13624

Entwicklungen und Perspektiven der Fachdidaktiken

wbv

E-Journal Einzelbeitrag
von: Rainer Mühlhoff, Sean Quägwer

Automatisierte Korrektur mit KI?

Wir testen zwei verbreitete Tools

aus: Entwicklungen und Perspektiven der Fachdidaktiken (SEM2602W)
Erscheinungsjahr: 2026
Seiten: 62 - 76
DOI: 10.3278/SEM2602W007

Die Studie untersucht die Gebrauchstauglichkeit KI-gestützter Korrektur- und Feedbacksysteme im schulischen Kontext anhand der Tools FelloFish und Edaira. Auf Basis kontrollierter Testreihen mit simulierten Schülertexten wird insbesondere die Konsistenz automatisierter Bewertungen und die Wirkung iterativer Überarbeitungen anhand des automatisierten Feedbacks analysiert. Die Ergebnisse zeigen erhebliche Bewertungsvolatilität bei identischen Eingaben, eine fehlende Reproduzierbarkeit von Urteilen sowie inkonsistente Effekte bei der Umsetzung von Verbesserungsvorschlägen. Zudem wird deutlich, dass die Systeme dazu tendieren, wortgetreue Übernahmen automatisch vorgeschlagener Formulierungen besser zu bewerten als inhaltlich gleichwertige, eigenständige Überarbeitungen. Insgesamt weisen die Befunde auf strukturelle Grenzen von Bewertungssystemen auf Basis großer Sprachmodelle hin. Die Autoren folgern, dass solche Tools derzeit nicht für eine eigenständige Leistungsbewertung geeignet sind, sondern allenfalls als unterstützende Werkzeuge unter kritischer menschlicher Kontrolle eingesetzt werden sollten.

Schlagworte: KI-gestützte Leistungsbewertung; Automatisiertes Feedback; Bewertungsvolatilität; Sprachmodelle im Bildungsbereich; Digitale Bildung; EdTech
Zitiervorschlag: *Mühlhoff, Rainer & Quägwer, Sean (2026). Automatisierte Korrektur mit KI? Wir testen zwei verbreitete Tools. SEMINAR, 32(2), 62-76. Bielefeld: wbv Publikation. https://doi.org/10.3278/SEM2602W007*

wbv

Automatisierte Korrektur mit KI?

Wir testen zwei verbreitete Tools

Rainer Mühlhoff & Sean Quägwer

1 Einleitung

1.1 KI für den Schulunterricht

In den letzten Jahren hat der Ruf nach dem Einsatz künstlicher Intelligenz (KI) im Schulunterricht in Deutschland deutlich zugenommen. Bildungspolitische Gremien wie die Ständige Wissenschaftliche Kommission (SWK) der Kultusministerkonferenz (KMK) empfehlen einen rechtssicheren Zugang zu KI-Systemen für Schulen (SWK 2024, S. 4, Brand 2024); auf KMK-Fachtagungen wird die „intelligente Auswertung von Lern- und Leistungsdaten" als Baustein einer Kultur der Digitalität diskutiert (KMK 2025). Mehrere Bundesländer führen spezialisierte KI-Chatbots und Lernplattformen ein, um Lehrkräfte bei der Unterrichtsorganisation und Lernende bei der Aufgabenbearbeitung zu unterstützen; eine entsprechende Plattform wird von den Bundesländern gemeinsam entwickelt (Schleswig-Holstein.de 2025, NDR 2026, Telli.Schule o. J.).

Flankiert wird dieser Diskurs von einer wachsenden Zahl inländischer Unternehmen und „Startups", die spezialisierte KI-Werkzeuge für den deutschsprachigen Schulkontext anbieten. Anbieter wie Fobizz (https://fobizz.com/), Edaira (https://edaira.ai/, vormals Notencopilot) oder FelloFish (https://www.fellofish.com/, vormals Fiete.ai) versprechen Lösungen für Unterrichtsplanung, Materialerstellung und Leistungsbewertung. Technisch greifen viele dieser webbasierten Dienste auf große Sprachmodelle wie GPT (OpenAI), Gemini (Google) oder Claude (Anthropic) internationaler Technologiekonzerne zurück und übersetzen deren Textfähigkeiten in schulnahe Anwendungsszenarien.

Auch strukturell ist KI im deutschen Schulwesen inzwischen breit verankert. Mit dem bundesweiten Rollout des Chatbots „Telli" seit dem Schuljahr 2025/26 verfügen neun Bundesländer über einen landesweiten KI-Zugang für Schulen; zuvor gestartete Pilotprojekte (etwa in Bremen) sind damit in eine flächendeckende Bereitstellung überführt worden (News4Teachers 2025, Telli.Schule o. J.). Umfragen zeigen zugleich eine ambivalente Praxis: Laut aktuellem Deutschen Schulbarometer nutzen 31 % der Lehrkräfte KI-Tools mehrmals monatlich oder häufiger, während 62 % sich im Umgang mit KI unsicher fühlen (Robert Bosch Stiftung 2025, S. 10). Die Verfügbarkeit von KI-Software dürfte sich in den nächsten Jahren ausweiten, denn Bund und Länder stellen im Rahmen des „DigitalPakts 2.0" insgesamt 5 Mrd. Euro über fünf Jahre bereit, die

auch zur Anschaffung von KI-Softwaresystemen genutzt werden dürfen (BMBFSFJ o. J.).

Ein wichtiges Anwendungsfeld von KI im Schulkontext ist der Bereich automatisierter Rückmeldungen und Leistungsbewertungen. Solche KI-Tools sollen Aufsätze kommentieren, individuelle Verbesserungsvorschläge geben und Bewertungen generieren – wahlweise unterstützend für Lehrkräfte oder im Austausch mit Lernenden. In Zeiten von Lehrkräftemangel und hoher Arbeitsbelastung wird diese Form der (Teil-)Automatisierung einer pädagogischen Kernaufgabe als Entlastung und als Hebel für stärkere Individualisierung präsentiert. Zugleich gehen damit eine Automatisierung und potenzielle Entwertung der pädagogisch hoch sensiblen Praxis des Beurteilens, Kommentierens und Bewertens einher. Genau an diesem Punkt setzt die vorliegende Untersuchung an.

