SCHORCHT, Sebastian & BUCHHOLTZ, Nils
Dresden, Hamburg

Wie verlässlich ist ChatGPT? Modellvalidierung als empirische Methode zur Untersuchung der mathematikdidaktischen Qualität algorithmischer Problemlösungen

GPT ist ein auf künstlicher Intelligenz basierendes Large Language Modell, das über die Schnittstelle ChatGPT menschliche Sprache und Bilder versteht. Durch den Einsatz stochastischer Prozesse kann das Modell Inhalte automatisch vervollständigen. Aktuell werden die Möglichkeiten und Herausforderungen von generativen KI-Modellen intensiv und kontrovers im Kontext der schulischen Bildung sowie der Hochschullehre diskutiert. Der folgende Beitrag beschäftigt sich mit der Modellvalidierung als empirischer Forschungsmethode zur Untersuchung mathematikdidaktischer Qualität algorithmischer Problemlösungen durch ChatGPT.

Modellvalidierungen generativer KI-Modelle

Trotz der Verbesserungen, die neuere Versionen von GPT in Bezug auf mathematische Fähigkeiten erlangt haben (Schorcht et al., 2023), wird die Leistungsfähigkeit von generativen KI-Sprachmodellen im Allgemeinen bislang fast ausschließlich über das Abschneiden in mathematischen Leistungstests ermittelt, die nur einen ungefähren Eindruck über Fehlerraten und Vergleichswerte gewähren (Open AI, 2023). Für mathematikdidaktische Anwendungen von GPT wie etwa beim halb-automatisierten Planen von Unterrichtsentwürfen (Huget & Buchholtz, im Druck) oder bei der Vermittlung von Problemlösekompetenzen im Mathematikunterricht (Schorcht & Baumanns, im Druck) erscheint dies jedoch nicht hinreichend, da die Ausgaben der Modelle funktionalen (mathematikdidaktischen) Qualitätskriterien genügen müssen, um in Bildungsprozessen valide einsetzbar zu sein. Bislang existieren jedoch kaum empirische Ansätze, um den Output generativer KI-Sprachmodelle kriteriell zu validieren. Eine Möglichkeit stellt aber die Evaluation des KI-Outputs durch menschliche Expertinnen und Experten dar (Küchemann et al., 2023; Qiu et al., 2017; Maroengsit et al., 2019). Bei Modellvalidierungen unter Berücksichtigung menschlicher Einschätzungen werden in einer Versuchsserie mehrere, oft kontrolliert modifizierte Prompts in das generative KI-Sprachmodell eingegeben, um durch diese Simulationen zu untersuchen, ob das Modell konsistent und sinnvoll wiederholt auf die gleiche Anfrage reagiert. Anschließend werden die Ausgaben einer kriterienorientierten vergleichenden Bewertung durch Expertinnen und Experten in Bezug auf ihre Qualität in Abhängigkeit zur Fragestellung unterzogen

In: P. Ebers, F. Rösken, B. Barzel, A. Büchter, F. Schacht & P. Scherer (Hrsg.),
Beiträge zum Mathematikunterricht 2024.
57. Jahrestagung der Gesellschaft für Didaktik der Mathematik. WTM.
https://doi.org/10.37626/GA9783959872782.0

(Qiu, et al., 2017), um die Nutzbarkeit der Ausgaben vorzunehmen.

Studie zur Qualität algorithmischer Problemlösungen

Unser Beitrag beruht auf einer Studie zur Untersuchung der Qualität algorithmischer Problemlösungen, in der Modellvalidierungen eingesetzt wurden. Dabei folgen wir der Forschungsfrage: Welche mathematikdidaktische Qualität weisen algorithmische Problemlösungen des Mathematikunterrichts durch ChatGPT auf?

Vier Modellvarianten von vorgegebenen Prompt-Techniken, wie beispielsweise Chain-of-Thought (Schorcht et al., 2023), wurden verwendet, um eine qualitativ verbesserte Ausgabe zu generieren. Alle vier Modellvarianten wurden an drei verschiedenen Problemlöseaufgaben in jeweils 30 Versuchen getestet. Die Modellvalidierung erfolgte je Aufgabe im Sprachmodell GPT-3.5, GPT-4 und GPT-4 unter Zuhilfenahme des Plug-in Wolfram.

Die Bewertung der (N=1080) Ausgaben erfolgte durch zwei unabhängige Experten. Alle Ausgaben wurden anhand der Kriterien Spezifität, Klarheit, Korrektheit, Strategiebewusstheit, Darstellungswechsel und Selbstregulation dichotom (0;1) geratet. Die Bewertungen wurden anschließend aufgabenspezifisch auf Interkoderreliabilität überprüft.

