

Automatisiertes Assessment zur Unterstützung reflexiver Prozesse

Theoretischer Hintergrund

In den letzten Jahren hat *Machine Learning* (ML) zunehmend Einzug in die Bildungsforschung gehalten (Hilbert et al., 2021). Besonders im Bereich der Sprachverarbeitung eröffnen Entwicklungen im *Natural Language Processing* (NLP) neue Möglichkeiten, komplexe Lernprozesse automatisiert zu erfassen und zu bewerten (Jurafsky & Martin, 2025). Dadurch lassen sich Kompetenzen untersuchen, die bislang nur durch aufwendige manuelle Analysen zugänglich waren – etwa die Reflexionskompetenz angehender Lehrkräfte. Reflexionskompetenz ist zentral für die Professionalisierung von Lehrkräften (Baumert & Kunter, 2006) und die Qualitätssicherung des Unterrichts (Abels, 2011; KMK, 2004). Sie umfasst die Fähigkeit, eigene Unterrichtserfahrungen zu reflektieren, kritisch zu bewerten und daraus Schlüsse für zukünftiges Handeln abzuleiten (Wyss, 2013). Schriftliche Reflexionen sind in der Lehrkräftebildung daher ein etabliertes Instrument zur Förderung dieser Kompetenz (von Aufschnaiter et al., 2019) und können in der fachdidaktischen Forschung als Bestandteil des *enacted Pedagogical Content Knowledge* (ePCK) betrachtet werden (Carlson et al., 2019). Traditionell werden Reflexionstexte mit validierten Kodiermanualen analysiert (Kobl, 2021; Reimer & Tepner, 2025), ein zeit- und ressourcenintensives Verfahren. Vor diesem Hintergrund untersucht das vorliegende Promotionsprojekt, inwieweit Large Language Models (LLMs) eine valide und effiziente Alternative darstellen. Ziel ist es, zu prüfen, welche Aspekte der Reflexionskompetenz diese Modelle erfassen und wie zuverlässig ihre Einschätzungen im Vergleich zu menschlicher Kodierung sind. Sprachbasierte Modelle wie BERT (Devlin et al., 2018; Latif et al., 2024) ermöglichen die systematische Analyse natürlicher Sprache, die Erkennung inhaltlicher Muster und die Klassifikation von Texten anhand definierter Kategorien. Im Kontext der Lehrkräftebildung eröffnen sie die Möglichkeit, Reflexionsmuster von Studierenden zu identifizieren und damit Rückschlüsse auf Qualität und Tiefe der Reflexionsprozesse zu ziehen (Wulff et al., 2022).

Ziele (Z) und Fragestellungen (F)

- Z1:** Training eines Klassifikators (Modell) zur Evaluation spezifischer Aspekte der Reflexionskompetenz von Studierenden des Chemielehramts
- Z2:** Weiterentwicklung des Klassifikators als Feedback-Tool für Dozierende
- F1:** Inwiefern kann ein Machine-Learning-Modell zur Erfassung von Reflexionskompetenz von Chemielehramtsstudierenden entwickelt werden, das eine vergleichbare Alternative zu klassischen Kodiermanualen bietet?
- F2:** Welche Aspekte von Reflexionskompetenz kann ein Machine Learning basiertes Tool valide erfassen?

Methodik

Das Modelltraining basiert auf 168 schriftlichen Reflexionen von Chemielehramtsstudierenden, die mithilfe eines etablierten Kodiermanuals in 20 Kategorien der Reflexionsbreite und -tiefe eingeordnet wurden (siehe Tab. 1; Kobl, 2021; Reimer &

Tepner, 2025). Ergänzend wurde eine Kategorie für Textpassagen eingeführt, die weder der Breite noch der Tiefe zugeordnet werden konnten. Insgesamt wurden 29.367 Textsegmente kodiert, jeweils ein Satz oder Teilsatz. Ein zentraler methodischer Aspekt war der Umgang mit der unbalancierten Datenverteilung. Da einige Kategorien häufiger vertreten waren als andere, kam Oversampling zum Einsatz (Mohammed et al., 2020; Ta et al., 2022), bei dem Beispiele unterrepräsentierter Kategorien mehrfach zufällig ausgewählt werden, bis ein angestrebtes Verhältnis erreicht ist. Dies kann Verzerrungen im Modelltraining reduzieren und die Klassifikationsleistung über alle Kategorien hinweg verbessern.