1.2 Zielsetzung und Untersuchungsgegenstand

Ziel der vorliegenden Studie ist es, die Leistungsfähigkeit und Zuverlässigkeit KI-gestützter Korrektur- und Feedbacksysteme im schulischen Kontext systematisch zu überprüfen. Methodisch knüpfen wir an unsere frühere Untersuchung zum KI-Korrekturtool von Fobizz an (Mühlhoff & Henningsen 2025), in der wir anhand kontrollierter Testszenarien die Bewertungsqualität und Konsistenz automatisierter Rückmeldungen analysiert haben. Dabei stehen weniger fachdidaktische Fragestellungen, als „simple" Aspekte der funktionalen Zuverlässigkeit und Gebrauchstauglichkeit der Werkzeuge im Mittelpunkt. Während sich die Vorgängerstudie auf ein einzelnes, im deutschsprachigen Raum weit verbreitetes Tool konzentrierte, erweitert die vorliegende Untersuchung den Blick auf zwei weitere Anbieter: FelloFish (vormals Fiete.ai) und Edaira (vormals NotenCopilot). Neben der Testung dieser Tools soll damit auch erhellt werden, ob die von uns identifizierten Unzulänglichkeiten des Fobizz-Tools singuläre Schwächen eines einzelnen Anbieters darstellen oder Ausdruck technisch bedingter Grenzen von KI-Korrektursystemen auf Basis großer Sprachmodelle sind.

FelloFish ist ein Angebot der FelloFish GmbH und wird als KI-gestützter „Schreibbegleiter" vermarktet. Das Tool bietet ein Interface direkt für Lernende, die sich während des Schreibprozesses für ihren Textentwurf individuelle Rückmeldungen ausgeben lassen können. Im typischen Gebrauch sollen die Lernenden ihre Texte anhand dieser Rückmeldungen einmal verbessern, bevor sie dann über das Interface eine „Abgabefassung" einreichen. Diese finale Textfassung wird einer automatisierten Bewertung unterzogen, die aufgeschlüsselt nach den Bewertungskriterien visuell kommuniziert wird. Neben der Ansicht für Lernende stellt das Interface außerdem eine Ansicht für Lehrkräfte bereit, in der die Bewertung der Texte entlang der einzelnen Kriterien zusätzlich numerisch (in Prozentangaben, volle Zehnerschritte) ausgegeben wird.

Edaira, hervorgegangen aus der NotenCopilot OHG, bietet ein KI-Tool als Benotungsassistenz für Lehrkräfte. Über das Webinterface können Lehrkräfte die für die Benotung zu verwendenden Kriterien definieren und die Texte ihrer Lernenden in das Tool

einspeisen. Die automatische Bewertung wird aufgeschlüsselt in Punkte für die einzelnen Bewertungskriterien, eine Bewertung der sprachlichen Leistung und eine finale Gesamtbewertung in Prozentpunkten. Unsere Testreihen für Edaira wurden für Testreihe A von Juni bis November 2025 erhoben, für Testreihe B im März 2026 (die Umbenennung des Tools von NotenCopilot zu Edaira fällt dazwischen).

2 Testbedingungen und Methodik

2.1 Zur Methodologie

Methodisch steht unsere Studie vor zwei Herausforderungen. Erstens: Prinzipiell ist die Bewertung von Lernleistungen sowie die Frage, was als „gutes", „angemessenes" oder „förderliches" Feedback gelten kann, ein normativ hoch aufgeladener Kernbereich pädagogischer Professionalität. Wir unternehmen nicht den Versuch, fachdidaktische Theorien gelingender Rückmeldung oder angemessener Bewertung zu operationalisieren. Stattdessen beschränken wir uns auf die Identifikation *offensichtlicher funktionale Unzulänglichkeiten der Tools*, also solcher Inkonsistenzen und Fehlfunktionen, die unabhängig von didaktischen Schulen als Defizite gelten müssen. Unser Prüfmaßstab lautet folglich: Erfüllt das System überhaupt basale Anforderungen an Konsistenz, Nachvollziehbarkeit und Gebrauchstauglichkeit, die für einen verantwortbaren Einsatz im Schulalltag unabdingbar sind? Zweitens: Eine technische Schwierigkeit ergibt sich aus der Architektur großer Sprachmodelle, auf denen auch die hier untersuchten Systeme basieren. Deren Ausgaben sind nicht-deterministisch: Identische Eingaben können zu unterschiedlichen Bewertungen, Punktzahlen oder Kommentaren führen. Sowohl methodisch als auch für die schulische Bewertungspraxis bedeutet dies ein fundamentales Reproduzierbarkeitsproblem. In unserer Untersuchung dokumentieren wir daher sämtliche Eingaben und Ausgaben und wiederholen Tests mehrfach, um die Streuung der Resultate sichtbar zu machen (siehe Materialanhang). Wie sich zeigen wird, liegt gerade in dieser Variabilität ein zentraler Einwand gegen die Verlässlichkeit automatisierter Korrektursysteme.

Dazu kommt erschwerend die Intransparenz der von den jeweiligen Herstellern verwendeten internen Systemprompts hinzu – also jener übergreifenden Instruktionen, mit denen die Sprachmodelle intern als Bewertungs- und Feedbacktools instruiert werden. Auch durch häufige, durch Anbieter oft nicht transparent kommunizierte Updates der Prompts oder Wechsel der verwendeten Sprachmodelle kommt es zu Schwierigkeiten der Reproduzierbarkeit des Verhaltens der betreffenden Systeme.

2.2 Aufgabenstellung, Testkorpus und Bewertungskriterien

Unsere Untersuchung orientiert sich methodisch an Mühlhoff & Henningsen (2025). Um Vergleichbarkeit zu gewährleisten, haben wir dieselbe Aufgabenstellung und dieselben zehn Versuchstexte (simulierte Schülerabgaben) verwendet wie in der früheren Studie zum Fobizz-Korrekturtool. Auf diese Weise lassen sich die Ergebnisse nicht nur

innerhalb der hier getesteten Tools, sondern auch im Vergleich zur Vorgängerstudie systematisch gegenüberstellen. Die verwendete Aufgabenstellung lautete:

> „Schreibe eine begründete Stellungnahme für oder gegen die Absenkung des Wahlalters auf 14 Jahre. Gehe dabei auf mindestens ein Argument für jede Seite ein und beziehe anschließend Position."

Als Testkorpus dienten dieselben zehn simulierten Schülertexte wie in der Vorgängerstudie. Die Texte decken ein breites Qualitätsspektrum ab (siehe Tabelle 1). Ziel dieser Bandbreite ist es, typische Bewertungssituationen im Schulalltag zu simulieren und zugleich Grenzfälle (z. B. inhaltliche Falschbehauptungen, Thema verfehlt, sehr kurze Texte) systematisch zu testen.

