Rebuttal zu unseren Studien über KI-Korrektur- und Feedbacktools für den Schulunterricht

Stand: 1. September 2026

Rainer Mühlhoff, Professor für Ethik der künstlichen Intelligenz, Universität Osnabrück

Worum es in unseren Studien ging

In zwei Studien (Mühlhoff und Quägwer, 2026; Mühlhoff und Henningsen 2024) haben wir drei kommerziell angebotene KI-Werkzeuge zur Automatisierung von Feedback und Notengebung in der Schule untersucht: 2024 die damalige „KI-Korrekturhilfe“ von fobizz, 2025/26 die Tools von FelloFish und Edaira.

Die Ergebnisse waren deutlich: Identische Texte erhielten bei wiederholter Eingabe und unveränderten Kriterien teils erheblich unterschiedliche Bewertungen – bei FelloFish betrug die größte beobachtete Spannweite 24 Prozentpunkte, bei Edaira 35 Prozentpunkte. Das damalige Fobizz-Tool übersah in sämtlichen fünf Durchläufen eine einfache Falschbehauptung und bewertete eine vollständige Themenverfehlung je nach Durchlauf über mehrere Schulnoten hinweg unterschiedlich. Edaira vergab für einen Manipulationstext, der keine Argumentation enthielt, 92,7 Prozent. Die Tools honorierten die Umsetzung ihrer eigenen Verbesserungsvorschläge nicht verlässlich; Bewertungen konnten trotz der Befolgung des Feedbacks sinken oder zwischen Formulierungen oszillieren. Auffällig war außerdem, dass der einzige KI-generierte Text in allen drei Produkten besonders hoch und relativ stabil bewertet wurde.

In der – mitunter emotionalen – Rezeption dieser Studien zeigen sich einige Missverständnisse: Das Ziel dieser Studien war *kein* Leistungsvergleich zwischen Mensch und Maschine und *keine* repräsentative statistische Auswertung darüber, wie häufig bestimmte Fehler der automatisierten Bewertung in deutschen Klassenzimmern vorkommen. Wir führten vielmehr gezielte, explorative Produkttests durch: Erfüllen die Tools *funktionale Mindestanforderungen*, und lassen sich konkrete Schwachstellen unter kontrollierten Bedingungen reproduzieren?

Dafür verwendeten wir ein bewusst konstruiertes Korpus von Texten, an denen wir die Tools testeten. Es enthielt zehn Texte, die gezielt unterschiedliche Prüf- und Grenzfälle abdecken sollten: unter anderem einen nach unserem Dafürhalten sehr guten Text, kurze und schwache Texte, eine Themenverfehlung, einen Text mit Falschbehauptungen, eine Arbeitsverweigerung und einen KI-generierten Text. Unser Test war sodann denkbar einfach: Jeder dieser Texte wurde mit jedem der Tools *mehrfach* bei gleichbleibenden Kriterien bewertet – wir wollten wissen, ob dabei immer die gleiche Bewertung herauskommt, und haben sehr schnell beobachtet, dass dies nicht der Fall war. Außerdem prüften wir, wie die

Systeme auf die Umsetzung ihrer eigenen Verbesserungsvorschläge und auf bestimmte Manipulationsversuche reagierten. Die Aufgabenstellung, das Textkorpus und die Ausgaben der Tools sind in umfangreichen Materialanhängen offengelegt.

Diese Zielsetzung unserer Studien ist für die Interpretation ihrer Ergebnisse entscheidend. Ein gezielter Schwachstellentest beantwortet keine Fragen zur didaktischen oder psychologischen Wirksamkeit von KI im Unterricht. Er dokumentiert aber sehr wohl, dass ein bestimmtes Produkt unter bestimmten Bedingungen einen bestimmten Fehler gemacht hat. Unser kleiner konstruierter Testkorpus mit nur 10 Texten begrenzt Aussagen über die statistische *Häufigkeit* der von uns beobachteten Fehlfunktionen, nicht jedoch über ihre *Existenz*.

Im Folgenden beantworten wir einige Einwände, die gegen unsere Studien in der Podcastfolge „Bewertungs- und Feedbacktools auf dem Prüfstand“ (Kompass KI, S2E1, 28. August 2026, Hendrik Haverkamp – FelloFish-Gründer – und Benedikt Wisniewski) sowie auf LinkedIn und BlueSky vorgebracht wurden. Dabei beschränken wir uns im Sinne einer produktiven wissenschaftlichen Diskussion auf die sachlichen Argumente.

Zu einzelnen Kritikpunkten

1. Wurden Feedback, Bewertung und Benotung vermischt?

Der Einwand: FelloFish sei ein Feedbacksystem, Edaira eine Benotungsassistenz. Unsere gemeinsame Untersuchung werfe diese unterschiedlichen Funktionen in einen Topf.

