Editorial

Large Language Models im Kontext formativen und summativen Feedbacks

Olaf Köller

Die rasante Entwicklung generativer Künstlicher Intelligenz (KI), insbesondere von Large Language Models (LLMs), hat in den vergangenen Jahren zu tiefgreifenden Veränderungen in Bildungskontexten geführt. Insbesondere im Bereich des Feedbacks – einer der zentralen Einflussgrößen für Lernen und Leistungsentwicklung – eröffnen sich neue Möglichkeiten, sowohl für die Forschung als auch für die Praxis. Dieses Sonderheft der *Psychologie in Erziehung und Unterricht* widmet sich der Frage, welches Potenzial LLMs für formative und summative Feedbackprozesse besitzen, welche Chancen sich daraus für Lehr-Lern-Prozesse ergeben und welche Grenzen sowie offene Forschungsfragen bestehen.

Feedback gilt seit Langem als einer der wirksamsten Hebel für Lernprozesse. Es unterstützt Lernende dabei, Diskrepanzen zwischen dem aktuellen Leistungsstand und den Lernzielen zu erkennen und zu schließen. Während formatives Feedback vor allem auf Lernprozesse abzielt und kontinuierlich zur Verbesserung beiträgt, dient summatives Feedback der Bewertung von Leistungen am Ende eines Lernprozesses. Beide Formen sind für erfolgreiche Bildungsprozesse essenziell, stellen Lehrkräfte jedoch vor erhebliche Herausforderungen: Die Generierung qualitativ hochwertigen Feedbacks ist zeitaufwendig, erfordert eine hohe diagnostische Kompetenz und ist häufig durch Ressourcenknappheit begrenzt.

Hier setzen LLMs an. Sie versprechen, Feedback in großer Skalierbarkeit, zeitnah und kontextsensibel bereitzustellen. Erste Studien zeigen, dass KI-generiertes Feedback in bestimmten Kontexten qualitativ mit menschlichem Feedback vergleichbar oder sogar überlegen sein kann (vgl. z.B. für summatives Feedback Choi, Tate, Ritchie, Nixon & Warschauer, 2025; und für formatives Feedback Schiller, Fleckenstein, Mertens, Horbach & Meyer, 2024). Gleichzeitig ist jedoch deutlich geworden, dass die Qualität und Wirkung dieses Feedbacks von einer Vielzahl von Faktoren abhängt – darunter das verwendete LLM, die Gestaltung der Prompts sowie individuelle Merkmale der Lernenden.

Die Beiträge dieses Sonderhefts beleuchten diese Aspekte aus unterschiedlichen Perspektiven und tragen so zu einem differenzierten Verständnis des Potenzials von LLMs im Bildungsbereich bei. Dabei werden sowohl formative als auch summative Einsatzszenarien adressiert, ebenso Fragen der Qualität, Stabilität, Fairness und Individualisierung von Feedback.

Der Beitrag von Jacobsen, Pargmann, Rohlmann und Weber untersucht, welche Faktoren die Qualität von KI-generiertem Feedback in der Lehrkräftebildung beeinflussen. Ausgangspunkt ist die zentrale Rolle des Feedbacks für den Kompetenzerwerb angehender Lehrkräfte, wobei zugleich ein Mangel an qualitativ hochwertigem Feedback besteht. LLMs bieten hier neue Möglichkeiten, da sie schnell und kontextsensibel Rückmeldungen erzeugen können. Allerdings hängt deren Qualität stark von zwei Faktoren ab: der Wahl des Modells und dem Promptdesign. In zwei quasi-experimentellen Studien analysierten die Autor:innen, wie unterschiedliche Promptmerkmale und verschiedene LLMs die Feedbackqualität beeinflussen. Lehramtsstudierende formulierten Lernziele, zu denen Feedback durch verschiedene Modelle

