Stefan Oltmanns[1] [1]Universität Bremen
Christoph Kulgemeyer[1]

Bewertung von Performanztests mithilfe großer Sprachmodelle

Die Erklärkompetenz ist eine zentrale Kompetenz von Lehrkräften im naturwissenschaftlichen Unterricht (Osborne und Patterson, 2011). Schüler:innen erachten insbesondere Lehrkräfte dann als gut, wenn diese aus gut erklären können (Wilson und Mant, 2011), wobei Erklären von Lehrkräften als schwierig empfunden wird (Merzyn, 2005). Folglich ist die Messung und Förderung der Erklärkompetenz von Lehramtsstudierenden von besonderem Interesse.
Im Rahmen der Projekte *ProfiLe-P* und *ProfiLe-P+* wurde ein Performanztest zur Erklärkompetenz entwickelt: In einer nachgestellten Realsituation mussten die Probanden einer Schüler:in (gespielt von geschulten Studierenden) in einem Zeitraum vom 10 Minuten physikalische Konzepte anhand einer vorgegebene Aufgabenstellung erklären. Die Aufgabenstellung beinhaltet ein spezifisches Anwendungsszenario, wie etwa eine Kurvenfahrt oder die Sprengung eines Asteroiden. In dieser Nachhilfesituation werden von der Hilfskraft systematisch gezielte Rückfragen gestellt. Die Bewertung der Erklärungen erfolgt mit einem validierten Kodiermanual, mit welchem sich die Erklärkompetenz bewerten lässt (Kulgemeyer und Tomczyszyn, 2015).
Da die Auswertung und Bewertung durch Analyse der videographierten Durchführungen erfolgt, sind diese ressourcenintensiv und damit ungeeignet für Large-Scale Assessments. In einem ersten Ansatz wurde der offene Performanztest in ein geschlossenes Format überführt (Bartels und Kulgemeyer, 2018), jedoch hat eine Schließung einen signifikanten Einfluss auf das Testinstruments, da die Proband:in verstärkt nicht mehr selber handelt, sondern zum Beobachter wird (Kulgemeyer et al., 2023).

Ziel

Um den Performanztest in einem offenen Format großflächig einsetzbar zu machen, ist eine automatisierte Auswertung erforderlich. In der Vergangenheit wurde für Sprachmodelle wie BERT (Devlin et al., 2018) gezeigt, dass mit ihnen eine Auswertung von offenen Tests möglich ist (Wulff et al., 2020). Dabei werden die vorhandenen Kodierungen verwendet, um das Sprachmodell zu trainieren.
Die ursprüngliche Auswertung des Erklärtests betrachtet zwei Aspekte: Die Sachgerechtheit und die Adressatengemäßheit der Erklärungen. Da die Sachgerechtheit im wesentlichen Fachwissen ist, welches sich durch andere Testmethoden gut prüfen lässt, wird sich an dieser Stelle auf die Adressatengemäßheit fokussiert: Zur Bewertung der Adressatengemäßheit liegen zwölf Kategorien vor, wobei jeweils nur das erste Auftreten der Kategorien kodiert wird. Der Gesamtscore entspricht der Anzahl der aufgetretenen Kategorien.

Vorbereitung des Datensatzes

Insgesamt liegen 287 videographierte Testdurchführungen inklusive Kodierung vor. Die vorliegende Kodierung teilt die Videos in Segmente anhand von Sprecher:innenwechseln ein. Das erste Auftreten einer Kategorie kann einem spezifischen Segment zugeordnet werden, wobei mehrere Kategorien gleichzeitig einem Segment zugeordnet sein können. Wie in Tabelle 1 ersichtlich, unterscheidet sich die Häufigkeit des Auftretens stark, während in nur

29 Durchführungen ein physischer Gegenstand mit dem Szenario verknüpft wird, werden in 276 Durchführungen Fachbegriffe umschrieben.

