorie am höchsten ist. Hier deutet sich eine Inkonsistenz zwischen quantitativer Bepunktung und qualitativer Beurteilung per textuellem Feedback an, die sich noch an weiteren Stellen zeigt: Bei fünf Texten lässt das qualitative Feedback des KI-Systems keine Verbesserungsoptionen erkennen, beurteilt die Texte also als perfekt. Nicht nur für die Schüler:innen ist es verwirrend, dass das System aber nur in zwei dieser fünf Fälle analog zum qualitativen Feedback auch quantitativ die Höchstpunktzahl ausgibt. Hinzu kommt, dass in beiden Fällen, in denen das KI-System auch elf Punkte erteilt, das menschliche Feedback die Texte deutlich kritischer (sieben bzw. acht Punkte) einschätzt: So fehlen den sowohl qualitativ als auch quantitativ als perfekt beurteilten Argumenten teilweise (konkrete) Beispiele und auch sprachlich sind die Argumente stellenweise zu unsachlich formuliert. Insofern böten beide Texte noch Verbesserungsoptionen, die durch das maschinelle Feedback aber unentdeckt bleiben.

Auch die Kategorie *Schluss* wurde vom KI-System sprachlich überaus homogen beurteilt. So formuliert die KI für jeden der 34 Texte, dass der Schluss (gut) an die Einleitung anschließe, einen Appell/Ausblick enthalte, jedoch vice versa noch Ausblick/Appell formuliert werden könnte, um den Text zu verbessern. Mit Blick auf das Beurteilungskriterium (s. Tab. 1) war jedoch nur ein Ausblick *oder* ein Appell gefordert, weshalb jedes einzelne Feedback inhaltlich mindestens als unpräzise zu bewerten ist. Nur in drei Fällen weicht das Feedback (jeweils bei der zweiten Textversion) ganz leicht von der benannten Struktur ab, indem es auf die Überarbeitung konkret Bezug nimmt (*Dein Schluss schließt immer noch gut an die Einleitung an und enthält einen Ausblick.*).

Der Schluss eines Schüler:innentextes (Herv. MF) rief eine besonders hohe quantitative Unstimmigkeit hervor und die entsprechende Rückmeldung deckt eine Herausforderung maschinell erstellten Feedbacks auf:

Obwohl der Beruf Vorteile und Nachteile hat, ist es die Entscheidung jedes Einzelnen, ob er oder sie sich für den Beruf entscheidet. Allerdings sollte man sich den möglichen Konsequenzen, die die körperliche und mentale Gesundheit beeinträchtigen können, bewusst sein.
Dein Schluss schließt gut an die Einleitung an und enthält einen Ausblick. Du könntest jedoch einen stärkeren Appell einbauen, um deine Position zu unterstreichen.

Abb. 5: Feedbackbeispiel I

Der Schluss wird vom Sprachmodell mit der Höchstzahl bewertet (11 Punkte), während der menschliche Rater lediglich vier Punkte vergibt. Das qualitative KI-Feedback unterstreicht zu Recht und positiv die Verbindung zur Einleitung, moniert aber (trotz Höchstpunktzahl) die Stärke des Appells in Form eines Überarbeitungsvorschlags. Das maschinelle Feedback weist hier erneut mangelnde Konsistenz in der Bewertung auf – in der Diskrepanz zwischen Maximalpunktzahl als Bestnote und der qualitativen Rückmeldung – und lässt inhaltliche Mängel außer Acht, die im Vergleich zum menschlichen Feedback deutlich werden, welches den Schluss ob der vielen Allgemeinplätze als unzureichend ausweist. Da die Bewertungskriterien aber nicht explizit ausweisen (können), was alles nicht gemacht werden sollte, und womöglich auch weil Sprachmodelle selbst qua Entwicklung (Müller/Fürstenberg 2023) einen Hang zu sehr allgemeinen Aussagen aufweisen, erkennt das Sprachmodell ein zentrales Problem dieses Schlusses nicht. Darüber hinaus wird hier auch ein bekanntes Problem kriteriengeleiteter Beurteilung von Texten offenbar, denen das Modell ausgeliefert ist: Über die Kriterien hinaus können Texte sehr individuelle Mängel aufweisen, die aber durch den Kriterienkatalog nicht abgedeckt sind. Hier kommt also eher die kriteriengeleitete Beurteilung als das KI-System an seine Grenzen.