Tabelle 1: Kategorien des Datensatzes für das Modelltraining

	Kategorie (Label)	Häufigkeit
Reflexionsbreite	Adaptivität	386
	Adressatenorientierung	615
	Chemisches Fachwissen	66
	Lehrerperformanz	606
	Lernförderliches Klima	197
	Medien	459
	Organisationsform	34
	Regeln im Chemieunterricht	51
	Schülerexperiment	20
	Sonstiges	113
	Sprachliche Verständlichkeit	378
	Sprech- & Körperausdruck	452
	Strukturiertheit	325
	Visualisierung	113
Reflexionstiefe	Alternative	513
	Beschreibung	4420
	Negative Bewertung	2021
	Perspektive	120
	Positive Bewertung	2989
	Verbesserungsvorschlag/Konsequen	3761
	z	
	Nicht kategorisiert	11737

Ergebnisse

Die Ergebnisse zeigen, dass das verwendete Sprachmodell (Latif et al., 2024) die Reflexionssegmente in die Kategorien der Reflexionsbreite und -tiefe mit hoher Genauigkeit klassifizieren kann. Die besten Ergebnisse erzielte die 5-Fold-Cross-Validation mit dem unbalancierten Datensatz, mit einer Accuracy von 0.68 und einem F1-Score von 0.66 (siehe Tab. 2). Der F1-Score, als harmonisches Mittel von Precision und Recall (Sokolova et al., 2006), liegt zwischen 0 und 1; höhere Werte zeigen eine stärkere Übereinstimmung zwischen Vorhersagen und tatsächlichen Kategorien. Die Balancierung des Datensatzes mittels Oversampling brachte keinen Leistungsgewinn; die balancierte Variante erzielte sogar leicht

niedrigere Werte ($F1 = 0.63$). Um Data Leakage zu vermeiden, wurde der Datensatz personenspezifisch geschachtelt, sodass ähnliche Reflexionen entweder im Trainings- oder Testdatensatz enthalten waren. Kategorien mit weniger als 100 Ausprägungen (Tab. 1, nicht fett markiert) wurden im Finetuning ausgeschlossen, da sie geringe Performanz und Relevanz aufwiesen. Der Ansatz, zusätzlich irrelevante Sätze zu klassifizieren, wurde verworfen; in der finalen Modellversion erfolgt die Abgrenzung solcher Segmente über einen Wahrscheinlichkeits-Threshold im Webtool, wodurch eine trennscharfe Zuordnung zu den relevanten Reflexionskategorien gewährleistet ist.

Tabelle 2: Ergebnisse der 5-Fold Crossvalidation

	Unbalanciert	Balanciert (oversampling)	Unbalanciert nach Kategorien- reduktion	Unbalanciert nach Kategorienreduktion und Personen- schachtelung
Accuracy	0.68	0.64	0.64	0.63
F1	0.66	0.63	0.63	0.62
Micro F1	0.68	0.64	0.64	0.63
Macro F1	0.43	0.43	0.43	0.47
Cohens κ	0.57	0.52	0.52	0.56

Diskussion

Die Ergebnisse legen nahe, dass LLMs eine aussichtsreiche Alternative zu herkömmlichen Kodiermanualen darstellen. Durch die automatisierte Klassifikation lassen sich umfangreiche Reflexionen in kurzer Zeit auswerten, ohne Einbußen allzu große in der Validität zu riskieren. Die unausgeglichene Datenverteilung zeigte in dieser Studie keinen Einfluss auf die Modellleistung. Ungeklärt bleibt jedoch, auf welchen sprachlichen oder semantischen Merkmalen die Kategorisierung basiert. Positiv hervorzuheben ist, dass die personenspezifische Schachtelung die Performanz nicht beeinträchtigte und die Ergebnisse somit eine hohe Stabilität aufweisen.

