Christian Urff, Pädagogische Hochschule Weingarten. Erstveröffentlichung: 29. November 2025; überarbeitet: 23. Juli 2026 und 6. September 2026

Zusammenfassung: *KI-gestützte Lernbegleitung kann mathematisches Lernen unterstützen, aber auch Denkarbeit abnehmen, an der Kinder lernen würden. Der Beitrag sichtet den Forschungsstand und leitet vier Gestaltungsprinzipien für den Mathematikunterricht der Grundschule ab: fachdidaktische Fundierung, adaptive Dosierung, metakognitive Aktivierung und unterrichtliche Einbettung. Er benennt, wann auf KI-Lernbegleitung verzichtet werden sollte, was Lehrkräfte an einem System prüfen können und was für die technische Umsetzung folgt. Da die große Mehrheit der empirischen Befunde aus Sekundarstufe, Hochschule oder Erwachsenenstichproben stammt, sind die Prinzipien für die Primarstufe als forschungsbasierte Gestaltungsannahmen zu verstehen, nicht als gesicherte Wirkungsaussagen.*

Inhalt

- Die Ausgangslage
- Theoretischer Rahmen
  - Cognitive Load Theory und Expertise-Umkehr
  - Zone der nächsten Entwicklung und Scaffolding
  - Prozedurales und konzeptuelles Wissen
  - Selbstreguliertes Lernen und Metakognition
  - Was ein System dafür braucht
- Das Risiko der Auslagerung von Denkarbeit
  - Cognitive Offloading: Auslagerung kognitiver Anstrengung
  - Höflichkeit und pädagogische Qualität von Sprachmodellen
  - Die Spannung zwischen wirksamer Hilfe und eigenständigem Denken
- Vier Gestaltungsprinzipien
  - Fundierung: Fachdidaktisches Wissen als Voraussetzung
  - Dosierung: Adaptives Scaffolding und das Prinzip „Erst du – dann ich“
  - Aktivierung: Metakognition und kritisches Prüfen von KI-Antworten
  - Einbettung: Hybride Arrangements und die Rolle der Lehrkraft
- KI als Anstoß für mathematische Aktivitäten mit Material
- Wann auf KI-Lernbegleitung verzichtet werden sollte
- Was du als Lehrkraft prüfen kannst
- Zentrale Belege im Überblick
- Konsequenzen für die technische Umsetzung
- Ethische Anforderungen an KI-Lernbegleitung
- Das SKILL-Modell als Orientierungsrahmen
- Bilanz und offene Fragen

Die Ausgangslage

Generative Künstliche Intelligenz wird im Bildungsdiskurs mit großen Erwartungen oder mit ebenso großen Befürchtungen verbunden. Inzwischen liegen zahlreiche, methodisch und inhaltlich sehr unterschiedliche Studien vor, die für die Gestaltung von KI-Lernbegleitung erste Folgerungen zulassen. Meta-Analysen berichten positive Effekte KI-gestützter Systeme auf Lernleistungen, Motivation und höhere Denkprozesse (Wang & Fan, 2025; Deng et al., 2025; Alemdag, 2025; Wu & Yu, 2024; Zheng, Niu, et al., 2023). Eine Meta-Analyse zu generativer KI in der Mathematik über 22 Studien mit 46 unabhängigen Stichproben findet einen mittleren Effekt von g = 0,53 (Liu et al., 2026). Zur Einordnung: g und d sind

standardisierte Effektstärken, bei denen 0,2 als klein, 0,5 als mittel und 0,8 als groß gilt; r ist ein Korrelationskoeffizient zwischen −1 und 1. Dieselbe Analyse berichtet allerdings größere Effekte bei kleineren Stichproben, ein bekanntes Anzeichen für Publikationsverzerrung. Die positiven Effekte treten zudem unter sehr unterschiedlichen Bedingungen auf, und einige Studiendesigns sind methodisch fragwürdig. Bei der Interpretation sind daher mögliche Verzerrungen und die Qualität der eingeschlossenen Studien zu berücksichtigen.

Die Wirkungsbilanz fällt zudem je nach Systemtyp und Vergleichsbedingung sehr unterschiedlich aus. Für intelligente Tutorsysteme, die regelbasierten Vorläufer heutiger KI-Lernbegleiter, fand die Meta-Analyse von Steenbergen-Hu und Cooper (2013) über 34 unabhängige Stichproben in der Mathematik der Klassenstufen 1 bis 12 im Vergleich zu regulärem Unterricht nur sehr kleine mittlere Effekte (g = 0,01 bis 0,09). Ma et al. (2014) berichten über alle Altersstufen hinweg g = 0,42 gegenüber lehrergeleitetem Klassenunterricht, aber keinen Vorteil gegenüber individuellem menschlichem Tutoring (g = −0,11); Steenbergen-Hu und Cooper (2014) finden für Studierende g = 0,32 bis 0,37. Diese Zahlen sind wegen der unterschiedlichen Vergleichsbedingungen nicht direkt vergleichbar. Ein einfaches Gefälle nach Alter lässt sich daraus nicht ableiten; die Befundlage ist heterogen, und gerade die methodisch strenge Schulanalyse dämpft die Erwartungen am stärksten.

Eine systematische Übersicht zu KI-gestützten Tutorsystemen im Schulbereich (Létourneau et al., 2025) zeigt zudem, dass nur 4 der 28 eingeschlossenen Studien (14 %) Grundschulkinder betreffen. Die weit überwiegende Mehrheit der vorliegenden Studien wurde mit älteren Schülerinnen und Schülern oder Studierenden durchgeführt (Liu et al., 2026; Son, 2024). Für die Primarstufe liegen durchaus Studien vor. Hwang (2022) hat 21 Studien mit 30 Stichproben aus den Jahren 2000 bis 2022 zur Grundschulmathematik meta-analytisch zusammengefasst und findet einen kleinen Effekt (g = 0,35), allerdings vor der Verbreitung generativer Sprachmodelle; die Studien betreffen überwiegend regelbasierte adaptive Systeme. Studien, die dialogische, sprachmodellbasierte Lernbegleitung mit Grundschulkindern über längere Zeit und mit Lernzuwachs als Ergebnismaß untersuchen, sind bisher selten (Kuo et al., 2026; Liu et al., 2025; Listyaningrum et al., 2024). Viele der folgenden Ableitungen stützen sich daher weniger auf direkte Studienergebnisse für diese Zielgruppe als auf theoretische Modelle und auf Befunde aus älteren Altersgruppen. Wo ein Befund aus Sekundarstufe, Hochschule oder Erwachsenenstichproben stammt, wird das im Text jeweils genannt. Dieser

Beitrag sichtet und ordnet die vorliegenden Erkenntnisse und macht sie für den Mathematikunterricht der Grundschule nutzbar.

Digitale Medien haben für sich genommen keinen Lerneffekt, und für KI gilt dasselbe. Ein „Lerneffekt mit KI“ ist keine Eigenschaft der Technologie. Zu fragen ist nach der jeweiligen Lernumgebung und ihrer didaktischen Vorstrukturierung (Dinsmore & Fryer, 2026; Weidlich et al., 2025). Der Beitrag fragt deshalb, unter welchen Bedingungen KI-gestützte Lernbegleitung mathematische Lernprozesse in der Primarstufe unterstützt und wann auf ihren Einsatz verzichtet werden sollte.

