

MÜLLER, Fabian Anton; SCHORCHT, Sebastian & BUCHHOLTZ, Nils
Hamburg, Dresden, Hamburg

KI-gestützte Problemlösungen in der Mathematik: Eine Untersuchung zu Prompt-Techniken und Sprachmodellen

In der vorliegenden Untersuchung wurde im Rahmen eines Mixed-Methods Designs analysiert, wie aktuelle generative KI-Sprachmodelle zur Bearbeitung mathematischer Problemlösungen eingesetzt werden können und welche mathematikdidaktische Qualität deren Ausgaben aufweisen. Im Fokus stehen vier leistungsfähige Modelle – ChatGPT-o3-mini, Claude 3.5 Sonnet, Gemini 1.5 Pro und DeepSeek-R1 –, die mithilfe unterschiedlicher Prompt-Techniken (Schorcht et al., 2023) systematisch getestet wurden. Insgesamt wurden 3.840 KI-generierte Lösungen zu acht Problemstellungen aus den Bereichen „Number“ und „Algebra“ untersucht. Dabei wurden vier Prompt-Techniken verwendet: Zero-Shot, Chain-of-Thought, Persona sowie ein literaturgestützter Naïve-RAG-Prompt.

Die Studie geht drei zentralen Forschungsfragen nach: (1) Wie beeinflussen unterschiedliche Prompt-Techniken die mathematikdidaktische Qualität der KI-gestützten Lösungen? (2) Wie variiert die Qualität der KI-generierten Problemlösungen zwischen den eingesetzten Modellen? Und (3) wie können KI-gestützte Problemlösungen didaktisch so optimiert werden, dass sie Lernprozesse im Unterricht gezielt unterstützen?

Die mathematikdidaktische Analyse erfolgte anhand von elf Qualitätskriterien, etwa zu Korrektheit, Klarheit, Heuristiken, Reflexion und Fehlerbewusstsein. Die Ergebnisse zeigen, dass Persona- und Chain-of-Thought-Prompts besonders zur Verbesserung der Transparenz und didaktischen Brauchbarkeit beitragen. Die Qualität der Lösungen variiert jedoch deutlich zwischen den Modellen. Trotz signifikanter Fortschritte in der mathematischen Problemlösefähigkeit bleibt die fehlende Erklärbarkeit ein zentrales Hindernis für den schulischen Einsatz. Insgesamt verdeutlicht die Studie, dass generative KI-Modelle ein vielversprechendes Potenzial für die mathematikdidaktische Praxis aufweisen – vorausgesetzt, ihre Nutzung erfolgt reflektiert, erklärbar und mit didaktisch fundierter Steuerung.

Literatur

Schorcht, S., Buchholtz, N., Baumanns, L., Huget, J., Peters, F. & Pohl, M. (2023). Ask Smart to Get Smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, 115, 12–23. <https://ojs.didaktik-der-mathematik.de/index.php/mgdm/article/view/1179/1356>

In: L. Schick, M. Platz & A. Lambert (Hrsg.),
Beiträge zum Mathematikunterricht 2025.

KI-GESTÜTZTE PROBLEMLÖSUNGEN IN DER MATHEMATIK

EINE UNTERSUCHUNG ZU PROMPT-TECHNIKEN UND SPRACHMODELLEN

THEORETISCHER HINTERGRUND

GENERATIVE KI IM BILDUNGSKONTEXT

In der vorliegenden Untersuchung wurden verschiedene generative KI-Sprachmodelle eingesetzt, um deren Potenziale und Herausforderungen im Bezug auf mathematische Problemlösungen zu evaluieren. Zu den verwendeten Systemen zählen **ChatGPT-o3-mini**, **DeepSeek-R1**, **Claude 3.5 Sonnet** und **Gemini 1.5 Pro (Advanced)**. Diese Modelle basieren auf fortschrittlichen, generativen KI-Technologien, die menschliche Sprache verstehen und – teils auch multimodale Eingaben – verarbeiten können.

