Textorial

Automatische Textbewertung mit wissenschaftlich fundierten Rückmeldungen für den Schreibunterricht

Warum wir Textorial entwickelt haben 3

Wie automatisiertes Feedback die Schreibkompetenz fördert 4

Textorial auf einen Blick 5

So funktioniert Textorial 6

Mehr als ein Sprachmodell 8

Wissenschaftlich geprüft 9

Datensouveränität in der Schweiz 14

Ausblick und Weiterentwicklung 15

Literatur 16

Warum wir Textorial entwickelt haben

Schreiben gehört zu den wichtigsten Kompetenzen, die Schülerinnen und Schüler während ihrer Schulzeit erwerben. Wer schreiben kann, kann Gedanken ordnen, Wissen strukturieren, argumentieren und sich verständlich ausdrücken. Schreibkompetenz ist eine zentrale Voraussetzung für erfolgreiches Lernen in allen Fächern sowie für Ausbildung, Beruf und gesellschaftliche Teilhabe.



Die Forschung zeigt seit vielen Jahren: Schreiben lernt man durch Schreiben. Entscheidend sind zahlreiche Schreibgelegenheiten sowie zeitnahe, konkrete Rückmeldungen, die Lernende dabei unterstützen, ihre Texte zu überarbeiten und ihre Schreibkompetenz weiterzuentwickeln.

Im Schulalltag ist dies eine grosse Herausforderung. Ein Klassensatz Texte bedeutet für Lehrpersonen oft mehrere Stunden Korrekturarbeit. So entsteht ein Zielkonflikt: Aus pädagogischer Sicht wäre häufigeres Schreiben wünschenswert, es lässt sich jedoch nur mit erheblichem zeitlichem Aufwand realisieren.

Künstliche Intelligenz eröffnet die Möglichkeit, diesen Zielkonflikt zu entschärfen. Sie kann Texte in kurzer Zeit analysieren und Rückmeldungen erstellen. Doch Geschwindigkeit allein genügt nicht. Eine automatische Bewertung von Texten muss fair, nachvollziehbar und datenschutzkonform sein.

Aus genau diesem Anspruch heraus haben wir Textorial entwickelt, eine wissenschaftlich fundierte Lösung des Instituts für Bildungsevaluation (IBE), die Psychometrie, Sprachtechnologie und künstliche Intelligenz verbindet.

Textorial bewertet Texte zuverlässig, gibt differenzierte und altersgerechte Rückmeldungen und schafft Lehrpersonen Freiräume für die Begleitung von Lernprozessen.

Unsere Vision ist, dass sich die Förderung der Schreibkompetenz an pädagogischen Überlegungen orientiert – und nicht am Korrekturaufwand. Wenn qualitativ hochwertige Rückmeldungen innerhalb kurzer Zeit verfügbar sind, erhalten Lernende mehr Gelegenheiten, Texte zu verfassen und zu überarbeiten und dadurch ihre Schreibkompetenz systematisch weiterzuentwickeln.

Dieses Whitepaper beschreibt die wissenschaftlichen Grundlagen und die technische Architektur von Textorial und stellt die empirischen Befunde zu Fairness und Zuverlässigkeit vor. Es zeigt, wie Textorial die Möglichkeiten künstlicher Intelligenz mit der pädagogischen Expertise von Lehrpersonen verbindet und so den Schreibunterricht unterstützt.

Wie automatisiertes Feedback die Schreibkompetenz fördert

Texte werden bereits seit Jahren automatisiert beurteilt. Die entscheidende Frage ist jedoch, ob automatisiertes Feedback Schülerinnen und Schüler dabei unterstützt, besser schreiben zu lernen. Diese Frage wurde in zahlreichen wissenschaftlichen Studien untersucht.

Eine Metaanalyse von 20 Studien mit insgesamt 2'828 Lernenden zeigt, dass automatisiertes Feedback die Schreibleistung im Durchschnitt verbessert. Bei Lernenden, die in einer Zweit- oder Fremdsprache schrieben, fiel der Effekt tendenziell grösser aus als bei Lernenden, die in ihrer Erstsprache schrieben (Fleckenstein, Liebenow & Meyer, 2023). Eine Langzeitstudie weist zudem darauf hin, dass wiederholtes automatisiertes Feedback nicht nur die sprachliche Korrektheit, sondern auch die allgemeine Schreibqualität verbessern kann (Shermis, Garvan & Diao, 2008).

Ein differenziertes Bild ergibt sich aus dem direkten Vergleich von menschlichem und automatisiertem Feedback. In einer Studie mit Studierenden, die Englisch als Fremdsprache lernten, erzielte die Gruppe mit Lehrpersonenfeedback bei der Überarbeitung desselben Textes bessere Ergebnisse. Bei einer anschliessenden neuen Schreibaufgabe schnitt dagegen die Gruppe mit automatisiertem Feedback besser ab. Als besondere Stärken des automatisierten Feedbacks nannten die Studierenden die Fehlererkennung und die Vorschläge für Synonyme (Wang & Han, 2022).

Systeme auf der Grundlage generativer Sprachmodelle können darüber hinaus differenzierte, auf den jeweiligen Text bezogene Rückmeldungen erzeugen. Eine Studie mit Schülerinnen und Schülern der Sekundarstufe zeigte, dass KI-generiertes Feedback im Vergleich zu einer Bedingung ohne Feedback die Textüberarbeitung und die Motivation verbesserte und positive Emotionen förderte (Meyer et al., 2024). In einer weiteren Studie erreichte das von ChatGPT erzeugte Feedback insgesamt ein beachtliches Qualitätsniveau. Rückmeldungen geschulter Fachpersonen wurden jedoch in vier der fünf untersuchten Qualitätsmerkmale besser bewertet (Steiss et al., 2024).