Auch die deskriptiven Daten (s. Abbildung 3) zur Übereinstimmung zwischen Mensch und Maschine innerhalb der Kategorie *Zitate* wiesen auf eine hohe Unstimmigkeit hin, die insofern weniger erwartbar war, als der Einsatz von Zitaten auf der sprachlichen Oberfläche recht eindeutig markiert wird. Das folgende Beispiel aus einem Schüler:innentext weist den zentralen Grund für die Diskrepanz innerhalb dieser Kategorie aus:

Daraufhin ist er dann zwei oder drei Wochen „in", aber danach interessiert sich keiner mehr für ihn und er hat kein gutes Einkommen mehr, da er nicht so viele Klicks bekommt.

```
Du hast aus dem Material zitiert, aber die Quellenangabe fehlt.
Achte darauf, die Quelle unter Angabe der Quelle zu zitieren.
```

Abb. 6: Feedbackbeispiel II

Neben der zitierten Textstelle wird nur in der Einleitung mit Verweis auf die Quelle zitiert. Das KI-System vergibt sieben Punkte, während der menschliche Rater den Einsatz von Zitaten über den gesamten Text hinweg mit lediglich zwei Punkten beurteilt. Die vermeintliche Eindeutigkeit der Markierung von Zitaten auf der sprachlichen Oberfläche scheint bei dem Modell zu einer fehlergenerierenden Übergeneralisierung geführt zu haben. So zeigt dieses Beispiel, dass teilweise modalisierende An- und Abführungszeichen als Zitate „fehlinterpretiert" wurden, was stark an die Urteile rein oberflächenorientierter Systeme erinnert, deren Beschränktheit große Sprachmodelle durch ihre hochdimensionalen Vektorräume und damit einen stärkeren bedeutungsorientierten Zugang zu überwinden versprechen.

Auch in dieser Kategorie zeigen sich wieder Konsistenzprobleme, da zweimal sechs Punkte mit der qualitativen Rückmeldung gepaart werden, dass überhaupt nicht aus dem Material zitiert wurde.

4.4 Fragebogen

Abschließend wurden die Schüler:innen ($N = 17$) zum Feedback durch die KI mithilfe eines Fragebogens befragt. Es werden an dieser Stelle lediglich wenige Auszüge präsentiert.

Elf Schüler:innen sind der Meinung, das Überarbeiten grundsätzlich habe ihren Text verbessert. Bis auf eine Testperson hatten damit auch alle recht. Das zeigt neben der guten Selbsteinschätzung dieser Schüler:innen, dass die Überarbeitung an sich schon wirksam sein kann, so sie denn angegangen wird. KI-Systeme könnten hier durch die Unmittelbarkeit des Feedbacks die Motivation steigern, diese Überarbeitung auch tatsächlich durchzuführen. Die qualitativen Rückmeldungen zum maschinellen Feedback fokussieren vor allem die teilweise fehlende Genauigkeit bzw. Richtigkeit des Feedbacks, aber auch den mangelnden persönliche Kontakt, den auch Rüdian et al. (2025) in ihren

Daten (Meinungen von Lernenden zu automatisch generiertem Feedback) finden. Auch die mangelnde Möglichkeit der Nachfrage sowie die fehlende Korrektur im Text werden kritisiert. Folgendes Feedback fasst diesen Kritikpunkt zusammen:

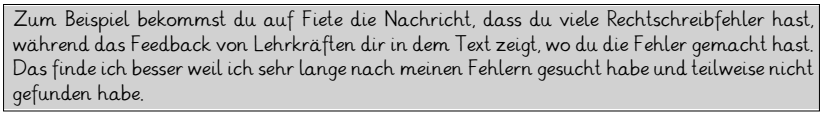

Abb. 7: Schüler:innenfeedback I

In diesem Feedback zeigt sich eine wichtige Herausforderung von KI-Systemen. Denn für die Anbindung an den Text bzw. die konkreten Textstellen wäre im Grunde doch wieder eine Art von Korrektur im Text notwendig, die wiederum ein eher deterministisches Vorgehen benötigen würde. Hier sind die Entwickler:innen dieser Systeme gefordert, Lösungen zu erarbeiten, die es den Schüler:innen möglich machen, das Feedback konkret an bestimmte Textstellen rückzubinden (s. auch das *Lokalisationsproblem* in Fürstenberg 2025).