In der vorliegenden Studie wurde unter anderem eine Aufgabe aus Polya (2010, S. 199) getestet: „Wie kann man vom Fluss genau 6 Liter Wasser heraufholen, wenn man nur zwei Gefäße hat, einen 4-l-Eimer und einen 9-l-Eimer, um damit zu messen?“

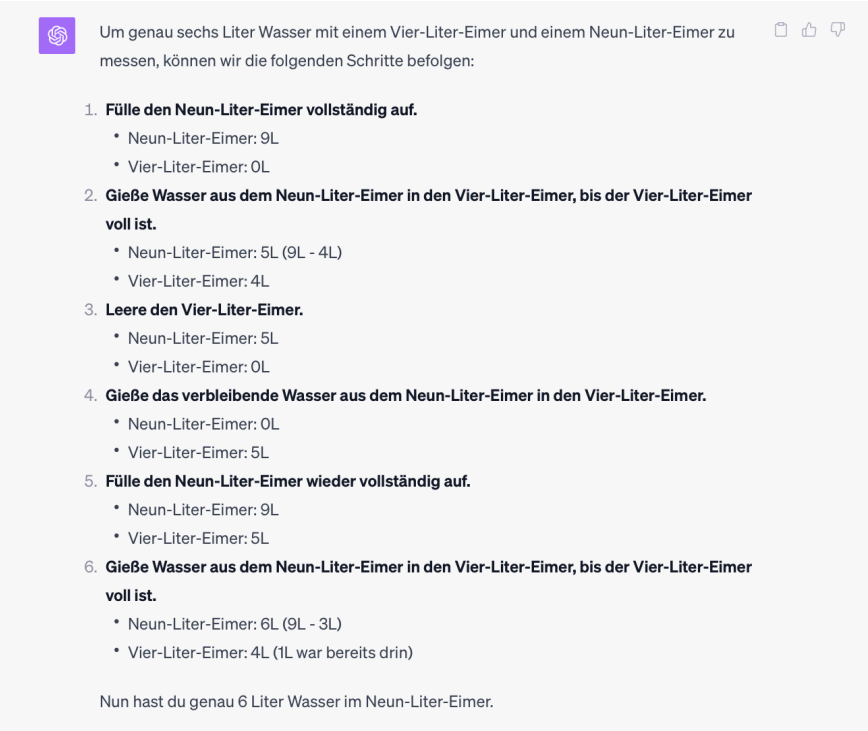
Die Aufgabe wurde in einem Chain-of-Thought-Szenario 30-mal in ChatGPT mit der Schnittstelle GPT-4 eingegeben. In diesem Szenario wird die ursprüngliche Aufgabenstellung um den Satz „Lass uns Schritt für Schritt vorgehen, um sicherzustellen, dass wir die richtige Antwort haben.“, ergänzt. Eine prototypische algorithmische Problemlösung durch ChatGPT ist in Abbildung 1 abgedruckt.

Kriteriengeleitete Untersuchung der mathematikdidaktischen Qualität

Unter dem Kriterium *Spezifität* wurde erfasst, inwiefern die Ausgabe alle relevanten Informationen der Aufgabenstellung für die Lösung des Problems verwendet. Die in Abb. 1 abgedruckte Lösung enthält alle benötigten Informationen („Vier-Liter-Eimer“, „Neun-Liter-Eimer“, „Wasser“ und „Sechs-Liter-Eimer“). Die Ausgabe ist somit spezifisch im Sinne der Aufgabenstellung.

Das Kriterium *Klarheit* erfasst, ob die Argumentation konsistent formuliert wurde und keine Schritte vergessen wurden oder zusätzliche Schritte auftauchen, die für die Lösung des Problems nicht relevant sind. So wurde in der

Ausgabe in Abbildung 1 zwischen Schritt 4 und 5 der Schritt des Umschüttens von 1 Liter im Neun-Liter-Eimer in den Vier-Liter-Eimer nicht benannt, weswegen die Ausgabe als nicht klar bewertet wurde.



Um genau sechs Liter Wasser mit einem Vier-Liter-Eimer und einem Neun-Liter-Eimer zu messen, können wir die folgenden Schritte befolgen:

1. **Fülle den Neun-Liter-Eimer vollständig auf.**
   - Neun-Liter-Eimer: 9L
   - Vier-Liter-Eimer: 0L
2. **Gieße Wasser aus dem Neun-Liter-Eimer in den Vier-Liter-Eimer, bis der Vier-Liter-Eimer voll ist.**
   - Neun-Liter-Eimer: 5L (9L - 4L)
   - Vier-Liter-Eimer: 4L
3. **Leere den Vier-Liter-Eimer.**
   - Neun-Liter-Eimer: 5L
   - Vier-Liter-Eimer: 0L
4. **Gieße das verbleibende Wasser aus dem Neun-Liter-Eimer in den Vier-Liter-Eimer.**
   - Neun-Liter-Eimer: 0L
   - Vier-Liter-Eimer: 5L
5. **Fülle den Neun-Liter-Eimer wieder vollständig auf.**
   - Neun-Liter-Eimer: 9L
   - Vier-Liter-Eimer: 5L
6. **Gieße Wasser aus dem Neun-Liter-Eimer in den Vier-Liter-Eimer, bis der Vier-Liter-Eimer voll ist.**
   - Neun-Liter-Eimer: 6L (9L - 3L)
   - Vier-Liter-Eimer: 4L (1L war bereits drin)

Nun hast du genau 6 Liter Wasser im Neun-Liter-Eimer.