Wie ihre Vorgänger beruhen auch diese KI-Sprachmodelle auf umfangreichen Trainingsdatensätzen, deren Analyse darauf abzielt, Muster und Zusammenhänge zu erkennen, um adäquate, menschenähnliche Antworten zu generieren (Hiemstra, 2009; Hadi et al., 2023). Während frühere Systeme, wie beispielsweise GPT-3 mit seinen 175 Milliarden Parametern (Floridi und Chiriatti, 2020), bereits beeindruckende Leistungen erbrachten, zeigen die aktuell eingesetzten Modelle insbesondere in der Bearbeitung komplexer Aufgaben und mathematischer Fragestellungen erhebliche Fortschritte.

BLACK-BOX PROBLEM

Besonders bei komplexen KI-Sprachmodellen werden Entscheidungen durch mehrschichtige, oft undurchschaubare Algorithmen getroffen, sodass für den Menschen nicht ersichtlich ist, wie genau ein bestimmtes Ergebnis zustande kommt (Herm et al., 2021; Franzoni, 2023). Diese mangelnde Erklärbarkeit führt dazu, dass wesentliche Informationen über die Funktionsweise der KI verborgen bleiben und die internen Prozesse nur schwer überprüfbar sind. Ohne Einsicht in den Entscheidungsprozess fällt es Lehrkräften und Schülerinnen und Schülern schwer, die Richtigkeit und Fairness der von der KI generierten Ergebnisse zu beurteilen. Dies kann das Vertrauen in die Technologie erheblich beeinträchtigen und zu einer ablehnenden Haltung gegenüber ihrer Nutzung führen.

Trotz der in den letzten Jahren erzielten Fortschritte in der mathematischen Problemlösung durch KI-Sprachmodelle – wie etwa Verbesserungen bei der Verarbeitung mathematischer Eingaben und der Lösung komplexer Aufgaben (Hendrycks et al., 2021; Lewkowycz et al., 2022; Frieder et al., 2023) – bleibt das Black-Box-Problem ein zentrales Hindernis.

EXPLAINABLE AI

Explainable AI (XAI) zielt darauf ab, komplexe KI-Systeme transparenter zu machen und die Black-Box-Modelle in „Gray-Box“-Modelle zu überführen (Gunning et al., 2019). Lokale Erklärbarkeit soll aufzeigen, warum eine bestimmte Entscheidung getroffen wurde (Adadi und Berrada, 2018; Arrieta et al., 2020). Unsere Studie stützt sich hierbei auf zwei Ansätze: den Einsatz menschlicher Raterinnen und Rater und Prompt Engineering.

Menschliche Expertinnen und Experten bewerten die Qualität der von KI-Sprachmodellen erzeugten Ausgaben, indem sie wiederholt eingegebene, kontrollierte Prompts analysieren (Qiu et al., 2017; Schoenfeld, 1992). Dabei fließen sowohl inhaltliche als auch prozessbezogene Kriterien mathematikdidaktischer Qualität von Problemlösungen ein (Pólya, 1957).

Durch gezielte Gestaltung der Eingaben – etwa mit Zero-Shot, Chain-of-Thought, Persona oder naïve RAG Prompts (Kojima et al., 2022; Yang, 2024; Olea et al., 2024; Gao et al., 2024) – können KI-Sprachmodelle nicht nur Antworten, sondern auch für den praktischen Unterricht nützliche Erklärungen liefern. Solche Techniken verbessern vor allem bei mathematischen Aufgaben die Nachvollziehbarkeit und Genauigkeit der Lösungen (Schorcht et al., 2023) und bieten somit vielfältige Möglichkeiten, KI-gestützte Problemlöseprozesse in Bildungskontexten produktiv einzusetzen.