Unsere Antwort: Der Einwand **trifft nicht zu.** Die Studie von 2026 unterscheidet die Produkte ausdrücklich. FelloFish wird als „Schreibbegleiter“ beschrieben, Edaira als „Benotungsassistenz“. Zugleich produziert FelloFish in seiner Ausgabe aber nicht nur freien Feedbacktext: In der Lehrkräfteansicht gibt das Tool für jedes Kriterium quantifizierte Werte aus und visualisiert auch für Lernende eine Bewertung der finalen Abgabe. Edaira gibt numerische Kriterienpunkte, Sprachpunkte und eine Gesamtbewertung aus. Diese quantifizierten Bewertungsfunktionen durften und mussten Gegenstand eines Tests ihrer Wiederholbarkeit sein.

Wir haben transparent gemacht, dass der für FelloFish ausgewiesene Gesamtwert von uns als arithmetisches Mittel der einzelnen Kriterienwerte berechnet wurde. Wir behaupten nicht, FelloFish vergebe selbst eine Schulnote. Ebenso bezeichnen wir Edaira nicht pauschal als „Notentool“, sondern beschreiben seine Funktion und die von ihm ausgegebenen Teil- und Gesamtbewertungen konkret. Die gemeinsame Klammer der Untersuchung ist daher eine klar benannte Funktion der Tools: Beide Systeme bewerten Schülertexte anhand von Kriterien numerisch und verbinden diese Bewertung mit Rückmeldungen. Für das 2024 getestete Fobizz-Tool gilt dasselbe.

2. Fehlte eine Theorie der Feedbackqualität?

Der Einwand: Ohne Definition von „gutem Feedback“ und ohne Einbezug der Feedbackforschung könnten unsere Studien die Qualität der Rückmeldungen nicht beurteilen.

Unsere Antwort: Dieser Einwand ist **überwiegend nicht gerechtfertigt.** Wir zitieren in den Studien tatsächlich keine umfassende fachdidaktische Feedbackforschung. Das war für die engere Untersuchungsfrage aber auch nicht erforderlich. Beide Studien grenzen eine umfassende Theorie und Messung von Feedbackqualität ausdrücklich aus. Der Fokus lag darauf, funktionale Mindestbedingungen der Gebrauchstauglichkeit zu prüfen: Gibt ein Tool die gleiche Rückmeldung oder Note aus, wenn derselbe Text nach denselben Kriterien mehrfach bewertet wird? Und bewertet es einen Text nach Umsetzung seiner eigenen Verbesserungsvorschläge anschließend konsistent?

Dafür braucht es zunächst keine Aussage darüber, ob Lernende durch das Feedback der Tools motivierter sind, das Feedback verstehen oder langfristig mehr lernen. Wenn ein System eine konkrete Änderung empfiehlt, deren korrekte Umsetzung später nicht honoriert oder sogar wieder die Ausgangsformulierung verlangt – also zwischen Vorschlägen hin und her oszilliert –, ist dies eine beobachtbare interne Inkohärenz *des Produkts*.

In den Diskussionsteilen beider Studien geht unsere Interpretation allerdings weiter als dieser enge Befund: Dort folgern wir, inkonsistente Rückmeldungen seien „ohne didaktischen Wert“ beziehungsweise das Feedback verliere seinen „didaktischen Wert“ (Mühlhoff/Henningsen 2024/25, S. 26; Mühlhoff/Quägwer 2026, S. 74). Diese Formulierungen beanspruchen streng genommen mehr, als unsere Daten allein zeigen; dafür wäre eine Einordnung anhand fachdidaktisch anerkannter Theorien guten Feedbacks erforderlich gewesen. Präziser wäre: Die beobachteten Inkohärenzen gefährden die Gebrauchstauglichkeit und stellen die Qualität des Feedbacks infrage. Ob sie tatsächlich die Lernwirkung mindern, sollte mit Lernenden und vor dem Hintergrund fachdidaktischer Feedbacktheorien untersucht werden. Von Unternehmen, die lernförderliches KI-Feedback anbieten, darf man erwarten, dass sie sich an unabhängigen Untersuchungen dieser Frage beteiligen und ihre Evidenz zugänglich machen.

3. Wurde FelloFish falsch konfiguriert, weil kein Begleitmaterial hinterlegt war?

Der Einwand: Wir hätten eine Vorlage zum „materialgestützten Argumentieren“ gewählt, aber kein Material hinterlegt; schlechte Resultate seien deshalb möglicherweise selbst verursacht.