Im Folgenden bezeichnet „KI-Lernbegleitung“ das Gesamtarrangement aus System, Aufgabe und Unterricht und „KI-Lernbegleiter“ das konkrete System im Einsatz mit Kindern. „Intelligente Tutorsysteme“ meint die ältere, regelbasierte Systemfamilie, auf die sich ein Teil der zitierten Forschung bezieht. Wo eine Quelle von „KI-Tutor“ oder „Chatbot“ spricht, wird ihre Bezeichnung übernommen.

Diese Überlegungen bilden die Grundlage für die Entwicklung und Erprobung appintegrierter KI-Lernbegleitung im Projekt PRIMA-KI. Sie sind ausdrücklich nicht endgültig, sondern werden mit neuen KI-Modellen, neuen empirischen Befunden und den darauf aufbauenden pädagogischen Anwendungen weiterentwickelt.

Theoretischer Rahmen

Vier theoretische Perspektiven strukturieren die folgenden Gestaltungsprinzipien: die Cognitive Load Theory, die Zone der nächsten Entwicklung mit dem Konzept des Scaffoldings, die Unterscheidung von prozeduralem und konzeptuellem Wissen sowie Modelle des selbstregulierten Lernens. Hinzu kommt eine Rahmenperspektive: Unterstützung wirkt nicht im einzelnen Medium, sondern im gesamten Lernarrangement.

Cognitive Load Theory und Expertise-Umkehr

Die Cognitive Load Theory (Sweller, 2020; Sweller et al., 2019) beschreibt die begrenzte Kapazität des Arbeitsgedächtnisses als Rahmenbedingung für die Gestaltung von Lernmedien. Für KI-gestützte Systeme hat diese Begrenzung unmittelbare Konsequenzen. Generative KI kann Informationen in hoher Geschwindigkeit und großem Umfang präsentieren: Texte, Bilder, Erklärungen, Visualisierungen, Animationen. Wenn Lernende mit Informationen und Reizen überflutet werden, ohne diese lernrelevant verarbeiten zu können, verfehlt auch die

inhaltlich beste Information ihr Ziel.

Diese Dosierung zu gestalten, ist eine der schwierigsten Aufgaben beim Entwurf von KI-Lernbegleitung. Gestufte Hilfe (Scaffolding), also die dosierte Bereitstellung von Hilfen und Erklärungen, ist aus Sicht der Cognitive Load Theory dafür das wichtigste Mittel. KI-Lernbegleiter müssen Impulse, Hilfen und Unterstützung so dosieren, dass die Belastung des Arbeitsgedächtnisses in einem Bereich bleibt, in dem Lernen möglich ist (Cosentino et al., 2025).

Wie viel Unterstützung angemessen ist, hängt vom Vorwissen ab. Der Expertise-Umkehr-Effekt beschreibt, dass Hilfen, die Anfängerinnen und Anfängern nützen, bei fortgeschrittenen Lernenden wirkungslos und schließlich hinderlich werden, weil das Abgleichen redundanter Erklärungen mit dem bereits vorhandenen Wissen selbst Arbeitsgedächtniskapazität bindet (Kalyuga et al., 2003; Kalyuga, 2007). Aus der Cognitive Load Theory folgt deshalb zweierlei. Unterstützung muss dosiert werden, und sie muss planmäßig zurückgenommen werden, sobald ein Kind sicherer wird. Für Tutorsysteme haben Koedinger und Aleven (2007) dieses Spannungsfeld als „assistance dilemma“ beschrieben. Zu frühe Hilfe verhindert die eigene Konstruktion, ausbleibende Hilfe lässt Lernende in unproduktiven Sackgassen stecken. Wo das Optimum liegt, verschiebt sich mit der Kompetenz.

Aus derselben Theorie stammt der Lösungsbeispiel-Effekt: Für Lernende ohne Vorerfahrung mit einem Aufgabentyp ist das Studieren durchgearbeiteter Lösungsbeispiele dem eigenständigen Problemlösen überlegen (Renkl, 2014; Renkl & Atkinson, 2003). Darauf ist beim Prinzip „Erst du – dann ich“ zurückzukommen.

Zone der nächsten Entwicklung und Scaffolding

Die Zone der nächsten Entwicklung (Vygotsky, 1978) beschreibt den Bereich zwischen dem, was ein Kind allein bewältigen kann, und dem, was es mit Hilfe einer kompetenteren Person bewältigen kann. Für den Entwurf von KI-Lernbegleitern folgt daraus eine harte Anforderung: Das System müsste fortlaufend einschätzen, was ein Kind selbstständig und was es mit Unterstützung bewältigen kann, und dann Impulse in genau diesem Bereich anbieten, weder zu leicht noch zu schwer.

Die Zone der nächsten Entwicklung ist dabei kein Schwierigkeitsmaß, das sich an einem Kind ablesen ließe, sondern eine Beschreibung dessen, was in der Interaktion möglich wird. Ein System kann sie nicht messen, sondern nur laufend abschätzen und diese Abschätzung an den Reaktionen des Kindes korrigieren. Die Übertragung

des Konzepts auf die Kalibrierung von Aufgabenschwierigkeit ist eine Deutung, keine Aussage Vygotskys.

Das zugehörige Handlungskonzept ist das Scaffolding. Wood, Bruner und Ross (1976) beschreiben damit die Unterstützung eines Kindes durch eine erwachsene Person beim Problemlösen und nennen als Funktionen unter anderem, das Interesse zu wecken, die Aufgabe zu vereinfachen, die Richtung zu halten, entscheidende Merkmale hervorzuheben und Frustration zu kontrollieren. Als Merkmal guten Scaffoldings gilt die Kontingenz: Die Unterstützung folgt einer Diagnose des aktuellen Stands, richtet sich nach ihr und wird zurückgenommen, sobald das Kind die Verantwortung selbst übernehmen kann (van de Pol et al., 2010). Adaptive Systeme müssen diese Einschätzung nicht einmalig, sondern kontinuierlich leisten (Wu et al., 2026; Kuo et al., 2026). Sie müssen aus den Eingaben der Lernenden ableiten, wo deren Verständnis liegt, welche Vorstellung zu einem Fehler geführt hat und welcher nächste Impuls produktiv wäre. Ob Adaptivität lernwirksam ist, hängt von der diagnostischen Grundlage und der Qualität der daraus abgeleiteten Hilfen ab. Zu bedenken bleibt, dass Scaffolding ursprünglich eine Interaktion zwischen zwei Personen beschreibt. Die Übertragung auf Software ist eine Analogie, deren Reichweite umstritten ist.

Prozedurales und konzeptuelles Wissen

Welche Unterstützung angemessen ist, hängt davon ab, welches Lernziel gerade verfolgt wird. Prozedurales Wissen (Wie löse ich diese Aufgabe?) braucht andere Unterstützung als konzeptuelles Wissen (Warum funktioniert das so?).

