

Bewertung von LLM-Abituraufgaben-Lösungen



Wie gut lösen Large Language Modells (LLM) die Physikaufgaben des ländergemeinsamen Abituraufgabenpools?
Armin Kunz (IQB), Andreas Borowski (Universität Potsdam)

Motivation und Theorie

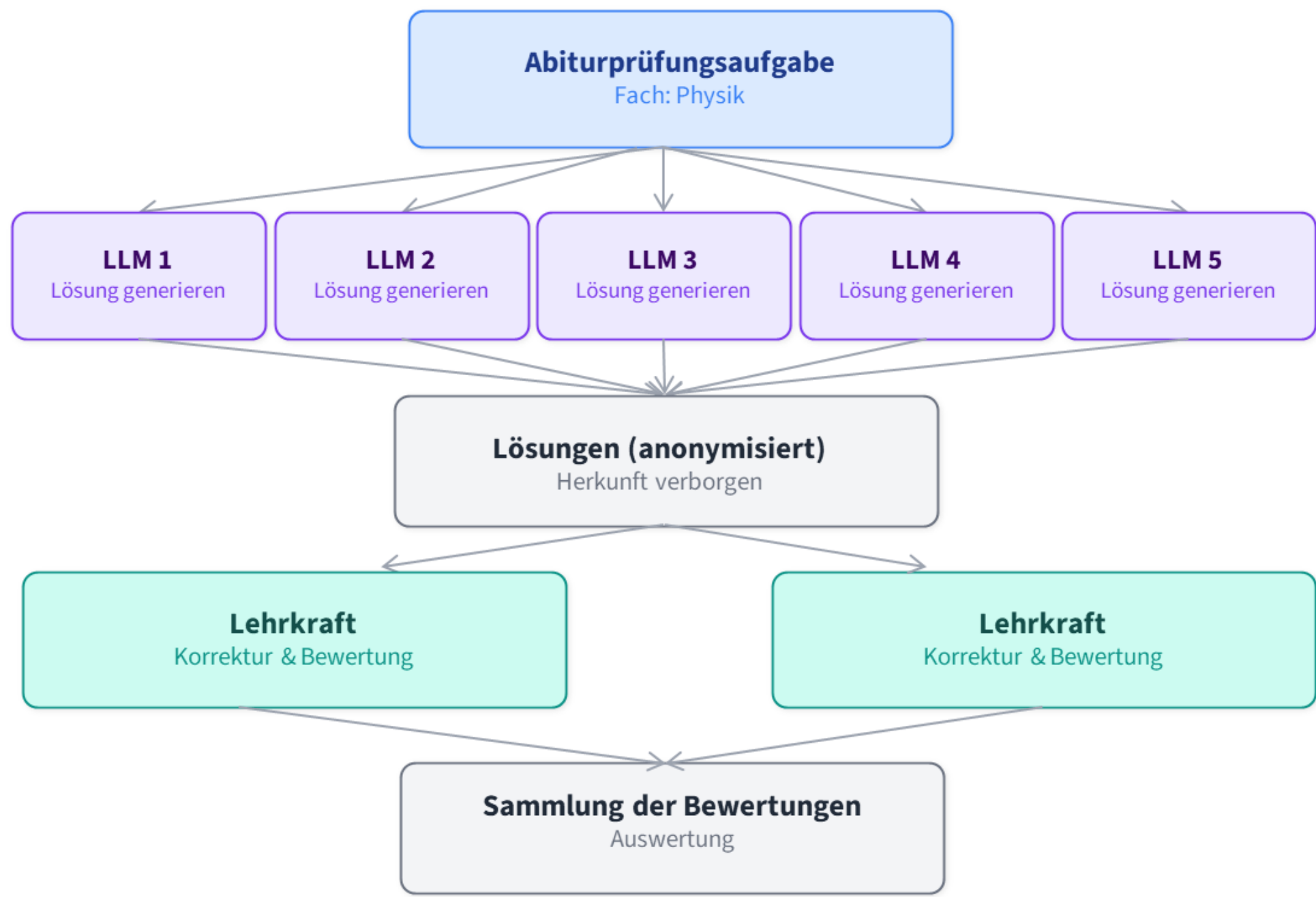
- 2025 findet erstmals die schriftliche Abiturprüfung mit Aufgaben im Fach Physik aus dem ländergemeinsamen Aufgabenpool statt (KMK, 2020a; KMK, 2020b).
- Die Aufgaben sind kontextualisiert und umfassen die Kompetenzbereiche der Sachkompetenz (S), der Erkenntnisgewinnungskompetenz (E), der Kommunikationskompetenz (K) und der Bewertungskompetenz (B) (KMK, 2020a).
- Künstliche Intelligenz in Form der Large Language Models (LLM) wird den Unterricht und die Prüfungskultur nachhaltig beeinflussen (Nehring et al., 2025; Kubsch et al. 2025).
- Aufgaben der ersten Runde der Physikolympiade werden von LLM bereits mehrheitlich besser bearbeitet als von menschlichen Teilnehmenden (Tschisgale et al., 2025; Huang, 2026).
- Im Fach Deutsch treten zum Teil große Differenzen in den Bewertungen derselben Aufgabenlösung durch unterschiedliche Bewertende auf (Schröter et al., 2022).