Unsere Antwort: Dafür gibt es **keinen Beleg.** Unsere Aufgabenstellung verlangte ausdrücklich eine begründete Stellungnahme zum Wahlalter, nicht die Auswertung vorgegebener Quellen. Wir orientierten uns an fünf Kriterien einer integrierten Vorlage, passten sie wegen des in unserer Aufgabe nicht vorhandenen Begleitmaterials aber ausdrücklich so an,

dass die Konfiguration keinen Bezug mehr auf Begleitmaterial nahm. Die tatsächlich verwendeten Kriterien sind im Materialanhang vollständig dokumentiert; keines verlangt die Nutzung eines nicht vorhandenen Materials. FelloFish erlaubt gerade die freie Anpassung solcher Kriterien.

Man kann diskutieren, ob eine andere Aufgabe einem bestimmten Standardworkflow des Produkts näher gewesen wäre. Daraus folgt jedoch nicht, dass unsere Konfiguration fehlerhaft war oder die dokumentierten Ausgaben ungültig wären. Für diesen Vorwurf müsste gezeigt werden, dass das System trotz der angepassten Kriterien zwingend auf nicht vorhandenes Quellenmaterial angewiesen war. Ein solcher Nachweis liegt nicht vor.

4. Ist das Korpus mit zehn konstruierten Texten zu klein und künstlich?

Der Einwand: Zehn simulierte Texte zu einer Aufgabe könnten keine belastbaren Aussagen über Schülertexte oder KI-Systeme erlauben.

Unsere Antwort: Der Einwand **verfehlt den Zweck der Untersuchung.** Unsere Studien beanspruchen keine statistische Repräsentativität für die Verteilung von Textqualitäten in einer Schulklasse. Das Korpus wurde absichtlich so konstruiert, dass verschiedene Fehlerarten, Texttypen und Strategien vorkommen. Dadurch konnten wir gezielt prüfen, wie die Produkte beispielsweise auf Themenverfehlung, Falschbehauptung, sehr kurze Abgabe, Arbeitsverweigerung, Manipulation und KI-generierten Text reagieren.

Solche Tests liefern **Existenzbelege** – und mehr wollten wir nicht leisten. Wenn eine vollständige Themenverfehlung in einem dokumentierten Lauf sehr gut bewertet wird oder derselbe Text bei identischer Eingabe stark verschiedene Werte erhält, ist dieser Produktfehler real. Für einen Existenzbeleg ist keine repräsentative Stichprobe erforderlich. Unbekannt bleibt, wie häufig der Fehler in einer zufällig gezogenen Population echter Schülerarbeiten vorkäme. Genau deshalb beschreiben wir die Untersuchung als qualitativ und explorativ und legen sämtliche Testfälle offen.

Einzelne Formulierungen über „typische“ Texte oder den „Großteil“ realer Abgaben sollten nicht als statistische Häufigkeitsaussagen gelesen werden; eine solche Aussage lässt das Korpus nicht zu. Der Kernbefund ist enger und zugleich praktisch relevant: Die Tests zeigen konkrete Fallstricke, nach denen in größeren, repräsentativen und fachspezifischen Studien systematisch gesucht werden sollte. Auch hier sollten sich die Unternehmen mit eigenen, wissenschaftlich publizierten Studien engagieren.

5. War eine menschliche Vergleichsgruppe erforderlich?

Der Einwand: Ohne Lehrkräfte als Vergleich könne man nicht beurteilen, ob die beobachteten Schwankungen in den Bewertungen desselben Textes ungewöhnlich groß seien.

Unsere Antwort: Für unsere Fragestellung war eine menschliche Vergleichsgruppe **nicht erforderlich.** Wir wählten als funktionalen Mindestmaßstab die Wiederholbarkeit bei identischem Input. Die Frage lautete: Erzeugt dasselbe Tool bei wiederholter Eingabe desselben Textes und derselben Kriterien dieselbe oder zumindest eine hinreichend ähnliche Ausgabe (Note oder Feedback)? Die beobachteten Spannweiten beantworten diese Frage direkt: Bei einigen Texten ist das nicht der Fall.

Die Studien behaupten *nicht*, die relative Zuverlässigkeit von Mensch und Maschine, ihre jeweilige demografische Fairness oder die Überlegenheit eines Systems gegenüber Lehrkräften zu bestimmen. Unsere dokumentierten Beobachtungen über das Verhalten der Produkte werden durch das Fehlen dieses Vergleichs nicht berührt. Wie die maschinellen Abweichungen im Verhältnis zur Streuung zwischen menschlichen Bewertenden einzuordnen sind, ist eine legitime und wichtige Anschlussfrage – aber eine andere Studie.

6. Wurden Variabilität und psychometrische Reliabilität unzulässig gleichgesetzt?

Der Einwand: Schwankende Einzelwerte seien lediglich „Variabilität“. Erst ein Reliabilitätskoeffizient könne zeigen, ob ein System zuverlässig sei.