Ein Kind, das Rechenflüssigkeit übt, etwa die schnelle Addition oder Subtraktion, profitiert von unmittelbarer, knapper Rückmeldung zu seinen Ergebnissen, vorausgesetzt, es löst die Aufgaben bereits ableitend und nicht zählend. Sie stützt die Automatisierung und verhindert, dass sich fehlerhafte Prozeduren einschleifen. Ausführliche Erklärungen mitten in einer Übungsserie unterbrechen die Automatisierung und belasten das Arbeitsgedächtnis mit Information, die das Kind in diesem Moment nicht verarbeiten kann. Ein Kind, das gerade Zahlverständnis entwickelt, muss hingegen durch gezielte Fragen und Handlungsmöglichkeiten dazu angeregt werden, sein Verständnis selbst zu konstruieren und weiterzuentwickeln. Es sollte eigene Lösungswege entwickeln, darstellen und begründen. Ein KI-System, das diese Unterscheidung nicht treffen kann, riskiert, dass seine Unterstützung kontraproduktiv wirkt, etwa indem es bei konzeptuellem Lernen zu schnell Lösungen liefert oder bei prozeduralem Üben mit unnötigen Impulsen das

Arbeitsgedächtnis belastet. Dass ein didaktisch vorstrukturiertes Modell konzeptuelles Verständnis stärker fördert als ein generisches Sprachmodell, zeigen Makransky et al. (2025); darauf kommt der Abschnitt zur Fundierung zurück.

Selbstreguliertes Lernen und Metakognition

Modelle des selbstregulierten Lernens beschreiben Lernen als Kreislauf aus Planen, Überwachen und Bewerten (Zimmerman, 2002). Lernende setzen sich ein Ziel, wählen eine Vorgehensweise, prüfen unterwegs, ob sie vorankommen, und beurteilen am Ende das Ergebnis. Für KI-Lernbegleitung ist dieses Modell in zwei Richtungen bedeutsam. Ein System kann diese Prozesse anregen, indem es nach dem Plan fragt, zum Überprüfen auffordert und die Bewertung des Ergebnisses dem Kind überlässt. Ein System kann diese Prozesse aber auch übernehmen und den Lernenden damit abgewöhnen. Molenaar (2022) beschreibt für adaptive Lernumgebungen, dass sie die Regulation des Lernens weitgehend selbst ausführen und die Entwicklung der Selbstregulation dadurch gefährden können, weshalb die Kontrolle schrittweise an die Lernenden zurückgegeben werden muss.

Für die Primarstufe kommt ein entwicklungspsychologischer Vorbehalt hinzu. Metakognitive Überwachung entwickelt sich im Grundschulalter erst. Jüngere Kinder überschätzen ihr Verständnis regelmäßig, und ausformulierte Begründungen fallen ihnen sprachlich schwer. Metakognitive Anregung kann in der Grundschule deshalb nicht bedeuten, Reflexion abzufragen; sie muss eingeübt werden, mit kurzen, konkreten Fragen zu einer gerade ausgeführten Handlung.

Was ein System dafür braucht

Diese Perspektiven bilden einen Rahmen, in dem die folgenden Gestaltungsprinzipien verortet sind. Sie machen zugleich deutlich, warum KI-Lernbegleitung kein einfaches technisches Problem ist. Sie erfordert Systeme, die über ein Modell des Lernenden, des Lerngegenstands und des Lernprozesses verfügen. In der Forschung zu intelligenten Tutorsystemen ist diese Anforderung seit Langem als Dreiteilung beschrieben: ein Domänenmodell des Lerngegenstands, ein Lernendenmodell des aktuellen Wissensstands und ein tutorielles Modell der Interventionsentscheidungen. Für das Lernendenmodell existieren erprobte Verfahren, etwa Knowledge Tracing (Corbett & Anderson, 1995), die den Wissensstand aus der Aufgabenbearbeitung fortlaufend schätzen. Große Sprachmodelle bringen ein solches Modell nicht mit. Sie verfügen über Sprachkompetenz, nicht über eine Repräsentation dessen, was ein bestimmtes Kind

über die Zeit gelernt hat. Diese Repräsentation muss die Anwendung außerhalb des Sprachmodells führen.

Schließlich entsteht Unterstützung nicht im einzelnen Medium. Puntambekar und Hübscher (2005) halten die Übertragung des Scaffolding-Begriffs auf Software nur dann für tragbar, wenn Diagnose, Kontingenz und Rücknahme erhalten bleiben, und beschreiben Scaffolding gemeinsam mit Tabak (2004) als verteilt: Software, Lehrkraft, Material und Mitschülerinnen und Mitschüler adressieren dieselbe Lernbedürftigkeit und können sich wechselseitig verstärken oder stören. Das ist die theoretische Grundlage für das vierte Prinzip, die Einbettung.

Das Risiko der Auslagerung von Denkarbeit

Die bisherige Darstellung der Studienlage könnte den Eindruck erwecken, KI-gestützte Lernbegleitung besitze ein überwiegend positives Potenzial, das nur noch technisch optimiert werden müsse. Das Bild ist ambivalenter. In mehreren Studien zeigte sich, dass der unkontrollierte Einsatz generativer KI beim Lernen nicht nur weniger wirksam sein kann als erhofft, sondern auch den Aufbau grundlegender Fähigkeiten abkürzen und Lernen dadurch behindern kann.

Cognitive Offloading: Auslagerung kognitiver Anstrengung

Kognitive Auslagerung (Cognitive Offloading) beschreibt das strukturell wichtigste Problem: Lernende lagern kognitive Anstrengung an die KI aus, statt sie selbst zu vollziehen. Gerlich (2025) berichtet in einer Querschnittsbefragung von 666 Erwachsenen negative Zusammenhänge zwischen der Häufigkeit der KI-Nutzung und den Werten in einem Test kritischen Denkens ($r = -0{,}68$) sowie zwischen dem Ausmaß kognitiver Auslagerung und kritischem Denken ($r = -0{,}75$). Jüngere Befragte nutzten KI stärker und erzielten niedrigere Werte. Die Studie zeigt Zusammenhänge, keine Wirkung. Sie beruht teilweise auf Selbstauskünften, die Zusammenhänge dieser Höhe überschätzen, und sie betrifft Erwachsene, nicht Schulkinder. Sie begründet eine Auslagerungshypothese, die experimentell und längsschnittlich zu prüfen ist.

Dieses Risiko zeigten Bastani et al. (2025) in einer kontrollierten Interventionsstudie mit rund 1000 Schülerinnen und Schülern der Klassen 9 bis 11 an türkischen Schulen. Wer während des Übens unregulierten Zugang zu einem Chatbot mit vollständigen Lösungen hatte, löste in der Übungsphase deutlich mehr Aufgaben

richtig. Im anschließenden Test ohne KI lag diese Gruppe 17 Prozent unter dem Niveau der Kontrollgruppe ohne KI. Die Autorinnen und Autoren sprechen von verringertem Kompetenzerwerb. Eine dritte Gruppe arbeitete mit einer Variante, die Hinweise statt Lösungen gab. Sie war in der Übungsphase noch erfolgreicher als die Gruppe mit freiem Zugang und erreichte im Test das Niveau der Kontrollgruppe, übertraf es aber nicht. Der Schaden entsteht also nicht durch KI an sich, sondern durch unbegrenzten Lösungszugang, und auch die gut gestaltete Variante erzeugte in dieser Studie keinen Lernvorteil. Der Befund betrifft diese Untersuchungsanordnung und belegt keinen dauerhaften Kompetenzverlust. Lernanalysen aus KI-basierten Nachhilfesystemen zeigen ergänzend, dass ein Teil der Lernenden versucht, Aufgabensequenzen „durchzuklicken“, ohne sich inhaltlich mit ihnen auseinanderzusetzen (Jančařík et al., 2023).

