Le|se|räu|me
ZEITSCHRIFT FÜR LITERALITÄT
IN SCHULE UND FORSCHUNG

MAURICE FÜRSTENBERG

KI-Feedback auf dem Prüfstand – eine explorative Untersuchung maschineller Rückmeldungen zu Schüler:innentexten

1 | Forschungsfrage

Der Beitrag geht explorativ der Frage nach, inwieweit ein didaktisch modelliertes, generatives Sprachmodell in der Lage ist, inhaltlich richtige Rückmeldungen zu Texten von Schüler:innen zu produzieren. Mit dieser Frage ist übergeordnet der Wunsch verbunden, generative KI-Systeme könnten in die Lage versetzt werden, Lehrkräfte beim Verfassen von (lernförderlichen) Rückmeldungen zu Schüler:innentexten zu unterstützen – einer Tätigkeit, die einen erheblichen Aufwand für Lehrkräfte darstellt (Mußmann et al. 2016). Zwei Vorteile maschineller Rückmeldungen zu Texten sind die zeitliche Unmittelbarkeit und die große Menge: KI-Systeme erstellen auf Anfrage direkt nach oder auch schon während des Schreibprozesses Feedback, und zwar beliebig oft und zu einer beliebig hohen Anzahl an Texten. Dies können Lehrkräfte nicht leisten. Zudem entspricht ihr Feedback nicht unbedingt Kriterien effektiven Feedbacks (Müller et al. 2023). Auch mit Blick auf den grassierenden Lehrkräftemangel und der damit einhergehenden Überlastung der Lehrkräfte ist davon auszugehen, dass Lehrer:innen in Zukunft eher noch weniger Zeit für qualitativ hochwertige Rückmeldungen aufbringen können. Die Forschungsfrage hat daher nicht nur eine didaktische Zielrichtung, sondern ist auch bildungspolitisch von enormer Bedeutung.

2 | Methodik

Mit *didaktisch modellierten, generativen Sprachmodellen* werden explizit solche Systeme adressiert, bei denen ein Foundation-Model wie GPT4 oder Mistral Large 2 mit gezielten System-Prompts (Fürstenberg/Müller 2024, 8) für einen speziellen Einsatzzweck vorbereitet werden, z. B. Rückmeldungen zu Schüler:innentexten zu geben (unter anderem *Fobizz, CornelsenAI*, *fiete.ai – jetzt FelloFish*). Das Alleinstellungsmerkmal dieser KI-Anwendungen liegt in der konkreten Ausgestaltung der System-Prompts, über die entsprechend wenig bekannt ist. Das macht die Erforschung dieser Systeme besonders schwierig. Die Forschungsfrage kann daher lediglich rekonstruktiv angegangen werden. Dazu werden die Produkte eines KI-Systems untersucht und an die jeweiligen Schüler:innentexte rückgebunden, um sich der Beantwortung der Frage nach der Qualität der Rückmeldungen anzunähern. Als eine gängige Methode der Qualitätsbestimmung hat sich für quantitativ ausgerichtete Forschung die Praxis durchgesetzt, die maschinellen Rückmeldungen mit menschlichen abzugleichen (Seßler et al. 2024). Das ist unter anderem aus forschungsmethodischer und didaktischer Sicht problematisch, insbesondere

dann, wenn es um komplexe Rückmeldungen zu komplexen Konstrukten, wie Texten, geht. Eine Herausforderung liegt im Vergleich mit den menschlichen Rückmeldungen, die durch diese Methode (meist unreflektiert) zum Gold-Standard erhoben werden. Zudem geht durch die Quantifizierung qualitativ hochwertiger Daten eine zu große Menge an Informationen verloren – das einzelne Feedback und der einzelne Text geraten aus dem Blick. Daher geht dieser Beitrag der Frage nach der Qualität maschineller Rückmeldungen zu Schüler:innentexten explorativ-qualitativ nach.