Zusammengenommen sprechen diese Befunde dafür, automatisiertes Feedback nicht als Ersatz, sondern als Ergänzung der Rückmeldungen von Lehrpersonen zu verstehen.

Weshalb wirkt Feedback überhaupt? Es hilft Lernenden zu erkennen, wo sie stehen und wie sie ihren Text weiterentwickeln können. Besonders wirksam ist Feedback, wenn es zeitnah erfolgt, konkret auf den jeweiligen Text bezogen ist und klare Hinweise für die Überarbeitung

enthält (Charalampous & Darra, 2023; Hattie & Timperley, 2007). Genau diese Bedingungen lassen sich im Schulalltag häufig nur schwer erfüllen.

Hier setzt Textorial an. Das System stellt kurz nach Abgabe des Textes eine differenzierte Rückmeldung bereit, die sich auf klar definierte Bewertungskriterien stützt und konkrete Hinweise zur Verbesserung enthält.

> **Schreiben lernt man durch Übung – und durch konkretes, zeitnahes Feedback.**

Textorial auf einen Blick

Textorial bewertet Texte kurz nach der Abgabe in den vier Bereichen Inhalt, Textzusammenhang, Sprachrichtigkeit und Stil. Die Lernenden erhalten dazu ein verständliches, altersgerechtes Feedback mit konkreten Verbesserungsvorschlägen.

Textorial kann für Texte in Deutsch, Englisch und Französisch eingesetzt werden und ist von der 3. Klasse der Primarstufe bis zum Ende der Sekundarstufe I sowie in der Berufsbildung nutzbar. Es verarbeitet sowohl digitale als auch handschriftlich verfasste Texte.

Im Unterschied zu allgemein einsetzbaren KI-Anwendungen wurde Textorial gezielt für die Bewertung von Schülertexten entwickelt und wissenschaftlich geprüft. Das System entstand im Rahmen eines Innosuisse-Projekts in Zusammenarbeit mit der Universität Zürich und der EPFL Lausanne. Grundlage bilden 3'238 von Sprachexpertinnen und Sprachexperten beurteilte Schülertexte.

Textorial verbindet speziell trainierte KI-Modelle, psychometrische Methoden und regelbasierte Sprachtechnologie in einem mehrstufigen Bewertungs- und Feedbacksystem. In der wissenschaftlichen Überprüfung stimmen die Bewertungen von Textorial in hohem Masse mit den menschlichen Bewertungen überein ($r = 0.91$ bis 0.99 auf der Ebene der acht Einzelkriterien).

Fairness ist bei Textorial kein Versprechen, sondern ein wissenschaftlich überprüftes Qualitätsmerkmal. Umfangreiche Analysen zeigen, dass die Bewertungen von Textorial unabhängig von Geschlecht, Erstsprache und Klassenstufe vergleichbar zuverlässig sind.

Neue Modelle werden vor der Einführung anhand der definierten Qualitäts- und Fairnesskriterien geprüft und auch im Betrieb kontinuierlich überwacht.

Ein ebenso zentraler Bestandteil des Systems ist der Datenschutz. Textorial arbeitet ausschliesslich mit offen verfügbaren Sprachmodellen (Mistral Small 3.1 und Gemma 4), die auf einer redundanten GPU-Infrastruktur in der Schweiz betrieben werden. Texte werden ausschliesslich in der Schweiz verarbeitet und nicht an externe KI-Anbieter weitergegeben.

Wissenschaftlich fundiert

Entwickelt und evaluiert mit der Universität Zürich und der EPFL Lausanne.

Der Mensch bleibt Massstab

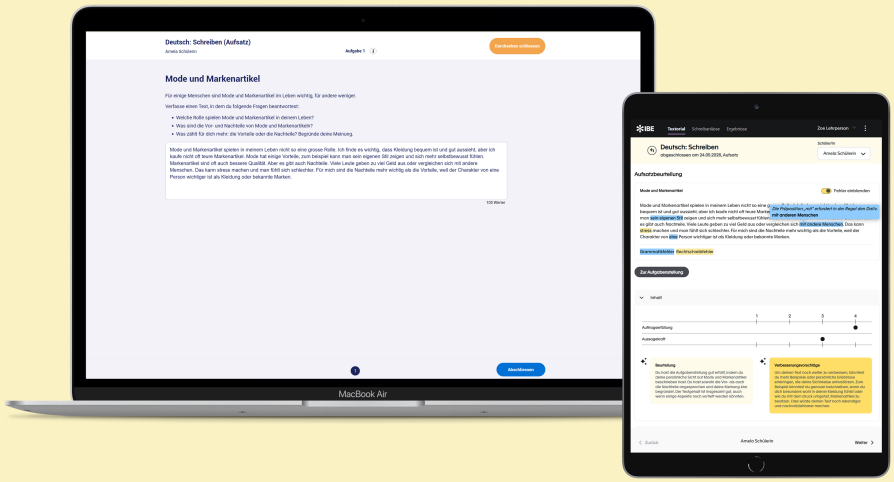
Die Ergebnisse unterstützen das professionelle Handeln von Lehrpersonen.

Datensicherheit aus der Schweiz

Die gesamte Verarbeitung erfolgt auf Schweizer Infrastruktur.

So funktioniert Textorial

Der Einsatz von Textorial beginnt mit einem neuen Schreibanlass. Die Lehrperson wählt Sprache und Aufgabenstellung. Dabei kann sie auf vorhandene Schreibanlässe zurückgreifen oder eigene erstellen. Ob Lernende von einem Erlebnis auf der Schulreise erzählen, auf Deutsch eine Position begründen, auf Französisch ein Ferienerlebnis schildern oder auf Englisch eine Person porträtieren – Textorial lässt sich bei vielfältigen Schreibanlässen flexibel einsetzen.