Psychologie in Erziehung und Unterricht, 2026, *73*, 134–136
DOI 10.2378/peu2026.art11d

generiert wurde. Die Ergebnisse zeigen, dass sowohl die Modellwahl als auch die Gestaltung des Prompts signifikante Prädiktoren für die Qualität sind. Besonders wirksam erwies sich der Einsatz von Fachsprache, während auch einfache Promptprinzipien bereits deutliche Verbesserungen bewirken können. Die Studie leistet einen wichtigen Beitrag, indem sie Promptstrategien in ein pädagogisch anschlussfähiges Modell (3K-Modell) überführt. Sie zeigt zudem, dass KI-Literacy – verstanden als Fähigkeit zur Auswahl geeigneter Modelle und zur Gestaltung effektiver Prompts – eine zentrale Kompetenz für Lehrkräfte darstellt. Insgesamt verdeutlicht die Arbeit, dass gezieltes Prompting und reflektierte Modellwahl entscheidend sind, um das Potenzial von KI-Feedback im Bildungskontext auszuschöpfen.

Die Studie von Jansen, Tanz, Pünjer, Schaller und Höft untersucht, für welche Schüler:innen KI-basiertes Feedback die Selbstwirksamkeitsüberzeugungen beim Schreiben von Texten besonders fördert. Ausgangspunkt ist die Bedeutung der Selbstwirksamkeitsüberzeugungen für den Erwerb von Schreibkompetenzen im Sekundarschulbereich. Feedback – sowohl selbstgeneriertes als auch externes, etwa durch KI – kann diese Überzeugungen stärken, wirkt jedoch nicht für alle Lernenden gleich. In einem experimentellen Design mit über 800 Schüler:innen verfassten die Teilnehmenden argumentative Texte und erhielten entweder KI-generiertes Feedback oder erzeugten eigenes Feedback durch erneutes Lesen ihrer Texte. Analysiert wurde, wie individuelle Unterschiede – insbesondere Zielorientierungen (Leistungs- oder Lernziele) und Schreibfähigkeiten – die Wirkung des Feedbacks moderieren. Die Ergebnisse zeigen differenzierte Effekte: Schüler:innen mit hohen Leistungszielen profitieren weniger von KI-Feedback als von Selbstfeedback, während bei niedrigen Leistungszielen der gegenteilige Effekt auftritt. Besonders deutlich sind negative Effekte bei Leistungs-Vermeidungszielen. Zudem zeigt sich, dass das KI-Feedback bei Lernzielorientierung in Kombination mit niedrigen Schreibfähigkeiten positive Effekte hat, bei hohen Schreibfähigkeiten jedoch negative. Die Studie verdeutlicht, dass KI-Feedback nicht universelle Effekte auf Selbstwirksamkeitsüberzeugungen hat, sondern diese stark von individuellen Voraussetzungen abhängen. Daraus folgt die Notwendigkeit, Feedback stärker zu personalisieren und an die Zielorientierungen sowie die Kompetenzen der Lernenden anzupassen, um optimale Lern- und Motivationseffekte zu erzielen.

Der Beitrag von Horbach, Melanchthon, Schaller, Keller, Meyer und Jansen analysiert die Leistungsfähigkeit automatisierter Essaybewertungssysteme (AES) und erweitert die bisherige Forschung um die wichtige Dimension der Stabilität. Während frühere Studien vor allem die Genauigkeit von Modellen betrachteten, fokussiert diese Arbeit zusätzlich auf die Konsistenz der Bewertungen über mehrere Durchläufe hinweg. Untersucht wurden drei Modelltypen: featurebasierte Regressionsmodelle, neuronale Transformer-Modelle sowie generative Large Language Models. Grundlage bildet ein Datensatz mit über 4.500 englischsprachigen Lernenden-Texten. Neben der Vorhersagegenauigkeit wurde analysiert, wie stark die Bewertungen zwischen wiederholten Modellläufen variieren und in welchem Maße verschiedene Modelle übereinstimmen. Die Ergebnisse zeigen, dass klassische featurebasierte Modelle weiterhin konkurrenzfähig sind, während auch LLMs eine hohe Genauigkeit erreichen. Allerdings weisen insbesondere generative Modelle wie GPT-5 eine deutliche Variabilität bei ihren Bewertungen auf, was ihre Zuverlässigkeit einschränken kann. Zudem wird deutlich, dass verschiedene Modelltypen unterschiedliche Stärken haben und dass sie dieselben Texte nicht gleich gut bewerten. Die Studie betont daher die Bedeutung der Stabilität als zentrale Voraussetzung für den Einsatz in Bildungskontexten. Sie empfiehlt eine sorgfältige Modellwahl sowie die Kombination verschiedener Modelle, um sowohl Genauigkeit als auch Konsistenz zu verbessern und damit vertrauenswürdige Bewertungssysteme zu entwickeln.

