Ein Kind rechnet 403 − 178 und schreibt 375 hin. Ein KI-Lernbegleiter kann darauf ganz verschieden reagieren. Er kann das Ergebnis korrigieren und gleich die richtige Lösung 225 nennen. Er kann den Fehler übersehen und das Kind fälschlicherweise loben (wie es viele KI-Modelle von Haus aus machen). Oder er erkennt, dass das Kind an jeder Stelle die kleinere von der größeren Ziffer abgezogen hat, und regt es mit einer gezielten Frage an, seine Rechnung selbst zu überprüfen. Alle drei Antworten klingen freundlich und kompetent. Für das Ermöglichen von mathematischem Lernen ist aber die letzte Rückmeldung am besten geeignet, da sie dem Kind unmittelbar eine Möglichkeit eröffnet, aus dem Fehler zu lernen.

Diese Unterschiede versucht **PRIMA-KI-Bench** sichtbar zu machen. Der Benchmark ist Teil des Forschungsprojekts PRIMA-KI an der PH Weingarten und untersucht, wie gut aktuelle KI-Modelle Kinder beim Mathematiklernen in der Grundschule begleiten. Rangliste, Situationen und sämtliche Antworten der Modelle sind öffentlich einsehbar: prima-ki-bench.urff.app

Warum ein eigener Benchmark?

KI-Modelle ändern sich schneller, als Lernstudien sie untersuchen können. Was eine Studie für ein bestimmtes Modell zeigt, kann mit der nächsten Modellversion schon überholt sein. Gängige KI-Ranglisten helfen jedoch für didaktische Zweckt nur begrenzt weiter, denn sie messen vor allem, ob ein Modell Aufgaben richtig löst. Für eine Lernbegleitung ist es zwar wichtig, dass das Modell die richtige Lösung weiß, aber sie nicht direkt ausgibt: Ein Modell, das jede Aufgabe sofort korrekt vorrechnet, nimmt dem Kind womöglich gerade das Denken ab, auf das es ankommt.

PRIMA-KI-Bench testet deshalb Merkmale lernwirksamer Lernbegleitung, die in echten Lernsettings anwendungsintegriert oder auch in einem Chat-Assistenten oder freien Chatbot realisierbar werden. Beispielsweise wird getestet: Antwortet die KI passend zu den Eingaben des Kindes und zu den mitgegebenen Kontextinformationen? Gibt sie dosierte Hilfe, statt die Lösung vorwegzunehmen wenn sie dazu instruiert wurde? Regt sie zum eigenen Nachdenken an durch gezielte Impulse? Ist die Rückmeldung fachlich korrekt, informativ und kindgerecht formuliert? Geprüft wird dabei immer das Zusammenspiel aus Modell und Anweisung, so wie es in einer Lern-App zum Einsatz käme.

Wie wird gemessen?

Der Benchmark umfasst 143 Situationen aus dem Mathematikunterricht der

Grundschule: typische Rechenfehler, Kinder, die nicht weiterwissen oder einfach nur die Lösung haben wollen, und Situationen aus Lern-Apps, in denen das Sprachmodell einen Bildschirminhalt einordnen und bei der Antwort berücksichtigen muss. 104 Situationen, die in dem Benchmark getestet werden, sind mit allen Antworten öffentlich einsehbar. 39 bleiben unveröffentlicht, damit die Modelle sie nicht aus ihren Trainingsdaten kennen können.

Zu jeder Situation sind vier bis acht genau beschriebene Kriterien formuliert, über die eine Ausgabe des Sprachmodells geprüft und bewertet wird. Manche Fehler sind als so schwerwiegend eingestuft, dass die Antwort insgesamt null Punkte erhält, etwa wenn die KI ein falsches Ergebnis bestätigt, selbst falsch rechnet oder dem Kind die Lösung verrät. Solche Situationen sollten nicht vorkommen und behindern das Lernen des Kindes. Die Antworten bewerten drei KI-Modelle verschiedener Anbieter, ohne zu erfahren, von welchem Modell eine Antwort stammt; je Kriterium entscheidet die Mehrheit. Zusätzlich werden Antworten mit menschlichen Ratern gegengecheckt. Ob ein bewertendes Modell die Modelle des eigenen Anbieters bevorzugt, wird extra kontrolliert. Zusätzlich werden Antwortzeit und Kosten erfasst, denn auch sie entscheiden darüber, ob sich ein Modell für den Einsatz in der Schule eignet. Bei unmittelbarem Feedback ist es wichtig, dass die Antwort in einer vertretbaren Zeit zurückkommt. Und Kosten spielen für einen breiten Einsatz ebenfalls eine Rolle, genaus etwa ob das Modell datenschutzkonform betrieben werden kann oder gar lokal ohne Cloud auskommt.

Jedes Modell durchläuft den Benchmark unter zwei Bedingungen: einmal mit einer ausführlichen fachdidaktischen Anweisung, wie sie eine Lern-App mitgeben würde, und einmal nur mit einem Satz zu Rolle und Zielgruppe (wie in einem Chat-Assistenten).