Tabelle 1: Übersicht Testkorpus. Die 10 Texte sind im →Materialanhang vollständig wiedergegeben.

Text	Umfang	Qualität	Text	Umfang	Qualität
1	250	gut; wenige Rechtschreibfehler	6	254	gut bis sehr gut; wenige Rechtschreibfehler
2	114	schlecht; nur Gegenargumente	7	41	nur drei Sätze; nahezu Arbeitsverweigerung
3	180	schlecht bis unsinnig; viele Zeichensetzungsfehler	8	69	kein Bezug zur Aufgabe, Thema verfehlt
4	280	sehr gut; wenige Rechtschreibfehler	9	170	mit ChatGPT generiert, kurz und prägnant
5	228	mittelmäßig; inhaltliche Falschbehauptungen	10	152	mittelmäßig; sehr knapp

Konfiguration von FelloFish

Im Interface von FelloFish kann die Lehrkraft die für eine Aufgabe relevanten Bewertungskriterien frei als Fließtext formulieren und eine Gewichtung angeben; zudem besteht die Möglichkeit, zusätzliches Material zur Aufgabenstellung bereitzustellen. Für unsere Tests haben wir uns an den fünf Bewertungskriterien orientiert, die in der integrierten Vorlage „Materialgestützt argumentieren" voreingestellt sind. Die Kriterien haben wir nur leicht angepasst, da unsere Aufgabenstellung kein Begleitmaterial enthielt (genauer Wortlaut siehe Materialanhang): (1) Einführung des Themas und Hinführung zur Fragestellung; (2) Klare Positionierung nach der Einleitung; (3) Mindestens ein ausgearbeitetes Argument für die eigene Position; (4) Nennung und Entkräftung eines Gegenarguments; (5) Schlüssiges Fazit ohne neue Argumente. Für jedes Kriterium haben wir die Voreinstellung „Normal" bei der Gewichtung beibehalten.

Konfiguration von Edaira

Bei Edaira werden zu einer eingegebenen Aufgabenstellung automatisch Bewertungskriterien generiert. Um uns auch hier an der Standard-Benutzungsweise zu orientieren, haben wir die vom System vorgeschlagenen Kriterien für unsere Testreihe un-

verändert übernommen. Diese umfassten (genauer Wortlaut siehe Materialanhang): (1) Einführung in Thema und Diskussionslage; (2) Mindestens ein umfassend erläutertes Pro-Argument; (3) Mindestens ein umfassend erläutertes Contra-Argument; (4) Klare Positionierung; (5) Logische Begründung der eigenen Position; (6) Auseinandersetzung mit Gegenargumenten; (7) Prägnantes Fazit mit Ausblick oder abschließender Bewertung. Zusätzlich flossen eine automatisch gewichtete Bewertung der sprachlichen Leistung (20 %) sowie eine integrierte Rechtschreibprüfung in die Gesamtnote ein. Diese Voreinstellungen haben wir nicht verändert.

2.3 Testdesign

Den Schwerpunkt der vorliegenden Untersuchung bildet ein Konsistenztest (Testreihe A), der für beide Tools die Stabilität von Bewertung und Feedback bei mehrfacher Bewertung identischer, unveränderter Abgaben prüft. Analog zur Vorgängerstudie geht es dabei um die Frage, ob numerische Bewertungen und qualitative Rückmeldungen reproduzierbar sind oder zufallsbedingten Schwankungen unterliegen. Ergänzend dazu haben wir zwei weitere Testreihen (B und C) durchgeführt, die den Umgang der Tools mit ihren eigenen Rückmeldungen und Verbesserungsvorschlägen untersuchen. Diese zusätzlichen Reihen zielen darauf ab, die didaktische Kohärenz der Feedback-Logik zu prüfen: Führt die Umsetzung von Rückmeldungen zu konsistent besseren Bewertungen?

Testreihe A: Konsistenztest

In Testreihe A wurde mit jedem der beiden Tools jeder der zehn Texte aus dem Testkorpus jeweils zehnmal unabhängig eingereicht und automatisiert bewertet. Ziel war es, die Streuung der numerischen Gesamtbewertung sowie die Variabilität des qualitativen Feedbacks bei identischem Input zu erfassen. **FelloFish** gibt eine auf volle Zehner gerundete Prozentbewertung für jedes der eingestellten Bewertungskriterien aus (Lehreransicht), die von uns in ein Tabellenkalkulationsprogramm übertragen wurden. Dort haben wir eine Gesamtbewertung (in Prozentpunkten) als arithmetisches Mittel dieser Einzelwertungen abgeleitet. Die Ausgabe von **Edaira** umfasst verschiedene Werte: Für jedes der eingestellten Kriterien wird ein Punktewert (hier: zwischen 0 und 5 oder 0 und 10, je nach Gewichtung) ermittelt, diese Punkte werden zu einer Gesamtpunktzahl für die Aufgabe summiert („Aufgabenpunkte“, max. 50 Punkte in unserem Fall). Die sprachliche Leistung bewertet das Tool in der Lehreransicht anhand von sechs schulnoten („sehr gut“, „gut“, …) für die Kategorien „Textsortenpassung und Textaufbau“, „Fachsprache“, „Ausdruck und Stil“ und „Standardsprachliche Normen“. Außerdem produziert es eine Gesamtpunktzahl zwischen 0 und 12 für die sprachliche Leistung („Sprachpunkte“, es ist nicht dokumentiert, wie das Tool die Sprachpunkte aus den vier Teilnoten für sprachliche Leistung ermittelt). Die Aufgabenpunkte und Sprachpunkte zusammen ergeben eine Gesamtpunktzahl (max. 62 Punkte in unserem Fall), die vom Tool zusätzlich in einen Prozentwert (0–100 %) übersetzt wird. Alle diese Werte wurden für jeden der Bewertungsdurchläufe in einer Tabellenkalkulation erfasst, um sie einer statistischen Auswertung zugänglich zu machen.

Testreihe B: Iterative Verbesserungen

In Testreihe B haben wir in Anlehnung an die iterative Reihenuntersuchung bei Mühlhoff & Henningsen (2025) untersucht, ob die sukzessive Umsetzung der von den Tools generierten Verbesserungsvorschläge – wie zu erwarten wäre – auch zu einer konsistenten Steigerung der Bewertung führt. Für die Testreihe wurden drei Texte ausgewählt, die unterschiedliche Qualitätsstufen repräsentieren: Text 1 (gute Qualität), Text 10 (mittlere Qualität), Text 7 (extrem kurzer „Minimalaufwand"-Text).