Unsere Antwort: Der Einwand ist **nicht berechtigt.** In keiner der beiden Studien und in keinem der geprüften Materialanhänge verwenden wir die Fachwörter „Reliabilität“ oder „reliabel“. Wir behaupten also gerade *nicht*, eine psychometrische Reliabilitätsanalyse vorzulegen. Wir sprechen von Konsistenz, Reproduzierbarkeit, Variabilität, Wiederholbarkeit, funktionaler Zuverlässigkeit beziehungsweise allgemeiner Verlässlichkeit und operationalisieren diese Begriffe als *Wiederholung identischer Eingaben*.

Die beobachtete Variabilität belegt unmittelbar, dass eine einzelne Ausgabe nicht exakt reproduzierbar ist. Wer denselben Text unter denselben Bedingungen einmal deutlich besser und einmal deutlich schlechter bewertet bekommt, erlebt eine reale Unsicherheit, auch wenn die Rangfolge aller Texte im Durchschnitt relativ stabil bleiben sollte. Dies als Problem der Wiederholbarkeit zu bezeichnen, ist wissenschaftlich zulässig und für den praktischen Einzeleinsatz relevant.

Psychometrische Reliabilität stellt eine andere Frage: Bleibt die Rangordnung verschiedener Texte nach den produzierten Bewertungen trotz der beobachteten Schwankungen ungefähr erhalten? Das ist ein mögliches zusätzliches Erkenntnisinteresse, aber nicht die Frage, die unsere Wiederholungstests beantworten sollten.

7. Fehlen ICC, Konfidenzintervalle und Inferenzstatistik?

Der Einwand: Ohne Intraklassenkorrelationen (ICC), Unsicherheitsintervalle und inferenzstatistische Tests seien Aussagen über die Zuverlässigkeit wertlos.

Unsere Antwort: Auch dieser Einwand ist bezogen auf das Erkenntnisinteresse der Studien **nicht berechtigt.** Für die dokumentierten Existenzbelege benötigen wir weder einen p-Wert noch eine Hochrechnung auf eine Grundgesamtheit. Wenn ein Tool eine objektiv falsche Behauptung übersieht oder denselben Text in mehreren Durchläufen deutlich verschieden bewertet, ist dieser Vorgang unabhängig von einer Inferenzstatistik bereits nachgewiesen.

Ein ICC lässt sich aus den veröffentlichten Daten durchaus nachträglich berechnen. Für die Fobizz-Ausgaben auf der in der Studie verwendeten Oberstufenpunkteskala ergibt sich beispielsweise `ICC(1,1) = 0,718` mit einem modellbasierten 95%-Intervall von `0,479–0,904`. Vereinfacht gesagt misst der ICC hier, ob die zehn Texte unseres konstruierten Korpus trotz schwankender Einzelbewertungen ungefähr in derselben Rangfolge bleiben. *Der ICC-Wert hängt deshalb nicht nur vom Tool, sondern stark davon ab, welche Texte in das Korpus aufgenommen wurden:* Ein bewusstes Gemisch aus sehr guten, sehr schwachen und extremen Fällen kann eine stabile Rangfolge erzeugen, obwohl einzelne Texte erheblich schwanken. Deshalb besitzt eine ICC-Analyse für unser Testdesign keine bedeutende Aussagekraft.

In einer anders angelegten Studie mit einem repräsentativen oder realitätsnah verteilten Korpus wäre ein ICC dagegen eine wichtige Kennzahl für die Zuverlässigkeit der Tools bei der Bestimmung beispielsweise eines Notenrankings. Für unser absichtlich konstruiertes Prüfkorpus beantwortet der ICC weder die Frage, wie häufig Fehler im Schulalltag vorkommen, noch entkräftet er die dokumentierte Abweichung im Einzelfall.

8. Misst die iterative Umsetzung von Feedback überhaupt Feedbackqualität?

Der Einwand: Eine ausbleibende monotone Scoresteigerung bei iterativer Einarbeitung der Verbesserungsvorschläge der Feedbacktools belege weder schlechtes Feedback noch fehlenden Lernzuwachs; dazu hätte es ein unabhängiges Qualitätsmaß und Daten von Lernenden gebraucht.

Unsere Antwort: Der Einwand richtet sich überwiegend gegen eine Behauptung, die wir nicht untersucht haben. Die Iterationsreihen zeigen valide eine **interne Inkohärenz**: Die Systeme honorieren die korrekte Befolgung ihrer eigenen Hinweise in der nächsten Bewertung nicht verlässlich. Teilweise entstehen Schleifen, in denen das Tool zwischen Formulierungen oszilliert. Das ist ein relevanter Produkttest, weil die Rückmeldung gerade mit dem Versprechen ausgegeben wird, den Text anhand der Kriterien zu verbessern.

