SEIBOLD, Moritz & KODWEIß, Jan
Schwäbisch Gmünd

Verfügen LLMs über mathematische Reasoningfähigkeiten?

Die Inhalts- und Konstruktvalidität von Reasoningfähigkeiten textgenerierender Large Language Modells (LLMs) ist nicht geklärt, auch wenn sie zunehmend in verschiedenen Disziplinen untersucht werden. Ziel dieses Beitrags ist es, über einen pragmatischen Ansatz Kriterien ableitbar zu machen, ab welcher Leistung aus mathematikdidaktischer Sicht von menschenähnlichen Reasoningfähigkeiten gesprochen werden kann. Anhand einer Aufgabe zur Untersuchung menschlicher Reasoningfähigkeiten des Kognitionspsychologen Peter Wason und einer Variation davon, die die Aufgabe in einen vertrauten Kontext situiert, werden Rückschlüsse auf die Vergleichbarkeit der Reasoningfähigkeiten von Mathematiklehramtsstudierenden ($n = 38$) und drei modernen LLMs ($n = 3 \cdot 150$) gezogen.

LLMs und Reasoning

Spätestens seit eine von Googles Forschungsgruppen mit ihrem Konzept des Chain of Thought Prompting LLMs die Fähigkeit zum logischen Denken zuschrieb (Wei et al., 2024), werden Reasoningfähigkeiten dieser Modelle auch in Disziplinen außerhalb der Informatik erforscht. Speziell auf mathematische Aufgaben bezogen, gibt es unter anderem bereits Untersuchungen darüber, wie sich verschiedene Möglichkeiten einen Prompt zu schreiben, auf die inhaltsbezogene und prozessbezogene didaktische Qualität der Antworten auswirken (Schorcht et al., 2024) und wie ChatGPT zur Erstellung geometrischer Beweise verwendet werden kann (Dilling & Herrmann, 2024). Keineswegs ist die Annahme, LLMs besäßen Reasoningfähigkeiten, unumstritten. Eine Forschungsgruppe von Apple antwortete jüngst darauf, dass LLMs überhaupt keine validen mathematischen Reasoningfähigkeiten besäßen und stattdessen nur Zwischenschritte aus ihren Trainingsdaten nachbildeten (Mirzadeh et al., 2024). Anstatt die Diskussion über das Vorhandensein von Reasoningfähigkeiten mit großen Datensätzen von tausenden mathematischen Textaufgaben wie dem GSM8K oder einer Abwandlung davon zu führen, verfolgen wir einen pragmatischen Ansatz: Wenn solche Datensätze aufgrund von Datenkontamination und Fragilität der Modelle die tatsächlichen mathematischen Reasoningfähigkeiten von Sprachmodellen sowieso nicht zuverlässig abbilden (Mirzadeh et al., 2024), ist es das Ziel, einen Schwellenwert bestimmbar zu machen, ab dem aus mathematikdidaktischer Sicht von menschenähnlichen Reasoningfähigkeiten gesprochen werden kann.

In: L. Schick, M. Platz & A. Lambert (Hrsg.),
Beiträge zum Mathematikunterricht 2025.
58. Jahrestagung der Gesellschaft für Didaktik der Mathematik. WTM.
https://doi.org/10.37626/GA9783959873307.0

Im 20. Jahrhundert lag ein besonderes Augenmerk verschiedener Disziplinen auf der Deduktion, dem Ziehen logisch gültiger Schlussfolgerungen. Eine der prominentesten Aufgaben zur Untersuchung menschlicher Fähigkeiten beim deduktiven Schließen ist die Wahlaufgabe von Wason, die darauf abzielt, logisches Denken und dabei auftretende Fehlermuster zu analysieren. Die Frage lautet: *Wenn auf einer Seite der Karte ein Vokal steht, dann befindet sich auf der anderen Seite eine gerade Zahl* (Abb. 1 oben). *Welche Karten muss man umdrehen, um die Regel zu überprüfen?*