Die Schülerinnen und Schüler verfassen ihren Text direkt im Browser oder schreiben ihn von Hand. Handschriftliche Texte werden von der Lehrperson eingescannt und vom System automatisch transkribiert.

Kurz nach der Abgabe steht die vollständige Auswertung zur Verfügung. Diese umfasst eine Punktzahl von 1 bis 4 in den Bereichen Inhalt, Textzusammenhang, Sprachrichtigkeit und Stil – basierend auf jeweils zwei Kriterien und bezogen auf

die Erwartungen der jeweiligen Klassenstufe. Für Texte der 4. Primarklasse gelten andere Bewertungsmassstäbe als für Texte der 3. Klasse der Sekundarstufe I.

Ergänzt wird die Bewertung durch direkt im Text markierte Fehler mit konkreten Verbesserungsvorschlägen sowie durch ein verständliches und altersgerechtes Feedback, das sich am sprachlichen Entwicklungsstand der jeweiligen Klassenstufe orientiert. Ein Kind der 3. Primarklasse erhält deshalb Rückmeldungen zu weniger und altersgerechteren Fehlertypen als eine Schülerin oder ein Schüler der 3. Klasse der Sekundarstufe I. Jede Auswertung kann als PDF gespeichert oder ausgedruckt werden.

Im Hintergrund nutzt Textorial eine klassenstufenübergreifende vertikale Skala. Jeder Text wird zunächst auf dieser Skala verortet; anschliessend wird das Ergebnis in eine klassenstufenbezogene Bewertung übersetzt. Lehrpersonen erhalten damit eine Grundlage, Lernfortschritte gezielt zu begleiten – von einzelnen Schülerinnen und Schülern bis hin zur gesamten Klasse.

Punktzahl

Je 1 bis 4 Punkte für Inhalt, Textzusammenhang, Sprachrichtigkeit und Stil

Fehlermarkierung

Im Text markierte Rechtschreib- und Grammatikfehler mit Korrekturvorschlägen

Feedback

Hinweise zu Stärken, Entwicklungsbereichen und zur Überarbeitung

Textorial bietet weit mehr als eine automatische Bewertung: Lehrpersonen sehen, welche Texte bereits eingereicht wurden, wie die Klasse in den einzelnen Bewertungsbereichen abschneidet und wo gemeinsame Stärken oder Ansatzpunkte für die Förderung liegen.

Wie jedes KI-System kann auch Textorial ungewöhnliche Ausdrucksweisen oder kreative sprachliche Lösungen nicht immer angemessen einordnen. Lehrpersonen können daher jede Bewertung einsehen und fachlich einordnen. Textorial unterstützt ihr professionelles Urteil, ersetzt es aber nicht.

Mehr als ein Sprachmodell

Viele KI-Anwendungen bewerten Texte, indem sie den vollständigen Text direkt an ein grosses Sprachmodell senden. Dieses erstellt anschliessend sowohl die Bewertung als auch das Feedback. Textorial verfolgt bewusst einen anderen Ansatz.

Moderne Sprachmodelle sind beeindruckend leistungsfähig. Sie können Texte verstehen, zusammenfassen und kommentieren. Für eine faire, nachvollziehbare und wissenschaftlich fundierte Leistungsbeurteilung genügt ein Sprachmodell allein jedoch nicht. Solche Modelle übersehen Fehler, erzeugen gelegentlich unzutreffende Rückmeldungen und ihre Bewertungen sind nicht immer reproduzierbar. Deshalb kombiniert Textorial künstliche Intelligenz, psychometrische Modelle und regelbasierte Sprachtechnologie zu einem gemeinsamen Bewertungssystem.

Im Zentrum steht ein für die Bewertung von Texten trainiertes Sprachmodell. Als Trainingsgrundlage dienten 3'238 Texte von Lernenden sowie die dazugehörigen Beurteilungen geschulter Fachpersonen. Die Innovation liegt nicht im Sprachmodell selbst, sondern in seinem Trainingsziel. Das Modell bildet nicht einzelne menschliche Bewertungen unmittelbar nach. Es sagt vielmehr psychometrisch geschätzte Werte für die beiden Dimensionen der Schreibkompetenz – Textqualität und Sprachrichtigkeit – vorher.

Diese Werte wurden aus jeweils zwei voneinander unabhängigen Beurteilungen abgeleitet. Unterschiede in der Bewertungsstrenge der Fachpersonen wurden dabei statistisch berücksichtigt. Dadurch orientiert sich Textorial nicht an der Strenge oder Milde einzelner Fachpersonen, sondern an einem gemeinsamen fachlichen Bewertungsmassstab.

Auf dieser Grundlage werden Bewertungen für acht Einzelkriterien abgeleitet (Auftragserfüllung, Aussagekraft; Kohärenz, Kohäsion; Rechtschreibung, Grammatik; Satzbau, Wortschatz), die anschliessend zu den vier Bereichen Inhalt, Textzusammenhang, Sprachrichtigkeit und Stil zusammengeführt werden, die Lehrpersonen und Lernende in der Rückmeldung sehen. Das Ergebnis sind objektivere, fairere und wissenschaftlich fundiertere Bewertungen der vier Bereiche, die in die Punktzahlen von 1 bis 4 für die jeweilige Klassenstufe übersetzt werden.