Umschreibung von Fachbegriffen	276
Unangemessenes Beispiel	41
Zahlenbeispiel	176
Nonverbal verknüpfen	168
Gegenstand als Hilfsmittel	142
Gegenstände mit Szenario verknüpfen	29
Experiment	73
Verständnisversicherung	58
Handlungsaufforderung	30
Rückblick	127
Zusammenfassung	87
Ermunterung/Lob	36

Tab. 1: Anzahl der Durchführungen, in denen eine Kategorie vorkommt

Um eine Verarbeitung der in der Vergangenheit videographierten Situation in einem Sprachmodell zu ermöglichen, wird eine automatische Transkription des Tons mittels Whisper (Radford et al., 2022) angefertigt. Der Ton der vorliegenden Daten hat in vielen Fällen nur eine geringe Qualität (verursacht durch Nebengeräusche während der Testdurchführung, weit von den Sprecher:innen entferntes Mikrofon und wiederholte verlustbehaftete Konvertierung der Dateiformate), weshalb eine manuelle Überarbeitung der automatisch erstellten Transkripte erforderlich ist.
Die Transkripte werden anschließend auf Grundlage der vorhandenen Kodierung der Sprecher:innenwechsel in Abschnitte segmentiert. Für das Zusammenstellen der Trainingsdaten sind zwei Aspekte unbedingt zu beachten: 1) Ein Abschnitt kann einer, keiner oder mehreren Kategorien zugeordnet sein, folglich ist ein Modell, welches einen Abschnitt einer von 12 Kategorien zuordnet ungeeignet. 2) Da nur das erste Auftreten einer Kategorie kodiert wurde, können die Abschnitte einer Durchführung, die zeitlich nach dem Auftreten einer Kategorie folgen, nicht mehr für das Training dieser Kategorie verwendet werden, da ab dieser Stelle unbekannt ist, ob die Kategorie in dem Abschnitt auftritt oder nicht.

Ergebnisse der Sprachmodelle
Das Training in BERT erfolgte über 20 Epochen, wobei 80 % der Daten zum Training und 20 % der Daten zum Testen verwendet wurden. Für jede Kategorie wurde ein individuelles Modell mit zwei Ausgangsklassen (Kategorie kommt vor: Ja bzw. Nein) trainiert. Für alle Kategorien liegen deutlich mehr Negativ- als Positiv-Beispiele vor. Um ein ausgewogenen Datensatz zu erhalten, wurden in jeder Epoche alle Positiv-Beispiele verwendet und die gleiche Anzahl an wechselnden Negativ-Beispielen.
Abbildung 1 zeigt exemplarisch die Ergebnisse der automatisierten Bewertung von vier Kategorien mit zwei verschiedenen Samplings (die zufällige Aufteilung in Test- und Trainingsdaten): Es wird deutlich, dass eine zuverlässige Zuordnung mit BERT nicht möglich ist. Auffällig ist insbesondere, dass das spezifische Sampling einen großen Einfluss hat, was darauf hindeutet, dass die Datenmenge zu gering ist. Diese Annahme wird auch dadurch

gestützt, dass dieses Problem bei den Kategorien mit niedriger Häufigkeit deutlich ausgeprägter ist.

