Hendrik Prause[1] [1]Freie Universität Berlin
Dennis Dietz[1]

LLM-gestützte Untersuchungen zum vernetzten Lernen des Energiekonzepts

1. Ausgangspunkt

Large Language Models (LLMs) ermöglichen die zeitökonomische Verarbeitung großer Datenmengen. Aus diesem Grund hat der Einsatz von LLMs insbesondere in qualitativen Forschungsarbeiten erheblich an Bedeutung gewonnen und zugenommen (Chew et al., 2023; Zhang et al. 2023). Ziel dieser Arbeit ist es daher, ein LLM zu spezifizieren, mit dem Vernetzungsleistungen von Schüler*innen in Bezug auf das fachdidaktisch bedeutsame Energiekonzept automatisiert und zuverlässig untersucht werden können.

2. Theorie

Aus lerntheoretischer Perspektive ist der Aufbau vernetzter Wissensstrukturen anzustreben (u.a. Ausubel, 1974; Gagné, 1970; Mandl, 2006). In erfolgreichen Lernprozessen werden Begriffe gebildet, indem die relevanten Begriffselemente möglichst vielfältig miteinander verknüpft werden (Aebli, 1981). Wenn Lernende einen komplexen Begriff – bspw. in Form eines Essays – erklären sollen, können die hierarchischen Strukturen zwischen den im Zuge der Begriffsbildung verknüpften Begriffselementen temporär aufgedeckt werden. So kann das latente Konstrukt „vernetztes Wissen“ operationalisiert untersucht werden (Aebli, 1981; Dietz, 2023; Dietz & Bolte, 2021; 2022). Untersuchungen dieser Art sind mit dem Modell zur Analyse der Vernetzung von Begriffselementen (Akronym: MAVerBE) möglich (Dietz, 2023; Dietz & Bolte, 2021; 2022).

Im MAVerBE werden drei Strukturierungsdimensionen unterschieden: 1. Das vertikale Vernetzungsniveau, das die Komplexität der Verknüpfung zwischen Begriffselementen beschreibt, 2. die horizontale Vernetzung, die auf die verknüpften Begriffselemente fokussiert und offenlegt, inwieweit Fachgrenzen überschritten werden und 3. die fachliche Richtigkeit der Verknüpfung (Dietz, 2023; Dietz & Bolte, 2021; 2022). In der Strukturierungsdimension „vertikales Vernetzungsniveau“ werden im MAVerBE sechs Kategorien definiert, die fünf verschiedene vertikale Vernetzungsniveaus beschreiben (s. Tab. 1 in Abschnitt 4).

Das MAVerBE wurde bereits im Rahmen einer Feldstudie im Kontroll- und Interventionsgruppendesign für die Untersuchung der Effekte eines integrierten naturwissenschaftlichen Unterrichtsansatzes in den Jahrgangsstufen 7 und 8 auf das vernetzte Erlernen des fächerübergreifenden Energiekonzepts genutzt (Dietz, 2023; Dietz & Bolte, im Druck). Da sich in diesem Zusammenhang die Rekonstruktion der Vernetzungsleistungen jedoch als sehr zeitaufwendig herausgestellt hat, sind large-scale assessments zum vernetzten Erlernen des Energiekonzepts unter Nutzung des MAVerBE-Analyseverfahrens zurzeit aus ökonomischen Gründen kaum zu realisieren (Dietz, 2023).

LLMs bieten hier potenziell Hilfe, da in Aussicht gestellt wird, dass zeitaufwendige qualitativ-inhaltsanalytische Untersuchungen automatisiert werden können (Chew et al., 2023). LLMs basieren auf künstlichen neuronalen Netzen, die durch informationstechnologische Verfahren, wie „Prompt Engineering“, „Fine-Tuning“, „Retrieval-Augmented Generation (RAG)“ oder „Agentic Workflows“ auf spezifische Aufgaben ausgerichtet werden können (Fan et al., 2024; White et al., 2023; Zhang et al. 2023). Beim „Prompt Engineering“ wird z.B. eine

Befehlseingabe entwickelt, um eine gewünschte Ausgabe des LLMs zu erzeugen (White et al., 2023). Mit Blick auf diese Vielfalt an Spezifikationsmöglichkeiten für LLMs gehen wir der folgenden Forschungsfrage nach:
Inwieweit stimmen die Ergebnisse aus den Analysen zu den vertikalen Vernetzungsleistungen von Schüler*innen in Bezug auf das Energiekonzept überein, wenn die Schüler*innen-Essays mit dem MAVerBE-Analyseverfahren
a) ad personam oder
b) automatisiert durch ein eigens spezifiziertes LLM kodiert werden?

