Michele Brott[1] [1]Universität Potsdam
Constantin Egerer[2] [2]Universität Potsdam

CUKI – Chemie Unterricht geplant durch Künstliche Intelligenz

Ausgangslage

Mit Veröffentlichung des Chatbots „ChatGPT“ der Firma OpenAI im November 2022 wurde eine gesellschaftliche Diskussion über Künstliche Intelligenz (KI) im Allgemeinen sowie im Bildungskontext entfacht. Das Large Language Model wurde mit unzähligen Textdokumenten trainiert und generiert auf dieser Grundlage nach Wahrscheinlichkeiten gewichtete Antworten auf die eingegebenen Fragen (Prompts) (Gozalo-Brizuela & Garrido-Merchan, 2023).
Das Tool bietet großes Potential für (angehende) Chemielehrkräfte und damit für die Lehramtsausbildung. Neben fachlicher Vertiefung oder Ideenfindung, erscheint der vor allem für Berufsanfänger:innen interessanteste Nutzen in der Unterrichtsplanung und Erstellung von Unterrichtsmaterialien zu liegen. Die eigene Erfahrung in der Chemielehramtsausbildung zeigt hierbei, dass die Studierenden oft zeitlich ineffizient bei der Planung vorgehen und die gelernten Grundsätze eines motivierenden und aktivierenden Chemieunterrichts nur begrenzt in die eigene Planung implementieren. Das vorliegende Forschungsprojekt hat daher zum Ziel zu untersuchen, wie gewinnbringend der Einsatz von ChatGPT in der Lehramtsausbildung und im Speziellen in der Unterrichtsplanung ist und inwiefern die damit generierten Materialien bestimmten Qualitätskriterien genügen. Außerdem soll untersucht werden, ob die Qualität der Ergebnisse von ChatGPT von der Planungskompetenz der Anwender:innen abhängig ist.

Studiendesign der Pilotierung

Die erste Phase des Projekts CUKI startete im Sommersemester 2023 im Rahmen eines Chemiedidaktikseminars für die Studierenden des Masterstudiengangs Lehramt Chemie an der Universität Potsdam. Dabei wurden Unterrichtsplanungen und -materialien mithilfe von ChatGPT erstellt und von den Studierenden anhand ausgewählter Kriterien reflektiert. Insgesamt nahmen 19 Studierende an dem Modul teil. Diese wurden zum einen in einem Pre- und Post-Test zu ihren Vorerfahrungen und Einstellungen zu ChatGPT sowie ihren selbst eingeschätzten Planungskompetenzen befragt. Daneben gab es eine Pre-Post-Erhebung zur Verlaufsplanerstellung, aus der Aussagen zur Entwicklung der Planungskompetenz der Studierenden und Zusammenhänge der Planungskompetenz der einzelnen Proband:innen mit der Qualität der von ihnen mit KI generierten Planung abgeleitet werden sollten (Abb. 1). Die Ergebnisse dieses zweiten Studienelements werden im Folgenden dargelegt.

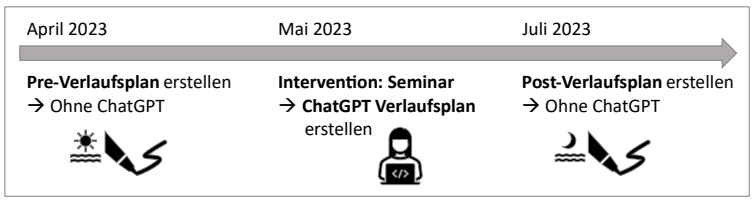

Abb. 1 Schematischer Ablauf der Pilotierung

In der zweiten Projektphase erfolgt im Rahmen eines Wahlpflichtmoduls im Wintersemester 2023/24 die Umsetzung von mit ChatGPT erstellten Planungen im Chemieunterricht. Außerdem wird die Eignung des Chatbots als Lernassistent in der universitären Begleitung der Praxissemesterstudierenden getestet. In der dritten Projektphase soll im Sinne des Design-Based-Research-Ansatzes (Reinmann, 2005) nach Adaption und Evaluation der Maßnahmen und Instrumente von Phase 1 eine Wiederholungsstudie stattfinden. Übergeordnetes Ziel ist außerdem eine Erstellung von Guidelines zur Formulierung zielführender Prompts für die Planung von Chemieunterricht und die zugehörige Materialerstellung mithilfe von ChatGPT.