Dinsmore und Fryer (2026) argumentieren aus lernpsychologischer Perspektive, dass generative KI in Form nicht vorstrukturierter Chatbots dazu verführt, den anstrengenden Prozess der eigenen Wissenskonstruktion abzukürzen. Wer eine fertige Lösung übernimmt, ohne sie zu durchdenken, spart Anstrengung, baut aber keine belastbaren Wissensstrukturen auf. Das ist nicht dasselbe wie das gezielte Studieren eines Lösungsbeispiels. Dort wird die Lösung mit Aufforderungen zum Erklären verbunden, und die Denkarbeit verlagert sich vom Suchen zum Nachvollziehen. Viele Anbieter von Chatbots bieten inzwischen einen „Lernmodus“ an, in dem der Chatbot keine Lösungen direkt ausgibt, sondern Rückfragen stellt. Ob diese Modi lernwirksam sind, ist bislang nicht untersucht. Technisch sind sie eine Anweisungsebene über demselben Modell und lassen sich im Gespräch meist umgehen.

Höflichkeit und pädagogische Qualität von Sprachmodellen

Zu diesem strukturellen Risiko tritt ein zweites. Abdulsalam und Aroyehun (2025) analysieren in einer noch nicht begutachteten Vorabveröffentlichung (Preprint) Mathematik-Nachhilfedialoge, in denen erfahrene Tutorinnen und Tutoren, Novizen und sieben Sprachmodelle auf dieselben Schülerfehler antworten. Größere Modelle erreichen in der Qualitätsbewertung annähernd Expertenniveau, unterscheiden sich aber im Profil: Sie antworten länger, vielfältiger und höflicher, nutzen aber typische Expertenstrategien wie das Nachfassen zu Begründung und Genauigkeit seltener. Höfliche Sprache hing negativ mit der eingeschätzten pädagogischen Qualität zusammen. Lernergebnisse wurden nicht erhoben.

Diese Tendenz wird als Sycophancy bezeichnet, als übermäßige Zustimmung. Sie ist

keine zufällige Eigenart, sondern eine Folge des Trainingsverfahrens: Sprachmodelle werden im Nachtraining darauf optimiert, Antworten zu geben, die Menschen gut bewerten, und Menschen bewerten Zustimmung systematisch besser als Widerspruch (Sharma et al., 2024). Für die Gestaltung bedeutet das, dass sich diese Tendenz durch Anweisungen im Systemprompt, also in der festen Grundanweisung an das Modell, abschwächen, aber nicht zuverlässig abstellen lässt.

Chudziak und Kostka (2025) beschreiben ein verwandtes Problem: Viele aktuelle KI-Systeme neigen zu einem anweisenden, vorschreibenden (preskriptiven) Interaktionsstil. Sie geben vor, leiten an und lösen, statt Raum für eigene Denkprozesse zu lassen. Systeme, die bei jeder Eingabe sofort Rückmeldung geben, riskieren, genau jene Denk- und Kontrollprozesse zu unterbinden, die sie fördern sollen. Eine computergestützte Analyse realer menschlicher Nachhilfedialoge (Wang et al., 2025) zeigt, dass Grundschulkinder auf aktivierende Fragen positiv reagieren, sich insgesamt aber weniger aktiv beteiligen als ältere Lernende. Die Studie erfasst Interaktionsmuster, keine Lernergebnisse, und sie beschreibt menschliches Tutoring. Auf KI-Systeme lässt sich das Muster nur als Hypothese übertragen.

Daraus folgt nicht, dass ein KI-Lernbegleiter unfreundlich sein sollte. Gerade bei jüngeren Kindern ist ein zugewandter Ton die Voraussetzung dafür, dass sie einen Fehler überhaupt zeigen. Problematisch ist nicht die Freundlichkeit, sondern die Freundlichkeit anstelle von Information: Zustimmung, die eine inhaltliche Rückmeldung ersetzt.

Die Spannung zwischen wirksamer Hilfe und eigenständigem Denken

Diese Befunde verdichten sich zu einer Spannung, die als Leitproblem für die Gestaltung von KI-Lernbegleitung verstanden werden kann: Je bereitwilliger ein System fertige Lösungen liefert, desto zuverlässiger nimmt es Lernenden genau die Denkarbeit ab, an der sie lernen würden. Nicht die Leistungsfähigkeit des Systems ist das Problem, sondern die Frage, wofür sie eingesetzt wird. Bei Bastani et al. (2025) war die Variante mit Hinweisen in der Übungsphase leistungsfähiger als der freie Zugang und schadete im Test dennoch nicht.

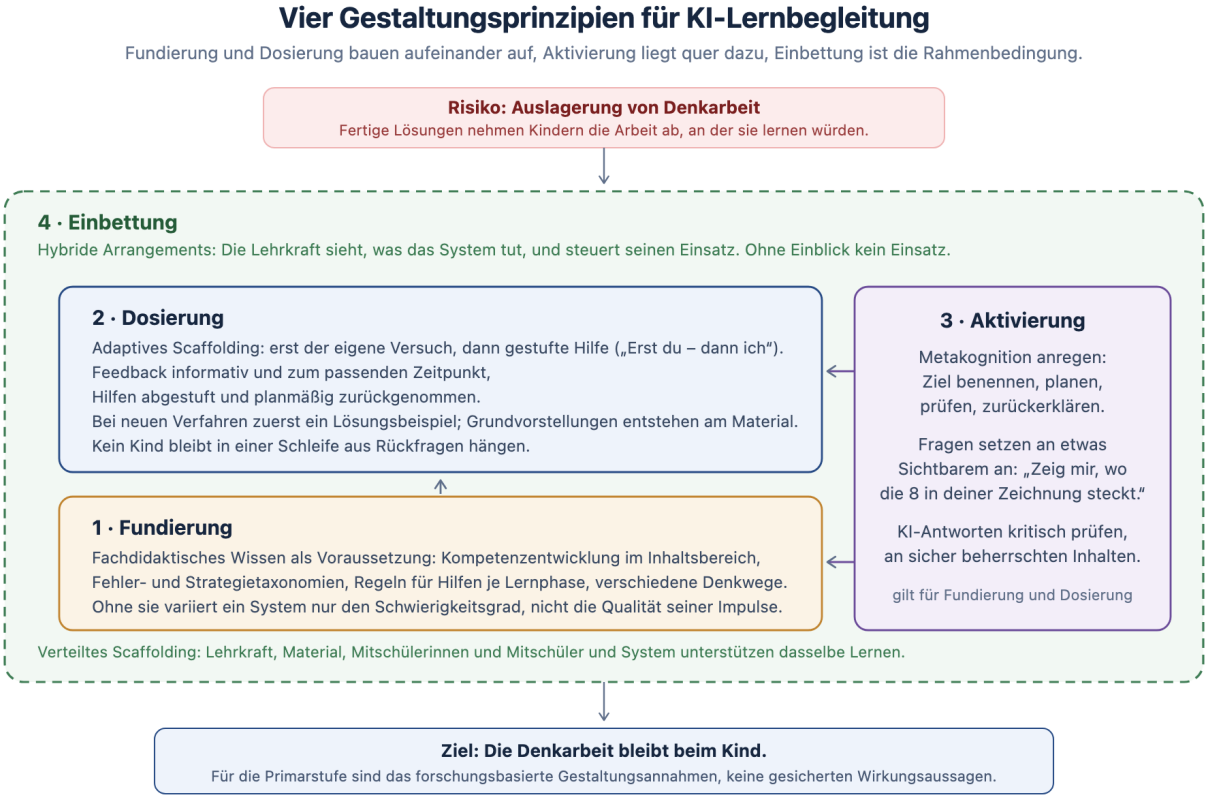
Ein guter KI-Lernbegleiter muss deshalb manchmal bewusst weniger tun, als er könnte. Er muss zulassen, dass ein Kind zeitweise nicht weiterkommt. Mason et al. (1982) beschreiben dieses Steckenbleiben („being stuck“) als normalen Bestandteil

mathematischen Arbeitens und verbinden es mit Strategien, es zu benennen und zu bearbeiten: festhalten, was man weiß und was man sucht, einen einfacheren Fall betrachten, eine Vermutung aufschreiben. Für ein KI-System heißt das nicht, Frustration entstehen zu lassen, sondern das Nichtweiterkommen zu benennen und mit einer Handlungsmöglichkeit zu beantworten. Es hält Lösungen zurück, gibt unvollständige Hinweise und lässt Bearbeitungszeit. Damit arbeitet es teilweise gegen die Mechanismen, die generative KI so eindrucksvoll machen.

