Sebastian Tassoti[1] [1]Universität Graz

Wie löst ChatGPT eine Aufgabe zur Säure-Base-Chemie?

Künstliche Intelligenz (KI) hat im letzten Jahr einen regelrechten Aufmerksamkeitshype erfahren. Im Bildungsbereich wurde vor allem die Sprachgenerierungssoftware ChatGPT von OpenAI seit ihrer Veröffentlichung medial kontrovers diskutiert, und auch erste wissenschaftliche Studien ergründen Stärken und Schwächen der KI (Adiguzel et al., 2023; Farrokhnia et al., 2023; Kasneci et al., 2023). Im Bereich der Chemie und des Chemieunterrichts wurde neben vielen möglichen Anwendungsbereichen (Emenike & Emenike, 2023; Exintaris et al., 2023; Humphry & Fuller, 2023) allerdings auch festgestellt, dass ChatGPT zu fachlichen Fehlern bei der Verwendung chemischer Konzepte neigt (Leon & Vidhani, 2023; Tyson, 2023). Um das Potential für mögliche Anwendungen im Bildungskontext im Fach Chemie erschließen zu können ist es notwendig unser Verständnis davon, wie die Software mit Aufgaben in der Chemie umgeht, zu schärfen. Dieser Beitrag gibt erste Einblicke, inwiefern ChatGPT eine Aufgabe zur Säure-Base-Chemie aus österreichischen Chemie-Schulbüchern der Sekundarstufe II lösen kann. Basierend darauf werden Anwendungsmöglichkeiten für die KI im Kontext Chemieunterricht abgeleitet.

Methode

Für die von ChatGPT zu lösende Aufgabenstellung wurde das Buch „ELMO - Elemente und Moleküle“ vom öbv-Verlag herangezogen, eines der meistverwendeten österreichischen Schulbücher im Chemieunterricht (Magyar et al., 2020). Als Thema wurde die Säure-Base-Chemie gewählt. Die Eingabe einer Aufgabenstellung erfolgt als Text, die Beantwortung verlangt allerdings neben Text auch Formelzeichen und Reaktionssymbole. Die genaue Eingabe lautete *„Erstelle die Säure-Base-Reaktion und die vollständige Gleichung (= Ergänzung der an der Reaktion unbeteiligten Gegenionen), markiere die Gleichgewichtslage [und berechne pK.] Stoff 1 und Stoff 2 reagieren in Wasser.“ (Magyar et al. 2020, S. 112)*. In acht Unterpunkten waren jeweils 2 Stoffe angegeben, die an der Reaktion beteiligt sein sollen. Da bereits gezeigt wurde, dass ChatGPT derzeit nicht zuverlässig rechnen kann (Tyson, 2023) wurde die Rechen-Teilfrage (in eckigen Klammern) nicht gestellt. Jeder der 8 Unterpunkte wurde 5-fach als Prompt an ChatGPT 3.5 gegeben (n = 40). Verwendet wurde die gratis Onlineversion, weil sie zugänglich für die Öffentlichkeit ist und daher am wahrscheinlichsten von Schüler:innen verwendet wird. Die Prompts wurden jeweils in einem neuen Chatfenster gestellt, da der Kontext innerhalb eines Chats die Antworten von ChatGPT verändert. Alle Antworten wurden hinsichtlich ihrer fachlichen Richtigkeit analysiert. Dafür wurde die Lösungserwartung des Schulbuchs mit zwei Lehrpersonen aus der Praxis und drei Fachchemiker:innen der Uni Graz diskutiert und validiert. Die Antworten wurden außerdem qualitativ in Anlehnung an Kuckartz analysiert und es wurden induktiv Kategorien gebildet.

Ergebnisse

Auf allgemeiner Ebene hat sich gezeigt, dass bei Eingabe der Fragestellung als Text ohne Formelsymbole trotzdem in keiner Antwort die Wortgleichung für die Säure-Base-Reaktion gegeben wurde. Stattdessen wurden immer die reagierenden Stoffe in Formelsymbole umgewandelt. Es konnte außerdem festgestellt werden, dass ChatGPT nur in circa 25 % aller

Antworten eine korrekte Indexierung verwendet hat. Stellt man eine Frage und verwendet Formelsymbole mit korrekter Indexierung, antwortet ChatGPT grundsätzlich auch mit richtig verwendeten Indices. Eine weitere Beobachtung war, dass nur rund die Hälfte aller Antworten einen Gleichgewichtspfeil anstelle eines gerichteten Reaktionspfeils enthielten, obwohl im Prompt das Wort „Gleichgewichtslage“ vorkommt. Insgesamt zeigten die meisten Antworten von ChatGPT auch Gemeinsamkeiten im Lösungsweg auf. Eine typische Antwort enthielt dabei die Bausteine Identifikation der Säure und der Base, eine Brutto- und eine Nettogleichung, eine Erklärung über den Verlauf der Reaktion und eine Aussage zum Gleichgewicht in der Reaktion. Der einzige Baustein, der in jeder Antwort der KI vorkam, war dabei die Reaktionsgleichung.