Für jeden dieser drei Texte wurden mit jedem Tool fünf Iterationsschritte durchgeführt: Der Text wurde eingespeist und eine Bewertung generiert. Verbesserungsvorschläge wurden eingearbeitet und das Resultat wurde in einer neuen Bewertungssitzung als neuer Text wiederum zur Bewertung eingespeist. Der Vorgang wurde so oft wiederholt, dass insgesamt fünf Iterationsstufen vollzogen wurden. Analysiert wurde, ob die Bewertung durch das iterative Einarbeiten der Verbesserungsvorschläge des jeweiligen Tools monoton steigt, oszilliert oder sich verschlechtert.

Da **FelloFish** zwischen der Einreichung als Entwurf und als Abgabe unterscheidet, war das Testverfahren hier wie folgt organisiert: Jeder Text wurde zunächst als „1. Entwurf" eingereicht. Anschließend wurde das vom Tool für die Lernenden generierte Feedback in den Text eingearbeitet, und die überarbeitete Version wurde als „Abgabe" eingereicht und bewertet (Lehreransicht). Für die nächste Iteration wurde dann über einen neuen Schülerzugang die zuletzt als Abgabe eingereichte Textversion wieder als „1. Entwurf" eingespeist und der Vorgang wurde wiederholt: Feedback auf den Entwurf wurde wiederum eingearbeitet, das Resultat als „Abgabe" eingereicht, usf. Änderungen am Text erfolgten also ausschließlich *innerhalb* eines Schülerzugangs zwischen „Entwurf" und „Abgabe". Dieses Detail ist entscheidend, da bei FelloFish der zuvor eingereichte Entwurf in die Bewertung der finalen Abgabe einfließt. Um diesen systeminternen Zusammenhang methodisch zu kontrollieren, wurde jede Iteration in einem neuen Schülerzugang gestartet. (Seit unseren Tests im März 2025 wurde der Funktionsumfang des Tools offenbar erweitert, so dass nun innerhalb eines Schülerzugangs eine beliebige Anzahl Anpassungen und Neubewertungen vor der finalen Abgabe möglich ist, vgl. FelloFish 2025).

Testreihe C: Copy-Paste-Übernahme des Feedbacks (FelloFish)

Testreihe C untersucht spezifisch für **FelloFish** die Frage, welche Art der Überarbeitung zwischen „Entwurf" und „Abgabe" hinsichtlich der automatisch generierten Bewertung besonders positiv ins Gewicht fällt. Insbesondere wird getestet, ob das System systematisch eine möglichst wortnahe Umsetzung seiner eigenen Formulierungsvorschläge, die in den Rückmeldungen enthalten sind, belohnt. Hierfür wurden für jeden der Texte aus dem Testkorpus drei unterschiedliche Überarbeitungsstrategien getestet (Text 8 wurde ausgenommen, da er konsistent mit 0 % bewertet wurde):

Strategie 1, wortgenaue Übernahme: Vom Tool vorgeschlagene Beispielsätze wurden möglichst unverändert in den Text übernommen. Falls keine expliziten Beispielsätze generiert wurden, wurde der Versuch neu gestartet. Bei Mangel an Beispielsätzen trotz

mehrfachem Neustart wurde das Feedback möglichst eng am Sprachstil des Tools orientiert eingebaut.

Strategie 2, sinngemäße Umsetzung mit abweichendem Wortlaut: Das Feedback wurde sinngemäß aufgegriffen, jedoch mit klar verändertem sprachlichem Ausdruck umgesetzt.

Strategie 3, unabhängige Überarbeitung: Das Feedback wurde ignoriert; stattdessen wurde der Text nach eigener inhaltlicher Einschätzung verbessert und erneut eingereicht.

Durch aggregierten Vergleich der aus der jeweiligen Verbesserungsstrategie resultierenden Änderung der vorgeschlagenen Bewertung lässt sich prüfen, ob das Tool primär semantische (also inhaltliche) Verbesserungen honoriert, oder ob es strukturell dazu tendiert, sein eigenes sprachliches Muster (genaue Formulierung) zu privilegieren.

3 Ergebnisse

3.1 Testreihe A: Konsistenztest

Im Konsistenztest wurde jeder der zehn Texte zehnmal unverändert durch das jeweilige Korrekturtool bewertet, um die Streuung der Gesamtbewertung sowie der einzelnen Bewertungskriterien bei identischem Input zu erfassen.

Für **FelloFish** zeigen die Ergebnisse erhebliche Schwankungen (Abb. 1). Je nach Text variiert die Gesamtbewertung zwischen minimaler und maximaler Prozentbewertung bei Wiederholung des Bewertungsvorgangs für denselben Text um 4 bis 24 Prozentpunkte (%-pkt). Die stärkste beobachtete Schwankung (24 %-pkt) betrifft Text 6, einen nach unserer Einschätzung sehr guten Text mit nur wenigen kleineren Fehlern. Die geringste Schwankung (4 %-pkt) zeigte der KI-generierte Text 9. Einzig Text 8, eine eindeutige Themenverfehlung ohne inhaltlichem Bezug zur Aufgabenstellung (Schilderung eines Freibadbesuchs) wurde konsistent mit 0 % bewertet und weist entsprechend keine Streuung auf. Die durchschnittliche Streubreite über alle von Menschen geschriebenen Texte hinweg liegt bei 12 %-pkt.

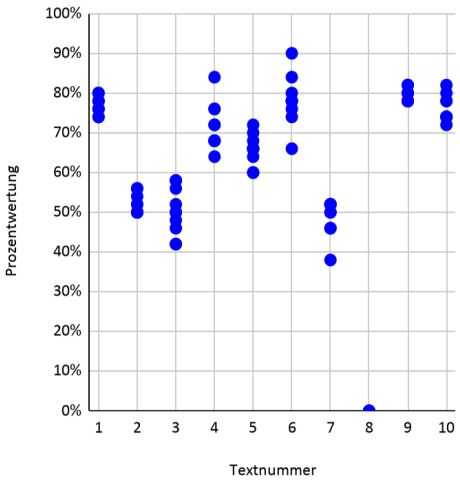

Abbildung 1: FelloFish, Volatilität der ermittelten Gesamtbewertungen für die einzelnen Texte des Testkorpus bei 10-facher unabhängiger Wiederholungen des Bewertungsvorgangs für jeden Text

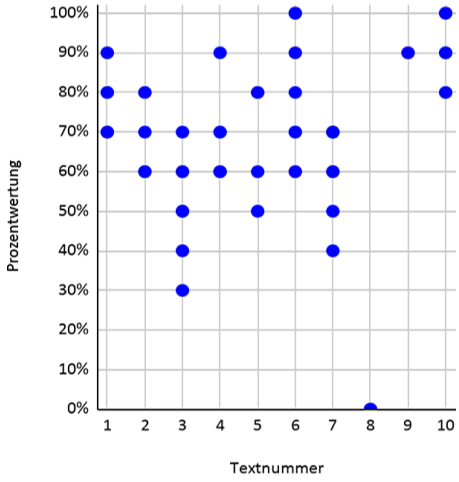

Abbildung 2: FelloFish, Volatilität der Bewertung für ein einzelnes Bewertungskriterium (Kriterium 2) bei 10-facher unabhängiger Wiederholung des Bewertungsvorgangs für jeden Text.