Hypothesen

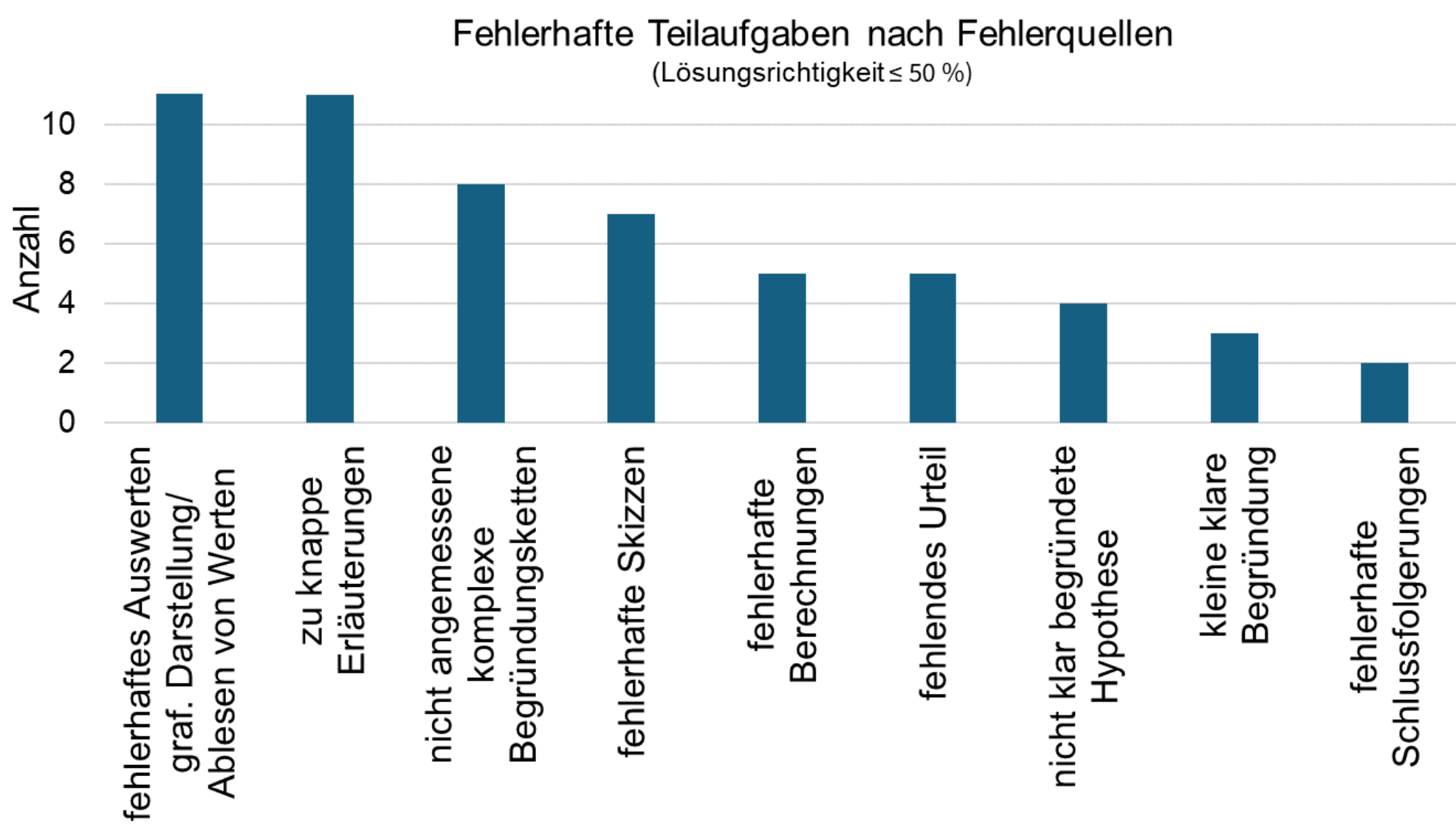
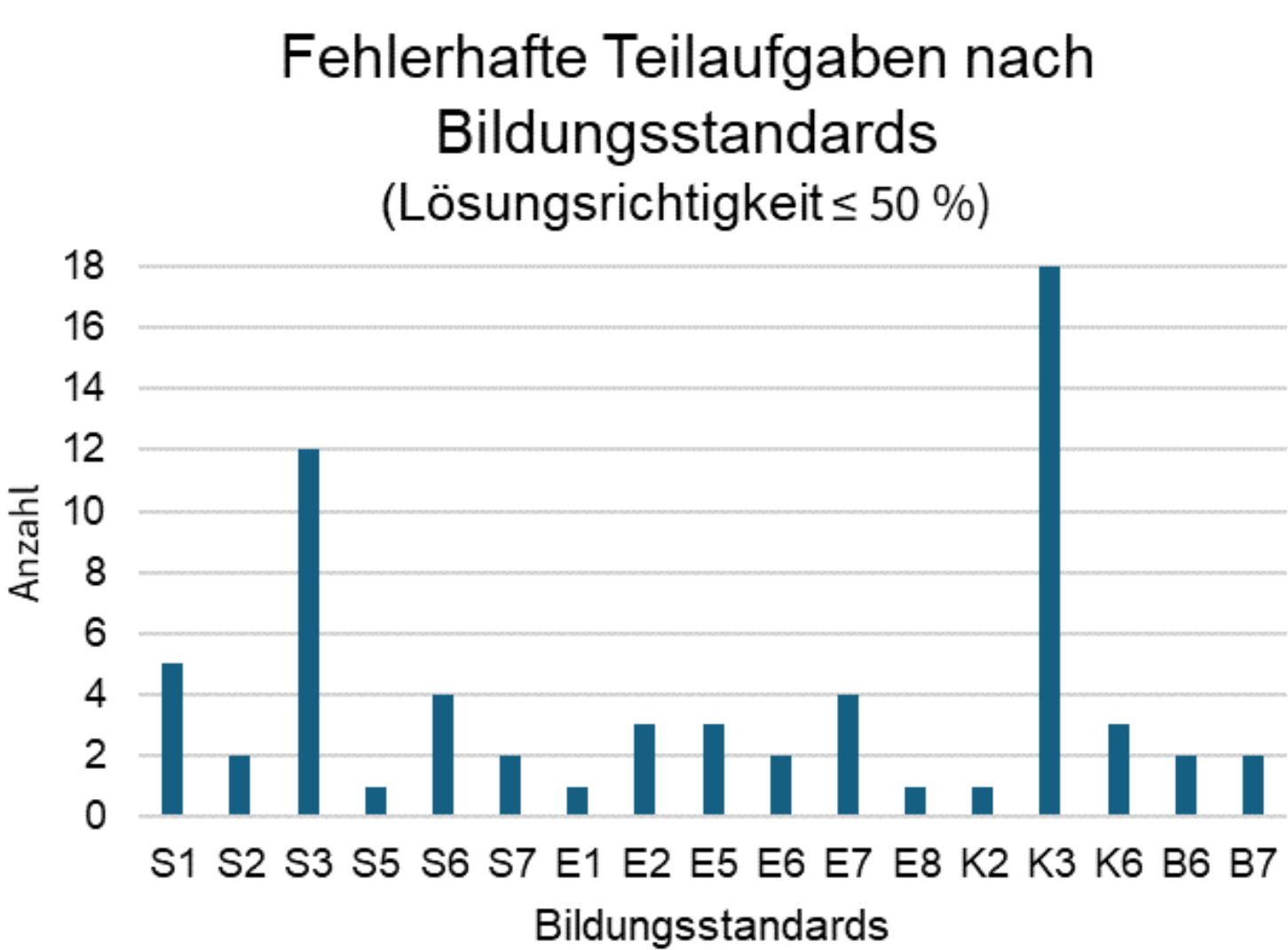
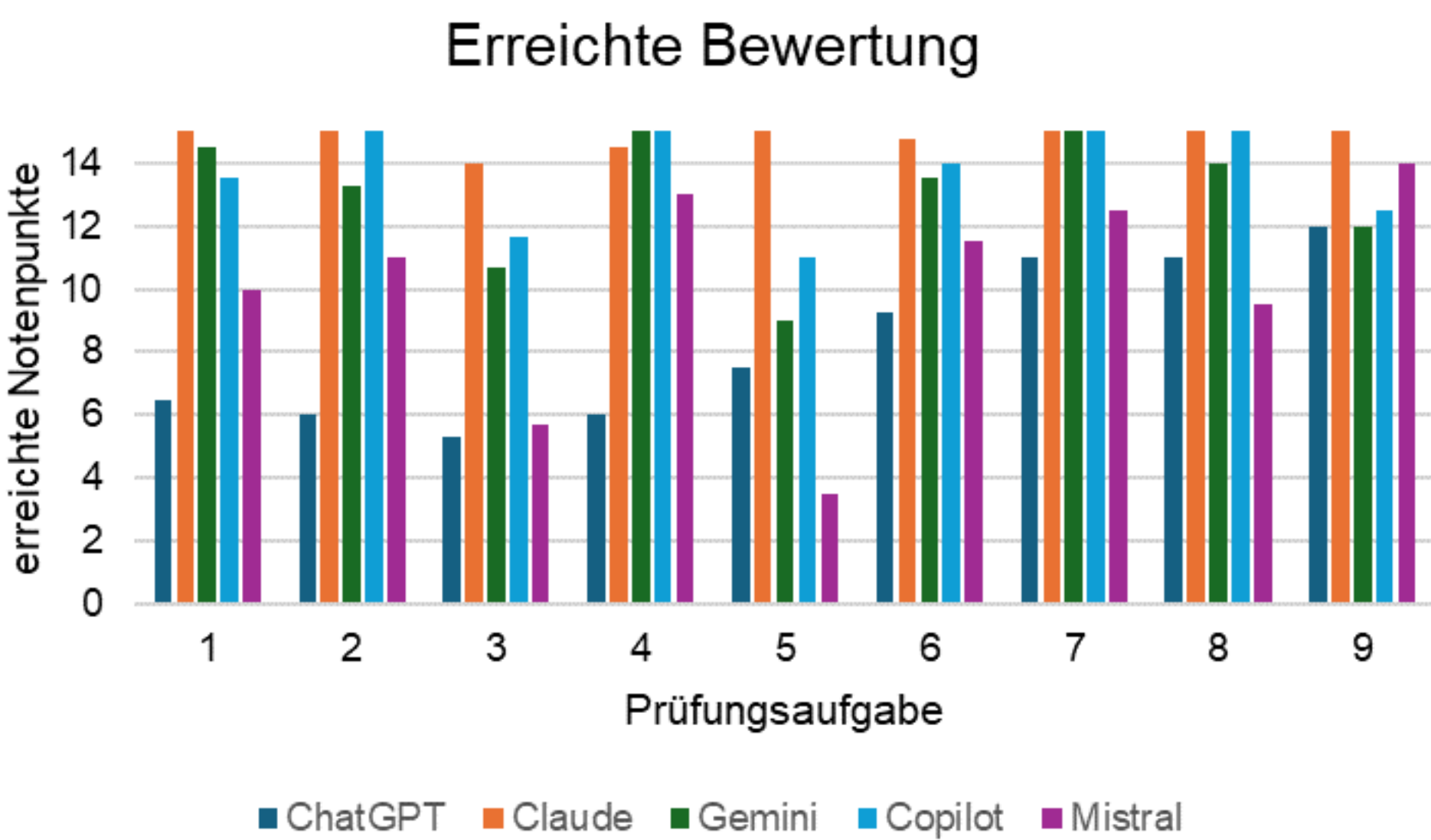
- H1: Die LLM können Aufgaben der schriftlichen Abiturprüfung des ländergemeinsamen Aufgabenpools im Fach Physik erfolgreich bearbeiten.
- H2: Es gibt keine Unterschiede in der Qualität der LLM-Lösungen zwischen den einzelnen Kompetenzbereichen. Das trifft auch auf die Teilaufgaben zu, die die Kompetenzbereiche Kommunikation und Bewertung adressieren.
- H3: Die unabhängig voneinander erstellten Bewertungen der LLM-Lösungen durch die einzelnen Gutachter weisen eine geringere Differenz als im Fach Deutsch auf, da der Struktur der Aufgaben gemäß die Beschreibungen der Erwartungen eindeutiger ausfallen.