Abb. 1: Ausgabe einer algorithmischen Problemlösung unter dem Chain-of-Thought-Szenario in ChatGPT mit der Schnittstelle zu GPT-4.

Unter *Korrektheit* haben wir die korrekte Durchführung des Lösungsprozesses erfasst. Hierbei wurden nur vollständig korrekte Lösungen positiv geratet, während kleine Fehler schon zu einem negativen Rating führten. In Abbildung 1 wird beispielsweise der Vier-Liter-Eimer mit fünf Litern befüllt. Dieser Fehler führt unter dem Kriterium Korrektheit zur Beurteilung mit 0.

Strategiebewusstheit umfasst alle Äußerung, die auf ein planvolles Vorgehen hinweisen. So taucht in einigen Ausgaben die Phrase „Wir müssen eine Reihe von Schritten durchführen“ auf, die ein explizit benanntes Vorgehen vorschlägt. Indirekte Strukturierungen, wie die Nummerierung von Schritten, wurden als negativ geratet. Auch die Ausgabe in Abb. 1 wurde als indirekt interpretiert und mit 0 geratet.

Der *Darstellungswechsel* wurde kodiert, wenn ein Wechsel der Darstellung vom geschriebenen Wort in eine andere Darstellung vollzogen wurde. Hierzu zählen der Wechsel vom geschriebenen Wort in Funktionsgraph oder vom geschriebenen Wort in Gleichung (in Abbildung 1 beispielsweise „9L - 4L“ unter Schritt 2). Der Wechsel vom geschriebenen Wort in Zahl wurde in der vorliegenden Studie als negativ geratet.

Das Kriterium *Selbstregulation* erfasst Sätze in den Ausgaben, in denen ChatGPT die Anpassung des Lösungsprozesses anführt oder eine Beurteilung der eigenen Lösung zeigt. Ein Beispielsatz ist unter anderem: „Es scheint, als gäbe es einen Fehler in der Rechnung." Im vorliegenden Beispiel konnte dieses Kriterium in der Ausgabe nicht erkannt werden.

Diskussion

Die Ergebnisse liefern einen ersten Einblick in die mathematikdidaktische Qualität der Ausgaben von ChatGPT unter Berücksichtigung verschiedener Prompt-Techniken in unterschiedlichen Versionen des KI-Sprachmodells. In Zukunft spielen Sprachmodelle eine große Rolle in der Anwendung im Mathematikunterricht und in der Lehrerausbildung. Wir konnten mit der Studie empirisch belegen, dass gewisse Prompt-Techniken die Qualität des Outputs erhöhen. Gleichzeitig erhielten wir einen Einblick in die Funktionsweise verschiedener Versionen von GPT und der Einflüsse auf Ausgaben unter dem Plug-in Wolfram.

Literatur

Huget, J. & Buchholtz, N. (im Druck). Gut gepromptet ist halb geplant - ChatGPT als Assistenten bei der Unterrichtsplanung nutzen. In A. König (Hrsg.), *Praxisratgeber: Künstliche Intelligenz, 2*.

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E. & Kuhn, J. (2023). *Can ChatGPT support prospective teachers in physics task development?*. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128.

Polya, G. (2010). *Schule des Denkens. Vom Lösen mathematischer Probleme*. Tübingen & Basel: Francke.

Maroengsit, Wari & Piyakulpinyo, Thanarath & Phonyiam, Korawat & Pongnumkul, Suporn & Chaovalit, Pimwadee & Theeramunkong, Thanaruk. (2019). *A Survey on Evaluation Methods for Chatbots*. 111-119. https://doi.org/10.1145/3323771.3323824.

OpenAI (2023). *GPT-4 Technical Report*. https://arxiv.org/pdf/2303.08774.pdf.

Schorcht, S. & Baumanns, L. (im Druck). Alles falsch?! Reflektiertes Problemlösen mit KI-Unterstützung im Mathematikunterricht. In A. König (Hrsg.), *Praxisratgeber: Künstliche Intelligenz, 2*.

Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F. & Pohl, M. (2023). Ask Smart to Get Smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. In *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 115*, S. 12–24.

Qiu, M., Li, F.-L., Wang, S., Gao, X., Chen, Y., Zhao, W., Chen, H., Huang, J. & Chu, W. (2017). AliMe Chat: A Sequence to Sequence and Rerank based Chatbot Engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, S. 498–503.