Bearbeitung der Buchangabe

ChatGPT identifizierte in 27 von 40 Antworten die an der Reaktion beteiligte Säure und Base korrekt, während in 13 Antworten keine Aussage dazu getroffen wurde. Dabei wurden in 25 Antworten sowohl Säure und Base korrekt erkannt. Einmal wurden beide Stoffe als Säure erkannt, und bei einer Antwort gab ChatGPT an, dass es sich nicht um eine Säure-Base-Reaktion handle. Der akzeptablen Lösungsrate von rund zwei Drittel (26 aus 40) korrekten Antworten bei der Nettogleichung im ersten Antwortteil stand eine deutlich schlechtere Rate an korrekten Bruttogleichungen im zweiten Antwortteil von circa einem Drittel (15 aus 40) gegenüber (siehe Abbildung 1).

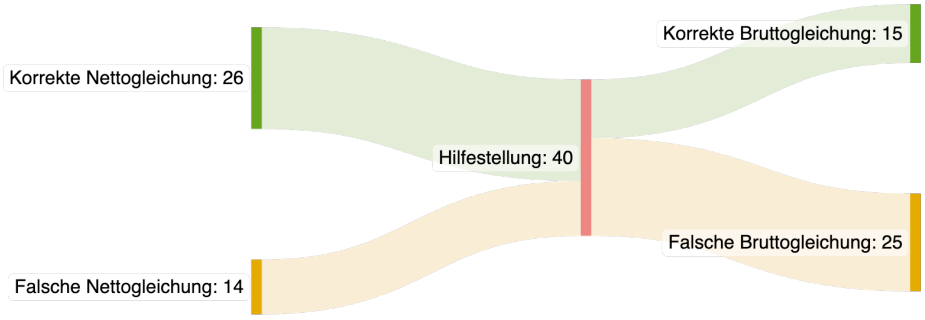

Abb. 1. Korrekte Nettogleichungen in der Antwort von ChatGPT ergeben nicht-korrekte Bruttogleichungen in derselben Antwort/Lösung.

Für die ablaufende Reaktion gab ChatGPT in 32 von 40 Fällen eine Erklärung, wobei davon 16-mal gar kein Säure-Base-Konzept verwendet wurde, sondern der Verlauf der Reaktion lediglich beschrieben wurde. Insgesamt 12-mal wurde das Brønsted-Konzept zur Erklärung verwendet, wobei es nur 9-mal korrekt angewendet wurde, und 4-mal wurde das Lewis-Konzept angewendet – ohne einen Fehler. Die Lage des Gleichgewichts wurde nur in 5 Antworten korrekt vorhergesagt. In weiteren 3 Antworten wurde die falsche Seite vorhergesagt und 4-mal behauptete die Software, es handle sich nicht um eine Gleichgewichtsreaktion. In den restlichen Antworten wurde entweder lediglich das chemische Gleichgewicht allgemein erklärt (15-mal), behauptet die Lage des Gleichgewichts hänge von anderen Faktoren wie Druck, Temperatur und Konzentration ab (11-mal) oder, dass das Vorliegen eines Gleichgewichtspfeils anzeige, dass es sich um eine Gleichgewichtsreaktion handle (15-mal). In manchen Antworten kamen auch mehrere dieser Erklärungen vor.

Diskussion

Um die Fragestellung aus dem Schulbuch zu beantworten, muss ChatGPT in mehreren Erklärungsbausteinen mit chemischem Wissen argumentieren. Während gezeigt wurde, dass die Erkennung von Säure und Base nahezu fehlerfrei funktioniert, sind die Ergebnisse in allen anderen Bereichen durchwachsen. Das Weglassen von Indexierung und Gleichgewichtspfeil bei einer Gleichgewichtsreaktion stellen Fehler dar, die bei Schüler:innen mindestens als Ungenauigkeit ausgelegt werden würden – wenn man ChatGPT allerdings zuvor mit einer korrekt indexierten Reaktionsgleichung anlernt, lässt sich zumindest dieser Fehler umgehen. Beim Aufstellen der Säure-Base-Reaktion zeigt sich dann aber, dass die KI kein „chemisches Verständnis" hat. Unter Einbeziehung der Hilfestellung *„Ergänzung der Gegenionen"* verschlechtert sich die Lösungsquote von circa zwei Drittel korrekten Antworten bei der Nettogleichung auf ein Drittel korrekte Antworten bei der Bruttogleichung innerhalb derselben Antwort. Fehler treten hier vor allem deswegen auf, weil ChatGPT anstelle von Ionen oft einfach Wasser ergänzt, wodurch sogar zwei zuvor falsche Antworten korrekt werden. Hier zeigt sich, dass die Software kein „echtes chemisches Verständnis" hat: der Versuch der Interpretation der Anweisung endete einmal sogar darin, dass in der symbolischen Gleichung das Wort Gegenionen ergänzt wird. Es hat sich auch gezeigt, dass ChatGPT anstatt der Verwendung eines Säure-Base-Konzeptes wortreich beschreibt, dass eine Säure mit einer Base reagiert, ohne in der Erklärung präzise zu werden. In den Antworten wird außerdem das Brønsted-Konzept häufiger zur Erklärung als das Lewis-Konzept verwendet. Hier wird nur die Formulierung „Protonen werden abgegeben" begrenzt verwendet, ohne zu überprüfen, ob nun die Säure oder die Base Protonen abgibt.