Noch deutlicher wird die Volatilität bei der Betrachtung der Teilnoten für einzelne Bewertungskriterien. Hier treten Abweichungen von bis zu 40 %-pkt innerhalb desselben Texts auf (z. B. Text 3, Kriterium 2 [Abb. 2]; Text 4, Kriterium 5 [→Materialanhang]; Text 6, Kriterium 2 [→Materialanhang]). Diese Streuung ist kein Ausnahmephänomen, sondern systematisch beobachtbar. Während Extremfälle stabil eingeordnet werden, zeigt sich insbesondere bei „normalen" Schülertexten eine ausgeprägte Bewertungsvolatilität. Gerade typische, solide oder leicht fehlerhafte Arbeiten, also jene, die den Großteil realer Abgaben ausmachen, werden inkonsistent bewertet. Auffällig ist zudem, dass der KI-generierte Text 9 unter allen Abgaben abseits von Text 8 am stabilsten beurteilt wird.

Für **Edaira** zeigt sich ein ähnliches Bild (Abb. 3). Die maximale Schwankung der Gesamtbewertung innerhalb eines identischen Textes beträgt bis zu 35 %-pkt. Die durchschnittliche Streubreite über alle von Menschen geschriebenen Texte hinweg liegt bei 19 %-pkt. Im Vergleich zu FelloFish ist die maximale Streubreite somit geringfügig höher; zugleich treten andere Inkonsistenzen auf: Besonders auffällig ist der Umgang mit Text 8 (eindeutige Themenverfehlung). Anders als bei FelloFish wird dieser Text nicht konsistent mit 0 % bewertet. Die Gesamtbewertung schwankt hier zwischen 6 % und 19 %, obwohl sämtliche inhaltlichen Kriterien (mit Ausnahme der Sprachpunkte) durchgehend mit 0 bewertet wurden. Die Variation entsteht somit aus der Sprachbewertung.

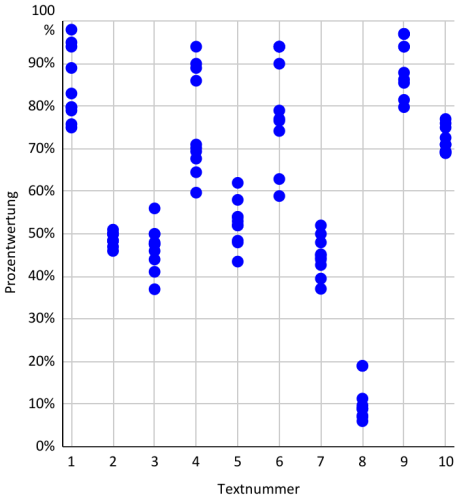

Abbildung 3: Edaira, Volatilität der ermittelten Gesamtbewertungen für die einzelnen Texte des Testkorpus bei 10-facher unabhängiger Wiederholungen des Bewertungsvorgangs für jeden Text.

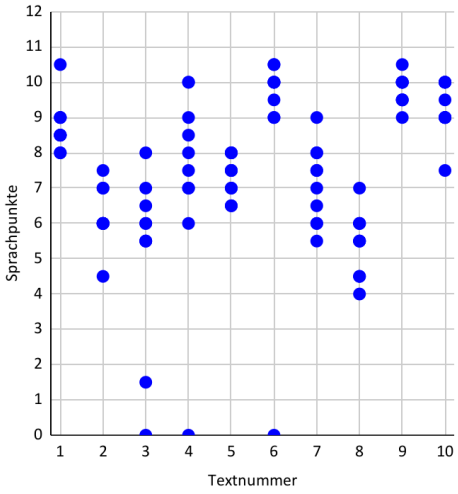

Abbildung 4: Edaira produziert für jede Abgabe eine Bewertung der sprachlichen Leistung zwischen 0 und 12 Punkten. Das Diagramm zeigt die Volatilität für die einzelnen Texte des Testkorpus bei 10-facher unabhängiger Wiederholungen des Bewertungsvorgangs für jeden Text.

Der konsistenteste Text ist bei Edaira nicht der KI-generierte Text 9 (Schwankung: 17 %-pkt), sondern Text 2, eine sehr kurze, unterkomplexe Minimalabgabe, mit einer Varianz von lediglich 5 %-pkt. Wie bei FelloFish schwanken auch bei Edaira die Einzelkriterien erheblich. So variiert beispielsweise die Sprachbewertung bei Text 3 zwischen 0 und 8 (von maximal 12) Punkten, bei Text 4 zwischen 0 und 10 Punkten (Abb. 4). Mehrere weitere Kriterien zeigen Abweichungen von bis zu 4 Punkten (siehe →Materialanhang).

Der KI-generierte Text 9 erzielte in beiden Systemen die höchste Durchschnittsbewertung und eine relativ geringe Schwankungsbreite der Bewertungen: Bei Edaira erreicht er im Durchschnitt 90 % bei einer Schwankung von 17 %-pkt, bei FelloFish 79 % bei einer Schwankung von 4 %-pkt. Die bestbewerteten menschengeschriebenen Texte liegen nach dem Durchschnitt der 10 unabhängigen Bewertungen jeweils darunter, bei Edaira ist es Text 1 mit durchschnittlich 85 % und bei FelloFish Text 6 mit durchschnittlich 78 %. Damit bestätigt sich analog zur Vorgängerstudie der strukturelle Befund, dass die Tools Texte, die strukturell und stilistisch nahe an LLM-generierten Mustern liegen, konsistenter einordnet und besser bewerten als menschliche Texte mit individueller Varianz.