Instrumentarium zur Bewertung der Unterrichtsplanung

Zunächst sollten die Studierenden in einem Pre-Test den Status Quo ihrer Planungskompetenz darlegen und innerhalb von 60 Minuten eine Verlaufsplanung für 90 Minuten Chemieunterricht erstellen. Um die Freiheitsgrade möglichst gering zu halten und die Vergleichbarkeit zu erhöhen, wurden das Thema, eine fiktive Bedingungsanalyse sowie Lernziele vorgegeben. Um einen möglichen Kompetenzzuwachs durch das ChatGPT gestützte Seminar abzuleiten, erfolgte der Post-Test auf dieselbe Weise zu einem anderen Stundenthema. Die Verlaufsplanungen zu beiden Testzeitpunkten wurden von den Studierenden ohne Hilfe einer KI erstellt. Da im ersten und letzten Seminar die Anwesenheit variierte, konnten nur die Planungen von 7 Studierenden ausgewertet werden, die sowohl am Pre- als auch am Post-Test teilgenommen haben. In der Mitte des Semesters erhielten die Studierenden die Aufgabe eine Stundenverlaufsplanung mithilfe von ChatGPT zu erstellen. Für die Auswertungen wurden die KI Planungen derselben 7 Proband:innen herangezogen.
Die Beurteilung der Qualität aller drei Entwürfe aus Pre- und Posttest sowie dem von ChatGPT erstellten Plan erfolgte mithilfe 17 theoriebasierter Kriterien (König, Krepf, Bremerich-Vos & Buchholtz, 2021; Rothland, 2022; Schröder, Riese, Vogelsang, Borowski, Buschhüter, Enkrott, Kempin, Kugelmeyer, Reinhold & Schecker, 2020). Deren Erfüllung wurde auf einer endpunktbestimmten Intervallskala eingeschätzt. Da die Bewertung bislang nur von einer Person erfolgte, wurde zur Erhöhung der Objektivität eine Handreichung mit Erläuterungen zu den Kriterien erstellt. Daraus konnten für die einzelnen Proband:innen und für alle Kriterien für jede der drei Planungen Mittelwerte (M) errechnet werden. Der Mittelwert der einzelnen Kriterien *(M_K)* gibt summative Hinweise, in welchen Aspekten die Studierendengruppe Veränderungen zeigt. Der Mittelwert der jeweiligen Studierenden *(M_S)* lässt Rückschlüsse auf die Qualität der Entwürfe und damit auf die Entwicklung der Planungskompetenz zu.

Ergebnisse der Pilotierung

Zunächst sollen die Ergebnisse der Proband:innen betrachtet werden. Person 2 *(M_S=2,59)* und Person 7 *(M_S=2,24)* erreichten im Pre-Test die schlechtesten Ergebnisse. Im Gegensatz dazu lieferten die Planungen von Person 3 *(M_S=3,71)* und Person 6 *(M_S=4,06)* die höchsten Mittelwerte. Die individuellen Ergebnisse der Post-Entwürfe zeigten für alle Studierenden einen höheren Mittelwert auf als die Pre-Entwürfe. Dabei stellen die Ergebnisse von Person 2 *(M_S=2,82)* und Person 7 *(M_S=2,65)* auch zu diesem Messzeitpunkt die schlechtesten dar. Ebenso lassen sich die Planungen von Person 3 *(M_S=4,65)* und Person 6 *(M_S=4,12)* erneut als die besten im Vergleich zur Studierendengruppe einordnen. Auch bei der Auswertung der mit ChatGPT generierten Entwürfe war auffällig, dass Person 2 *(M_S=2,27)* und 7 *(M_S=2,07)* die schlechtesten und Person 6 *(M_S=2,80)* die besten Ergebnisse lieferten. Darüber hinaus ist festzustellen, dass die Pläne von Person 1, 4 und 5 *(M_S=2,53)* den gleichen Mittelwert und damit das zweitbeste Ergebnis erreichten.

In den einzelnen Kriterien zeigte sich die Heterogenität der Studierendengruppe sowie ihre Stärken und Schwächen. So fanden die Kriterien „Handlungsalternative bzw. didaktische Reserve“ *(M_K=1,43)*, „Differenzierungsmaßnahmen“ *(M_K=1,71)* und „Anpassung an die Bedingungen der fiktiven Klasse“ *(M_K =2,00)* nur ungenügende Berücksichtigung in den Planungen. Im Gegensatz dazu wurden die Kriterien „fachliche Richtigkeit“ *(M_K=4,43)*, „Handlungsorientierung“ *(M_K=4,14)* und „roter Faden“ *(M_K=3,86)* häufig sehr gut umgesetzt. Bei der Analyse der Post-Entwürfe fällt auf, dass die Studierenden in 15 von 17 Bewertungskriterien der Planungsqualität gleich oder besser im Vergleich zu den Pre-Entwürfen abschnitten. Dabei wurden besonders Zuwächse bei den Kriterien „Stimmigkeit mit dem Lehrplan“ *(Pre: M_K =3,57, Post: M_K =4,71),* „angemessene didaktische Reduktion“ *(Pre: M_K =3,43, Post: M_K =4,43),* „Anpassung an die Bedingungen der fiktiven Klasse“ *(Pre: M_K =2,00, Post: M_K =2,86)* gemessen.
Bei der Analyse der KI-Planungen wurde die „Anpassung an die Bedingungen der fiktiven Klasse“ nicht analysiert, da diese nicht vorgegeben war. Außerdem wurde die fachliche Richtigkeit nicht bewertet, da die Entwürfe für eine Beurteilung meist zu oberflächlich waren. Die Studierenden waren in der Lage mit dem KI-Tool einen „roten Faden“ (*M_K=4,00),* eine „gute Stimmigkeit zum Lehrplan“ (*M_K=3,86)* und „Handlungsorientierung“ (*M_K=3,29)* in den Entwürfen zu realisieren. Die „Steuerung des Lernprozesses durch Impulse und Aufgabenstellungen“ (*M_K=1,14),* eine „Passgenauigkeit des Verlaufsplans mit den Lernzielen“ (*M_K=1,29)* oder „Differenzierungsmaßnahmen“ (*M_K=1,43)* konnten nur ungenügend umgesetzt werden. Insgesamt wurden die KI-Entwürfe in 14 von 15 Kriterien schlechter bewertet als die Post-Verlaufspläne.