Ein nicht vorstrukturierter Chatbot, etwa ein frei zugängliches Sprachmodell ohne pädagogische Rahmung, wird zur Antwortmaschine und kann verständnisförderndes Lernen eher verhindern als ermöglichen. Die folgenden Prinzipien zielen darauf, Unterstützung zu dosieren und die Denkarbeit beim Kind zu halten.

Vier Gestaltungsprinzipien

Aus der Spannung zwischen dem Potenzial adaptiver KI-Lernbegleitung und den beschriebenen Risiken der kognitiven Auslagerung ergeben sich vier Gestaltungsebenen, die zusammenwirken müssen. Die fachdidaktische Fundierung des Systems ist die Voraussetzung für Diagnostik und Förderung. Darauf baut die adaptive Dosierung von Hilfen auf. Die metakognitive Aktivierung liegt quer zu beiden und beschreibt, welche Denkprozesse die Interaktion anregen soll. Die hybride Einbettung mit menschlicher Aufsicht ist die Rahmenbedingung: Ohne sie werden die ersten drei Ebenen im Unterricht nicht wirksam.



Fundierung: Fachdidaktisches Wissen als Voraussetzung

Die erste Gestaltungsebene betrifft die Wissensbasis des Systems. Generative KI erzeugt sprachlich und formal beeindruckende Ausgaben, aber fachdidaktisch sind diese häufig oberflächlich, wenig problemorientiert oder fehlerhaft, wenn sie nicht gezielt gesteuert werden. Literaturanalysen zum Einsatz generativer KI in der Mathematikbildung weisen immer wieder auf die geringe theoretische Fundierung und die Fehleranfälligkeit vieler Systeme hin (Almheiri et al., 2025; Awang et al., 2025; Holmes & Tuomi, 2022; Opesemowo & Adewuyi, 2024). Das ist keine Implementierungsschwäche, die sich beiläufig beheben ließe, sondern eine Folge dessen, worauf diese Modelle optimiert wurden. Die Trainingsdaten werden zwar gefiltert und für Mathematik gezielt zusammengestellt, das Optimierungsziel ist dabei aber die richtige Lösung und im Nachtraining die von Menschen als hilfreich bewertete Antwort. Fachdidaktische Angemessenheit, also die Frage, welcher Impuls einem Kind an dieser Stelle weiterhilft, ist kein Trainingsziel.

Cárdenas et al. (2025) identifizieren in ihrer systematischen Analyse die fehlende theoretische Rahmung als eines der Haupthindernisse für lernwirksame KI-

Tutorsysteme. Fachdidaktische Fundierung betrifft sowohl die inhaltliche Korrektheit von Aufgaben und Erklärungen, in der die Systeme besser werden, als auch die Art, wie ein System auf Lernende reagiert. Fachexpertise und Tutoring-Expertise sind nicht dasselbe. Macina et al. (2025) zeigen mit einem standardisierten Prüfverfahren (Benchmark) für pädagogische Fähigkeiten von Sprachmodellen, dass Modelle, die Aufgaben korrekt lösen, nicht automatisch erkennen, wo die Schwierigkeit für Lernende liegt, welche Fehlvorstellung einem Fehler zugrunde liegt oder welcher Impuls in einer Lernsituation produktiv wäre. Fachliches Wissen in der Anweisung an das Modell (Prompt) reicht daher nicht aus. Das System muss über fachdidaktisches Wissen verfügen, also darüber, wie Kinder mathematische Konzepte aufbauen, welche typischen Irrwege es gibt und welche Interventionen an welcher Stelle im Lernprozess wirksam sind.

Dass dieser Unterschied praktisch bedeutsam ist, zeigen Makransky et al. (2025) in einem präregistrierten Experiment (Auswertung vorab festgelegt) mit 175 Studierenden und einer Replikation mit 234 Schülerinnen und Schülern der Sekundarstufe II. Sie verglichen einen theoriegeleitet gestalteten KI-Tutor (ChatTutor) mit einem generischen Sprachmodell. Der didaktisch vorstrukturierte Tutor erzielte in beiden Studien unmittelbar höheres konzeptuelles Wissen sowie mehr Vertrauen und Freude; im verzögerten Test blieb der Vorteil nur in der zweiten Studie bestehen, Unterschiede in der Selbstwirksamkeit zeigten sich nicht. Auch große Sprachmodelle wenden fachdidaktisches Wissen also nicht von selbst an. Mehrere aktuelle Rahmenwerke für pädagogische Agenten nehmen deshalb explizit auf pädagogische Theorien Bezug; Cohn et al. (2026) leiten die Rückmeldungen ihres Agenten aus Zone der nächsten Entwicklung, Selbstwirksamkeit und Zielsetzung ab und erproben ihn mit 104 Lernenden der Mittelstufe. Ob diese Bezugnahme zu besserem Lernen führt, ist für die meisten Systeme noch nicht untersucht. Eine Studie zu GeoGebra- und KI-gestützten Lernumgebungen deutet darauf hin, dass konzeptuelles Verständnis und Selbstvertrauen nur dann steigen, wenn Fach- und Fachdidaktik in die Systemgestaltung einfließen und Technik nicht bloß aufgesetzt wird (Canonigo, 2024).

Ein KI-Lernbegleiter für den Mathematikunterricht der Primarstufe benötigt mindestens die folgenden fachdidaktischen Wissensgrundlagen. Wie sie technisch bereitgestellt werden können, behandelt der Abschnitt zur Umsetzung.

Erstens ein Modell der mathematischen Kompetenzentwicklung für den jeweiligen Inhaltsbereich. Das System muss „wissen“, dass Zahlverständnis nicht durch Auswendiglernen von Fakten entsteht, sondern durch den Aufbau von

Vorstellungen, und dass es für diesen Aufbau typische Entwicklungsverläufe und Vorläuferfähigkeiten gibt. Ohne dieses Wissen bleibt jede Adaptation oberflächlich. Das System kann dann allenfalls den Schwierigkeitsgrad variieren, aber nicht die Qualität seiner Impulse an den Verständnisstand anpassen.