Auch bei der sprachlichen Analyse verfolgt Textorial einen anderen Ansatz als viele KI-Systeme. Rechtschreib- und Grammatikfehler werden nicht ausschliesslich durch ein Sprachmodell erkannt, sondern mit einer regelbasierten Sprachtechnologie auf der Grundlage von LanguageTool. Diese arbeitet deterministisch und reproduzierbar: Derselbe Fehler wird heute wie morgen gleich erkannt und gleich beschrieben. Das IBE hat das System für den schulischen Einsatz erweitert – unter anderem mit eigenen Fehlerkategorien, Filtern für Eigennamen, Schweizer Sprachbesonderheiten sowie einer klassenstufengerechten Auswahl der zurückgemeldeten Fehlertypen. Mit einer zusätzlichen KI-gestützten Qualitätskontrolle wird die Fehlerliste überprüft, bevor sie für das Feedback verwendet wird (König, Stahlhut, Berger & Debelak, 2026).

Für die Formulierung von Verbesserungsvorschlägen und Feedbacks kommt ein zweites Sprachmodell zum Einsatz. Die rund 23'000 Feedbacktexte, auf denen das Sprachmodell trainiert wurde, wurden von Fachpersonen entwickelt, geprüft und redaktionell überarbeitet. Sie orientieren sich an didaktischen Grundsätzen, verwenden eine altersgerechte Sprache und geben konkrete Hinweise für den nächsten Lernschritt.

Das Modell erhält den Text, die Aufgabenstellung, die Bewertungsergebnisse sowie die überprüfte Fehlerliste und formuliert daraus verständliche, altersgerechte Rückmeldungen. Es erklärt die Ergebnisse und zeigt konkrete Verbesserungsmöglichkeiten auf, ohne die vergebenen Punktzahlen zu verändern. Dasselbe Modell übernimmt auch die Transkription handschriftlich verfasster Texte.

Die gewählte Systemarchitektur wurde nicht nur theoretisch begründet, sondern auch empirisch überprüft. Die wissenschaftliche Überprüfung zeigte, dass die Kombination aus regelbasierter Fehleranalyse und KI-gestützter Bewertung insbesondere bei der Rechtschreibung die Übereinstimmung mit menschlichen Beurteilungen erhöht und systematische Verzerrungen reduziert. Somit wird sowohl die Qualität als auch die Nachvollziehbarkeit der Bewertung von Textorial verbessert (Stahlhut, König, Berger & Debelak, 2026).

Gerade diese Transparenz ist ein entscheidender Vorteil gegenüber vielen KI-Anwendungen. Lehrpersonen erhalten neben der Punktzahl eine dokumentierte Fehlerliste, ein darauf abgestimmtes Feedback sowie Verbesserungsvorschläge. Diese getrennten Bestandteile machen die Auswertung transparenter und ermöglichen eine fachliche Einordnung. Die pädagogische Beurteilung bleibt Aufgabe der Lehrperson.

Wissenschaftlich geprüft

Textorial entstand im Rahmen des Innosuisse-Projekts «Innovative AI Approaches for Language Proficiency Testing in Switzerland», das vom IBE gemeinsam mit der Universität Zürich und der EPFL Lausanne durchgeführt wurde. Ziel des Projekts war es, KI-Verfahren zu entwickeln, die Schülertexte nicht nur effizient, sondern auch wissenschaftlich fundiert, fair und nachvollziehbar bewerten.

Von Beginn an wurden deshalb klare Qualitätsziele definiert. Während der zweijährigen Projektphase wurden diverse Modelle anhand wissenschaftlicher Kriterien evaluiert. Massgebend für die Freigabe zum produktiven Einsatz war das Erreichen der festgelegten Anforderungen. Dies gilt auch für zukünftige Modellversionen.

Die wissenschaftliche Evaluation stützte sich auf die beschriebenen 3'238 Schülertexte in Deutsch, Englisch und Französisch. Überprüft wurden insbesondere folgende Qualitätsmerkmale:

- die Übereinstimmung mit menschlichen Bewertungen (Korrelation zwischen Modell und Fachperson $r \geq 0.75$)
- der systematische Bias (Cohen's $|dz| \leq 0.30$)
- die Fairness gegenüber verschiedenen Schülergruppen (Differential Item Functioning $|DIF| \leq 0.20$)
- die Stabilität der Bewertungen bei wiederholter Auswertung desselben Textes
- die Plausibilität der vertikalen Skala über die Klassenstufen hinweg (Anstieg der Bewertung von der 3. Klasse der Primarstufe bis zur 3. Klasse der Sekundarstufe I)

Die Qualität des Systems wird nicht behauptet, sondern fortlaufend gemessen, dokumentiert und anhand wissenschaftlicher Kriterien überprüft.

Psychometrische Modelle führen zu präziseren Bewertungen

Die erste Forschungsfrage betraf die Genauigkeit automatischer Bewertungen: Liefert ein Sprachmodell genauere Ergebnisse, wenn es psychometrisch geschätzte Fähigkeitswerte statt einzelner menschlicher Bewertungen vorhersagt?

Die Ergebnisse zeigen einen deutlichen Vorteil dieses Ansatzes: Das für Textorial eingesetzte Modell erreicht Korrelationen zwischen r = 0.91 und r = 0.99 mit den menschlichen Beurteilungen (Debelak & Vignoud, 2026). Alle acht Kriterien liegen in allen drei Sprachen über dem gesetzten Qualitätsziel von r ≥ 0.75.

Im direkten Vergleich erwies sich dieser Ansatz als genauer und fairer als ein ansonsten identisches Modell, das unmittelbar auf einzelnen menschlichen Bewertungen trainiert wurde (Debelak, Vignoud & Tomasik, 2026).

Abbildung 1 zeigt die hohe Übereinstimmung des eingesetzten Bewertungsmodells mit den menschlichen Beurteilungen. Der Vergleich mit einem direkt auf Bewertungen trainierten Modell findet sich bei Debelak, Vignoud & Tomasik (2026).