Durchaus erfreulich ist der Umstand, dass sich Schüler:innen der bereits diskutierten Fehler des KI-Systems bewusst waren und diese auch im Fragebogen monierten:

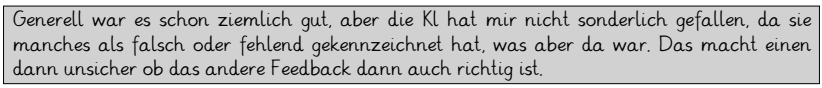

Abb. 8: Schüler:innenfeedback II

Der Umstand, dass auch didaktisch systemgepromptete KI-Systeme eindeutige Fehler produzieren und auch quantitativ kaum mit menschlicher Bewertung übereinstimmen, zeigt überdeutlich, dass bei der Entwicklung dieser Modelle Expert:innen aus Gebieten wie der Didaktik, Psychologie, Linguistik und Informatik gemeinsam an Lösungen arbeiten müssten. Denn was die Schüler:in hier noch recht milde mit „nicht sonderlich gefallen" umschreibt, hat wohl eine der größten Herausforderungen maschinell erstellten Feedbacks in Zukunft zur Folge: das Vertrauen in die KI-Systeme und deren Ausgaben. Entwickeln

Lehrkräfte und Schüler:innen kein Zutrauen in die Modelle, werden sie erst gar nicht eingesetzt oder ihr Feedback wird nicht ernst genommen. Gerade jetzt, in einer Zeit, in der Modelle schon breit rezipiert werden, aber womöglich noch nicht ausreichend gut entwickelt sind, ist eine überaus sensible Phase für dieses Vertrauensverhältnis. Bereits in diesem frühen Stadium das Vertrauen der betroffenen Personen (Schüler:innen und Lehrkräfte) in KI-generiertes Feedback zu verlieren, wäre fatal.

Das letzte Fragebogenitem erfragte die Vorteile KI-gestützten Feedbacks im Vergleich zu Feedback durch die Lehrkraft und vice versa. Tabelle 4 fasst die Ergebnisse zusammen:

Vorteile Lehrerfeedback	**N**	**Vorteile KI-Feedback**	**N**
Korrektheit	9	Unmittelbarkeit	8
Genauigkeit	9	Nichts	4
Rückfragemöglichkeit	2	Visualisierung	4
persönlich	2	Neutralität & Wiederholbarkeit	1

Tab. 4: Zusammenfassung Schüler:innenfeedback zu Vorteilen von Feedback durch Lehrkräfte/KI

Die Schüler:innen heben vor allem einen zentralen Vorteil des Feedbacks durch KI-Systeme hervor: die Unmittelbarkeit des Feedbacks. Dies wird auch auf lange Sicht der größte Vorteil automatisch generierten Feedbacks bleiben. Des Weiteren wird die Visualisierung bzw. Übersichtlichkeit der KI-generierten Rückmeldungen betont, womit auf die Balkenanzeigen (s. Abbildung 1) abgehoben wird, welche es den Schüler:innen theoretisch erleichtert, auf die Schnelle besonders gut oder schlecht erfüllte Kriterien zu erfassen. Vor dem Hintergrund der quantitativen Ergebnisse, die zeigen, dass die Balkenanzeige und das qualitative Feedback nicht immer konsistent sind, bleibt dieser zweite Vorteil noch fragwürdig zurück. Beim Vorteil des Feedbacks durch Lehrkräfte wird einerseits die Genauigkeit betont, womit die Schüler:innen die Rückbindung an ihren konkreten Text durch die Korrektur meinten. Andererseits wird die Korrektheit des Feedbacks durch Lehrkräfte von den Schüler:innen besonders betont. Dies ist in dieser Untersuchung durchaus auch als Reaktion auf die bereits thematisierten Fehler zurückzuführen, welche die KI-Systeme

zeigten, und verstärkt noch einmal die Relevanz von inhaltlich richtigem Feedback mit Blick auf die Vertrauenswürdigkeit des Feedbackgebenden, die von zentraler Bedeutung für die Feedbacknehmenden ist.