Zweitens Fehler- und Strategietaxonomien, die typische Denkwege und Irrtümer von Kindern abbilden (Padberg & Benz, 2021; Gaidoschik, 2010). Die Diagnose von Fehlvorstellungen gehört zum fachdidaktischen Handwerk und muss einem System eigens beigebracht werden, wenn es über die Feststellung „richtig/falsch“ hinausgehen soll. Beispiele sind ein fehlendes Bündelungsverständnis, bei dem ein Kind die 2 in der 24 als „zwei“ und nicht als „zwei Zehner“ deutet, oder das zählende Rechnen als verfestigte Strategie. Die Zuordnung einer konkreten Kinderäußerung zu einer Fehlerkategorie kann zudem misslingen, und eine falsche Diagnose erzeugt keinen unspezifischen, sondern einen präzise falsch zielenden Impuls. Ein System sollte deshalb seine Sicherheit mitführen und unterhalb einer Schwelle nachfragen, statt zu deuten. In eigenen Erprobungen im Projekt PRIMA-KI schlugen auch große Sprachmodelle ohne fachdidaktischen Kontext Kindern bei Fehlern wiederholt vor, die Aufgabe zählend zu lösen, also genau die Strategie, die im Unterricht abgebaut werden soll.

Drittens explizite Regeln, welche Hilfen in welcher Lernphase und bei welcher Art von Schwierigkeit sinnvoll sind. Die Unterscheidung zwischen prozeduralem und konzeptuellem Lernen ist eine solche Leitlinie: Routineförderung erfordert andere Interventionen als der Aufbau begrifflichen Verständnisses. Ein System sollte zudem zwischen einem einmaligen Fehlgriff und einem zugrunde liegenden Verständnisproblem unterscheiden. Aus einer einzelnen Aufgabe ist das nicht zu entscheiden, sondern erst aus dem Muster über mehrere Aufgaben hinweg, etwa aus der Wiederholung derselben Fehlerart oder aus der Bearbeitungsdauer. Das setzt voraus, dass die Anwendung diese Informationen führt.

Viertens muss das System nicht nur den Schwierigkeitsgrad anpassen, sondern auch unterschiedliche Denkwege erkennen, mit denen Kinder zum Ergebnis gelangen. Kinder lösen mathematische Probleme auf vielfältige Weise, und das ist wünschenswert. Ein fachdidaktisch fundierter KI-Lernbegleiter erkennt alternative Strategien, prüft, ob sie auch bei größeren Zahlen und anderen Aufgaben funktionieren, fragt gegebenenfalls nach und kann Kinder darin bestärken oder behutsam zu effizienteren Wegen anregen. Für Kinder mit Rechenschwierigkeiten ist dabei entscheidend, woran ein System seine Anpassung festmacht. Die naheliegende Anpassung, kleinere Zahlen und leichtere Aufgaben, kann

kontraproduktiv sein: Ein Kind, das zählend rechnet, löst kleinere Aufgaben zählend weiterhin erfolgreich und wird in dieser Strategie bestätigt. Notwendig ist eine Anpassung der Impulse, nicht des Niveaus (Gaidoschik, 2010; Wartha & Schulz, 2012). Aus dem Ergebnis allein ist die Strategie allerdings nicht erkennbar; 8 + 6 = 14 sieht in beiden Fällen gleich aus. Hinweise liefern die Bearbeitungsdauer, typische Fehlermuster wie Abweichungen um eins und die direkte Nachfrage, wie das Kind gerechnet hat.

Fachdidaktische Fundierung ist die Bedingung dafür, dass adaptive Unterstützung überhaupt als solche funktionieren kann. Ohne sie bleibt ein KI-Lernbegleiter, um eine pointierte Formulierung aus einem unveröffentlichten Manuskript von Schneider (o. J.) aufzugreifen, ein „sprachbegabter Zufallsgenerator“ (o. S.), der gelegentlich hilfreiche, häufig aber unspezifische oder irreführende Impulse produziert.

Dosierung: Adaptives Scaffolding und das Prinzip „Erst du – dann ich“

Auf der Grundlage fachdidaktischen Wissens wird die zweite Gestaltungsebene möglich: die adaptive Dosierung von Impulsen und Hilfen. Sie unterscheidet KI-Lernbegleitung von einer Antwortmaschine. Die Befundlage deutet darauf hin, dass sorgfältig gestaltete KI-Lernumgebungen, die auf Lernstand, Fehler und Strategien reagieren, statt richtige Antworten auszugeben, Lern- und Verstehensprozesse anregen. Ob der adaptive Anteil dafür verantwortlich ist, lässt sich aus den vorliegenden Studien nicht isolieren, weil Adaptivität dort mit sorgfältiger Gestaltung insgesamt einhergeht; den einzigen sauberen Vergleich adaptiver und fester Hilfen liefern Salden et al. (2010) für Lösungsbeispiele.

Im Folgenden wird zwischen Feedback und Hilfe unterschieden. Feedback ist die Rückmeldung auf ein Ergebnis oder einen Lösungsweg; Hilfe ist die Unterstützung während der Bearbeitung, also Hinweise, Fragen und Teilschritte. Beide werden unterschiedlich dosiert: Feedback soll informativ sein und zum passenden Zeitpunkt kommen, Hilfe abgestuft und zurückhaltend.

In der Studie zum System TALPer profitierten leistungsschwächere Fünftklässlerinnen und Fünftklässler in Taiwan besonders stark von adaptiver Unterstützung, während leistungsstärkere Kinder komplexere Interaktionsmuster mit dem KI-Lernbegleiter entwickelten (Kuo et al., 2026). Ein System konnte damit unterschiedliche Lernvoraussetzungen abdecken. Liu et al. (2025) berichten für eine

generative Lernumgebung zu Textaufgaben in der Grundschule Leistungsgewinne, wobei die wahrgenommene Qualität der Unterstützung und nicht deren bloße Verfügbarkeit den stärksten Einfluss auf Motivation und Lernwirkung hatte. Die systematische Übersicht von Son (2024) bestätigt positive Effekte gut konzipierter intelligenter Tutorsysteme auf mathematische Lernleistungen, besonders wenn diese adaptiv auf individuelle Lernbedürfnisse reagieren. Generative Systeme können Rückmeldungen an Aufgabe, Eingabe und bisherigen Lernverlauf des einzelnen Kindes anpassen.

Arten von Feedback. Die Feedbackforschung unterscheidet Rückmeldungen zunächst nach ihrem Informationsgehalt (Shute, 2008). Die einfachste Rückmeldung teilt nur mit, ob eine Antwort richtig oder falsch war. Die zweite nennt zusätzlich die richtige Lösung. Die dritte, elaborierte Form gibt Auskunft darüber, worin der Fehler besteht, welche Vorstellung ihm zugrunde liegen könnte und welcher nächste Schritt sinnvoll ist. Die Wirkungen unterscheiden sich deutlich. Wisniewski et al. (2020) fassen in einer Meta-Analyse 435 Studien mit 994 Effektstärken zusammen. Der mittlere Effekt von Feedback liegt bei d = 0,48 und verteilt sich sehr ungleich: Informationsreiche Rückmeldungen erreichen d = 0,99, korrektive Rückmeldungen d = 0,46, Verstärkung und Sanktionierung, also Lob, Belohnung und Tadel, nur d = 0,24. Der niedrige Wert betrifft also Lob und Tadel, nicht die sachliche Richtig-falsch-Rückmeldung, die unter die korrektiven Formen fällt. Kluger und DeNisi (1996) zeigten in einer Meta-Analyse überwiegend mit Erwachsenen, dass über ein Drittel aller Feedbackinterventionen die Leistung verschlechtert, vor allem dann, wenn die Rückmeldung die Aufmerksamkeit von der Aufgabe weg auf die eigene Person lenkt.

