

Neue Perspektiven und Befunde zum Einsatz von KI in der Planung von Chemieunterricht (CUKI)

Bisheriger Forschungsstand

Nach der Veröffentlichung des frei zugänglichen Large Language Models ChatGPT im Winter 2022 wurde die Studie CUKI – Chemieunterricht geplant durch Künstliche Intelligenz – initiiert. Ziel der Studie ist es, zu untersuchen, inwieweit der Chatbot dazu in der Lage ist, qualitativ hochwertige Unterrichtsplanungen im Fach Chemie zu erstellen und ob er Lehramtsstudierende sowie Berufseinsteiger:innen bei der Unterrichtsplanung sinnvoll unterstützen kann. Im Mittelpunkt stehen dabei folgende Forschungsfragen: Welche Kriterien eignen sich zur sinnvollen Messung der Qualität von Unterrichtsplanungen? Können durch den Einsatz von ChatGPT didaktisch und methodisch hochwertige Chemiestunden entworfen werden? In welchem Maße hängt die Qualität der erstellten Planung von der individuellen Planungskompetenz der Anwender:innen ab? Und wie beurteilen die Studierenden ihre eigene Planungskompetenz sowie die Arbeit mit ChatGPT?

Eine Pilotierung mit sieben Masterstudierenden an der Universität Potsdam im Sommersemester 2023 lieferte erste Hinweise darauf, dass die Planungskompetenz der Teilnehmenden signifikant und mit hoher Effektstärke mit der Qualität der durch die KI generierten Planungen korreliert (Brott & Egerer, 2024a). Diese Ergebnisse sollten im Sinne des Design-Based-Research-Ansatzes in einer Wiederholungsstudie überprüft werden (Reinmann, 2005). Diese Ergebnisse werden im Folgenden vorgestellt und diskutiert.

Studiendesign der Wiederholungsstudie

Aufgrund anderer Rahmenbedingungen musste das Studiendesign für die Wiederholungserhebung angepasst werden. Diese fand im Sommersemester 2024 an der Universität Jena mit 21 Chemielehramtsstudierenden des vierten Semesters im Rahmen eines Seminars statt. Bei der Pilotierung handelte es sich um fortgeschrittene Studierende im achten Semester. Die Studie erstreckte sich über den ganzen Semesterverlauf und wurde in einem Pre-Post-Design angelegt. Dies war in der erneuten Erhebung nicht umsetzbar. In einer Seminarsitzung zur Grundlagenvorlesung Chemiedidaktik fand die Erhebung statt und die Studierenden erhielten die Aufgabe, innerhalb von 40 Minuten eine selbstständige Unterrichtsplanung für eine 90-minütige Doppelstunde zum Thema „Leitfähigkeit von Salzen und Salzlösungen“ zu erstellen. Hierzu wurden der Lehrplanausschnitt, eine fiktive Bedingungsanalyse, Lernziele mit zugehörigen Kompetenz- und Anforderungsbereichen sowie eine vorstrukturierte Planungstabelle vorgegeben. Durch die engen Vorgaben sollten der Zeitbedarf für die Planung und die Freiheitsgrade der Entwürfe reduziert werden, um eine bessere Vergleichbarkeit zu gewährleisten. Im Anschluss daran sollten sie mithilfe von ChatGPT für die gleiche fiktive Lerngruppe eine weitere Stunde, diesmal zum Thema „Flammenfärbung“ planen. Dafür hatten die Studierenden 20 Minuten Zeit und haben wiederum Lernziele vorgegeben bekommen. Zum Abschluss der Seminarsitzung wurden die Studierenden zu ihren Einschätzungen der eigenen Planungskompetenz und der von ChatGPT sowie ihren Vorerfahrungen und Haltungen zum Chatbot befragt.