3.2 Testreihe B: Iterative Verbesserungen

Bei der iterativen Anwendung der Bewertungstools nach dem in Abschnitt 2 als Testreihe B beschriebenen Verfahren ergibt sich der in Abb. 5 und 6 gezeigte Bewertungsverlauf. Im Fall **FelloFish** fällt auf, dass es zu einem Unterschied zwischen der Bewer-

tung eines Textes als „Entwurf“ und als „Abgabe“ kommt. Derselbe Text erhält als finale Abgabe regelmäßig eine deutlich höhere Bewertung denn als Entwurf. Dass dieser Effekt systematisch auftritt, lässt sich nicht durch die in Testreihe A nachgewiesene zufallsbedingte Streuung erklären. Vielmehr deutet sich ein struktureller positiver Bias des Tools zugunsten von Abgaben an: Sobald zwischen Entwurf und Abgabe eine (minimale) Veränderung vorgenommen wurde, steigt die Bewertung an. Wird derselbe Text anschließend unverändert über einen neuen Schülerzugang erneut als erster Entwurf eingereicht, fällt die Bewertung wieder niedriger aus. Dadurch entsteht über mehrere Iterationen hinweg ein charakteristisches „Sägezahn“-Muster (vgl. Abb. 5).

Bei Text 7 handelt es sich um einen besonders drastischen Fall. Der ursprüngliche Text besteht nur aus drei kurzen Sätzen und bewegt sich inhaltlich an der Grenze zur Arbeitsverweigerung. Das Feedback des Tools enthielt jedoch mehrere konkrete Beispielsätze, die zur Verbesserung eingefügt werden sollten. Durch das bloße Abtippen dieser Vorschläge stieg die Bewertung bereits in der ersten Iteration stark an. Nach nur einer Überarbeitungsrunde enthält der Text damit anteilig mehr vom Tool generierte Textbausteine als eigenständig formulierte Passagen.

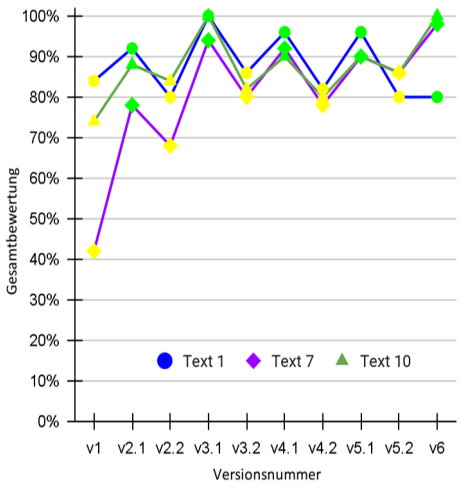

Abbildung 5: FelloFish, Entwicklung der Gesamtbewertung bei iterativer Einarbeitung der Verbesserungsvorschläge über 5 Iteraktionsstufen. Der Test wurde für die Texte (1, 7, 10) durchgeführt. Gelbe Punkte = Bewertung bei Einreichung als „Entwurf“, grüne Punkte = Bewertung bei Einreichung als „Abgabe“. Innerhalb einer Versionsnummer, also etwa von v2.1 zu v2.2, wurden *keine* Änderungen am Text vorgenommen, lediglich der vormals als Abgabe eingereichte Text in einer neuen Schülersitzung als Entwurf eingespeist.

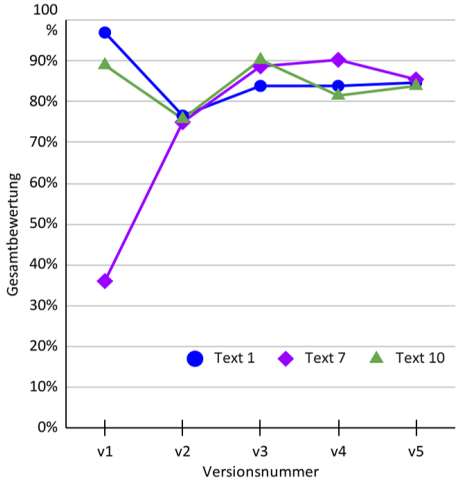

Abbildung 6: Edaira, Entwicklung der Gesamtbewertung bei iterativer Einarbeitung der Verbesserungsvorschläge über 5 Iterationsstufen. Der Test wurde für drei Texte (1, 7, 10) durchgeführt.

Auch bei **Edaira** zeigen sich diese Effekte (Abb. 6): Die initial als sehr gut bewerteten Texte 1 und 10 verschlechtern sich langfristig durch die Einarbeitung der Verbesserungsvorschläge. Die Minimalabgabe, Text 7, verbessert sich zunächst deutlich, bis auch hier ein „schwankendes Plateau" erreicht wird, von dem aus eine Maximalbewertung durch Einarbeitung der Vorschläge nicht erreichbar ist.

3.3 Testreihe C: Copy-Paste-Übernahme des Feedbacks (FelloFish)

In vielen Fällen enthält das von FelloFish generierte Feedback konkrete Beispielsätze, die zur Verbesserung des Textes vorgeschlagen werden. Die wörtlichen Verbesserungsvorschläge lassen sich von Lernenden ohne weitere Anpassung direkt in den eigenen Text übernehmen. Testreihe C untersucht deshalb, in welchem Maße das Tool unterschiedliche Überarbeitungsstrategien honoriert: die wortwörtliche Übernahme solcher Beispielsätze (Strategie 1); eine sinngemäße Umsetzung des Feedbacks mit verändertem Wording (Strategie 2); eine eigenständige Überarbeitung ohne Bezug auf das Feedback (Strategie 3).

Die Ergebnisse zeigen eine deutliche Präferenz des Tools für die wortwörtliche Übernahme seiner eigenen Formulierungen (siehe Abb. 7). Im Durchschnitt werden Texte der Strategie 1 („wörtliche Übernahme") 7,6 %-pkt höher bewertet als Texte, in denen die vorgeschlagenen Beispielsätze nur sinngemäß übernommen wurden (Strategie 2). Im Extremfall beträgt die Differenz 18 %-pkt (Text 1). In keinem der untersuchten Fälle wurde eine sinngemäße Überarbeitung besser bewertet als das Copy-Pasting der Vorschläge; der bestmögliche Fall ist eine identische Bewertung (Text 3).