Dazu werden die Rückmeldungen einer didaktisch ausgerichteten Anwendung, *fiete.ai* (Haverkamp et al. 2024), zu argumentativen Texten von 17 bayerischen Gymnasiast:innen einer 9. Klasse genauer untersucht. Die Texte wurden von den Schüler:innen an iPads in einer Doppelstunde verfasst, Materialanalyse und ein Schreibplan wurden in Stunden zuvor bereits erarbeitet. Hilfsmittel waren lediglich die Hefteinträge, in denen unter anderem die gemeinsam erarbeiteten Feedbackkriterien festgehalten sind. Das KI-System erhielt neben den Schüler:innentexten auch die Aufgabenstellung[1], in die Aufgabe integrierte Materialien sowie sieben Feedbackkriterien, auf denen die Rückmeldungen basieren (siehe unten). Neben dem hier im Fokus stehenden qualitativen Feedback erstellte die KI zu den Kriterien auch ein quantitatives Feedback (Likert-Skala [0,10]), das zum Vergleich auch von der Lehrkraft der Klasse gegeben wurde.

In der vorliegenden Untersuchung werden die qualitativen, maschinellen Rückmeldungen dargestellt und in Rückbindung an den konkreten Text des:der jeweiligen Schüler:in insofern beurteilt, als bewertet wird, wie inhaltlich zutreffend diese Rückmeldung ist.

3 | Ergebnisse und Diskussion

Aus Platzgründen werden die meisten Ergebnisse der Untersuchung an anderer Stelle publiziert (Fürstenberg i. V.). Es werden in der Folge exemplarisch einige der Textteile (orthografisch unverändert) gemeinsam mit den Urteilen des KI-Systems dargestellt und besprochen, um Thesen zu generieren, wo mögliche Herausforderungen für den Einsatz maschineller Rückmeldungen zu Texten liegen. Der vorliegende Beitrag fokussiert Rückmeldungen zu den Textteilen *Einleitung* und *Synthese*.

3.1 | Feedback zum Textteil *Einleitung*

Die Anforderungen für die Einleitung lauteten:

> Die Einleitung steigt mit einer sprachlichen Auffälligkeit (Frage, These, Zitat) ein, skizziert die Thematik und stellt kurz die Kontroverse dar. Sie führt dann knapp zur Themafrage hin, mit der in den Hauptteil übergeleitet wird.

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

Abbildung 1 zeigt zuerst die Einleitung eines:einer Schüler:in und das KI-generierte Feedback dazu:

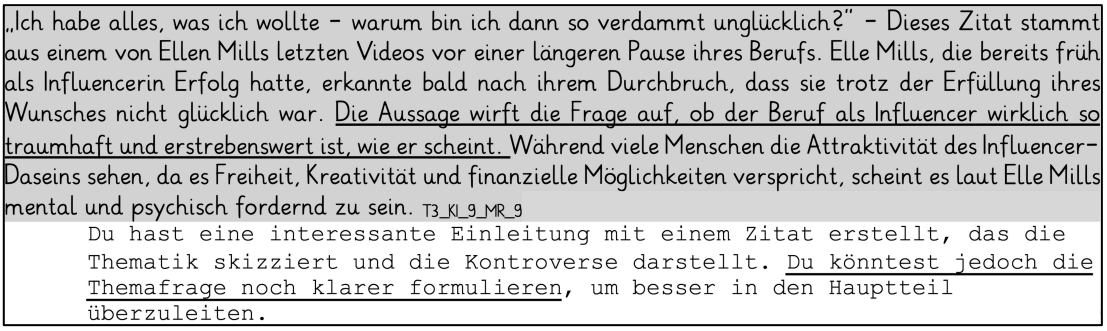

Abbildung 1: Einleitung und KI-generiertes Feedback I

Die Einleitung ist mit Blick auf die Kriterien gut gelungen und quantitativ liegt sowohl das menschliche als auch das maschinelle Rating bei 8/10 Punkten. Die qualitative Kritik der Maschine trifft zwar den richtigen Aspekt (Themafrage als Überleitung), ist jedoch in der Anweisung unpräzise, denn die Abweichung des Schüler:innentextes vom Feedbackkriterium ist nicht die Klarheit der Formulierung, sondern die Positionierung der Themafrage, die nicht ans Ende der Einleitung gestellt wurde.