Studiendesign

- Auswahl von fünf LLM, die auch für Schülerinnen und Schüler frei verfügbar sind: ChatGPT 5.3, Claude Sonnet 4.6, Google Gemini 3 fast, Microsoft Copilot smart, Mistral Large 2.
- Entwicklung des verwendeten Prompts (→ QR-Code unten auf der Seite) zur Lösung der Aufgaben und Ausgabe der Ergebnisse.
- Lösung von $N_A = 9$ genutzten materialgestützten Abiturprüfungsaufgaben mit je 7-10 Teilaufgaben im Fach Physik des Prüfungsjahres 2026 durch die LLM (insgesamt $N_L = 45$ LLM-Lösungen).
- Korrektur durch Physik-Lehrkräfte: Pro Aufgabe wenigstens zwei unabhängige Bewertungen der anonymisierten Lösungen.
- Sammlung der $N_B = 115$ Bewertungen und deren Auswertung.



Ergebnisse



- Zu H1: Die LLM können Aufgaben der schriftlichen Abiturprüfung des ländergemeinsamen Aufgabenpools über alle Kompetenzbereiche im Fach Physik erfolgreich bearbeiten. Dabei bestehen Varianzen sowohl zwischen den LLM als auch innerhalb einzelner LLM bezüglich unterschiedlicher Aufgaben. Nur in einem von 115 Fällen, lag die Bewertung unter „ausreichend“.
- Zu H2: Kein Kompetenzbereich stellt sich als insgesamt fehlerträchtig heraus. Größte Fehlerquellen liegen bei den Standards S3 („Die Lernenden wählen aus bekannten Modellen bzw. Theorien geeignete aus, um sie zur Lösung physikalischer Probleme zu nutzen.“) und K3 („Die Lernenden entnehmen unter Berücksichtigung ihres Vorwissens aus Beobachtungen, Darstellungen und Texten relevante Informationen und geben diese in passender Struktur und angemessener Fachsprache wieder.“). Diese treten auf bei der fehlerhaften Nutzung gegebener grafischer Darstellungen und bei nicht hinreichend gegebenen Erläuterungen.
- Zu H3: Statistisch zeigt sich maximal eine moderate Interrater-Reliabilität. In Notenpunkten liegen die Gutachten aber nur in vier Fällen weiter als drei Notenpunkte auseinander, was in den meisten Bundesländern eine Drittkorrektur erfordern würde.

Diskussion und Ausblick

- LLM können schriftliche materialgestützte Aufgaben des zentralen Aufgabenpools schon jetzt mit z. T. hervorragenden Resultaten lösen.
- Fehlerhafte Bearbeitungen könnten technischer Natur sein (fehlerhaftes „Erkennen“ von Abbildungen bzw. mangelndes Verständnis der Operatoren) und bei folgenden Generationen der LLM und angepasstem Prompt mit Verweis auf die IQB-Operatorenliste (IQB, 2022) eventuell bereits nicht mehr auftreten.
- Die geringe statistische Interrater-Reliabilität spielt in der Schulpraxis nur eine untergeordnete Rolle, da die Bewertungen in den meisten Fällen keine weitere (Dritt-)Korrektur erfordern würde. Die Bewertungen derselben Aufgabe durch unterschiedliche Bewerter liegen näher beieinander als eine Studie für das Fach Deutsch ergeben hat (Schröter et al., 2022).
- Eine Weiterentwicklung der Prüfungskultur könnte stärker auf eine Prozessorientierung statt einer Ergebnisorientierung ausgerichtet sein (Kortemeyer, 2026).

Kontakt
Armin Kunz
+49 30 2093 46988
Armin.Kunz@iqb.hu-berlin.de

Verwendeter Prompt



Referenzen



Poster-Download