Noch deutlicher wird dieser Effekt beim Vergleich mit den eigenständigen inhaltlichen Überarbeitungen (Strategie 3). Im Durchschnitt werden diese weitere 6 %-pkt niedriger bewertet als die sinngemäße Umsetzung der Verbesserungen (Strategie 2). Zwischen der wortwörtlichen Übernahme des Feedbacks (Strategie 1) und der freien Überarbeitung (Strategie 3) ergibt sich somit im Mittel eine Bewertungsdifferenz von 13,6 %-pkt. Im Extremfall (Text 2) wurde die frei überarbeitete Version sogar 38 %-pkt niedriger bewertet als die Version mit wörtlich übernommenem Feedback.

Die Ergebnisse der Strategie 3 sind dabei methodisch mit größerer Vorsicht zu interpretieren, da die eigenständigen Verbesserungen naturgemäß nicht nach vollständig standardisierbaren Kriterien erfolgen können. Der Vergleich zwischen Strategie 1 und 2 ist hingegen eindeutig und zeigt bereits für sich genommen einen strukturellen Effekt: Das Tool bewertet die Übernahme seiner eigenen sprachlichen Formulierungen systematisch höher als inhaltlich gleichwertige, aber anders formulierte Verbesserungen.

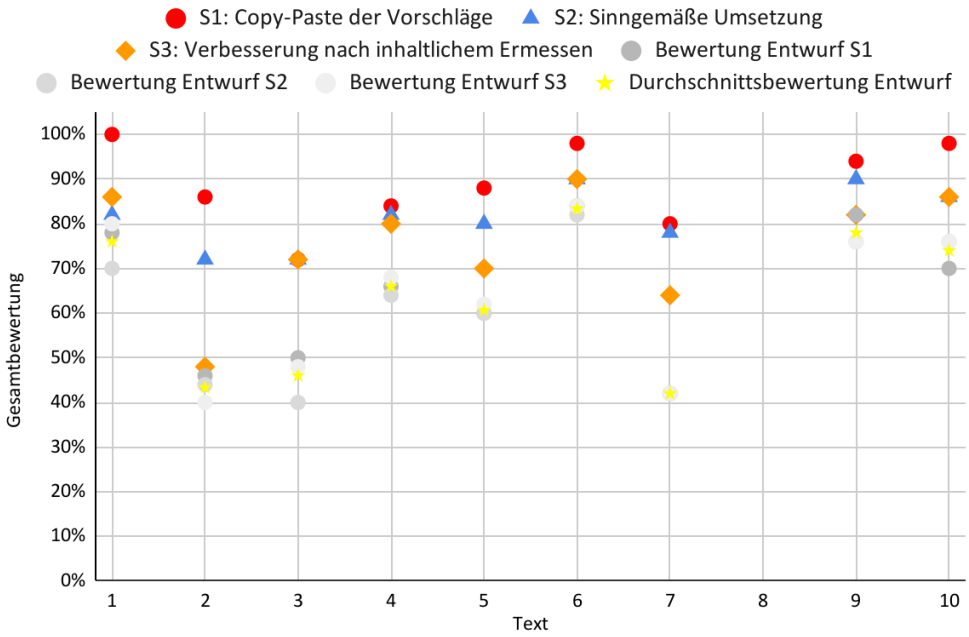

Abbildung 7: FelloFish, Bewertung nach der Einarbeitung von Verbesserungsvorschlägen nach den Strategien S1–S3. Als Baseline wird in grau/gelb die Bewertung des Texts im Entwurfsstadium mit abgebildet – da die drei Verbesserungsstrategien in drei unabhängigen Schülersitzungen getestet werden mussten, wurde jeder Text unverändert dreimal als Entwurf eingereicht. Text 8 (Themenverfehlung) wurde ausgeschlossen.

Weitere Beobachtungen: Neben den Testreihen A–C traten weitere Auffälligkeiten im Verhalten der getesteten Systeme auf, die im →Materialanhang dokumentiert sind.

4 Diskussion

Ziel der vorliegenden Untersuchung ist die Analyse der funktionalen Gebrauchstauglichkeit von KI-Tools für automatisierte Bewertung und Rückmeldung zu textbasierten Aufgaben. Es wurde geprüft, ob die getesteten Tools FelloFish und Edaira grundlegende Anforderungen an Konsistenz und Nachvollziehbarkeit erfüllen, die für einen verantwortbaren Einsatz im schulischen Kontext – unabhängig einer detaillierten fachdidaktischen Bewertung – erforderlich wären. Die Ergebnisse lassen sich auf drei zentrale Befunde verdichten.

Bewertungsvolatilität als strukturelles Problem: Der Konsistenztest (Testreihe A) zeigt, dass identische Texte bei wiederholter Durchführung des automatischen Bewertungsvorgangs deutlich unterschiedliche Bewertungen erhalten. Diese Volatilität bleibt bei der realen Verwendung der Tools unsichtbar, weil Nutzende normalerweise nicht auf die Idee kommen würden, ein und dieselbe Abgabe mehrfach zu bewerten. Damit fehlt eine zentrale Voraussetzung fairer Leistungsbewertung: die Reproduzierbarkeit

des Urteils. Bei diesem Befund handelt es sich nicht um ein spezifisches Problem einzelner Tools, sondern um eine systemische Eigenschaft von Bewertungssystemen auf Basis großer Sprachmodelle.

Problematisch ist dabei auch, dass gerade Texte *mittlerer* Qualität besonders starken Bewertungsschwankungen unterliegen. Extremfälle (Themenverfehlungen, KI-Texte) werden dagegen vergleichsweise stabil eingeordnet. Pädagogisch ist dies besonders heikel, da der Großteil schulischer Bewertungssituationen weder perfekte noch völlig misslungene Arbeiten betrifft, sondern Texte mittlerer Qualität, bei denen differenzierte Urteile erforderlich sind.

Texte, die stilistisch und strukturell stark an typische Muster generativer Sprachmodelle erinnern, werden von den Tools tendenziell konsistenter und besser bewertet als individuell formulierte Schülertexte. Die Systeme orientieren sich implizit an den sprachlichen Mustern, die KI-Systemen selbst produziert würden. Eine Bestbewertung ist fast nur erreichbar, wenn man KI für die Hausaufgaben nutzt.

Inkonsistente Feedback-Logik: Neben einer Bewertung generieren die getesteten Tools Feedback und Verbesserungsvorschläge. Pädagogisch wäre zu erwarten, dass die Umsetzung dieser Rückmeldungen zu einer besseren Bewertung führt. In unseren Iterationstests (Testreihe B) zeigte sich jedoch kein solcher monotoner Zusammenhang. Die Bewertung schwankt vielmehr und kann sich sogar verschlechtern. Feedback verliert so seinen didaktischen Wert.