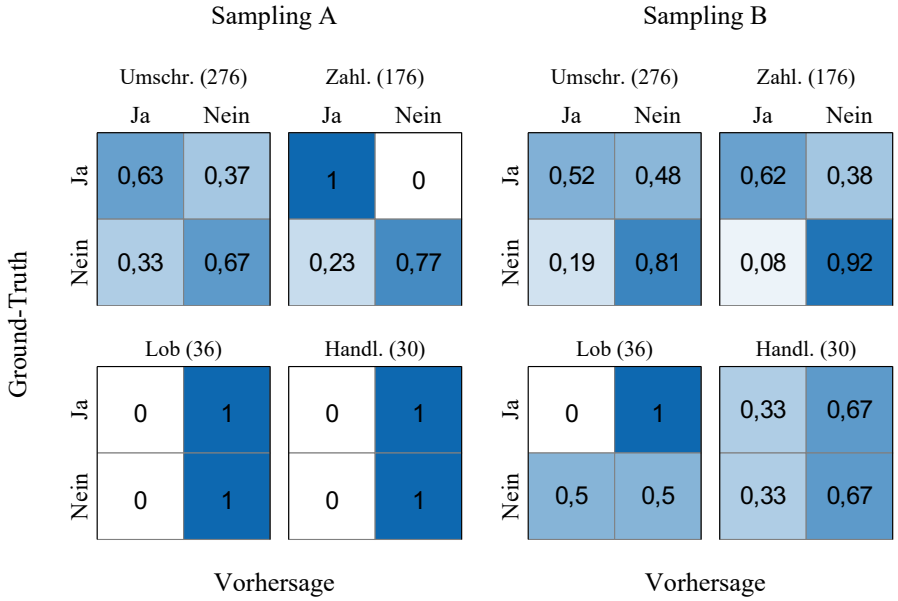

Abb. 1: Confusion-Matrizen des trainierten BERT-Modells mit zwei verschiedenen Samplings, je für zwei Kategorien mit hohem und niedrigem Auftreten

Eine weitere Problematik ist, dass bei manueller Kontrolle der ursprünglichen Kodierung mit dem Kodiermanual eindeutige Fehlzuordnungen festgestellt werden konnten. Die Interraterreliabilität in der ursprünglichen Untersuchung weist ein zwar ein Cohens κ von 0,80 ± 0,03 auf, jedoch wurde dies zu Beginn ermittelt, es ist unklar ob dieser Wert mit später hinzugekommenen Hilfskräften reproduziert werden kann. Über alle Kategorien und mehrere Samplings gemittelt beträgt die Übereinstimmung des Modells mit den menschlichen Ratern Cohens $\kappa = 0{,}20 \pm 0{,}05$.

Da die Datenmenge zum Trainieren von BERT zu gering erscheint, wurden auch wesentlich neuere Sprachmodelle der LLaMA-Familie (Touvron et al., 2023) getestet. Es zeigte sich, dass das LLaMA-Instruct Modell 3.1 mit 8 Milliarden Parametern (Dubey et al., 2024) nicht geeignet ist, was möglicherweise daran liegt, dass insbesondere nicht-englischsprachige Texte bei der reduzierten Modellgröße nur noch mit verminderter Qualität verarbeitet werden können. Das Modell mit 70 Milliarden Parametern konnte auf Grund begrenzter Hardware-Ressourcen nicht mehr trainiert werden, jedoch zeigte sich hier, dass allein durch Prompting bereits ein Cohens κ von 0,15 über alle Kategorien erreicht werden kann, wobei der höchste Wert einer Kategorie hierbei mit $\kappa = 0{,}42$ in der Kategorie Handlungsaufforderung auftritt.

Literatur

Bartels, H., & Kulgemeyer, C. (2018). Ein Videovignettentest zur Messung der Erklärfähigkeit von Lehrkräften. In C. Maurer (Hrsg.), Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Regensburg (S. 162–165). Universität Regensburg.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., … Zhao, Z. (2024). The Llama 3 Herd of Models. arXiv.

Kulgemeyer, C., Riese, J., Vogelsang, C., Buschhüter, D., Borowski, A., Weißbach, A., Jordans, M., Reinhold, P., & Schecker, H. (2023). How authenticity impacts validity: Developing a model of teacher education assessment and exploring the effects of the digitisation of assessment methods. Zeitschrift für Erziehungswissenschaft, 26(3), 601–625.

Kulgemeyer, C., & Tomczyszyn, E. (2015). Physik erklären – Messung der Erklärensfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation. Zeitschrift für Didaktik der Naturwissenschaften, 21, 111–126.

Merzyn, G. (2005). Junge Lehrer im Referendariat. Der mathematische und naturwissenschaftliche Unterricht, 58(1), 4–7.

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? Science Education, 95(4), 627–638.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv.

Wilson, H., & Mant, J. (2011). What makes an exemplary teacher of science? The pupils' perspective. School Science Review, 93(342), 121–125.

Wulff, P., Buschhüter, D., Westphal, A., Nowak, A., Becker, L., Robalino, H., Stede, M., & Borowski, A. (2020). Computer-Based Classification of Preservice Physics Teachers' Written Reflections. Journal of Science Education and Technology, 30(1), 1–15.