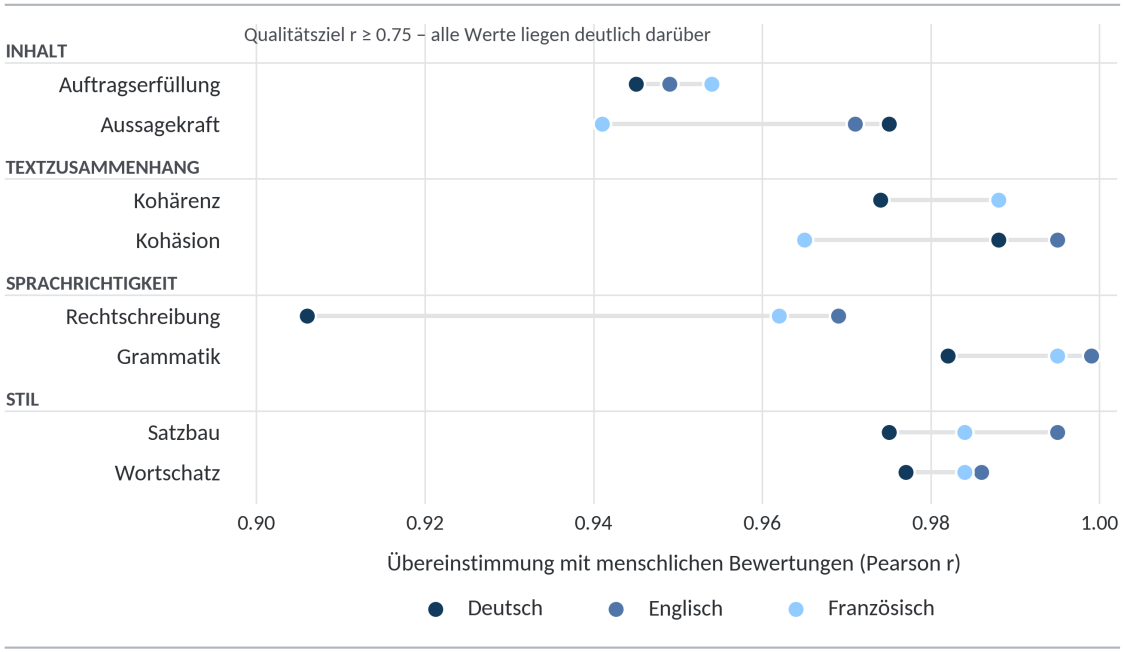


Abbildung 1: *Übereinstimmung von Textorial mit menschlichen Bewertungen nach Kriterien und Sprache (Pearson-Korrelationen), Debelak & Vignoud (2026)*

Eine klassenstufenübergreifende Skala macht Entwicklungen sichtbar

Die zweite Forschungsfrage betraf die Generalisierbarkeit des Verfahrens über verschiedene Klassenstufen hinweg. Durch die zugrunde liegende psychometrische Skalierung ist es möglich, Schreibkompetenzen auf einer klassenstufenübergreifenden Kompetenzskala abzubilden und Entwicklungen über mehrere Schuljahre hinweg zu verfolgen.

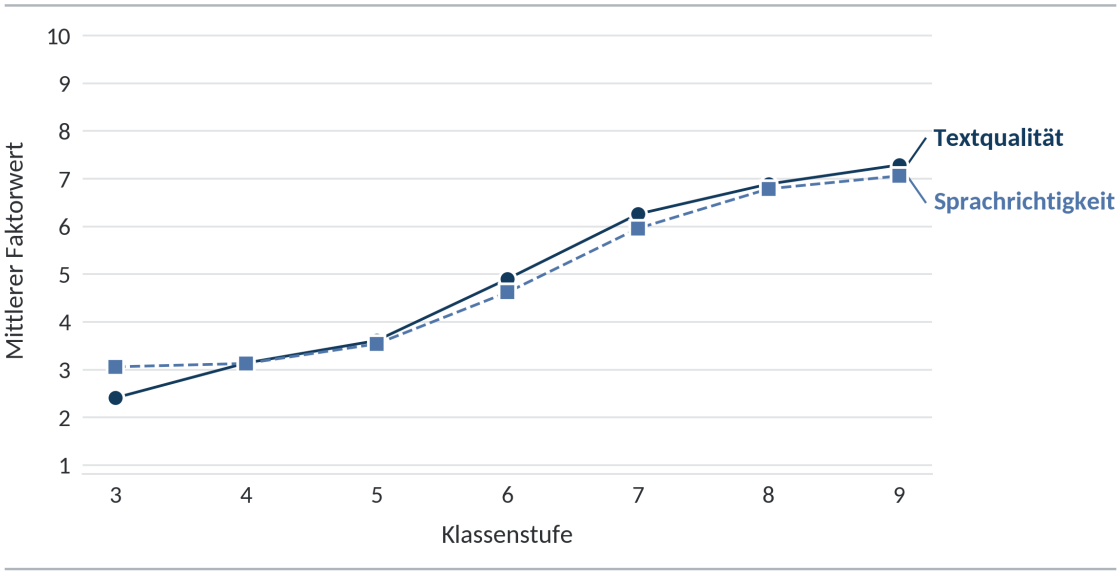


Abbildung 2: *Entwicklung der Schreibkompetenz über die Klassenstufen 3–9 auf der klassenstufenübergreifenden vertikalen Skala, IBE (2026)*

Erwartungsgemäss erzielten Schülerinnen und Schüler höherer Klassenstufen höhere Skalenwerte als jene tieferer Klassenstufen. In beiden Dimensionen – Textqualität und Sprachrichtigkeit – stiegen die mittleren Skalenwerte über die Klassenstufen hinweg kontinuierlich an.