Anpassung des Untersuchungsinstrumentariums

Nach der Pilotierung ergaben sich einige Defizite im Forschungsinstrumentarium. Der grundlegende Aufbau und die theoretische Fundierung, die bereits im letzten Jahr vorgestellt wurden, blieben aber auch in dieser Erhebung erhalten (Brott & Egerer, 2024b). Demnach wurden die bereits beschriebenen 17 Kriterien für die Analyse der selbstständigen und der KI-generierten Planungen wiederum zur Bewertung der Planungsqualität herangezogen. Ein zusätzliches Kriterium „Gegenseitige Rückmeldung zum Lernprozess“ wurde in dieser Erhebung noch aufgenommen und bezieht sich auf die von Hattie (2008) herausgearbeiteten Einflussfaktoren für erfolgreiches Unterrichten. Aufgrund der geringen Proband:innenzahl und der semesterbegleitenden Durchführung war in der Pilotierung keine Anonymisierung bei der Erstellung der Verlaufsplanungen möglich, wohingegen die Befragung anonym erfolgte, eine Zuordnung zwischen Selbsteinschätzung und Bewertungsergebnis war damit nicht möglich. Die Wiederholungsstudie erfolgte dagegen komplett anonymisiert über die randomisierte Zuteilung einer Fallnummer für jede/n Teilnehmende/n. Die Studierenden gaben diese Nummer dann in einen personalisierten Online-Link ein und gelangten zum Dokument zur Erstellung der Planungen. Sie wählten außerdem das entsprechend nummerierte Chatfenster bei ChatGPT aus und gaben ihre Fallnummer in der abschließenden Befragung ein. Somit konnte eine eindeutige und gleichzeitig anonyme Zuteilung der verschiedenen Untersuchungsinstrumente erfolgen. In der Analyse der Daten aus der Pilotierung zeigte sich daneben, dass die Auswahl auf der gewählten 5-stufigen Skala zur Einschätzung jedes Kriteriums nicht trennscharf genug erfolgen konnte. Es wurde daher für die erneute Erhebung ein detailliertes Bewertungsmanual erstellt, das jeden Skalenwert jedes Kriteriums verbalisiert darstellt und damit die Auswahl wie die Bewertung der Planung objektiv ermöglicht. Tabelle 1 zeigt die Spezifizierung beispielhaft für eines der Kriterien.

Kriterium	1 <i>Nicht erfüllt</i>	2 <i>Weniger erfüllt</i>	3 <i>Teilweise erfüllt</i>	4 <i>Eher erfüllt</i>	5 <i>Voll erfüllt</i>
Berücksichtigung Lehrplanvorgaben	Kein Lehrplanbezug erkennbar	Inhaltliche Vorgaben beachtet, keine weiteren geforderten Kompetenzen, Experimente, vorige Einheiten o. v.v.	Inhalt und Experiment einbezogen, keine weiteren geforderten Kompetenzen, vorige Einheiten o.v.v.	Einbindung geforderter Kompetenzen, Fachbegriffe, Experimente, Vorwissen nicht beachtet o. v.v.	Einbindung geforderter Kompetenzen, Fachbegriffe, Experimente, Berücksichtigung des Vorwissens

Tab. 1 Bewertungsmanual für die Einschätzung der Planungsqualität

Schließlich wurde auch die Befragung noch einmal überarbeitet und die Einschätzung der eigenen Planung sowie der ChatGPT Planung genau an die Kriterien des Manuals angepasst, um Selbst- und Fremdeinschätzung passender abgleichen zu können. Auf Grundlage des TAM-Modells (Davis, 1989) wurde außerdem die Haltung zum KI-Tool erfasst.

Diskussion der Ergebnisse

Zunächst sollen die Vorerfahrungen und Haltungen der Stichprobe gegenüber ChatGPT dargelegt werden. Von den 21 Proband:innen kannten 20 den Chatbot bereits, 16 hatten ihn schon im privaten, 18 im universitären und zwei im beruflichen Kontext verwendet. Im universitären Bereich gaben 17 an, es für Recherchezwecke nutzen, 14 zur Vertiefung des eigenen Wissens, zwei zur Erstellung von Mails oder Briefen, sechs zur Erstellung wissenschaftlicher Texte, vier zur Erstellung von Präsentationen, 15 zur Bearbeitung von Seminaraufgaben, acht zur Erstellung von Unterrichtsplanungen, zwei für die Erstellung von Schüler:innenaufgaben und eine Person zur Kontrolle von Schüler:innenaufgaben. Von den acht Studierenden, die es bereits für die Unterrichtsplanung verwendet hatten, haben alle damit einen Verlaufsplan erstellt, fünf haben sich Lernziele formulieren lassen, zwei haben Unterrichtsmaterial damit entwickelt und jeweils eine Person gab an, eine Bedingungsanalyse, Sachanalyse und didaktische Reduktion damit erstellt zu haben. Die Proband:innen gaben an, dass ihnen die Arbeit mit ChatGPT Spaß macht ($M=4,38$) und ihnen die Arbeit mit dem Tool leicht fällt ($M=4,57$). Außerdem sind sie der Ansicht, dass sie zielführende Prompts für den Chatbot formulieren können ($M=4,24$). Die Studierenden stimmten hingegen eher nicht zu, dass man die Antworten der KI immer unverändert übernehmen kann ($M=2,48$). Einige haben Bedenken, was die Entwicklungen auf dem Gebiet der KI betrifft ($M=3,24$), aber ethisch verwerflich finden die meisten den Einsatz nicht ($M=2,09$).

