

Paul Tschisgale¹
 Katja Plicht²
 André Meyer³
 Fabian Kieser⁴

¹IPN Kiel
²Universität Paderborn
³Leibniz Universität Hannover
⁴Freie Universität Berlin

KI trifft Problemlösen: Chancen und Herausforderungen für die Physikbildung

Künstliche Intelligenz (KI) – insbesondere in Form leistungsfähiger Large Language Models (LLMs) – verändert den Bildungsbereich in rasantem Tempo und wirft grundlegende Fragen zu ihrer Rolle in Lehr- und Lernprozessen auf. Im Physikunterricht und -studium nimmt der Erwerb von Problemlösefähigkeiten eine zentrale Rolle ein, da diese Fähigkeiten für Karrieren in der Physik von entscheidender Bedeutung sind (Mulvey & Pold, 2020). Wie Lernende in der Entwicklung solcher Fähigkeiten bestmöglich unterstützt werden können, stellt eine zentrale Fragestellung der physikdidaktischen Forschung dar. Dabei ist bekannt, dass insbesondere regelmäßige Übung und gezieltes Feedback wesentliche Beiträge zum Lernerfolg leisten (Gaigher et al., 2007).

In den letzten Jahren hat die Diskussion über die Rolle von KI – insbesondere generativer KI in Form von Large Language Models (LLMs) – im Bildungsbereich stark an Bedeutung gewonnen. Diese Entwicklung ist nicht zuletzt darauf zurückzuführen, dass die Fähigkeiten solcher Modelle in nahezu allen Anwendungsbereichen stetig zunehmen. Die potenziellen Einsatzmöglichkeiten in der naturwissenschaftsdidaktischen Forschung sind vielfältig (vgl. Abb. 1).

Vor diesem Hintergrund wurden im Rahmen des vorliegenden Symposiums die Chancen und Herausforderungen des KI-Einsatzes im Kontext physikalischen Problemlösens aus unterschiedlichen Perspektiven beleuchtet. Ausgehend von einer historischen Einordnung der Verflechtungen von KI- und Problemlöseforschung wurden die vermeintlichen Problemlösefähigkeiten von LLMs analysiert sowie der Einsatz von KI als automatisiertes Assessment-Tool und als individueller Tutor untersucht.

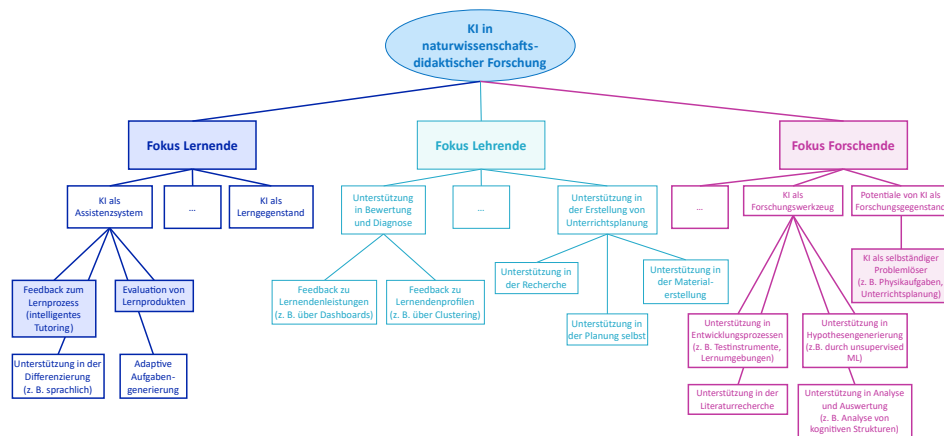


Abb. 1. Einsatzbereiche künstlicher Intelligenz in der naturwissenschaftsdidaktischen Forschung.