Es war nicht der Anspruch dieser Testreihe, zu messen, ob Lernende das Feedback verstehen, ob die Texte nach einem unabhängigen fachlichen Urteil durch das Feedback besser werden oder ob langfristiger Lernzuwachs entsteht. Dafür wären Lernendenstudien, unabhängige Ratings und andere Designs notwendig. Ihr Fehlen widerlegt den von uns tatsächlich erhobenen Befund nicht. Wie bereits unter Punkt 2 eingeräumt, sollte der Schritt von

interner Inkohärenz zu einer umfassenden Aussage über den „didaktischen Wert“ vorsichtiger formuliert und in Anschlussstudien untersucht werden.

9. Ist der Copy-Paste-Test methodisch mehrdeutig?

Der Einwand: Wörtlich übernommene, sinngemäß umgesetzte und frei überarbeitete Fassungen der Verbesserungsvorschläge der Tools seien nicht unabhängig als gleich gut validiert worden. Höhere Werte für Copy-and-paste könnten deshalb reale Qualitätsunterschiede oder zufällige Bewertungsschwankungen abbilden.

Unsere Antwort: Dieser Einwand ist **grundsätzlich berechtigt.** Die Studie macht transparent, wie die drei Fassungen entstanden (Mühlhoff/Quägwer 2026, S. 67–68), und weist ausdrücklich darauf hin, dass die freie Überarbeitung nicht vollständig standardisierbar ist (S. 72). Der Materialanhang legt die verglichenen Fassungen vollständig offen (S. 38–53). Der Test war als explorative Prüfung einer auffälligen Hypothese angelegt, nicht als abschließender Kausalnachweis.

Zwei Grenzen hätten wir noch deutlicher benennen sollen. Erstens wurden die wortgetreue und die sinngemäße Fassung nicht durch unabhängige, verblindete Fachpersonen auf inhaltliche Gleichwertigkeit geprüft. Zweitens liegt pro Text und Strategie nur eine resultierende Bewertung vor, obwohl Testreihe A gerade die Schwankung von Einzelbewertungen gezeigt hatte. Unterschiedliche Sitzungen und die Auswahl verwertbarer Beispielsätze schaffen zusätzliche Unsicherheit.

Der Befund bleibt deshalb ein ernstzunehmender Hinweis, aber kein eindeutiger Nachweis eines allgemeinen Copy-Paste- oder Selbsterkennungs-Bias. Gerade dafür sind explorative Tests da: Sie machen ein mögliches Problem sichtbar und motivieren ein kontrolliertes Anschlussdesign – mit qualitätsgematchten Fassungen, Blindrating, identischen Ausgangsrückmeldungen und vielen Wiederholungen. Eine Kooperation mit den Unternehmen könnte insbesondere das Sitzungsproblem kontrollieren: Die Anbieter könnten einen stabilen Testzugang oder einen exportier- und wiederherstellbaren identischen Sitzungszustand bereitstellen, sodass alle Überarbeitungsvarianten von exakt derselben Ausgangsbewertung, demselben Feedback und derselben Produktversion ausgehen.

10. Belegen unsere Beobachtungen die systematische Bevorzugung KI-generierter Texte?

Der Einwand: Dass der von ChatGPT generierte Text, der Teil unseres Testkorpus war, häufig sehr hohe Bewertungen erhalten habe, beweise nicht, dass die Tools KI-Texte generell bevorzugen.

Unsere Antwort: Dieser Einwand ist in Bezug auf unsere Studie **berechtigt.** Die Studien legen offen, dass das gemeinsame Korpus nur *einen* ausdrücklich mit KI erzeugten Text

enthält. Korrekt ist die Beobachtung, dass dieser Text in allen getesteten Produkten besonders hoch und relativ stabil bewertet wurde. Daraus allein folgt aber statistisch noch kein allgemeiner Nachweis, dass KI-generierte Texte als *Klasse* von den Tools bevorzugt werden. Unsere Diskussion formuliert an dieser Stelle zu weit; der Befund sollte als Hypothese für eine systematische Anschlussstudie verstanden werden.

Die Hypothese ist allerdings nicht aus der Luft gegriffen. Eine 2026 veröffentlichte ETS-Studie verglich in großem Umfang KI- und menschengeschriebene GRE-Essays. Nach der Zusammenfassung des ETS Research Institute bewertete der automatische Scorer „e-rater“ die KI-Texte höher als menschliche Rater; insbesondere sprachliche Oberflächenmerkmale konnten schwächere Argumentation überdecken (Zhong et al. 2026, DOI 10.1111/emip.70013). Daneben dokumentiert Forschung zu LLMs als Bewertungsinstanzen einen „Self-Preference Bias“, also eine Bevorzugung eigener oder stilistisch vertrauter Modellausgaben unter kontrollierter Berücksichtigung von Qualitätsunterschieden (Chen et al. 2025, EMNLP).