Bewertungen sind über verschiedene Gruppen hinweg fair

Faire Bewertungen setzen voraus, dass für alle Lernenden derselbe Massstab gilt. Deshalb wurde als dritte Forschungsfrage untersucht, ob das Bewertungsmodell die Schreibkompetenz unabhängig von Erstsprache und Geschlecht der Schülerinnen und Schüler nach demselben Massstab erfasst. In der psychometrischen Forschung wird diese Eigenschaft als Messinvarianz bezeichnet. Sie ist eine grundlegende Voraussetzung dafür, Leistungen fair miteinander vergleichen zu können.

Die Fairness des Bewertungsmodells wurde auf zwei Ebenen geprüft. Psychometrische Analysen zeigen, dass das Modell Textqualität und Sprachrichtigkeit über Geschlecht, Erstsprache und Klassenstufe hinweg vergleichbar erfasst. Für die Klassenstufen 8 und 9 wurde diese Vergleichbarkeit zusätzlich nach Kanton bestätigt (Debelak, Vignoud & Tomasik, 2026).

Darüber hinaus wurde untersucht, wie stark die Bewertungen von Textorial in verschiedenen Gruppen mit den Beurteilungen geschulter Fachpersonen übereinstimmen. Wie Abbildung 3 zeigt, war diese Übereinstimmung für die nach Geschlecht und Erstsprache gebildeten Gruppen nahezu gleich hoch. Die Befunde geben somit keinen Hinweis darauf, dass einzelne Gruppen systematisch anders beurteilt werden.

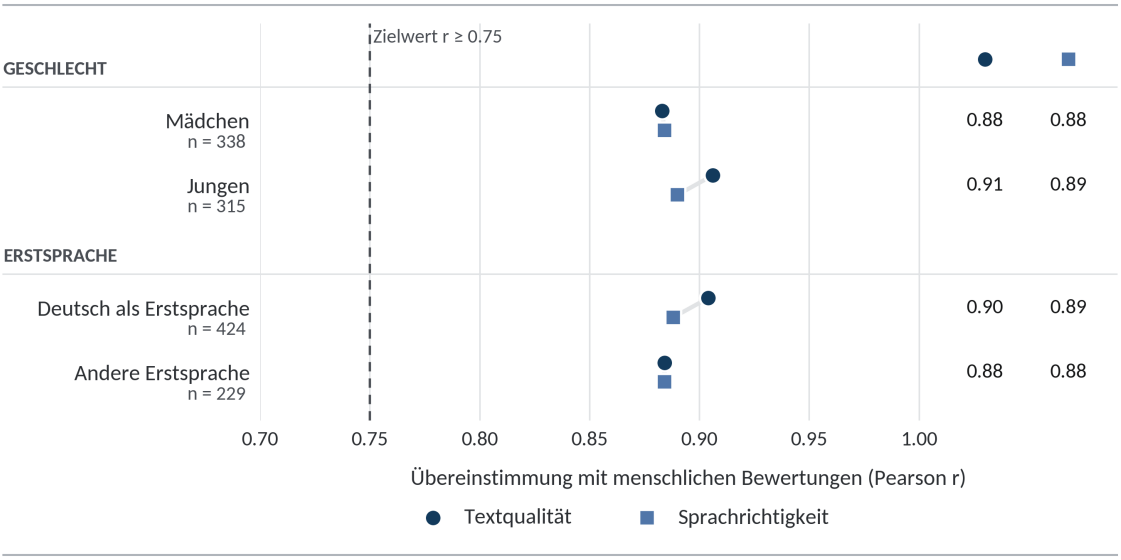


Abbildung 3: *Übereinstimmung von Textorial mit menschlichen Urteilen nach Geschlecht und Erstsprache (Pearson r), Debelak & Vignoud (2026)*

Robustheit bei Textveränderungen und Manipulationsversuchen

Eine automatische Bewertung muss nicht nur unter normalen Bedingungen zuverlässig sein. Sie muss auch angemessen auf ungewöhnliche oder fehlerhafte Texte reagieren und gezielten Manipulationsversuchen standhalten. Im Rahmen der Evaluation wurde deshalb untersucht, wie das Bewertungsmodell auf systematisch veränderte Schülertexte reagiert und ob sich seine Bewertung durch eingeschleuste Anweisungen beeinflussen lässt.

Die Untersuchung ging zwei Fragen nach: Wie verändert sich die Bewertung, wenn einzelne Merkmale eines Textes gezielt verändert werden? Und lässt sich das System dazu bringen, eine höhere Bewertung zu vergeben, als es der Text rechtfertigt?

Zur Beantwortung der ersten Frage wurden fünf Veränderungen geprüft: durchgehende Kleinschreibung, eine ausgetauschte Aufgabenstellung, vertauschte Buchstaben bei 20 Prozent der Wörter, entfernte Konjunktionen sowie 15 Prozent zufällig entfernte Wörter.

Das Bewertungsmodell reagierte auf diese Eingriffe insgesamt plausibel: Geringfügige Veränderungen beeinflussten die Bewertung nur wenig, stärkere Beeinträchtigungen senkten sie deutlich.

Durchgehende Kleinschreibung reduzierte den Mittelwert der Textqualität lediglich von 5.3 auf 5.1. Wurden dagegen 15 Prozent der Wörter entfernt, sank er auf 3.4. Besonders aufschlussreich ist das Ergebnis bei den vertauschten Buchstaben: Die Sprachrichtigkeit sank von 5.3 auf 3.9, die Textqualität dagegen nur auf 4.9. Das Modell reagierte somit unterschiedlich auf Beeinträchtigungen der Sprachrichtigkeit und der Textqualität – wie es das zweidimensionale Bewertungsmodell vorsieht.