Fazit und Ausblick

Abschließend lässt sich feststellen, dass Machine-Learning-Ansätze, insbesondere der Einsatz von Large Language Models (LLMs), neue Möglichkeiten für die Analyse von Unterrichtsreflexionen in der Lehrkräftebildung eröffnen. Die Ergebnisse verdeutlichen, dass diese Modelle valide Aussagen über Reflexionskompetenz erlauben und zugleich eine effiziente, ressourcenschonende Alternative zu herkömmlichen Kodierv Verfahren bieten. Ihr größter Vorteil liegt in der Skalierbarkeit: LLMs können große Mengen an Reflexionstexten instantan automatisiert auswerten, wodurch der Analyseaufwand erheblich sinkt und unmittelbares Feedback an Studierende ermöglicht wird – ein zentraler Aspekt für die Förderung ihrer professionellen Entwicklung. Im weiteren Forschungsvorhaben wird das Modell im Sinne des *interpretable machine learning* weiterentwickelt, um die Transparenz der Kategorienz Zuordnung zu erhöhen. Zudem sollen im Rahmen der Webtool-Testungen weitere Daten erhoben werden, die zur weiteren Modelloptimierung und Validierung dienen. Langfristig ist vorgesehen, das Webtool in die chemiedidaktische Hochschullehre zu integrieren, um Lehrende zu entlasten und Studierende durch automatisiertes Feedback gezielt in ihren Reflexionsprozessen zu unterstützen.

Literatur

- Abels, S. (2011). *LehrerInnen als „Reflective Practitioner“: Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht*. VS Verlag für Sozialwissenschaften GmbH.
- Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, 9(4), Article 4. <https://doi.org/10.1007/s11618-006-0165-2>
- Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Kam Ho Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., ... Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In A. Hume, R. Cooper, & A. Borowski (Hrsg.), *Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science* (S. 77–94). Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_2
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. <https://doi.org/10.48550/ARXIV.1810.04805>
- Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D. & Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, 9.
- Jurafsky, D., & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd Aufl.). <https://web.stanford.edu/~jurafsky/slp3/>
- KMK (Hrsg.). (2004). Standards für die Lehrerbildung: Bildungswissenschaften. (Beschluss der Kultusministerkonferenz vom 16.12.2004 i. D. F. vom 07.10.2022). https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf
- Kobl, C. (2021). *Förderung und Erfassung der Reflexionskompetenz im Fach Chemie*. Logos Verlag Berlin. <https://doi.org/10.30819/5259>
- Latif, E., Lee, G.-G., Neuman, K., Kastorff, T. & Zhai, X. (2024). G-SciEdBERT: A contextualized LLM for science assessment tasks in German. <https://doi.org/10.48550/ARXIV.2402.06584>
- Mohammed, R., Rawashdeh, J. & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, 243–248.
- Reimer, S. & Tepner, O. (2025). Development of preservice science teachers' reflection skills: Does reflecting on one's own or other people's instructional explanations lead to different developments? *International Journal of Science Education*. Vorab-Onlinepublikation. <https://doi.org/10.1080/09500693.2025.2579575>
- Sokolova, M., Japkowicz, N. & Szpakowicz, S. (2006). Beyond accuracy, F-Score and ROC: A family of discriminant measures for performance evaluation. In A. Sattar & B. Kang (Hrsg.), *AI 2006: Advances in Artificial Intelligence* (Bd. 4304, S. 1015–1021). Springer Berlin Heidelberg. https://doi.org/10.1007/11941439_114
- Ta, T., Butt, S., Angel, J., Sidorov, G. & Gelbukh, A. (2022). The combination of BERT and data oversampling for relation set prediction.
- von Aufschnaiter, C., Fraij, A. & Kost, D. (2019). Reflexion und Reflexivität in der Lehrerbildung. *Herausforderung Lehrer_innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion*, 144-159. <https://doi.org/10.4119/UNIBI/HLZ-144>
- Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—A case for pretrained models-based clustering. *Journal of Science Education and Technology*, 31(4), 490–513. <https://doi.org/10.1007/s10956-022-09969-w>
- Wyss, C. (2013). *Unterricht und Reflexion: Eine mehrperspektivische Untersuchung der Unterrichts- und Reflexionskompetenz von Lehrkräften*. Waxmann.