Nun sollen die Selbst- und Fremdeinschätzungen zur Bewertung der Unterrichtsplanungen ausgewertet werden. Es hat sich gezeigt, dass die Studierenden sowohl ihre eigenen Planungen als auch die von ChatGPT generierten deutlich besser einschätzten, als dies die Fremdbewertung über das Manual ergab. Hier lässt sich mutmaßen, dass dies auf das Alter und die bislang wenige Unterrichtserfahrung und damit geringere Reflexionskompetenz zurückzuführen sein könnte. Außerdem zeigte sich, dass die Mehrzahl der mit ChatGPT geplanten Stunden schlechter abschneidet, sowohl nach der Selbst- als auch nach der Fremdeinschätzung. Anders als in der Pilotierung gab es aber diesmal durchaus Fälle, wo die KI-generierte Stunde besser abschnitt, sowohl in der Selbst- als auch in der Fremdeinschätzung jeweils bei 7 von den 21 Proband:innen. Auch dies lässt sich einerseits über die fehlenden Erfahrungen in der Unterrichtsplanung der Stichprobe im Vergleich zu den fortgeschrittenen Studierenden der Pilotierung begründen. Ein weiterer Grund für das unterschiedliche Ergebnis dürfte die Nutzung der aktuelleren Version des Chatbots gewesen sein. In der Pilotierung wurde mit GPT 3.5, in dieser Erhebung mit GPT 4o gearbeitet.

In Hinblick auf die Forschungsfragen werden abschließend Gruppenunterschiede und Zusammenhänge besprochen. Zunächst wurde überprüft, ob sich Unterschiede in der Planungsqualität ergeben, wenn man die Kriterien der vorigen Unterrichtserfahrung, der Intensität der Nutzung von ChatGPT oder auch der Vorerfahrung bei der Erstellung von Verlaufsplänen mit ChatGPT heranzieht. Allerdings ergaben sich in keiner der Berechnungen signifikante Unterschiede. Dagegen zeigten sich auch in dieser Erhebung wieder signifikante Korrelationen mittlerer bis hoher Effektstärke zwischen den Mittelwerten der Planungsqualität der selbstständigen und der mit ChatGPT generierten Stundenplanung, sowohl bei der Berechnung aus den Daten der Selbsteinschätzung der Studierenden (*Pearson* $r=0,457$; $\alpha=0,037$; *Spearman* $\rho=0,654$; $\alpha=0,004$) als auch bei denen der Fremdbewertung (*Pearson* $r=0,585$; $\alpha=0,009$; *Spearman* $\rho=0,653$; $\alpha=0,002$ unter Ausschluss zweier Ausreißerfälle). Insgesamt kann nun die in der Pilotierung formulierte Hypothese, dass die Qualität der mit ChatGPT erstellten Planungen umso höher ist, je größer die Planungskompetenz der Anwender:innen ist, anhand dieser Wiederholungsstudie bestätigt werden.

Literatur

- Brott, M., Egerer, C. (2024a): Kann ein Chatbot die Vorbereitung für den Chemieunterricht übernehmen? Eine Analyse der Qualität von Unterrichtsplanungen von ChatGPT und Chemielehramtsstudierenden. In: MNU Journal 77(2), S.104-110.
- Brott, M., Egerer, C. (2024b): CUKI – Chemie Unterricht geplant durch künstliche Intelligenz. In: van Vorst, H. (Hrsg.): Frühe naturwissenschaftliche Bildung. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Hamburg 2023.
- Davis, F. D. (1989): Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology, MIS Quarterly, 13(3), S. 319-340.
- Hattie, J. (2008): Visible Learning. A Synthesis of over 800 Meta-Analyses relating to Achievement. London: Routledge.
- Reinmann, G. (2005). Innovation ohne Forschung? - Ein Plädoyer für den Design-Based Research-Ansatz in der Lehr-Lernforschung. In: Unterrichtswissenschaft 33 (1), S. 52-67.