Von GPS zu GPT? Zur Verschränkung von KI und Problemlöseforschung

In diesem Beitrag werden historische Verschränkungen der KI- und Problemlöseforschung betrachtet, um über gemeinsame Konzepte aus der Kognitionspsychologie einen Zugang zum Konzept Künstlicher Intelligenz aus Perspektive der Naturwissenschaftsdidaktik anzubieten. Bereits zu Beginn der KI-Forschung in den 1950er Jahren gab es eine Vielzahl an Arbeiten, die sich zur programmatischen Bearbeitung von Problemen an den kognitiven Prozessen des Problemlösens orientierten (Buchanan et al., 2018). Es entstanden eine Reihe von Problemlöse-Programmen, wie dem Logic Theorist (LT, Newell & Simon 1956) oder dem General Problem Solver (GPS, Ernst & Newell, 1969), deren Prozesse sich dem algorithmischen und heuristischen Problemlösen (Funke & Zumbach, 2006) zuordnen lassen. An dieser Stelle werden sie dem Maschinellen Lernen im Sinne des beispielbasierten Problemlösens gegenübergestellt und hinsichtlich des Problemlöseprozesses und ihrer Erklärbarkeit und Interpretierbarkeit (Bauckhage et al., 2021) verglichen.

Während algorithmische und heuristische Systeme eine hohe Ähnlichkeit zu menschlichen kognitiven Prozessen abbilden sollen (Buchanan et al., 2018), sodass naturgemäß eine hohe Transparenz und damit Erklärbarkeit vorausgesetzt werden kann, ergibt sich bei der Entscheidung für Verfahren des Maschinellen Lernens i. d. R. die Hinnahme einer Black-Box (Bauckhage et al., 2021) für die Lösung des Problems, welche keine erklärenden und insbesondere keine kausalen Rückschlüsse mehr ermöglicht. Gleichzeitig bieten Methoden des Maschinellen Lernens die Möglichkeit der Bearbeitung sehr großer Datenmengen und somit auch umfangreicher Probleme, sodass die Nutzung dieser verschiedenen Methoden ebenso wie deren Interpretation stets vor den beiden Aspekten des Problemumfangs und der Erklärbarkeit reflektiert werden sollte (weiterführend s. Einzelbeitrag von Plicht in diesem Tagungsband).

LLMs vs. PhysikOlympiade-Teilnehmende: Wer kann besser Probleme lösen?

LLMs sind inzwischen breit verfügbar – auch für Lernende aller Bildungsstufen. Damit wächst die Sorge, dass durch ihren Einsatz erwünschte Lernprozesse umgangen und die Integrität bestimmter Prüfungsformate beeinträchtigt werden könnte. Für die Physikdidaktik ist es daher zentral, die sich rasch entwickelnden fachspezifischen Fähigkeiten dieser Modelle zu verstehen, um ihren Einsatz didaktisch sinnvoll und verantwortungsvoll zu gestalten.

In dieser Studie wurden anhand anspruchsvoller Wettbewerbsprobleme der deutschen PhysikOlympiade die physikalischen Problemlösefähigkeiten eines general-purpose LLMs (*GPT-4o*) und eines reasoning-optimierten Modells (*o1-preview*) systematisch mit denen von Teilnehmenden der PhysikOlympiade verglichen. Dabei kamen verschiedene Prompting-Techniken zum Einsatz. Beide Modelle erzielten im Durchschnitt bessere Leistungen als die Teilnehmenden, wobei *o1-preview* sowohl *GPT-4o* als auch die Teilnehmenden im Durchschnitt übertraf. Der Einfluss der Prompting-Techniken war insgesamt gering. Qualitative Analysen zeigten typische Fehlermuster in LLM-generierten Lösungen, etwa fehlende Herleitungen nicht-trivialer Zusammenhänge oder Schwächen bei der Modellierung physikalischer Systeme.

Die Ergebnisse verdeutlichen, dass die Integrität unüberwachter Prüfungsformate (z. B. Hausaufgabenrunden in Wettbewerben, Take-Home-Prüfungen, bewertete Hausaufgaben) durch den Einsatz von LLMs zusätzlich gefährdet ist – über die ohnehin bestehende Möglichkeit externer Unterstützung hinaus. LLM-generierte Lösungen lassen sich derzeit nicht zuverlässig erkennen, und ungleicher Zugang zu diesen Systemen schafft Fairnessprobleme. Kurzfristig erscheint eine Abkehr von unüberwachten Formaten oder die gezielte Nutzung aktueller Schwächen der Modelle sinnvoll; langfristig sollte jedoch eine Integration von KI im Sinne einer *Human-AI-Collaboration* angestrebt werden.

Für die vollständige Studie und eine weiterführende Diskussion der Befunde siehe Tschisgale et al. (2025) und Tschisgale (2025).