Quer dazu liegt die Frage, worauf sich eine Rückmeldung richtet. Hattie und Timperley (2007) unterscheiden vier Bezugspunkte: die Aufgabe („Das Ergebnis stimmt nicht“), den Bearbeitungsprozess („Du hast den Übertrag vergessen, prüfe die Zehner“), die Selbstregulation („Woran könntest du selbst merken, ob das stimmen kann?“) und die Person („Du bist gut in Mathe“). Wirksam sind vor allem die beiden mittleren; Rückmeldung auf die Person ist die am wenigsten wirksame Form. Für einen KI-Lernbegleiter ist die einfache Richtig-falsch-Rückmeldung beim Automatisieren bereits verstandener Rechenfertigkeiten angemessen; dort soll sie knapp und unmittelbar erfolgen. Beim Aufbau von Verständnis reicht sie nicht aus. Hier braucht es Feedback, das den Denkweg benennt oder eine Selbstprüfung anstößt, nicht das Ergebnis bewertet und nicht das Kind. Für die Grundschule entsteht daran ein Zielkonflikt: Elaboriertes Feedback ist informationsreich, und

Informationsreichtum belastet das Arbeitsgedächtnis eines Kindes, das noch mühsam liest. Elaboriert heißt hier deshalb nicht ausführlich, sondern präzise: ein Satz, der auf den konkreten Schritt zeigt, verbunden mit einer Handlungsaufforderung („Schau noch einmal auf die Zehner. Was passiert beim Übergang?“), und nicht die vollständige Erklärung des Verfahrens.

Neben der Art ist der Zeitpunkt bedeutsam, und zwar in zwei Hinsichten. Die erste betrifft die Frage, ob das System nach jedem Teilschritt eingreift oder erst, wenn ein Versuch abgeschlossen ist. Beim Einüben bereits verstandener Fertigkeiten, etwa Kopfrechnen, Einmaleins oder Zahlzerlegungen, spricht die Befundlage für schrittweise, unmittelbare Rückmeldung: Sie verhindert, dass sich Fehler einschleifen, und hält die Belastung des Arbeitsgedächtnisses gering. Beim Problemlösen und beim Aufbau von Verständnis sollte die Rückmeldung dagegen erst nach einem abgeschlossenen eigenen Versuch kommen, weil sie dem Kind sonst die Gelegenheit nimmt, selbst zu prüfen und zu revidieren. Das ist eine Gestaltungsregel, kein gesicherter Befund; Shute (2008) berichtet für schwierige Aufgaben und leistungsschwächere Lernende auch Vorteile unmittelbarer Rückmeldung, weshalb der „abgeschlossene Versuch“ bei Grundschulkindern eng zu fassen ist. Die zweite Hinsicht betrifft die eigentliche Verzögerung, also Rückmeldung erst nach Minuten oder Tagen. Hier ist die Befundlage uneinheitlich (Shute, 2008); für die Grundschule ist sie von geringerer praktischer Bedeutung, weil ein Kind die eigene Bearbeitung dann nicht mehr präsent hat.

Erst du – dann ich. Bedeutsam ist die Reihenfolge von Denken und Unterstützung. Kumar et al. (2025) prüften in zwei präregistrierten Online-Experimenten mit insgesamt 1818 Erwachsenen an Mathematikaufgaben im Multiple-Choice-Format, wie Erklärungen eines Sprachmodells auf spätere Testleistungen wirken. Erklärungen wirkten besser als die bloße Ausgabe der richtigen Lösung, und der Vorteil war am größten, wenn die Teilnehmenden die Aufgabe zuvor selbst versucht hatten. Er zeigte sich allerdings auch bei umgekehrter Reihenfolge. Im zweiten Experiment wurden Erklärungen mit Rechenfehlern und falschem Endergebnis eingespielt; die Lernzuwächse lagen dann zwischen denen der reinen Lösungsanzeige und denen fehlerfreier Erklärungen. Fehlerhafte Erklärungen waren in diesem Design also weniger schädlich als erwartet. Auf Grundschulkinder, die einen Fehler seltener bemerken, lässt sich das nicht übertragen. Ruan et al. (2020) verglichen mit 72 Kindern der Klassenstufen 3 bis 5 eine erzählbasierte Lernumgebung ohne Rückmeldung, mit statischen Hinweisen und mit einem tutoriellen Chatbot. Das Narrativ steigerte das Engagement, Lernzuwächse zeigten

sich in der Chatbot-Bedingung. Der Chatbot wurde allerdings von einer Person gesteuert (Wizard-of-Oz-Aufbau); die Studie belegt den Nutzen dialogischer Rückfragen, nicht die Leistungsfähigkeit eines KI-Systems.

Im Bereich des selbstregulierten Lernens weisen einige Studien darauf hin, dass adaptives, KI-gestütztes Scaffolding im Unterschied zu statischen Hilfesequenzen die Qualität von Selbststeuerungsprozessen verbessern kann und Vorteile gegenüber einem Konzept „gleiche Hilfen für alle“ hat (Liu et al., 2025; Wu et al., 2026). Generative KI muss also so eingebunden und vorstrukturiert werden, dass sie adaptiv auf die Eingaben der Lernenden eingeht und nicht nur vorgefertigte Impulse anbietet.

Das Prinzip hat eine Grenze, die sich aus dem oben genannten Lösungsbeispiel-Effekt ergibt: Ohne Vorerfahrung mit einem Aufgabentyp füllt die Suche nach einem Lösungsweg das Arbeitsgedächtnis mit Prozessen, die nicht zum Lernen beitragen; erst mit wachsender Vorerfahrung wird der eigene Versuch lernwirksamer als das Beispiel (Renkl, 2014). In Studien mit Tutorsystemen der Sekundarstufe hat sich eine Stufung bewährt: ein vollständiges Lösungsbeispiel mit Frage zum Nachvollziehen, dann ein unvollständiges Beispiel, bei dem das Kind die letzten Schritte ergänzt, dann der eigene Lösungsversuch mit Rückmeldung (Renkl & Atkinson, 2003). Entscheidend ist, woran das Ausblenden der Schritte hängt: Salden et al. (2010) zeigen, dass ein an den Erklärungen des Lernenden ausgerichtetes Ausblenden dem Ausblenden nach festem Schema überlegen ist. Das ist die Kontingenzforderung des Scaffoldings, angewandt auf Lösungsbeispiele. „Erst du – dann ich“ beschreibt also den Regelfall für Aufgabentypen, mit denen ein Kind bereits Erfahrung hat, nicht den Einstieg in einen neuen Inhalt.

Hilfen begrenzen und Hilfesuche beobachten. Hilfen müssen nicht nur individualisiert, sondern auch begrenzt sein. Dafür gibt es zwei unterschiedliche Mechanismen. Fading, das planmäßige Zurücknehmen von Unterstützung, richtet sich nach dem Kompetenzzuwachs: Unterstützung wird in dem Maß zurückgenommen, in dem das Kind die entsprechenden Schritte allein bewältigt, nicht danach, wie oft es um Hilfe bittet. Davon zu unterscheiden ist die Begrenzung von Hilfen bei Missbrauch. Die Forschung zu intelligenten Tutorsystemen zeigt seit Langem, dass Lernende Hilfefunktionen nicht so nutzen, wie es der Systementwurf vorsieht. Ein Teil klickt sich durch Hinweisketten bis zur Lösung, ohne die Zwischenschritte zu verarbeiten; ein anderer Teil vermeidet Hilfe genau dann, wenn sie nötig wäre, und rät oder arbeitet still falsch weiter (Aleven et al., 2016; überwiegend Sekundarstufe). Adaptives Hilfesuchen ist selbst eine metakognitive

Fähigkeit, die Grundschulkinder erst entwickeln. Ein KI-Lernbegleiter für diese Altersgruppe kann sich deshalb nicht darauf verlassen, dass Kinder Hilfe anfordern, wenn sie sie brauchen. Er sollte beide Muster erkennen und unterschiedlich beantworten: bei sehr schnellen, wiederholten Hinweisanfragen ohne Bearbeitungszeit mit einer Rückfrage statt mit dem nächsten Hinweis, bei wiederholten Fehlversuchen ohne jede Hilfeanfrage mit einem aktiv angebotenen Hinweis. Ziel ist nicht, Hilfe zu verknappen, sondern Hilfe an den richtigen Stellen nutzbar zu machen.