Auswertung

Die erhobenen Daten dieser Pilotierung lassen vermuten, dass die Intervention in Form eines Seminars, welches ChatGPT zur Planung und Gestaltung von Unterricht eingebunden hat, die Qualität der Unterrichtsplanungen steigern konnte. Die höheren Mittelwerte in den Kriterien „Stimmigkeit mit dem Lehrplan“ *(Pre: M_K =3,57, Post: M_K =4,71),* „angemessene didaktische Reduktion“ *(Pre: M_K =3,43, Post: M_K =4,43)* und „Anpassung an die Bedingungen der fiktiven Klasse“ *(Pre: M_K =2,00, Post: M_K =2,86)* könnten auf das beschriebene Seminar zurückzuführen sein. Jedoch lässt sich nicht ausschließen, dass parallel belegte fachdidaktische Veranstaltungen anderer Fächer ebenfalls positive Effekte hatten. Aus der Analyse der Pre-, Post- und KI-Entwürfe lässt sich die Hypothese ableiten, dass Studierende, die qualitativ hochwertige Unterrichtsentwürfe anfertigen, auch in der Lage sind, gute KI-Entwürfe zu produzieren. Eine Korrelationsanalyse der Stichprobe *(N=7)* offenbarte, dass der Pearson-Korrelationskoeffizient *r* zwischen den Pre- und den KI-Entwürfen eine höchst signifikante Korrelation *(r=0,917; a=0,004)* anzeigt. Um die Hypothese abschließend bestätigen zu können, muss die Stichprobe noch vergrößert werden. Zwischen den Post- und KI-Entwürfen lag keine signifikante Korrelation *(r=0,718; a=0,069)* vor. Auch wenn die Ergebnisse von ChatGPT eine schlechtere Planungsqualität aufweisen als die von den Studierenden selbst erstellten Entwürfe, ist auffallend, dass dabei trotzdem basale Planungsvorlagen entstanden sind. Dies gilt insbesondere auch für die Personen, die zuvor keine guten Ergebnisse im Vergleich zur Studierendengruppe erzielten. Somit können die Ergebnisse der KI zwar nicht unmittelbar als Vorlage für qualitativ hochwertigen Unterricht gelten, aber eine erste Grundlage für Unterrichtsplanungen sein. Gleichwohl lässt sich vermuten, dass zur Nutzung des Tools didaktische Grundlagen vorherrschen müssen, um die Prompts so zu steuern, dass brauchbare Stundenentwürfe generiert werden können.

Literatur

Gozalo-Brizuela, R.; Garrido-Merchan, E. C. (2023). ChatGPT ist not all you need. A State of the Art Review of large Generative AI models. [https://doi.org/10.48550/arXiv.2301.04655, zuletzt geprüft am 07.02.2023]

König, J.; Krepf, M.; Bremerich-Vos, A.; Buchholtz, C. (2021). Meeting Cognitive Demands of Lesson Planning: Introducing the CODE-PLAN Model to Describe and Analyze Teachers' Planning Competence. The Teacher Educator, 56(4), 466-487.

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.

Rothland, M. (2022). Anmerkungen zur Modellierung und Operationalisierung (allgemeindidaktischer) Unterrichtsplanungskompetenz. Unterrichtswissenschaft 2022(50), 347–372

Schröder, J.; Riese, J.; Vogelsang, C.; Borowski, A.; Buschhüter, D.; Enkrott, P.; Kempin, M.; Kulgemeyer, C.; Reinhold, P.; Schecker, H. (2020). Die Messung der Fähigkeit zur Unterrichtsplanung im Fach Physik mit Hilfe eines standardisierten Performanztests. *ZfDN* 26, 103–122.