3. Design und Methode

Zur Beantwortung unserer Forschungsfrage galt es, im ersten Schritt aus der enormen Vielzahl unterschiedlicher LLMs ein geeignetes zu identifizieren. Neben den forschungsethischen und rechtlichen (Gewährleistung des Datenschutzes) sowie ökonomischen (lizenzfreie und damit kostenlose Nutzung im universitären Kontext) Voraussetzungen sind bei der Auswahl eines geeigneten LLMs auch technische Kriterien zu beachten (wie bspw. die benötigte Rechenleistung und die mögliche zu verarbeitende Textlänge). Unter Abwägung der Vor- und Nachteile verschiedener LLMs haben wir uns schlussendlich für das *Llama-3.1-8B-instruct-q8_0* von Meta (2024) entschieden.
Für die Spezifikation des ausgewählten LLMs bestand die Aufgabe darin, einen geeigneten „Prompt“ zu entwickeln, durch den Schüler*innen-Texte zum Energiekonzept in Analyseeinheiten unterteilt und diese anschließend einem vertikalen Vernetzungsniveau im MAVerBE zugeordnet werden können. Hierfür haben wir unter Anwendung des Ansatzes von Chew und Kolleg*innen (2023) zur Entwicklung eines LLM-gestützten qualitativ-inhaltsanalytischen Untersuchungsverfahrens (dem LACA-Ansatz; Akronym für: LLM-Assisted Content Analysis) in einem ersten Schritt den Kodierleitfaden zur Nutzung des MAVerBE adaptiert. In diesem Zusammenhang haben wir uns zusätzlich an theoriebasierten Empfehlungen zum „Prompt Engineering“ orientiert (Eager & Brunton, 2023; Wei et al., 2022; White et al., 2023). Der auf Grundlage dieser theoretischen Überlegungen heraus formulierte „Prompt“ wurde anschließend in einem iterativen Prozess überarbeitet, indem die ad personam vorgenommenen und die LLM-gestützten Kodierungen von insgesamt 32 randomisiert ausgewählten Schüler*innen-Texten zum Energiekonzept (aus dem Datensatz von Dietz (2023)) miteinander verglichen wurden. Das auf diese Weise spezifizierte LLM wurde anschließend an 82 weiteren randomisiert ausgewählten Schüler*innen-Texten zum Energiekonzept final erprobt.

4. Ergebnisse

Die Ergebnisse zu den jeweiligen Kodierungen der vertikalen Vernetzungsleistungen in den 82 Schüler*innen-Texten sind in der Tabelle 1 dargestellt.
Tabelle 1 legt offen, dass das LLM insgesamt deutlich weniger Analyseeinheiten ausweist als dies im Zuge der persönlichen Untersuchung der Schüler*innen-Texte zum Energiekonzept erfolgt ist (s. Zeile „gesamt“ in Tab. 1). Darüber hinaus wurden die von dem LLM festgelegten Analyseeinheiten lediglich in 21 von 789 Fällen einem höheren vertikalen Vernetzungsniveau zugeordnet (s. N_1 in Tab. 1). Insgesamt wurde eine prozentuale Übereinstimmung zwischen LLM-gestützter und ad personam Kodierung von 56 % ermittelt (s. Tab. 1).

Tabelle 1. Ergebnisse in Bezug auf die Anzahl der LLM-gestützten (N_1) bzw. ad personam vorgenommenen Kodierungen (N_2, ap) zum vertikalen Vernetzungsniveau in 82 Schüler*innen-Texten zum Energiekonzept sowie Befunde zur prozentualen Übereinstimmung (LLM und ap); in Klammern: relativer Anteil an Kodierungen in Bezug auf die Gesamtanzahl an Kodierungen

Vertikales Vernetzungsniveau nach MAVerBE (Dietz, 2023)	N_1: LLM (relativer Anteil)	N_2: ad personam (ap, relativer Anteil)	Übereinstimmung (LLM und ap)
1a) Erfahrungswissen	137 (17 %)	126 (10 %)	41 %
1b) wissenschaftlicher Fakt	228 (29 %)	539 (44 %)	69 %
2) Zusammenhang ohne Begründung	403 (51 %)	403 (33 %)	65 %
3) verbundener Zusammenhang	18 (2 %)	136 (11 %)	2 %
4) Zusammenhang mit Begründung	0 (0 %)	18 (1 %)	0 %
5) multiperspektivische Verallgemeinerung	3 (0 %)	11 (1 %)	0 %
gesamt	789	1233	56 %

Normiert man die Kodierungen von LLM-gestützter und ad personam Kodierung in Bezug auf die Gesamtanzahl der jeweils gebildeten Analyseeinheiten, dann werden Unterschiede mit Blick auf die relative Gewichtung der von den Schüler*innen erbrachten vertikalen Vernetzungsleistungen deutlich erkennbar (s. Werte in Klammern für N_1 und N_2 in Tab. 1).