Diese Studien stützen die Plausibilität unserer weitergehenden Interpretation, beweisen aber nicht formal denselben Mechanismus für FelloFish, Edaira oder das Fobizz-Korrekturtool.

11. Ist ein einzelner Manipulationsversuch nur eine belanglose Anekdote?

Der Einwand: Der im Materialanhang unserer 2026er Studie dokumentierte Manipulationstest mit einem Metatext (Mühlhoff/Quägwer 2026, Materialanhang, S. 67) sei offenbar nur einmal durchgeführt worden und sage deshalb wenig aus.

Unsere Antwort: Der Einzelfallcharakter wurde **klar kommuniziert** und entspricht dem Erkenntnisinteresse dieses Tests. Der Manipulationsversuch bestand darin, dass der eingereichte Text keine geforderte Argumentation enthielt, sondern auf einer Metaebene lediglich beschrieb, wie eine gute Argumentation aufgebaut sein sollte. Die getestete Edaira-Version vergab dafür 57,5 von 60 Punkten beziehungsweise 92,7 Prozent. Bei FelloFish ließ sich dieser Effekt nicht zeigen.

Ein erfolgreicher Test genügt als Existenzbeleg dafür, dass die damalige Edaira-Version in der dokumentierten Konfiguration auf diese Weise manipulierbar war. Wir leiten daraus weder eine Fehlerhäufigkeit noch die Behauptung ab, jedes abgesicherte System wäre für diesen Manipulationsversuch empfänglich. Dass ein Anbieter die Lücke später schließen könnte, macht den für die getestete Version dokumentierten Befund nicht falsch.

12. Sind Aussagen über strukturelle Grenzen großer Sprachmodelle zu weit?

Der Einwand: Aus drei Black-Box-Produkten könne nicht auf alle gegenwärtigen und zukünftigen LLM-basierten Bewertungssysteme geschlossen werden.

Unsere Antwort: Eine Allaussage über jedes denkbare, auch noch gar nicht entwickelte System lässt sich mit Produkttests selbstverständlich nicht beweisen. Das ist aber kein sinnvoller Maßstab für die Interpretation einer technologischen Entwicklung. Wir haben nacheinander drei im deutschen Schulmarkt sichtbare, viel genutzte und von ihren Anbietern als leistungsfähige Lösungen präsentierte Systeme getestet. Bei allen drei fanden wir gravierende Unzulänglichkeiten. Unsere Einordnung struktureller Ursachen ist eine fachlich begründete Interpretation dieser wiederkehrenden Befunde und bekannter Eigenschaften generativer Sprachmodelle, kein mathematischer Unmöglichkeitsbeweis für alle Zukunft.

Bemerkenswert ist, dass inzwischen auch fobizz selbst zu fast derselben Einschätzung gelangt ist. Das Unternehmen erklärte am 2. August 2026, das KI-Korrekturtool nicht mehr standardmäßig anzubieten, mit folgender Begründung:

> „Große Sprachmodelle (LLMs) liefern bei der Bewertung von Texten keine ausreichend zuverlässigen Ergebnisse. Trotz intensiver Weiterentwicklung über anderthalb Jahre hat sich daran grundlegend nichts geändert. [...] Bei gleicher Eingabe, gleichen Kriterien und wiederholter Auswertung liefern Sprachmodelle unterschiedliche Bewertungen. Das ist kein Prompt-Problem, das sich mit der nächsten Iteration lösen lässt. Es ist eine strukturelle Einschränkung der Technologie in ihrer aktuellen Form.“ (Erklärung von fobizz).

Damit bestätigt gerade einer der hier getesteten Anbieter unsere strukturelle Beobachtung und ihre technologische Relevanz.

Gleichzeitig bleiben wir ergebnisoffen. Sorgfältig kalibrierte Spezialarchitekturen können in anderen Aufgaben deutlich bessere Resultate erzielen; etwa zeigt die Studie von Huang, Palermo und Wilson (2026), dass ein feinabgestimmtes GPT-4o-System in einem großen englischsprachigen Korpus menschlicher Übereinstimmung nahekommen kann. Solche Ergebnisse widerlegen unsere Produkttests jedoch nicht – und es handelt sich hier um Produkte, die von echten Lehrkräften in echten Schulen und für echte Schüler:innen eingesetzt werden. Wir werden neue oder grundlegend überarbeitete Werkzeuge weiterhin offen und anhand ihrer tatsächlichen Ausgaben testen. Bis belastbare produktspezifische Evidenz vorliegt, ist bei folgenreichen schulischen Bewertungen jedoch Vorsicht geboten.