FORSCHUNGSFRAGEN

- 1. WIE BEEINFLUSSEN UNTERSCHIEDLICHE PROMPT-TECHNIKEN DIE MATHEMATIKDIDAKTISCHE QUALITÄT DER KI-GESTÜTZTEN LÖSUNGEN?**
- 2. WIE VARIERT DIE QUALITÄT DER KI-GENERIERTEN PROBLEMLÖSUNGEN ZWISCHEN DEN MODELLEN?**
- 3. WIE KÖNNEN KI-GESTÜTZTE PROBLEMLÖSUNGEN DIDAKTISCH SO OPTIMIERT WERDEN, DASS SIE LERNPROZESSE IM UNTERRICHT GEZIELT UNTERSTÜTZEN?**

METHODIK

KI-SPRACHMODELLE:

- **OpenAI:** ChatGPT-o3-mini (Jan 2025)
- **Anthropic:** Claude 3.5 Sonnet (June 2024)
- **Google:** Gemini Advanced 1.5 Pro (Feb 2024)
- **DeepSeek:** DeepSeek-R1 (Jan 2025)

PROMPT-TECHNIKEN (SCHORCHT ET AL., 2023):

- **Zero-Shot:** direkte Eingabe der Problemstellung ohne zusätzlichen Input
- **Chain-of-Thought:** Eingabe der Problemstellung mit Aufforderung zur Schritt-für-Schritt Lösung
- **Persona:** rollenbasierte Eingabe; „Act as a mathematics teacher providing worked-out solutions for mathematical problems in the classroom. Break down complex problems, highlight key concepts, and use age-appropriate language for students. Discuss alternate approaches and address common mistakes. Connect mathematics to real-world examples when possible. Focus on building understanding, not just getting answers. Here is your task: [Aufgabe]“
- **Naïve RAG:** literaturbezogene Eingabe; „Please answer the above question based on the provided file.“

PROBLEMSTELLUNGEN (BEISPIELE):

PROBLEMBEREICH: NUMBER

- **Orchard Problem:** “Three boys agree to divide a bag of marbles in the following manner. The first boy takes one more than half the marbles. The second takes a third of the number remaining. The third boy finds that he is left with twice as many marbles as the second boy.” (Leonardo de Pisa, 1202)

PROBLEMBEREICH: ALGEBRA

- **Car Problem:** “A car travels at the speed of 10 kph. Four hours later a second car leaves to overtake the first car, using the same route and going 30 kph. In how many hours will the second car overtake the first car?” (Cooper and Sweller, 1987, p. 361)

DATENGEWINNUNG:

Für jede Kombination aus KI-Sprachmodell, Prompt-Technik und Problemstellung wurden systematisch und kontrolliert 30 Antworten generiert. Daraus resultiert ein Datensatz von insgesamt 3.840 Antworten (4 Modelle x 4 Prompt-Techniken x 8 Problemstellungen (aus den zwei Bereichen Number und Algebra) x 30 Wiederholungen), die anhand von 11 mathematikdidaktischen Kriterien kodiert wurden.

DATENAUSWERTUNG

QUALITÄTSKRITERIEN (SCHORCHT ET AL., 2024):

- **Spezifität:** Dieses Kriterium gibt an, ob die Lösung alle relevanten Informationen enthält, wie Variablen und Beschreibungen, die zur Lösung des Problems erforderlich sind.
- **Klarheit:** Dieses Kriterium gibt an, ob die Lösung klar und prägnant formuliert ist.
- **Korrektheit:** Dieses Kriterium gibt an, ob die Lösung korrekt ist.
- **Strategien:** Dieses Kriterium gibt an, ob die Lösung heuristische Beschreibungen des Vorgehens enthält.
- **Darstellungen:** Dieses Kriterium gibt an, ob die Lösung eine Umwandlung zwischen verschiedenen Darstellungen enthält (zum Beispiel von Worten in eine andere Form, wie einen Funktionsgraphen oder eine Gleichung).
- **Reflexion:** Dieses Kriterium gibt an, ob die Lösung einen Rückblick darauf enthält, wie die KI vorgegangen ist.
- **Handlungsaufforderung:** Dieses Kriterium gibt an, ob die Lösung auf-fordernde, imperative oder motivierende Elemente enthält.
- **Situationsbewusstsein:** Dieses Kriterium gibt an, ob die Lösung einen Verweis auf einen plausiblen Anwendungsbezug enthält.
- **Häufige Fehler:** Dieses Kriterium gibt an, ob die Lösung eine Beschreibung möglicher Missverständnisse oder Fehler enthält.
- **Alternativansätze:** Dieses Kriterium gibt an, ob die Lösung eine Beschreibung von mehr als einem Ansatz oder einen Hinweis auf einen anderen Ansatz enthält.
- **Literaturbezug:** Dieses Kriterium gibt an, ob die Lösung einen plausiblen Verweis auf Pólya's "How to solve it" oder einen ähnlichen Verweis enthält.