Die Regel wird als Implikation $P \rightarrow Q$ präsentiert. Zwei mögliche logische Schlussfiguren darauf sind Modus tollendo tollens (MT) und Modus ponendo ponens (MP). Aus dem MT folgt, dass aus Gültigkeit der Implikation ($P \rightarrow Q$) und $\neg Q$ stets auch $\neg P$ folgt (Kontraposition), aus dem MP folgt, dass aus den beiden Prämissen $P \rightarrow Q$ und P, Q folgt (Sanford, 2011). Die Erhebung von Johnson-Laird und Wason (1970) zeigt, dass fast die Hälfte der Probanden korrekterweise den Schluss MP vollzogen und folglich die Karte A (Buchstabe U) wählten. Lediglich vier Prozent der Probanden berücksichtigten auch MT und wählten zusätzlich die Karte D (Zahl 13) aus. Allerdings gibt es deutlich weniger Fehlentscheidungen, wenn die Aufgabe in einem vertrauten Kontext situiert ist, z. B. bei einer Implikation, die das gesetzliche Mindestalter für Alkoholkonsum in Florida miteinbezieht (Griggs & Cox, 1982).

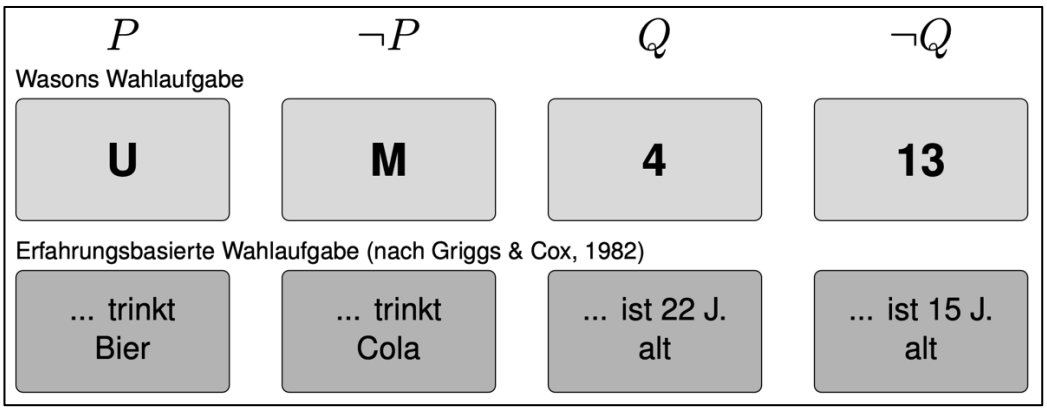

Abb. 1: Wasons Wahlaufgabe und erfahrungsbasierte Wahlaufgabe

Um die mathematischen Reasoningfähigkeiten testen zu können, sondieren wir zunächst die Möglichkeit, Entscheidungen von LLMs als Resultat einer inhärenten Reasoningfähigkeit zu bewerten, um so eine Annäherung an einen pragmatischen Schwellenwert zu ermöglichen.

RQ1: Lassen sich bei LLMs ähnliche Effekte der Kontextualisierung feststellen, die zur Suppression falscher Schlussfolgerungen führen?

RQ2: Ist es (trotzdem) möglich, LLMs systematisch auf mathematische Reasoningfähigkeiten hin zu untersuchen?

Untersuchungsdesign

Im Rahmen der vorliegenden Untersuchung wurden 38 Lehramtsstudierende (Primarstufe, Fachschwerpunkt Mathematik) befragt und drei weitverbreitete LLMs untersucht. Die Studierenden beantworteten die Fragen nacheinander: zunächst Wasons Wahlaufgabe und dann die erfahrungsbasierte Variante (Abb. 1). Die LLMs wurden jedes Mal erneut aufgerufen und hatten keine Möglichkeit, „sich an vorangegangene Anfragen zu erinnern“. Insgesamt wurden alle drei LLMs jeweils 150-mal mit beiden Versionen der Wahlaufgabe konfrontiert. Die Aufrufe wurden automatisiert mit OpenAI's Python-Bibliothek an die Chat Completions API durchgeführt, die Anfragen an die häufig verwendeten LLMs von OpenAI ermöglicht.