5 Fazit

Die meisten Schüler:innen konnten durch die Überarbeitung ihre Texte verbessern, was allerdings aus Gründen des Forschungsdesigns nicht eindeutig auf das Feedback der KI zurückzuführen ist. Der Vergleich von menschlicher und maschineller Bewertung deutet an, dass es zwischen dem Rater und der eingesetzten KI keine allzu hohen Übereinstimmungen bei kriteriengeleiteter Bewertung gibt. Allerdings muss auch dieses Ergebnis in weiteren Studien an einer breiteren Proband:innengruppe geprüft werden. Ein Problem des maschinellen Feedbacks liegt in der mangelnden Konsistenz, dass also qualitatives Feedback in Form von Text und die quantitative Bewertung durch das Sprachmodell nicht übereinstimmen. Hohe Unstimmigkeitswerte zwischen Mensch und Maschine sind meist auf fehlerhaftes Feedback durch die KI rückführbar, die jedoch neben den hier zitierten Stellen durchaus auch passendes Feedback produzierte, was die deskriptiven Ergebnisse zu den quantitativen Beurteilungen auch durchaus eindrücklich zeigen. In der Befragung loben die Schüler:innen die Unmittelbarkeit des maschinell erstellten Feedbacks. Ein großer Nachteil sind die inhaltlich falschen Feedbacks, welche die Schüler:innen zwar stellenweise als solche erkennen, aber dennoch bleibt es eine aktuell nicht überschreitbare Grenze des Einsatzes Künstlicher Intelligenz für das Feedback zu Schüler:innentexten. Denn der Umstand des inhaltlich falschen Feedbacks durch beispielsweise Halluzinationen bzw. Bullshit (Müller/Fürstenberg 2023; Hicks/Humphries/Slater 2024) ist ein architekturinhärentes Problem (Müller/Fürstenberg 2023: 335–341), das sich nur durch überaus aufwändiges Training oder Feintuning (Simbeck i. d. B.) einschränken, aber vorerst nicht gänzlich beheben lässt.

Die inhaltliche Qualität KI-generierten Feedbacks benötigt dringend weitere qualitative und fachdidaktisch fundierte Forschung, um den aufgezeigten

Herausforderungen zu begegnen und Lehrkräften weitere Möglichkeiten sowie Schüler:innen schnelles und lernförderliches Textfeedback zu ermöglichen.

Literatur

Averintseva-Klisch, Maria (2018): *Textkohärenz*. 2., aktualisierte Aufl. Heidelberg: Winter.

Bachmann, Thomas/Becker-Mrotzek, Michael (2010): Schreibaufgaben situieren und profilieren. In: Pohl, Thorsten/Steinhoff, Torsten (Hg.): *Textformen als Lernformen*. Duisburg: Gilles & Francke, S, 191–209.

Batzlen, Christian (13.1.2023): *ChatGPT: Das Ende der Hausarbeit?* URL: https://www.swr.de/swr2/programm/chatgpt-in-der-uni-schreiben-sich-wissenschaftliche-arbeiten-bald-von-alleine-100.html (letzter Zugriff: 7.2.2024).

Bender, Emily M/Gebru, Timnit/McMillan-Major, Angelina/Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada,* S. 610–623. https://doi.org/10.1145/3442188.3445922.

Birkel, Peter/Birkel, Claudia (2002): Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. In: *Psychologie in Erziehung und Unterricht* 49 (3), S. 219–224.

Böttcher, Ingrid/Wagner, Monika (1993): Kreative Texte bearbeiten. In: *Praxis Deutsch*, 20 (199), S. 24–27.