EXEMPLARISCHE ANTWORT (CLAUDE 3.5 SONNET; ORCHARD PROBLEM; PERSONA-PROMPT):

Let me help you solve this interesting problem by **working backwards**. This is a great example of **using reverse reasoning and algebraic thinking**! Let's break it down step by step. Let's first understand what happens:

The man starts with some apples (let's call this number x); He goes through 7 guards; At each guard, he gives away half of his remaining apples plus 1 more; He ends up with 1 apple

Let's solve it starting from the end:

1. After the 7th guard, he has 1 apple:
Before meeting the 7th guard, he had some amount (let's call it n_7); After giving half of n_7 plus 1 apple, he had 1 apple left; We can write: $n_7/2 - 1 = 1$; Therefore: $n_7 = 4$ apples
 2. Let's work backwards for each guard:
Before 7th guard: 4 apples; Before 6th guard: 10 apples (because $10/2 + 1 = 6$, leaving 4); Before 5th guard: 22 apples (because $22/2 + 1 = 12$, leaving 10); Before 4th guard: 46 apples (because $46/2 + 1 = 24$, leaving 22); Before 3rd guard: 94 apples (because $94/2 + 1 = 48$, leaving 46); Before 2nd guard: 190 apples (because $190/2 + 1 = 96$, leaving 94); Before 1st guard: 382 apples (because $382/2 + 1 = 192$, leaving 190)
- Therefore, he must have gathered 382 apples in the orchard! **Let's verify this is correct by working forward:**
Started with 382 apples; After 1st guard: $382/2 + 1 = 192$ taken, leaving 190; After 2nd guard: $190/2 + 1 = 96$ taken, leaving 94; After 3rd guard: $94/2 + 1 = 48$ taken, leaving 46; After 4th guard: $46/2 + 1 = 24$ taken, leaving 22; After 5th guard: $22/2 + 1 = 12$ taken, leaving 10; After 6th guard: $10/2 + 1 = 6$ taken, leaving 4; After 7th guard: $4/2 + 1 = 3$ taken, leaving 1 ✓

Common mistakes to avoid:

1. Don't try to solve forward - it's much harder!
2. Don't forget to add the extra apple at each step
3. Remember that each new amount must be even (why? because we need to divide it by 2)

Real-world connection: This type of problem solving (working backwards) is used in many real situations, like:

Planning a project schedule from a deadline; Calculating savings needed for a future purchase; Recipe scaling when you know the final amount needed

Would anyone like to try creating a similar problem with different numbers? Or shall we explore why working backwards was easier than working forwards in this case?

FABIAN ANTON MÜLLER¹, PROF. DR. SEBASTIAN SCHORCHT² & PROF. DR. NILS BUCHHOLTZ¹

¹UNIVERSITÄT HAMBURG – ARBEITSBEREICH MATHEMATIKDIDAKTIK SEKUNDARSTUFE

²TECHNISCHE UNIVERSITÄT DRESDEN – PROFESSUR FÜR GRUNDSCHULPÄDAGOGIK/MATHEMATIK