Automatische Bewertung physikalischer Problemlösungen mit und ohne KI

In der Beforschung von KI Feedback wird überwiegend auf proprietäre LLMs wie GPT zurückgegriffen, wohingegen die Nutzung von open-source Lösungen noch eine Forschungslücke darstellt. Datenschutz und die Unabhängigkeit von Konzernen können aber für die Nutzung solcher Lösungen sprechen. In diesem Beitrag zum Symposium wurden zwei Studien vorgestellt, die auf unabhängige Techniken für die Bewertung von Problemlöseaufgaben zurückgreifen.

Mithilfe des open-source LLMs *OpenChat 3.6* wurde Feedback auf Freitextantworten generiert, in denen Schülerinnen und Schüler vollständige Problemlösungen evaluieren. Die Studie konnte zeigen, dass dieses LLM ohne Verbindung zu externen Servern unmittelbares Feedback erzeugen kann. Das Feedback wurde im Hinblick auf korrekte Bewertung, fachliche Richtigkeit und Angemessenheit analysiert. Dabei konnten zufriedenstellende Ergebnisse erzielt werden, die vergleichbar mit *GPT-4o* sind. Probleme traten auf, wenn das LLM falsche Fachbegriffe aus den Antworten reproduzierte oder bei der Interpretation von Zahlenwerten. Für die vollständige Studie und eine weiterführende Diskussion der Befunde siehe Meyer et al. (2025).

In der zweiten Studie wurde die Bewertung von Gleichungen mithilfe regelbasierter Verfahren exploriert. Die Bewertung von Gleichungen stellt eine große Herausforderung dar, weil semantische Aspekte wie die Benennung von Variablen sowie logische Aspekte wie mathematische Korrektheit betrachtet werden müssen. Das explorierte Verfahren konnte logische Aspekte überprüfen und erzeugt keine falsch-positiven Bewertungen, hat allerdings noch Probleme mit den semantischen Aspekten. Häufig werden die Gleichungen nicht richtig erkannt, wodurch die Bewertung fehlschlägt. In Zukunft kann dieses Verfahren als effiziente Vorprüfung für ein komplexeres Auswertungsverfahren dienen.

Individuelles Tutoring durch generative KI beim physikalischen Problemlösen

Um Lernenden das Verbessern ihrer physikalischen Problemlösefähigkeiten zu ermöglichen, ist es unerlässlich in den verschiedenen Phasen des Problemlöseprozesses zeitnahes und spezifisches Feedback zu geben (Maries et al, 2023). Die Gestaltung und Auswertung von Aufgabenformaten, die über reine Punktzahlen hinausgehende Einblicke in das Verständnis der Lernenden ermöglichen, stellt in großen Lernszenarien wie Vorlesungen oder Schulklassen eine besondere Herausforderung dar (Wilcox & Pollock, 2014). Zahlreiche Studien zeigen, dass sofortiges Feedback deutlich wirksamer ist als verzögertes Feedback (vgl. Opitz et al., 2011). Vor diesem Hintergrund entstand das Ziel, Intelligente Tutoring-Systeme (ITS) zu entwickeln, die Lernende individuell und automatisch im Lernprozess begleiten. Ein wesentlicher Vorteil solcher Systeme liegt in ihrer Automatisierung, die eine konsistente und jederzeit verfügbare Unterstützung ermöglicht (Gertner & vanLehn, 2000).

ITS müssen auf unterschiedlichen Ebenen adaptiv agieren, um Lernprozesse optimal zu unterstützen. Dabei wird zwischen einer Makroebene, die sich mit der Auswahl von Lernaufgaben auf Basis eines kontinuierlichen Lernendenmodells befasst (Shute, 1993), und einer Mikroebene, die Hilfestellungen während der Aufgabenbearbeitung anpasst (Rus et al., 2013), unterschieden. Vor allem auf Mikroebene eröffnen LLMs neue Potenziale für adaptive Lernsysteme. Sie können auf ein breites Spektrum von Antworten der Lernenden flexibel reagieren (Yin et al., 2024), während ältere, dialogbasierte Tutoring-Systeme noch deutlich aufwendiger zu entwickeln waren und meist auf stichwortbasierte Verfahren der computerbasierten Sprachverarbeitung zurückgreifen mussten (Litmann et al., 2004). Im Einklang mit aktuellen Forschungsergebnissen weisen die Ergebnisse dieses Beitrags insbesondere bei Sonderfällen auf Grenzen der Adaptivität im LLM-generierten Feedback hin.