Chiang, Cheng-Han/Hung-yi Lee (2023): Can large language models be an alternative to human evaluations? In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, S. 15607–15631.

Dax, Simon (2023): Werteorientiertes Argumentieren im Deutschunterricht. In: *Mitteilungen des Deutschen Germanistenverbandes* (3), S. 254–269.

Emmerich, Nadine (25.1.2023): *ChatGPT in der Bildung. „Hausaufgaben sind tot"*. URL: https://www.gew.de/aktuelles/detailseite/hausaufgaben-sind-tot (letzter Zugriff: 15.2.2024).

Feilke, Helmuth/Tophinke, Doris (2017): Materialgestütztes Argumentieren. In: *Praxis Deutsch*. 44 (262), S. 4–13.

Fleckenstein, Johanna/Meyer, Jennifer/Jansen, Thorben/Keller, Stefan/Köller, Olaf/Möller, Jens (2024): Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays. In: *Computers and Education: Artificial Intelligence* (6), S. 100209. https://doi.org/10.1016/j.caeai.2024.100209.

Fürstenberg, Maurice (2025): Zur Qualität von KI-generiertem Feedback – Ergebnisse einer explorativen Untersuchung. In: *Leseräume* (11).

Fürstenberg, Maurice/Matz, Daniela (2025): Künstliche Intelligenz als Diskussionspartner? Materialgestütztes Argumentieren mit und reflektieren über KI. In: *Praxis Deutsch* (311), S. 23–30.

Fürstenberg, Maurice/Müller, Hans-Georg (2024): KI im Deutschunterricht. Funktionsprinzipien und kompetenzbezogene Einsatzmodelle. In: *Der Deutschunterricht* 76 (5), S. 2–13.

Haverkamp, Hendrik/Hecht, Malte/Schindler, Kirsten (2024): Lernförderliches Feedback KI-basiert vermitteln. In: *Der Deutschunterricht* 76 (5), S. 60–71.

Hedderich, Jürgen/Sachs, Lothar (2020): *Angewandte Statistik Methodensammlung mit R*. 17., überarbeitete und ergänzte Aufl.. Berlin: Springer.

Herbold, Steffen/Hautli-Janisz, Annette/Heuer, Ute/Kikteva, Zlata/Trautsch, Alexander (2023): A large-scale comparison of human-written versus ChatGPT-generated essays. In: *Scientific Reports* 13 (1), S. 18617. https://doi.org/10.1038/s41598-023-45644-9.

Hicks, Michael/Humphries, James/Slater, Joe (2024): ChatGPT is bullshit. In: *Ethics and Information Technology* (26), S. 38. https://doi.org/10.1007/s10676-024-09785-3.

Jansen, Thorben/Horbach, Andrea/Meyer, Jennifer (2025): Feedback from Generative AI: Correlates of Student Engagement in Text Revision from 655 Classes from Primary and Secondary School. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25),* S. 831–836. https://doi.org/10.1145/3706468.3706494.

Mizumoto, Atsushi/Eguchi, Masaki (2023): Exploring the potential of using an ai language model for automated essay scoring. In: *Research Methods in Applied Linguistics* 2 (2), S. 100050.

Mußmann, Frank/Hardwig, Thomas/Riethmüller, Martin (2017): *Arbeitszeit und Arbeitsbelastung von Lehrkräften in Niedersachsen: Ergebnisbericht der*

Arbeitsbelastungsstudie 2016. Georg-August-Universität Göttingen, Kooperationsstelle Hochschulen und Gewerkschaften.

Mußmann, Frank/Hardwig, Thomas/Riethmüller, Martin/Klötzer, Stefan/Peters, Stefan (2020): *Arbeitszeit und Arbeitsbelastung von Lehrkräften an Frankfurter Schulen 2020: Ergebnisbericht*. Georg-August-Universität Göttingen, Kooperationsstelle Hochschulen und Gewerkschaften. https://doi.org/10.3249/ugoe-publ-7.

Müller, Hans-Georg/Fürstenberg, Maurice (2023): Der Sprachgebrauchsautomat. Sieben Thesen, die aus den technischen Grundlagen von GPT folgen. In: *Mitteilungen des Deutschen Germanistenverbandes* (4), S. 327–345.