Ein besonders aufschlussreicher Befund ergibt sich aus der Untersuchung verschiedener Strategien, das Feedback der KI-Tools einzuarbeiten (Testreihe C). Wenn Lernende die vom Tool vorgeschlagenen Beispielsätze *wortwörtlich* in ihren Text übernehmen, führt dies systematisch zu besseren Bewertungen als eine *eigenständige* oder nur *sinngemäße* Umsetzung der Rückmeldungen. Das System honoriert damit nicht primär die inhaltliche Verbesserung eines Textes, sondern die Übernahme seiner eigenen sprachlichen Muster.

Technische Grenzen der zugrundeliegenden Systeme: Die beobachteten Effekte lassen sich weitgehend auf grundlegende Eigenschaften großer Sprachmodelle zurückführen. Diese Systeme erzeugen ihre Ausgaben anhand stochastischer Algorithmen. Sie verfügen nicht über ein internes Bewertungsmodell im didaktischen Sinne, sondern generieren Bewertungen durch Mustererkennung im Text. Viele der beobachteten Probleme sind daher nicht als Softwarefehler zu verstehen, sondern als strukturelle Grenzen der zugrundeliegenden Technologie großer Sprachmodelle.

Konsequenzen für den schulischen Einsatz: Aus diesen Befunden ergibt sich eine klare Schlussfolgerung: KI-gestützte Korrektur- und Feedbacksysteme können derzeit nicht als eigenständige Instanzen für Leistungsbewertung eingesetzt werden. Allenfalls eignen sich solche Systeme als unterstützende Werkzeuge für Lehrkräfte, etwa zur Generierung erster Feedbackentwürfe. Die Vertretbarkeit dieses Einsatzes hängt wesentlich davon ab, ob die Lehrkraft um die inhärenten Mängel (insbesondere Volati-

lität) der Tools weiß und die Möglichkeit hat, sich durch mehrfache Wiederholungen des Bewertungsvorgangs in jedem Einzelfall einen Überblick darüber zu verschaffen.

Gleichzeitig zeigen die Ergebnisse, dass der verbreitete Diskurs über KI als Lösung für zu hohe Arbeitsbelastung im Bildungssystem mit Vorsicht zu betrachten ist. Aktuelle KI-Systeme können diese sozial und politisch gewachsenen Probleme nicht lösen.

Materialanhang mit ergänzenden Daten, Messungen und Informationen

https://go.uos.de/ECTAI-material-automatisierte-korrektur-seminar-2026

Literaturverzeichnis

BMBFSFJ. (o. J.). *Digitalpakt 2.0*. Das Bundesministerium für Bildung, Familie, Senioren, Frauen und Jugend. Abgerufen 16. März 2026, von https://www.digitalpaktschule.de/de/digitalpakt-2-0-1874.html

Brand, A. (2024). SWK-Bildungsforscher empfehlen Einsatz von ChatGPT an Schulen. *Deutsches Schulportal*. Abgerufen 16. März 2026, von https://deutsches-schulportal.de/unterricht/swk-bildungsforscher-empfehlen-einsatz-von-chatgpt-an-schulen/

FelloFish. (2025). *FelloFish Feature Friday: Neue Schüler:innen Ansicht–YouTube*. Abgerufen 16. März 2026, von https://youtu.be/oKWhZ2n2-Y0

Fobizz. (o. J. a). Fobizz Landeslizenz für Mecklenburg-Vorpommern. *fobizz*. Abgerufen 16. März 2026, von https://fobizz.com/de/lehrerfortbildung-mecklenburg-vorpommern/

Fobizz. (o. J. b). Lehrerfortbildung Rheinland-Pfalz. *fobizz*. Abgerufen 16. März 2026, von https://fobizz.com/de/lehrerfortbildung-rheinland-pfalz/

Fobizz. (o. J. c). Lehrerfortbildung Sachsen 6 Monate Fortbildungs-Flatrate für Lehrkräfte aus Sachsen. *fobizz*. Abgerufen 16. März 2026, von http://fobizz.w4p.tech/lehrerfortbildung-sachsen/

KMK. (2025). „Dimension Digitalisierung – Für alle mehr drin“: Fachtagung zeigt Wege für Schule in der Kultur der Digitalität. Abgerufen 16. März 2026, von https://www.kmk.org/aktuelles/pressearchiv/mitteilung/dimension-digitalisierung-fuer-alle-mehr-drin-fachtagung-zeigt-wege-fuer-schule-in-der-kultur-der-digitalitaet.html

Mühlhoff, R., & Henningsen, M. (2025). *Chatbots im Schulunterricht: Wir testen das Fobizz-Tool zur automatischen Bewertung von Hausaufgaben* (arXiv:2412.06651). arXiv. https://doi.org/10.48550/arXiv.2412.06651

NDR. (2026). *Unterricht mit KI: Chatbot "Telli" an allen Schulen einsetzbar*. Abgerufen 16. März 2026, von https://www.ndr.de/nachrichten/niedersachsen/unterricht-mit-ki-chatbot-telli-an-allen-schulen-einsetzbar,telli-102.html

News4Teachers. (2025). „Feedback jederzeit und zu jedem Schüler“–Wie Künstliche Intelligenz Lehrkräften konkret beim Unterrichten helfen kann. https://www.news4teachers.de/2025/02/feedback-jederzeit-und-zu-jedem-schueler-wie-kuenstliche-intelligenz-lehrkraeften-konkret-beim-unterrichten-helfen-kann/

Robert Bosch Stiftung. (2025). Deutsches Schulbarometer.

Schleswig-Holstein.de. (2025). *Lernen mit KI-Chatbot „telli“*. Abgerufen von https://www.schleswig-holstein.de/DE/landesregierung/ministerien-behoerden/III/_startseite/Artikel_2025/11_November_2025/20251110_telli

SWK. (2024). Large Language Models und ihre Potenziale im Bildungssystem. Impulspapier der Ständigen Wissenschaftlichen Komission der Kultusministerkonferenz.
Telli.Schule. (o. J.). *Der KI-Chatbot für die Schule*. Abgerufen 16. März 2026, von https://telli.schule/

Autoren

Prof. Dr. Rainer Mühlhoff
Professor für Ethik und kritische Theorien der künstlichen Intelligenz, Universität Osnabrück.
rainer.muehlhoff@uni-osnabrueck.de

Sean Quägwer
Wissenschaftlicher Mitarbeiter, Arbeitsgruppe Ethik und kritische Theorien der künstlichen Intelligenz, Universität Osnabrück
sequaegwer@uni-osnabrueck.de