Zuletzt deuten die Probleme der KI mit der korrekten Vorhersage des Gleichgewichts auf Probleme beim „Verständnis" hin. Eine typische Antwort war hier *„Der Gleichgewichtspfeil markiert das Gleichgewicht*." – hier kann man die wörtliche Interpretation des Operators *„markiere"* als Auslöser für die Lösungsschwierigkeiten sehen. Bei der Validierung der Aufgaben mit Lehrpersonen aus der Praxis wurde bereits bemängelt, dass Schüler:innen mit dieser Aufgabe auch aufgrund dieser Formulierung Probleme haben. Probleme der KI bei der Beantwortung einer Fragestellung könnten also eventuell Rückschlüsse darauf zulassen, wo Schüler:innen ebenfalls Schwierigkeiten haben könnten, und damit Lehrpersonen in der Praxis ein nützliches Tool zur Vorüberprüfung von Aufgabenstellungen sein.

Zusammenfassend lässt sich sagen, dass die KI ChatGPT kein „echtes Verständnis" (inwiefern man von Verständnis reden kann, muss noch diskutiert werden!) chemischer Vorgänge im Kontext der Säure-Base-Chemie hat und die Schulbuchaufgaben nicht zuverlässig lösen kann. Die Antworten der Software werden mithilfe von Wahrscheinlichkeitsmodellen aufgrund einer nicht exakt bekannten Datenbasis ausgegeben, weswegen einige Fehler gehäuft auftreten können. Wenn Antworten der KI nicht zufriedenstellend sind, kann man aber auch über sogenanntes Prompt Engineering bessere Antworten erhalten. Die präsentierten Ergebnisse legen nahe, dass Prompt Engineering für die produktive Verwendung von ChatGPT im Kontext des Faches Chemie notwendig ist, um eher fachlich korrekte Antworten zu erhalten. Wir konnten bereits erste Erfolge mit einem wissenschaftlich geleiteten Prompt Engineering erzielen. Dieses Vorgehen wird insbesondere für die Arbeit mit Schüler:innen förderlich sein. Erste Daten aus unserer Arbeit mit Schüler:innen zeigen, dass rund die Hälfte der Lernenden zur Lösung von Aufgaben mit ChatGPT die Angabe als solche als Prompt abtippt (Publikation in Vorbereitung). Ein moderierender Umgang mit der KI könnte Schüler:innen dabei auf eine transformierte Arbeits- und Lebenswelt, in der Unterstützung durch und Arbeit mit KI-Assistenten immer mehr Einzug erhalten werden, vorbereiten.

Literatur

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. In *Contemporary Educational Technology* (Vol. 15, Issue 3). Bastas. https://doi.org/10.30935/cedtech/13152

Emenike, M. E., & Emenike, B. U. (2023). Was This Title Generated by ChatGPT? Considerations for Artificial Intelligence Text-Generation Software Programs for Chemists and Chemistry Educators. *Journal of Chemical Education*, *100*(4), 1413–1418. https://doi.org/10.1021/acs.jchemed.3c00063

Exintaris, B., Karunaratne, N., & Yuriev, E. (2023). Metacognition and Critical Thinking: Using ChatGPT-Generated Responses as Prompts for Critique in a Problem-Solving Workshop (SMARTCHEMPer). *Journal of Chemical Education*, *100*(8), 2972–2980. https://doi.org/10.1021/acs.jchemed.3c00481

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2195846

Humphry, T., & Fuller, A. L. (2023). Potential ChatGPT Use in Undergraduate Chemistry Laboratories. *Journal of Chemical Education*, *100*(4), 1434–1436. https://doi.org/10.1021/acs.jchemed.3c00006

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Leon, A. J., & Vidhani, D. (2023). ChatGPT Needs a Chemistry Tutor Too. *Journal of Chemical Education*, *100*(10), 3859–3865. https://doi.org/10.1021/acs.jchemed.3c00288

Magyar, R., Liebhart, W., Jelinek, G., Faber, W., & Strnad, A. (2020). *EL-MO Elemente und Molekül, Schülerbuch*. öbv.

Tyson, J. (2023). Shortcomings of ChatGPT. *Journal of Chemical Education*, *100*(8), 3098–3101. https://doi.org/10.1021/acs.jchemed.3c00361