Müller, Nora/Utesch, Till/Busse, Vera (2023): Qualität statt Quantität? Zum Zusammenhang von Schreibförderungs- und Feedbackpraktiken mit Textqualität unter Berücksichtigung von migrationsbedingter Mehrsprachigkeit. In: *Unterrichtswissenschaft* (51), S. 169–198. https://doi.org/10.1007/s42010-023-00173-2.

Naismith, Ben/Mulcaire, Phoebe/Burstein, Jill (2023): Automated evaluation of written discourse coherence using gpt-4. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, S. 394–403.

Neff, Tina (2023): Einsatz digitaler Korrekturhilfen im Rechtschreibunterricht. Erste Erkenntnisse einer Pilotstudie in der Primarstufe und Sekundarstufe I. In: *Medien im Deutschunterricht* (1), S. 1–17.

Padó, Ulrike/Eryilmaz, Yunus/Kirschner, Larissa (2023): Short-Answer Grading for German: Addressing the Challenges. In: *International Journal of Artificial Intelligence in Education* 34 (4), S. 1321–1352. https://doi.org/10.1007/s40593-023-00383-w.

Page, Ellis (1966): The Imminence of… Grading Essays by Computer. In: *The Phi Delta Kappan* 47 (5), S. 238–243. http://www.jstor.org/stable/20371545.

Philipp, Maik (2015): *Schreibkompetenz. Komponenten, Sozialisation und Förderung.* Tübingen: A. Francke.

Philipp, Maik (2023): Formatives Feedback aus der Sicht des selbstregulierten Lernens. Grundlagen und Grundsätze förderlicher Rückmeldungen. In: *ide* 47 (2), S. 8–17.

Ramesh, Dadi/Sanampudi, Suresh (2022): An automated essay scoring systems: a systematic literature review. In: *Artificial Intelligence Review* 55 (3), S. 2495–2527.

Reichardt, Anke/Kruse, Norbert/Lipowsky, Frank (2014): Textüberarbeitung mit Schreibkonferenz oder Textlupe. Zum Einfluss der Schreibumgebung auf die Qualität von Schülertexten. In: *Didaktik Deutsch.* 19 (36), S. 65–85.

Rüdian, Sylvio/Podelo, Julia/Kužílek, Jakub/Pinkwart, Niels (2025): Feedback on Feedback: Student's Perceptions for Feedback from Teachers and Few-Shot LLMs. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, S. 82–92. https://doi.org/10.1145/3706468.3706479.

Sawatzki, Jörg/Schlippe, Tim/Benner-Wickner, Marian (2022): Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. Cheng, Eric/Koul, Rekha/Wang, Tianchong/Yu, Xinguo (Hg.): *Artificial Intelligence in Education: Emerging Technologies, Models and Applications. Proceedings of 2021 2nd International Conference on Artificial Intelligence in Education Technology* (Lecture Notes on Data Engineering and Communications Technologies: 104). Singapore: Springer, S. 65–75.

Schröter, Pauline/Söldner, Hannelore/Hoffmann, Lars/Riemenschneider, Anja/Jost, Jörg/Wieser, Dorothee (2022): Wie vergleichbar sind die Bewertungen von Abiturarbeiten im Fach Deutsch? Empirische Studien zu verschiedenen Bewertungsmodellen. In: Schröter, Pauline/Groß, Alexander/Schmid-Kühn, Svenja/Stanat, Petra/Hoffmann, Lars (Hg.): *Das unvergleichliche Abitur: Entwicklungen-Herausforderungen-Empirische Analysen.* Bielefeld: wbv, S. 213–250.

Schwarze, Cordula (2016): Was ist ein gutes Argument? – Zu Analyse, Reflexion und Beurteilung mündlichen Argumentierens. In: Hinger, Barbara (Hg.): *Zweite „Tagung der Fachdidaktik" 2015. Sprachsensibler Sach-Fach-Unterricht – Sprachen im Sprachunterricht* (Innsbrucker Beiträge zur Fachdidaktik: 2). Innsbruck: innsbruck university press, S. 161–190. URL: https://www.uibk.ac.at/iup/buch_pdfs/zweite-fachdidaktik/10.152033122-51-2.pdf (letzter Zugriff: 5.2.2024).