5. Diskussion

Die ermittelte Übereinstimmung über alle Analyseebenen zwischen LLM-gestützter und ad personam vorgenommener Kodierung der vertikalen Vernetzungsleistungen von Schüler*innen in Höhe von 56 % könnte in Anlehnung an Altman (1991) als „moderat“ bezeichnet werden. Der Blick auf die einzelnen vertikalen Vernetzungsniveaus offenbart allerdings, dass es dem eigens spezifizierten LLM gegenwärtig nicht ausreichend gelingt, Analyseeinheiten als Grundlage für die Kodierung des vertikalen Vernetzungsniveaus zu bilden. Das liegt vor allem daran, dass das von uns spezifizierte LLM nahezu ausschließlich ganze Sätze als Analyseeinheiten identifiziert. In Folge dessen werden in vielen Fällen mehrere Analyseeinheiten, die sprachlich in Form einer Aufzählung verdichtet sind, ebenso, wie Sinnzusammenhänge, die sich über mehrere Sätze erstrecken, vom LLM nicht erkannt. Aus diesem Grund sind für Kategorien, die ein niedrigeres vertikales Vernetzungsniveau beschreiben, höhere Übereinstimmungsraten zwischen LLM-automatisierter und ad personam Kodierungen zu finden; die Übereinstimmungsrate nimmt demgegenüber mit steigendem vertikalen Vernetzungsniveau rapide ab (s. Tab. 1). Die Probleme in Bezug auf die Verarbeitung komplexer semantischer Strukturen durch LLMs konnten auch von anderen Forschenden bisher nicht gelöst werden (He et al., 2024). Inwieweit diese Ergebnisse durch die Auswahl des LLMs oder dem konkret in dieser Arbeit entwickelten „Prompt“ limitiert sind, ist zum jetzigen Zeitpunkt eine offene Frage.

6. Ausblick

LLMs bieten neben dem „Prompt Engineering“ weitere Methoden und Techniken zur Spezifizierung für bestimmte Aufgaben (s. Abschnitt 2). Im nächsten Schritt werden wir daher die Möglichkeiten des „Fine-Tuning“ zur aufgabenspezifischen Adaption des LLMs erproben. Wir erhoffen uns auf diese Weise, die Probleme bei der automatisierten Bildung von Analyseeinheiten, die insbesondere bei Aufzählungen und komplex(er)en Schüler*innen-Aussagen aufgetreten sind, begegnen zu können.

Literatur

Aebli, H. (1981). *Denken: das Ordnen des Tuns. Band 2: Denkprozesse.* Klett-Cotta.

Altman, D. G. (1991). *Practical Statistics for Medical Research.* Chapman and Hall.

Ausubel, D. P. (1974). *Psychologie des Unterrichts. Band 1.* Beltz.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J. & Kim, A. (2023). LLM-assisted content analysis: *Using large language models to support deductive coding.* arXiv:2306.14924.

Dietz, D. (2023). *Vernetztes Lernen im fächerdifferenzierten und integrierten naturwissenschaftlichen Unterricht aufgezeigt am Basiskonzept Energie. Eine Studie zur Analyse der Wirksamkeit der Konzeption und Implementation eines schulinternen Curriculums für das Unterrichtsfach „Integrierte Naturwissenschaften 7/8"*. Logos.

Dietz, D. & Bolte, C. (2021). Mehrdimensionale Analyse zur Vernetzung von Begriffselementen des Basiskonzepts Energie. In: V. Nordmeier & H. Grötzebauch (Hrsg.), *Phydid B: Beiträge zur DPG-Frühjahrstagung. Digitale Frühjahrstagung 2021* (S. 233-241), Berlin: DPG.

Dietz, D. & Bolte, C. (2022). Multidimensional Analysis of Knowledge-Linking within the Concept of Energy in Student Essays. *NorDiNa, 18(3),* 353-368.

Dietz, D. & Bolte, C. (im Druck). Revisioning science education: Fostering content knowledge-linking within the energy concept using the integrated science teaching approach. In: *Proceedings of the 14th Nordic Research Symposium on Science Education.* University of Iceland.

Eager, B., & Brunton, R. (2023). Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice. *Journal of University Teaching & Learning Practice, 20(5)*, 1–19.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). *A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models*. ArXiv (Cornell University), *24*, 6491–6501.

Gagné, R. M. (1970). *Die Bedingungen des menschlichen Lernens (2. Auflage).* Hermann Schroedel Verlag.

He, Q., Zeng, J., Huang, W., Chen, L., Xiao, J., He, Q., Zhou, X., Liang, J., & Xiao, Y. (2024). *Can Large Language Models Understand Real-World Complex Instructions? Proceedings of the AAAI Conference on Artificial Intelligence*, *38(16),* 18188–18196.

Mandl, H. (2006). Wissensaufbau aktiv gestalten. *SCHÜLER Wissen für Lehrer.* 28–30.

Meta (2024). Introducing Llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/ (letzter Zugriff: 11.09.2024)

Wei, J. Z., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* arXiv:2201.11903

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with chatgpt*. arXiv preprint arXiv:2302.11382.

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2023). *Redefining qualitative analysis in the AI era: utilizing ChatGPT for efficient thematic analysis.* arXiv (Cornell University).