Die folgende Einleitung erhält von der KI 7/10, vom menschlichen Rater hingegen 3/10 Punkte:

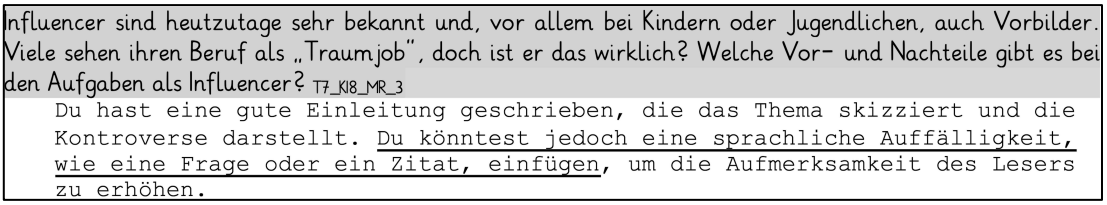

Abbildung 2: Einleitung und KI-generiertes Feedback II

Die Einleitung beginnt mit einer allgemeinen Information, die als Hinführung zum Thema interpretiert werden kann, und stellt im Anschluss die Themafrage sowie eine verwandte Frage, welche den antithetischen Charakter des restlichen Textes ankündigt. Damit fehlen neben einer sprachlichen Auffälligkeit zu Beginn, welche durch das Sprachmodell auch kritisiert wird (allerdings wieder, ohne auf die Position hinzuweisen), die Skizzierung der Thematik, die Kontroverse sowie die Hinführung zur Themafrage. Daher ist die maschinelle Bepunktung nicht nachvollziehbar. Das Sprachmodell moniert folglich korrekterweise die fehlende sprachliche Auffälligkeit, geht aber nicht auf die restlichen fehlenden Bestandteile ein und hat bei beiden Einleitungen ein *Lokalisationsproblem*, da Kritik nicht an konkrete Stellen im Text rückgebunden wird, was ein Kriterium lernförderlichen Feedbacks darstellt (Haverkamp et al. 2024).

3.2 | Feedback zum Textteil *Synthese*

Der (Bewertung der) Synthese sollten folgende Kriterien zugrunde gelegt werden:

> Die Synthese wägt vor dem Schluss nochmal die stärksten Argumente gegeneinander ab und begründet die eigene Position abschließend.

Alles in allem ist der Beruf des Influencers für mich persönlich ein Traumjob, da man ja sowieso zweimal überlegen sollte, ob man etwas postet, also sollten keine peinlichen Videos im Internet erscheinen und außerdem sollte man sich sowieso einen Videotypen aussuchen, der auch in Jahren noch Menschen begeistern kann.T15_KI1_MR7

```
Du hast keine Synthese geschrieben, in der du die stärksten Argumente
gegeneinander abwägst und deine eigene Position begründest. Dies ist ein
wichtiger Teil der Aufgabe.
```

Abbildung 3: Synthese und KI-generiertes Feedback I

Das KI-System erteilt dem Textteil in Abbildung 3 nur 1/10 und der menschliche Rater 5/10 Punkten, was einen der 45 Fälle (von 266 Ratings insgesamt) beschreibt, in denen der Mensch mehr Punkte zuweist als die KI. Der Versuch einer Synthese gelingt in dem Schüler:innentext mit Blick auf die Feedbackkriterien insofern, als die eigene Position dargestellt und zumindest auf sprachformaler Ebene abschließend begründet wird. Was fehlt (siehe auch qualitative Kritik), ist die Abwägung der beiden stärksten Argumente dieses Aufsatzes (fehlende Sicherheit vs. hoher Verdienst). Stattdessen wird im Text auf das schwächere der beiden Argumente eingegangen (Gefahr der öffentlichen Bloßstellung) und mithilfe eines Lösungsvorschlags versucht, die Eigenposition zu begründen. Die Begründung passt hier nicht unmittelbar zur These, ist aber zumindest ein Ansatz, die eigene Position zu erklären. Vor diesem Hintergrund ist die quantitative Bewertung durch die Maschine als zu streng einzuschätzen. Das qualitative Feedback hingegen trifft insofern zu, als das Abwägen der stärksten Argumente fehlt. Problematisch ist, dass das Feedback so formuliert ist, als hätte das Sprachmodell die Synthese gar nicht erst erkannt und auch die minimale Punktzahl lässt den Schluss zu.