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hg.) (KMK) (2005): *Beschlüsse der Kultusministerkonferenz. Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4). Beschluss vom 15.10. 2004.* URL: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Deutsch-Primar.pdf (letzter Zugriff: 5.2.2024).

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hg.) (KMK) (2014): *Bildungsstandards im Fach*

Deutsch für die Allgemeine Hochschulreife (Beschluss der Kultusministerkonferenz vom 18.10.2012). URL: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Deutsch-Abi.pdf (letzter Zugriff: 5.2.2024).

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hg.) (KMK) (2022): *Bildungsstandards für das Fach Deutsch Primarbereich (Beschluss der Kultusministerkonferenz vom 15.10.2004, i. d. F. vom 23.06.2022)*. URL: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2022/2022_06_23-Bista-Primarbereich-Deutsch.pdf (letzter Zugriff: 5.2.2024).

Seßler, Kathrin/Fürstenberg, Maurice/Bühler, Babette/Kasneci, Enkelejda (2025): Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In: *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, S. 462–472. https://doi.org/10.1145/3706468.3706527.

Spitta, Gudrun (1992): *Schreibkonferenzen in Klasse 3 und 4. Ein Weg vom spontanen Schreiben zum bewussten Verfassen von Texten*. Frankfurt, Main: Cornelsen Scriptor.

Stahl, Maja/Biermann, Leon/Nehring, Andreas/Wachsmuth, Hennig (2024): Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In: *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*. S. 283–298. URL: https://aclanthology.org/2024.bea-1.23.pdf (letzter Zugriff: 13.02.2025).

Sturm, Afra/Schneider, Hansjakob (2021): Flüssiges Formulieren in der Textproduktion (Klasse 4/5). In: *Didaktik Deutsch* 26 (51), S. 28–49.

Wendt, Charlotte (2023): Schreiben lernen mit intelligenter Hilfe. Wie computergestütztes Feedback Schreiblernprozesse verändern kann. In: *ide* (2), S. 38–47.

Große generative Sprachmodelle (LLMs) haben die Exotik der Anfangstage verloren. Sie sind längst Alltag geworden, auch im Lehren und Lernen. In dieser neuen Phase des sprach- und literaturdidaktischen Umgangs mit ChatGPT & Co. muss sich der Deutschunterricht in Theorie und Praxis neuen Fragen stellen: Wie funktioniert KI und was ist von dieser Technik in Zukunft noch zu erwarten? Wie wirkt sich die Verwendung von KI auf den sprachlichen Kompetenzerwerb aus? Wie gehen Schüler:innen mit digitalen Textgeneratoren um und welche Konsequenzen hat das für ihre Bildungsgeschichte? Wie lässt sich KI sinnvoll in den Deutschunterricht integrieren? Wo liegen Risiken, wo Chancen für die Gestaltung der Schule von morgen? Diesen und anderen Fragen gehen die Beiträge des Bandes nach. Sie präsentieren zudem eine Auswahl aktueller Forschungsprojekte rund um den Einsatz von KI im Deutschunterricht. Der Band bildet somit den gegenwärtigen Forschungsstand zum Thema Deutschunterricht und KI in seiner ganzen Breite ab.

Prof. Dr. Hans-Georg Müller lehrt Sprachdidaktik am Institut für Germanistik der Universität Potsdam. Seine Forschungsschwerpunkte liegen in den Bereichen Kognitionswissenschaft, empirische Bildungsforschung und (Schrift-)Spracherwerb.

Dr. Maurice Fürstenberg lehrt und forscht als Akademischer Rat an der LMU München in der Germanistischen Linguistik an der Schnittstelle zur Sprachdidaktik. Er promovierte zum Gebrauch des Kommas durch Schüler:innen und habilitiert zu KI im Deutschunterricht.

ISBN 978-3-7329-1120-2

www.frank-timme.de