Erste Ergebnisse

Von den 38 Studierenden drehten 31 (81,6%) korrekterweise die Karte mit dem Vokal um (P) und 12 (31,6%) wählten korrekterweise die mit der ungeraden Zahl ($\neg Q$). Nur zwei Studierende (5,3%) wählten ausschließlich die beiden notwendigerweise umzudrehenden Karten mit dem Vokal und der ungeraden Zahl (P und $\neg Q$). Indessen erzielen moderne LLMs korrekte Antworten: Gpt-4o-2024-08-06 (GPT-4o) generierte keine fehlerhafte Entscheidung, gpt-3.5-turbo-0125 (GPT-3.5t) hingegen unterliefen Fehler beim Umdrehen der Karten: 35-mal (23,3%) wurde neben der Karte mit dem Vokal (P) fälschlicherweise die mit der geraden Zahl (Q) ausgewählt. Nur in 16 Fällen (10,7%) wählte GPT-3.5t korrekterweise P und $\neg Q$ zum Umdrehen aus. Dagegen entschied sich gpt-4-turbo-2024-04-09 (GPT-4t) sogar in 120 Fällen (80 %) für die Karten mit dem Vokal und der geraden Zahl (P und Q).

Wie erwartet wählten bei der Aufgabe im Kontext des gesetzlichen Mindestalters für Alkoholkonsum deutlich mehr Studierende korrekterweise ausschließlich P (Person trinkt Alkohol) und $\neg Q$ (Die Person ist nicht mind. 16 Jahre alt). Die Ergebnisse deuten darauf hin, dass dieser Effekt auch bei LLMs auftreten kann.

	Wasons Wahlaufgabe								Erfahrungsbasierte Wahlaufgabe							
	LAS		GPT-3.5t		GPT-4t		GPT-4o		LAS		GPT-3.5t		GPT-4t		GPT-4o	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
$P \wedge \neg Q$	2	5.3	16	10.7	30	20	150	100	24	63.2	98	65.3	150	100	150	100
P	4	10.5	0	0.0	0	0.0	0	0.0	2	5.3	0	0.0	0	0.0	0	0.0
$P \wedge Q$	16	42.1	35	23.3	120	80	0	0.0	1	2.6	0	0.0	0	0.0	0	0.0
Andere	16	42.1	99	66	0	0.0	0	0.0	11	28.9	52	34.7	0	0.0	0	0.0

n=38 für Lehramtsstudierende (LAS), n=150 für LLMs.

Abb. 2: Antwortverhalten von Lehramtsstudierenden und LLMs im Vergleich

Diskussion und Perspektiven

Bei den getesteten LLMs waren Effekte der Kontextualisierung feststellbar, die, wie auch bei Menschen, zur Suppression falscher Schlussfolgerungen führen. Das Argument des übertrainierten Inhalts kann die, im Vergleich zu dessen Nachfolgern, schlechten Ergebnisse von GPT-3.5t nicht vollständig erklären, da experimentelle Forschungsergebnisse mit Wasons Wahlaufgabe bereits ab der zweiten Hälfte der 1980er Jahre weit verbreitet waren. Da LLMs und andere neuronale Netze hauptsächlich aus Wahrscheinlichkeiten und anderen Zahlen bestehen, verfügen sie auch nicht über einen „Kontext" im eigentlichen Sinne. Deshalb stellt sich die Frage nach der Kontextsensitivität von LLMs der im Vergleich zur symbolischen, kontextfreien Wahlaufgabe besseren Leistung, insbesondere von GPT-3.5t.

Nachdem es auf diese Weise möglich scheint, LLMs systematisch auf Reasoningfähigkeiten zu untersuchen, kann nun im nächsten Schritt eine Auswahl an mathematischen Reasoningaufgaben zusammengestellt werden, deren Lösbarkeit durch den Kontext, in dem die jeweiligen Aufgaben präsentiert werden, beeinflusst wird.

Literatur

Dilling, F., & Herrmann, M. (2024). Using large language models to support pre-service teachers mathematical reasoning—An exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence, 7*. https://doi.org/10.3389/frai.2024.1460337

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology, 73*(3), 407–420. https://doi.org/10.1111/j.2044-8295.1982.tb01823.x

Johnson-Laird, P. N., & Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology, 1*(2), 134–148. https://doi.org/10.1016/0010-0285(70)90009-5

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models.* https://arxiv.org/abs/2410.05229

Sanford, D. (2011). If P, Then Q: Conditionals and the Foundations of Reasoning (2nd ed). Taylor and Francis.

Schorcht, S., Buchholtz, N., & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Frontiers in Education, 9*. https://doi.org/10.3389/feduc.2024.1386075

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2024). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.