13. Auch die Bewertungen menschlicher Lehrkräfte schwanken

Der Einwand: Auch Lehrkräfte bewerteten variabel, subjektiv und sozial verzerrt; die Variabilität von KI-Bewertungen könne deshalb kaum schlimmer sein.

Unsere Antwort: Als Einwand gegen unsere Studien ist dies nicht gerechtfertigt, weil wir keinen Vergleich der menschlichen und maschinellen Bewertungsqualität beanspruchen. Vor allem folgt aus menschlichen Fehlern logisch nicht, dass ein automatisiertes System hinreichend gut oder fair ist. Zwei mögliche Fehlerquellen heben einander nicht auf. Ob

ein konkretes System menschliche Bewertung verbessert, kann nur ein direkter, kontrollierter Vergleich zeigen.

Das diesem Einwand zugrunde liegende Argument begegnet uns allerdings in vielen Automatisierungsdebatten: Auch Menschen entscheiden bei Kreditvergabe, medizinischen Diagnosen oder Schulnoten nicht fehlerfrei; daher erscheinen die Fehler automatischer Systeme relativ betrachtet weniger gravierend. Diese Denkfigur ist verführerisch, aber sachlich falsch und sozial gefährlich. Sie prüft nicht, welche Fehler das neue System hinzufügt, welche Gruppen betroffen sind, ob Betroffene die Entscheidung nachvollziehen und anfechten können und ob menschliche Kontrolle in der Praxis tatsächlich stattfindet.

Maschinelle Fehler können insbesondere leichter skaliert werden: Derselbe systematische Bias kann bei einem KI-System in kurzer Zeit sehr viele Menschen betreffen, denn es gibt nur wenige Anbieter von KI-Modellen, die weltweit operieren – und auch die hier getesteten Tools greifen im Hintergrund auf die großen Sprachmodelle multinationaler Konzerne zu. Hinzu kommt der sogenannte Automation Bias – die psychologisch nachgewiesene Tendenz, computergenerierten Ergebnissen eine Objektivität zuzuschreiben, die sie nicht besitzen, und sie deshalb seltener kritisch zu prüfen als menschliche Einschätzungen und Bewertungen. Darum kann eine ungenaue oder verzerrte KI-Bewertung praktisch und sozialstrukturell folgenreicher sein als menschliche Fehlurteile. Das bedeutet nicht, dass wir hier menschliche Bewertung idealisieren möchten. Es bedeutet, dass ein neues Bewertungssystem einen eigenständigen Qualitäts- und Fairnessnachweis braucht und nicht durch den bloßen Hinweis auf menschliche Unvollkommenheit legitimiert werden kann.

14. Widerlegen positive Vergleichsstudien unsere Ergebnisse?

Der Einwand: Forschung, etwa die Metaanalyse von Yun (2023) und die Studie von Huang, Palermo und Wilson (2026), zeige bereits eine hohe Übereinstimmung automatischer und menschlicher Essaybewertungen; damit seien unsere kritischen Befunde überholt oder widerlegt.

Unsere Antwort: Nein. Die angeführten Arbeiten untersuchen andere Systeme, Sprachen, Textkorpora und Kalibrierungsverfahren. Yun (2023) fasst überwiegend klassische, spezialisierte Essay-Scorer aus den Jahren 1999 bis 2014 zusammen und berichtet trotz eines hohen mittleren Zusammenhangs eine extreme Heterogenität der Einzelergebnisse. Huang, Palermo und Wilson (2026) testen ein sorgfältig kalibriertes beziehungsweise feinabgestimmtes GPT-4o-System an 1.768 englischen Texten der Klassen 3 und 4. Das zeigt, dass gute automatisierte Bewertung unter bestimmten Bedingungen möglich sein kann. Es zeigt jedoch *nicht*, dass fobizz, FelloFish oder Edaira in den von uns getesteten Versionen ebenso gut funktionierten.

Positive Forschung zu einem spezialisierten System widerlegt keine vollständig dokumentierte Ausgabe eines anderen Produkts. Umgekehrt beweisen unsere Tests nicht, dass jede

zukünftige Architektur scheitern muss. Genau deshalb sind präzise, produktspezifische Evaluationen nötig. Die Verantwortung hierfür liegt vor allem bei den Anbietern, die solche Produkte auf den Markt bringen, und bei den Schulträgern beziehungsweise Bildungsverwaltungen, die sie für den schulischen Einsatz lizenzieren. Die pauschale Behauptung, KI bewerte heute bereits „genauso zuverlässig“ wie Erst- und Zweitkorrektoren, ist für die drei getesteten Produkte durch keine der angeführten Vergleichsstudien belegt – und durch unsere Studien deutlich in Zweifel gezogen.