Was zeigt sich bisher?

Derzeit sind 25 Modelle erfasst (Stand: 26. November 2026). Mit fachdidaktischer Anweisung erreichen die besten Modelle rund 80 von 100 Punkten, an der Spitze Claude Opus 5.5, GPT-6.1 Sol und Gemini Neon. Auch Modelle mit offenen Gewichten, die sich grundsätzlich auf eigenen Servern betreiben lassen, schneiden gut ab, etwa Qwen 3.8 Flash und DeepSeek V4.1 Flash. Das Sprachmodell, das Apple direkt auf dem Gerät ausführt, kommt dagegen nur auf rund 30 Punkte.

Mindestens so aufschlussreich wie die Rangfolge ist der Vergleich der beiden Bedingungen. Ohne didaktische Anweisung (Fundierung) sinkt der Score im Mittel

um rund 25 Punkte, und der Anteil der Antworten, die wegen eines schweren Fehlers durchfallen, steigt von 10 auf 41 Prozent. Meist liegt das daran, dass die Modelle die Lösung verraten. Wie stark die Modelle nachlassen, ist dabei sehr unterschiedlich: Gemini 4 Neon und Claude Opus 5.5 verlieren 8 bzw. 11 Punkte, GPT-6.1 Sol fällt von 80 auf 50 Punkte. Über die Qualität einer KI-Lernbegleitung entscheidet also nicht allein das Modell, sondern ebenso die didaktische Rahmung, in die es eingebettet ist.

Ob der Benchmark misst, was er messen soll, lässt sich unter anderem an KI-Konfigurationen aus veröffentlichten Lernstudien prüfen, deren Wirkung auf das echte Lernen bekannt ist. In einer Studie von Bastani und Kollegen (2025) etwa schnitten Lernende, die mit einem GPT-4-Chat ohne besondere Anweisung geübt hatten, in der anschließenden Prüfung ohne KI schlechter ab als die Kontrollgruppe; bei einer Variante mit Tutor-Anweisung blieb dieser Effekt weitgehend aus. Der Benchmark ordnet die beiden Varianten deutlich unterschiedlich ein (21 bzw. 62 Punkte). Insgesamt haben sich 10 von 11 vorab festgelegten Hypothesen dieser Art bestätigt.

Wie belastbar sind die Ergebnisse?

Die Ergebnisse sind vorläufig und eher eine praktikable Möglichkeit, neue Modelle auf ihre (fachdidaktische) Eignung in Lernsituationen zu beurteilen. Dazu lassen sich auch die Schwerpunkt der Bewertung je nach Anwendungsfall beliebig anpassen. Erste menschliche Fachurteile stimmen mit den KI-Bewertungen weitgehend überein, was darauf hindeutet dass eine Bewertung über mehrere KI-Rater (abgesichert durch menschliche Rater) durchaus eine praktikable Lösung ist. Der Benchmark ersetzt allerdings keine Lernstudie: Er zeigt, ob ein Modell die Merkmale guter Lernbegleitung umsetzt, nicht, ob Kinder damit besser lernen. Er kann aber helfen, Befunde aus solchen Studien auf neue Modelle zu übertragen und zu sehen, ob die dort gemachten Wirkungen mit hoher Wahrscheinlichkeit auch für andere Modelle gelten und bietet damit die Grundlage für eine begründete Auswahl für eigene Lern-Apps.

Wie kannst du mitmachen?

Wenn du in der Fachdidaktik, in der Schule oder in der Bildungsforschung tätig bist, kannst du die Bewertung unterstützen. Im Kurzvergleich siehst du zwei anonyme KI-Antworten auf dieselbe Situation und entscheidest, welche didaktisch besser ist. Ein Vergleich dauert etwa eine Minute. Aus vielen solchen Urteilen entsteht eine

Rangfolge aus Sicht der Fachleute, die mit der Rangliste des Benchmarks abgeglichen wird und so das Rating fachlich absichert. Wenn du genauer hinsehen möchtest und genauer prüfen möchtest, kannst du Antworten auch nach denselben Kriterien bewerten wie die KI-Modelle.

Der Benchmark wird laufend aktualisiert, sobald neue Modelle erscheinen. Fehlt dir ein Modell, kannst du es im Portal vorschlagen. Über Rückmeldungen und Kooperationen freue ich mich: christian.urff@ph-weingarten.de

Wie kann der Benchmark zitiert werden?

Weil der Benchmark laufend aktualisiert wird, gib beim Zitieren bitte immer Version und Datenstand an. Für den Stand dieses Beitrags lautet der Vorschlag (APA 7):

Urff, C. (2026). *PRIMA-KI-Bench: KI-Lernbegleitung in der Grundschulmathematik* (Version 2.3, Datenstand 26. November 2026) [Benchmark und Datensatz]. Pädagogische Hochschule Weingarten. https://prima-ki-bench.urff.app

Den jeweils aktuellen Vorschlag, auch als BibTeX, findest du im Portal unter Daten & Zitieren.