Die Arbeit von Föste-Eggers, Schmidt, List, Glüsing und Fleckenstein untersucht die Potenziale und Grenzen von LLMs bei der Bewertung deutschsprachiger argumentativer Texte von Schüler:innen der Sekundarstufe I. Ausgangspunkt sind die hohe Komplexität und der große zeitliche Aufwand menschlicher Textbewertung, die zudem anfällig für Verzerrungen und Inkonsistenzen ist. LLMs bieten hier eine potenzielle Entlastung; ihre Leistungsfähigkeit im deutschsprachigen Kontext ist jedoch bislang nur wenig erforscht. Im Rahmen der Studie wurden 1.000 Schüler:innentexte sowohl von menschlichen Rater:innen als auch von verschiedenen LLMs anhand von Bewertungsrubriken beurteilt. Analysiert wurden insbesondere die Übereinstimmung zwischen menschlichen und KI-basierten Bewertungen, Urteilstendenzen sowie Fairnessaspekte. Die Ergebnisse zeigen, dass LLMs bei holistischen Gesamtbewertungen eine gute Übereinstimmung mit menschlichen Urteilen erreichen. Schwierigkeiten bestehen hingegen bei analytischen Dimensionen, insbesondere bei sprachlichen Feinheiten. Zudem zeigen sich Unterschiede in den Bewertungsmustern der eingesetzten Modelle. Hinweise auf systematische Verzerrungen gegenüber bestimmten Schüler:innengruppen wurden jedoch nicht gefunden. Insgesamt wird das Potenzial von LLMs für summative Bewertungen hervorgehoben, zugleich aber auf bestehende Grenzen hingewiesen. Die Studie unterstreicht die Notwendigkeit weiterer Forschung, insbesondere zur Verbesserung analytischer Bewertungsfähigkeiten und zur Sicherstellung fairer und reliabler Beurteilungen im schulischen Kontext.

Die Beiträge dieses Sonderhefts zeigen eindrucksvoll, dass LLMs ein erhebliches Potenzial für die Unterstützung von Feedbackprozessen im Bildungsbereich besitzen. Sie können sowohl im formativen als auch im summativen Kontext eingesetzt werden und bieten neue Möglichkeiten der Skalierung, Individualisierung und Effizienzsteigerung. Gleichzeitig wird deutlich, dass ihr Einsatz sorgfältig gestaltet werden muss. Die Qualität von KI-Feedback hängt maßgeblich von der Kompetenz der Nutzenden ab, insbesondere im Hinblick auf Prompting und Modellwahl. Darüber hinaus sind individuelle Unterschiede der Lernenden zu berücksichtigen, um wirksames Feedback zu gewährleisten. Im summativen Bereich stellen Fragen der Stabilität, Fairness und Validität zentrale Herausforderungen dar.

Für die zukünftige Forschung ergeben sich mehrere zentrale Desiderate: die Entwicklung adaptiver Feedbacksysteme, die Integration von LLMs in didaktische Konzepte, die Verbesserung der Stabilität automatisierter Bewertungen sowie die Förderung von KI-Literacy bei Lehrkräften und Lernenden. Nur durch eine enge Verzahnung von technologischer Entwicklung und bildungswissenschaftlicher Forschung kann das Potenzial von LLMs verantwortungsvoll und effektiv genutzt werden.

Literatur

Choi, J., Tate, T., Ritchie, D., Nixon, N. & Warschauer, M. (2025). *Anchor is the key: Toward accessible automated essay scoring with Large Language Models through prompting.* https://doi.org/10.35542/osf.io/cbhgz_v1

Schiller, R., Fleckenstein, J., Mertens, U., Horbach, A. & Meyer, J. (2024). Understanding the effectiveness of automated feedback: Using process data to uncover the role of behavioral engagement. *Computers & Education, 223,* 105163. https://doi.org/10.1016/j.compedu.2024.105163