Fazit

Die öffentliche Debatte vermischt häufig drei verschiedene Fragen:

1. Ist ein konkreter Fehler dokumentiert? – Das beantworten unsere Materialien für zahlreiche Fälle mit Ja.
2. Wie häufig tritt dieser Fehler in allen realen Anwendungen auf? – Das lässt sich aus unserem gezielten Korpus nicht schätzen und war nicht Ziel der Studien.
3. Können KI-Systeme grundsätzlich besser oder genauso gut sein wie menschliche Bewertende? – Auch das haben wir nicht untersucht.

Die zweite und dritte Frage sind wichtig, aber sie entwerten die erste nicht. Wer KI-Systeme für schulische Rückmeldung oder Leistungsbewertung anbietet oder einkauft, sollte vor dem breiten Einsatz belastbar zeigen oder überprüfen, ob die konkrete Produktversion reproduzierbar, nachvollziehbar, fachlich valide und gegenüber typischen wie kritischen Eingaben robust arbeitet. Dass dies nicht systematisch erfolgt, sondern die schulische Realität als Testfeld für ungetestete Softwareversionen von Start-up-Unternehmen herhalten muss, ist ein gravierender Missstand. Unsere explorativen Tests zeigen, warum solche Prüfungen dringend erforderlich sind und welche Fehlerklassen dabei berücksichtigt werden sollten.

Unsere praktische Schlussfolgerung bleibt daher bestehen: Solange für konkrete Produkte keine überzeugende, unabhängige und öffentlich prüfbare Evidenz vorliegt, sollten sie nicht in der schulischen Leistungsbewertung eingesetzt werden. Anbieter sollten ihre Evaluationsdesigns, Testdaten und Ergebnisse offenlegen und sich unabhängigen Vergleichstests stellen.

Ein strukturelles Problem der Qualitätssicherung

Bei alledem sollte das institutionelle Umfeld nicht aus dem Blick geraten, in dem unsere Studien entstanden sind. Sie wurden weder von den untersuchten Unternehmen noch von Schulträgern oder Bundesländern in Auftrag gegeben und erhielten keine externe Projektförderung. Es waren unabhängige, ungefragte Untersuchungen aus dem Wissenschaftsbetrieb. Dass solche elementaren Produkttests von einzelnen Forschenden aus eigener Initiative vorgenommen werden müssen, ist selbst ein bildungspolitisch relevanter Befund.

Schulträger und Bundesländer sind gegenwärtig in erheblichem Umfang bereit, öffentliche Mittel für Digitalisierung und KI-Angebote in Schulen auszugeben. Eine vergleichbar starke, dauerhaft institutionalisierte Förderung unabhängiger Forschung zur Qualität, Tauglichkeit und Sicherheit der konkret beschafften Werkzeuge ist dagegen nicht erkennbar. Produktversionen werden lizenziert, in Schulen eingeführt und fortlaufend verändert, ohne dass vor ihrem Einsatz regelmäßig unabhängige, veröffentlichte und produktspezifische Eignungsnachweise vorliegen.

Damit entsteht ein problematisches Innovationsmodell, in dem Schulen zum Echtzeitlabor für noch nicht ausgereifte Lösungen junger Unternehmen werden. Die Anbieter erhalten Umsätze, Marktzugang, Nutzungserfahrungen und einen realen Erprobungsraum; Lehrkräfte und Schüler:innen tragen dagegen den zusätzlichen Prüfaufwand sowie die pädagogischen und sozialen Folgen möglicher Fehlfunktionen. Ob ihnen ein entsprechender Nutzen entsteht, ist gerade das, was unabhängig untersucht werden müsste. Öffentliche Beschaffung von KI für Schulen sollte deshalb an transparente Qualitätsanforderungen und Standards, unabhängige Evaluationen konkreter Produktversionen, die Veröffentlichung der Ergebnisse und wirksame Ausstiegsmöglichkeiten geknüpft werden.

Unsere Studien und die Materialangänge

- Rainer Mühlhoff und Sean Quägwer (2026): „Automatisierte Korrektur mit KI? Wir testen zwei verbreitete Tools“, *SEMINAR – Lehrerbildung und Schule* 2/2026, S. 62–76; Materialanhang.
- Rainer Mühlhoff und Marte Henningsen (2024/2025): „Chatbots im Schulunterricht: Wir testen das Fobizz-Tool zur automatischen Bewertung von Hausaufgaben“; Materialanhang.