Auch bei der folgenden Synthese waren sich Maschine und Mensch uneins:

Auf der einen Seite ist Influencer ein Beruf mit schnellen Aufstiegsmöglichkeiten und hohem Einkommen. Auf der anderen Seite ein sehr stressiger Job mit keiner abgesicherten Arbeitsstelle. Aus meiner Sicht ist ein sicherer Arbeitsplatz wichtiger als die Reichweite und das hohe Ansehen in der Video-Branche.T1_KI5_MR9

```
Die Synthese fehlt in deinem Text. Versuche, vor dem Schluss die
stärksten Argumente gegeneinander abzuwägen und deine eigene Position
abschließend zu begründen.
```

Abbildung 4: Synthese und KI-generiertes Feedback II

Das Sprachmodell moniert hier wieder die vermeintlich fehlende Synthese, gibt aber dennoch 5/10 Punkten (menschlicher Rater: 9/10). Die Begründung der eigenen Position könnte noch expliziter sein, aber rein strukturell ist die Synthese für einen ersten Versuch durchaus gelungen und nimmt auch die beiden stärksten Argumente des Textes auf.

Das KI-System meldet 9 Texten eine fehlende Synthese zurück, dies ist aber nur in einem dieser Fälle korrekt. Darüber hinaus fehlt in 2 weiteren Texten tatsächlich die Synthese, diese moniert die KI aber nicht. Nachdenklich stimmt auch, dass das vermeintliche Fehlen von Synthesen mit einer mittleren Bewertung von 5 Punkten einhergeht. Hier gehen qualitative und quantitative Bewertung der Maschine nicht Hand in Hand, was auf ein *internes Konsistenzproblem* hinweist. Solche Fehler haben in der Summe eine fatale Folge: Menschen schenken dem maschinellen Feedback *kein Vertrauen* – eine zentrale Herausforderung (Benk et al. 2024). Das Misstrauen zeigt sich auch in den Antworten zum Fragebogen (Fürstenberg i. V.), in denen mehrere Schüler:innen das KI-System wie folgt kritisieren:

> Generell war es schon ziemlich gut, aber die KI hat mir nicht sonderlich gefallen, da sie manches als falsch oder fehlend gekennzeichnet hat, was aber da war. Das macht einen dann unsicher ob das andere Feedback dann auch richtig ist.
> Item_8_Antwort_4

4 | Fazit

Ein didaktisch modelliertes, generatives Sprachmodell ist in der Lage, inhaltlich richtige Rückmeldungen zu Texten von Schüler:innen zu produzieren. Allerdings zeigte das KI-System Probleme bei der internen Konsistenz und der textuellen Lokalisation. Auch daraus folgte ein grundsätzliches Misstrauen seitens der Schüler:innen. Zudem bleibt fraglich, wie reliabel inhaltlich richtiges Feedback durch KI ist. Da zur Qualität maschineller Rückmeldungen zu menschlichen Texten, insbesondere im deutschsprachigen Raum, bisher kaum Daten vorliegen (Seßler et al. 2024), tragen die dargestellten Ergebnisse zur ersten Etablierung dieses Forschungsstandes bei.

5 | Literaturverzeichnis

Benk, Michaela / Kerstan, Sophie / von Wangenheim, Florian / Ferrario, Andrea (2024): Twenty-four years of empirical research on trust in AI. A bibliometric review of trends, overlooked issues, and future directions. In: AI & society. https://doi.org/10.1007/s00146-024-02059-y

Fürstenberg, Maurice (i. V.): Zur Qualität KI-generierten Feedbacks. Ein explorativer Vergleich menschlicher und künstlicher Intelligenzen. In: Müller, Hans-Georg / Fürstenberg, Maurice (Hg.): DeutschGPT 2.0. Deutschunterricht im Dialog mit Künstlicher Intelligenz. Berlin: Frank & Timme.

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

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

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

Mußmann, Frank / Riethmüller, Martin / Hardwig, Thomas / Peters, Stefan / Parciak, Marcel / Ohms, Ilka C. / Klötzer, Stefan (2016). Niedersächsische Arbeitszeitstudie Lehrkräfte an öffentlichen Schulen 2015/2016. Ergebnisbericht. Online unter https://www.gew-nds.de/fileadmin/media/sonstige_downloads/nds/Mehrarbeit/Niedersaechsische-Arbeitszeitstudie2015-2016-Endbericht.pdf (letzter Aufruf 12. Mai 2025).

Seßler, Kathrin / Fürstenberg, Maurice / Bühler, Babette / Kasneci, Enkelejda (2024): Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. arXiv. https://doi.org/10.48550/arXiv.2411.16337

Maurice Fürstenberg

Ludwig-Maximilians-Universität München
m.fuerstenberg@lmu.de