Zur Beantwortung der zweiten Frage wurden vier gezielte Manipulationsversuche durchgeführt. Bei den ersten beiden wurde das Modell direkt zur Vergabe der Höchstpunktzahl beziehungsweise zur Offenlegung der Bewertungsvorgaben aufgefordert. Zwei weitere Varianten ver-

suchten mithilfe manipulierter technischer Textmarken, die Trennung zwischen Schülertext und Systemanweisung zu umgehen.

Die ersten beiden Manipulationsversuche blieben erfolglos: Das Modell behandelte die Eingaben als unzureichende Texte und vergab Werte nahe dem Minimum. Die beiden Angriffe über manipulierte Textmarken waren dagegen nahezu durchgehend erfolgreich und führten in beiden Dimensionen zur Höchstbewertung. Damit wurde eine sicherheitsrelevante Schwachstelle der untersuchten Modellversion identifiziert.

Die offene Dokumentation dieses Befunds ist Teil der wissenschaftlichen Qualitätssicherung. Aus der Analyse wurden konkrete Anforderungen an die technische Absicherung abgeleitet und implementiert. Dazu gehören die Prüfung und Bereinigung eingereichter Texte, die Erkennung verdächtiger Textmarken und manipulativer Anweisungen, zusätzliche Trainingsdaten mit gezielten Angriffen sowie ein unabhängiges Sicherheitssystem. Robustheit wird deshalb – ebenso wie Genauigkeit und Fairness – nicht als einmalig erreichte Eigenschaft, sondern als fortlaufende Aufgabe der Qualitätssicherung verstanden.

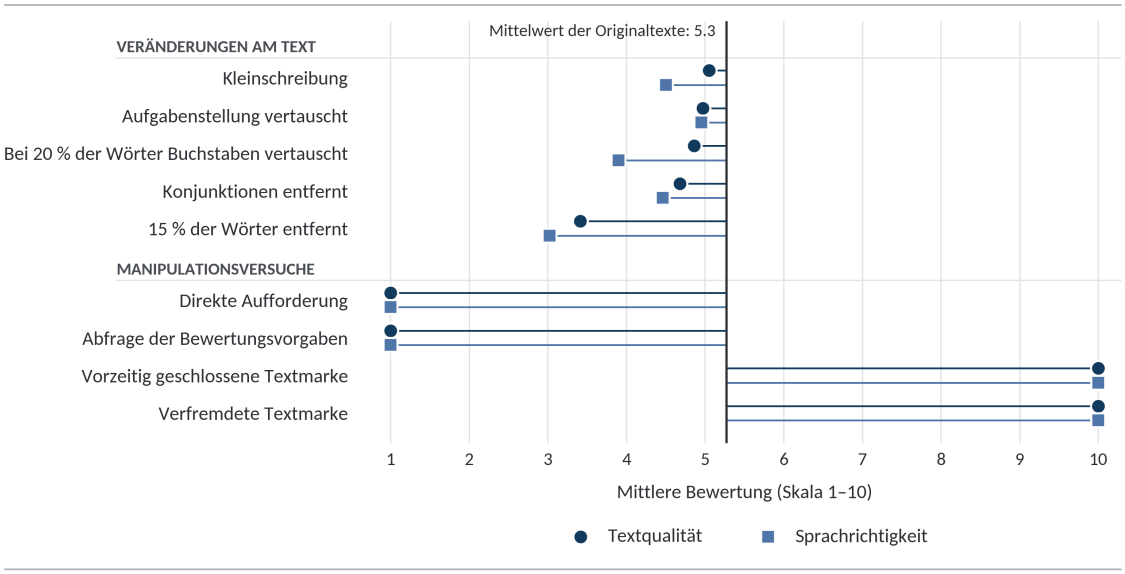


Abbildung 4: *Mittlere Bewertung je Veränderung und Manipulationsversuch im Vergleich zum Mittelwert der unveränderten Texte (Skala 1–10, dreisprachige Evaluation), Debelak & Vignoud (2026)*

Zuverlässig

Hohe Übereinstimmung mit den Beurteilungen geschulter Fachpersonen

Fair

Vergleichbare Genauigkeit über Geschlecht und Erstsprache hinweg

Entwicklungsorientiert

Sichtbare Entwicklung der Schreibkompetenz über mehrere Klassenstufen hinweg

Datensouveränität in der Schweiz

Schülerinnen und Schüler formulieren in ihren Texten teilweise sehr persönliche Gedanken und Ideen. Schulen und Bildungsbehörden müssen deshalb darauf vertrauen können, dass diese Daten sorgfältig verarbeitet und vor unbefugten Zugriffen geschützt werden. Deshalb wurde Datenschutz bei Textorial von Beginn an als grundlegendes Entwicklungsprinzip verankert.

Die gesamte Verarbeitung erfolgt auf einer redundanten GPU-Infrastruktur in der Schweiz. Sowohl die Transkription handschriftlicher Texte als auch die Bewertung durch Textorial und die Generierung der Rückmeldungen werden unter Verantwortung des IBE auf Infrastruktur der Schweizer Anbieter Exoscale und oriented.net ausgeführt. Die Texte von Lernenden werden ausschliesslich in der Schweiz verarbeitet und nicht an externe KI-Anbieter übermittelt.