Literatur

- Bauckhage, C., Fürnkranz, J. & Paaß, G. (2021). Vertrauenswürdiges, transparentes und robustes maschinelles Lernen. In G. Görz, U. Schmid, & T. Braun (Hrsg.), *Handbuch der Künstlichen Intelligenz* (6. Aufl., S. 429–508). Walter de Gruyter. <https://doi.org/10.1515/9783110659948-013>
- Buchanan, B. G., Davis, R., Smith, R. G. & Feigenbaum, E. A. (2018). Expert systems: A perspective from computer science. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt & A. M. Williams (Hrsg.), *The Cambridge Handbook of Expertise and Expert Performance* (2. Aufl., S. 84–104). Cambridge University Press.
- Ernst, G. W. & Newell, A. (1969). *GPS: A case study in generality and problem solving*. Academic Press.
- Funke, J. & Zumbach, J. (2006). Problemlösen. In H. Mandl & H. F. Friedrich (Hrsg.), *Handbuch Lernstrategien* (S. 206–220). Hogrefe Verlag GmbH & Company KG.
- Gaigher, E., Rogan, J. M. & Braun, M. W. H. (2007). Exploring the development of conceptual understanding through structured problem-solving in physics. *International Journal of Science Education*, 29(9), 1089–1110. <https://doi.org/10.1080/09500690600930972>
- Gertner, A. S. & VanLehn, K. (2000). Andes: A coached problem solving environment for physics. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (pp. 133–142). Springer-Verlag. <https://doi.org/10.5555/648030.745825>
- Litman, D. J. & Silliman, S. (2004). ITSPKE: An intelligent tutoring spoken dialogue system. In *Proceedings of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Association for Computational Linguistics. <https://aclanthology.org/N04-3002/>
- Maries, A. & Singh, C. (2023). Helping students become proficient problem solvers part I: A brief review. *Education Sciences*, 13(2), 156. <https://doi.org/10.3390/educsci13020156>
- Meyer, A., Bleckmann, T. & Friege, G. (2025). Automatic feedback on physics tasks using open-source generative artificial intelligence. *International Journal of Science Education*, 1–26. <https://doi.org/10.1080/09500693.2025.2499220>
- Mulvey, P. & Pold, J. (2020). *Physics doctorates: Skills used and satisfaction with employment*. American Institute of Physics.
- Newell, A. & Simon, H. A. (1956). The logic theory machine: A complex information processing system. *IRE Transactions on Information Theory*, IT-2, No. 3, 61–79.
- Opitz, B., Ferdinand, N. K. & Mecklinger, A. (2011). Timing matters: The impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience*, 5, 8. <https://doi.org/10.3389/fnhum.2011.00008>
- Rus, V., Baggett, W., Gire, E., Franceschetti, D., Conley, M. & Graesser, A.C. (2013). Towards learner models based on learning progressions in DeepTutor. In Sottolare, R. (Eds.), *Learner Models*, (pp. 185–196) Army Research Lab.
- Shute, V. J. (1993). A macroadaptive approach to tutoring. *Journal of Artif. Intel. in Education*, 4(1), 61–93.
- Tschisgale, P. (2025). Olympiad-level AI performance is here - What now? *Physics*, 18(147). <https://doi.org/10.1103/Physics.18.147>
- Tschisgale, P., Maus, H., Kieser, F., Kroehs, B., Petersen, S. & Wulff, P. (2025). Evaluating GPT- and reasoning-based large language models on physics olympiad problems. *Physical Review Physics Education Research*, 21(2), 020115. <https://doi.org/10.1103/6fmx-bsnl>
- Wilcox, B. R. & Pollock, S. J. (2014). Coupled multiple-response versus free-response conceptual assessment: An example from upper-division physics. *Physical Review Special Topics – Physics Education Research*, 10(2), 020124. <https://doi.org/10.1103/PhysRevSTPER.10.020124>
- Yin, J., Goh, T.-T. & Hu, Y. (2024). Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance. *IEEE Transactions on Learning Technologies*, 17, 1404–1415. <https://doi.org/10.1109/TLT.2024.3364015>