Diese Unabhängigkeit ist möglich, weil Textorial offen verfügbare Sprachmodelle verwendet. Für die Bewertung kommt ein für Textorial speziell feinabgestimmtes (engl. fine-tuned) Modell der Familie Mistral Small 3.1 zum Einsatz. Für die Generierung der Rückmeldungen und die Transkription handschriftlicher Texte wird ein Modell der Familie Gemma 4 verwendet. Die für Textorial angepassten Modellversionen werden vom IBE selbst betrieben und für ihre jeweiligen Aufgaben weiterentwickelt. Textorial ist deshalb nicht auf Schnittstellen internationaler Technologieunternehmen angewiesen. Die gesamte KI-Infrastruktur steht unter der Kontrolle des IBE.

Die für die Entwicklung verwendeten Texte wurden vor dem Training anonymisiert. Namen von Personen, Ortsbezeichnungen, Datumsangaben und weitere personenbezogene Informationen wurden entfernt beziehungsweise maskiert.

Die gewählte Architektur bietet Schulen und Bildungsbehörden nicht nur ein hohes Datenschutzniveau, sondern auch langfristige digitale Souveränität. Änderungen der Nutzungsbedingungen oder Datenschutzrichtlinien internationaler KI-Anbieter haben keinen Einfluss auf den Betrieb von Textorial.

Auch für Schulen in Deutschland und Österreich bietet dieser Ansatz Vorteile. Die Schweiz verfügt über ein von der Europäischen Union als angemessen anerkanntes Datenschutzniveau.

Zusammen mit der transparenten Systemarchitektur schafft dies gute Voraussetzungen für den datenschutzkonformen Einsatz von Textorial ausserhalb der Schweiz. Die erforderlichen vertraglichen Regelungen, beispielsweise zur Auftragsverarbeitung, werden den jeweiligen nationalen Anforderungen entsprechend bereitgestellt.

> **Geschützt**
>
> Texte von Lernenden bleiben in der Schweiz und werden nicht an externe KI-Anbieter übermittelt.

> **Souverän**
>
> Vom IBE betriebene Modelle schaffen Unabhängigkeit von externen KI-Schnittstellen.

Ausblick und Weiterentwicklung

Seit August 2026 steht Version 1.0 von Textorial Schulen in der Schweiz, Deutschland und Österreich zur Verfügung. Dieser Schritt markiert zugleich den Beginn einer kontinuierlichen wissenschaftlichen Weiterentwicklung.

Im Zentrum steht der weitere Ausbau der Datenbasis, um die Genauigkeit von Textorial über die gesamte obligatorische Schulzeit hinweg weiter zu erhöhen. Parallel dazu werden neue Sprachmodelle, Verfahren der künstlichen Intelligenz und sprachtechnologische Komponenten laufend wissenschaftlich evaluiert und – wenn sie nachweislich zu besseren Ergebnissen führen – schrittweise in das System integriert.

Die modulare Architektur von Textorial ermöglicht es, einzelne Komponenten weiterzuentwickeln, ohne die Stabilität und Vergleichbarkeit der Bewertungen zu beeinträchtigen.

Mit dem Einsatz in weiteren Sprachregionen und Ländern entstehen gleichzeitig neue Möglichkeiten für Forschung und Qualitätsentwicklung.

Bei der Weiterentwicklung bleibt der pädagogische Nutzen leitend: Textorial soll Lehrpersonen entlasten und ihnen mehr Raum für die gezielte Begleitung von Schreibprozessen eröffnen.

Schulen können Textorial während 30 Tagen kostenlos erproben und seine Möglichkeiten im Unterricht kennenlernen.

Literatur

Charalampous, A. & Darra, M. (2023). The Contribution of Teacher Feedback to Learners' Work Revision: A Systematic Literature Review. World Journal of Education, 13(3), 40–63.

Debelak, R. & Vignoud, J. T. T. (2026). Evaluation report: Fine-tuned Mistral Small 3.1 automated essay scoring pipeline. Technischer Bericht, Innosuisse-Projekt 119.542 IP-SBM. Universität Zürich / EPFL.

Debelak, R., Vignoud, J. T. T., & Tomasik, M. J. (2026). Using psychometric modeling to improve the effectiveness of automated essay scoring systems. Manuskript zur Publikation eingereicht.

Fleckenstein, J., Liebenow, L. W. & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. Frontiers in Artificial Intelligence, 6:1162454.

Hattie, J. & Timperley, H. (2007). The Power of Feedback. Review of Educational Research, 77(1), 81–112.

IBE (2026). Vertikale Skala 2026 [Unveröffentlichte Rohdaten]. Institut für Bildungsevaluation, Zürich.

König, N., Stahlhut, L., Berger, S. & Debelak, R. (2026, Juni). From ratings to formative feedback: An AI-based system for automated essay scoring [Poster]. 21st EARLI SIG Writing Conference, Winterthur.

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A. & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: Artificial Intelligence, 6, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Shermis, M. D., Garvan, C. W. & Diao, Y. (2008). The Impact of Automated Essay Scoring on Writing Outcomes. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Stahlhut, L., König, N., Berger, S. & Debelak, R. (2026, Juni). Enhancing automated essay scoring with rule-based error detection [Vortrag]. 21st EARLI SIG Writing Conference, Winterthur.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M. & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894.

Wang, Z. & Han, F. (2022). The Effects of Teacher Feedback and Automated Feedback on Cognitive and Psychological Aspects of Foreign Language Writing: A Mixed-Methods Research. Frontiers in Psychology, 13, 909802.

Gute Schreibförderung beginnt dort, wo häufiges Schreiben und hochwertiges Feedback selbstverständlich werden.

Testen Sie Textorial während 30 Tagen kostenlos im Unterricht.

Zum 30-Tage-Test



KONTAKT

Nina König
Data Analyst & Software Manager

nina.koenig@ibe-edu.ch
www.ibe-edu.ch · 043 268 39 60