Anstrengung mit Untergrenze. Ein gewisses Maß an eigener Anstrengung ist für das Lernen bedeutsam, besonders beim Aufbau von Verständnis. Schwierigkeiten, die Lernende mit Anstrengung bewältigen können, etwa verteiltes statt geblocktes Üben oder das Abrufen statt des Wiederlesens, verschlechtern die Leistung während des Übens und verbessern gerade dadurch Behalten und Transfer (Bjork & Bjork, 2011). Diese Unterscheidung von Übungsleistung und Lernzuwachs ist für die Bewertung von KI-Lernbegleitung wichtig: Was während der Bearbeitung nach Erfolg aussieht, ist kein Beleg für Lernen. Der Befund von Bastani et al. (2025) ist genau dieser Fall. Lernförderlich sind Schwierigkeiten zudem nur unter zwei Bedingungen. Erstens muss die Schwierigkeit im Bereich des mit Unterstützung Erreichbaren liegen. Zweitens braucht die Phase des eigenen Ringens einen Abschluss: In der Forschung zum produktiven Scheitern entsteht der Lernvorteil nicht durch das Scheitern selbst, sondern durch die anschließende Klärungs- und Systematisierungsphase (Kapur, 2016). Diese Position steht in einer bekannten Spannung zur Cognitive Load Theory, die früh und stark anleitende Instruktion favorisiert (Kirschner et al., 2006); die Gegenposition hält dem entgegen, dass problemorientierte Formate keineswegs unangeleitet sind, sondern anders gestützt werden (Hmelo-Silver et al., 2007). Die Spannung ist nicht aufgelöst, und die Befunde zum produktiven Scheitern stammen überwiegend aus der Sekundarstufe.

Für die Primarstufe folgt daraus eine vorsichtige Fassung mit Untergrenze. Sechs- bis Zehnjährige erleben ungelöste Schwierigkeit ohne Ansprechpartner schnell als Misserfolg, nicht als Herausforderung, und ein wartender Bildschirm ist für ein Kind keine Denkpause, sondern eine defekte App. In der ursprünglichen Beschreibung von Scaffolding gehört die Kontrolle der Frustration ausdrücklich zu den Aufgaben der unterstützenden Person (Wood et al., 1976). Die Phase des eigenen Versuchs muss deshalb kurz und erkennbar begrenzt sein, sie braucht eine sichtbare Ausstiegsmöglichkeit, und die Unterstützung sollte in fester Folge zunehmen: nach dem ersten erfolglosen Versuch ein aktiv angebotener, konkreterer Hinweis statt

einer weiteren Frage, nach dem zweiten oder dritten ein Wechsel der Aufgabe oder der Darstellung oder ein Hinweis an die Lehrkraft. Kleinere Zahlen allein sind kein Wechsel; sie bestätigen, wie oben beschrieben, eine zählende Strategie. Ein Kind darf nie in einer Schleife aus Rückfragen hängen bleiben. Dahinter steht eine Anforderung, die sich erst aus der Häufung ergibt: Jede einzelne Fehlerrückmeldung kann angemessen sein und die Serie trotzdem eine Botschaft über das Kind erzeugen, die niemand gestaltet hat. Eine Anwendung muss deshalb die Folge der Rückmeldungen mitführen, nicht nur die einzelne. Welche Wartezeit angemessen ist, ist empirisch nicht geklärt; im Projekt PRIMA-KI wird das als offene Gestaltungsfrage behandelt.

Hinzu kommt eine praktische Schwierigkeit. Kinder im Grundschulalter können oft nicht formulieren, was sie nicht verstehen, und viele lesen und schreiben noch mühsam. Ein System, das einen eigenen Lösungsversuch nur als Text oder gesprochene Erklärung entgegennimmt, schließt genau die Kinder aus, die es erreichen soll. Der eigene Versuch muss deshalb auch nonverbal möglich sein: durch Handeln mit Material, eine Skizze oder das Auswählen aus mehreren dargestellten Wegen.

Ein KI-Lernbegleiter sollte daher nach dem Prinzip „Erst du – dann ich“ arbeiten, sobald ein Kind mit einem Aufgabentyp vertraut ist. Zunächst wird ein eigener Lösungsversuch eingefordert. Die KI kommt erst bei Bedarf hinzu, auf Anfrage oder automatisch bei Fehlern, und fragt dann nach Ideen, Teilstrategien und Beobachtungen. Erklärungen werden an vorhandene Verständnisgrundlagen angedockt. Vollständige Musterlösungen bleiben die Ausnahme und dienen der Reflexion, nicht als primäres Lernformat. Für das Lob gilt eine Regel aus der Feedbackforschung: Rückmeldung soll sich auf das beziehen, was das Kind getan hat, nicht auf das, was es ist. „Du hast zuerst die Zehn voll gemacht, das war hier ein guter Weg“ ist eine Rückmeldung, aus der ein Kind etwas mitnimmt; „Super, du bist ein Mathe-Ass“ ist keine. Zuschreibungen an die Person enthalten keine Information über die Aufgabe und gehören zu den am wenigsten wirksamen Formen der Rückmeldung (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Bei Fünftklässlerinnen und Fünftklässlern führte Lob für Intelligenz nach einem Misserfolg zu geringerer Ausdauer und schlechteren Leistungen als Lob für Anstrengung (Mueller & Dweck, 1998). Ein KI-Lernbegleiter, der auf jede Eingabe mit Begeisterung reagiert, ist deshalb nicht besonders freundlich, sondern besonders inhaltsleer.

Aktivierung: Metakognition und kritisches Prüfen von KI-Antworten

Die dritte Gestaltungsebene geht über die Dosierung von Hilfen hinaus und zielt auf die Qualität der kognitiven Prozesse, die durch die Interaktion angeregt werden. KI-Systeme können nicht nur Hinweise geben und Aufgaben erklären. Sie können Prozesse des Planens, Überwachens und Reflektierens im Problemlöseprozess anregen, wenn sie darauf ausgelegt werden. Die systematische Übersicht von Wu et al. (2026) zeigt, dass Chatbots selbstreguliertes Lernen unterstützen können, sofern ihre Scaffolds an Modelle des selbstregulierten Lernens gekoppelt sind. Guo et al. (2025) berichten in einer noch nicht begutachteten systematischen Übersicht, dass KI-Systeme bei Grundschulkindern die psychologischen Bedürfnisse nach Autonomie, Kompetenz und sozialer Eingebundenheit ansprechen können, die Motivation und Engagement beeinflussen.

Voraussetzung für all das ist Zielklarheit. Feedback kann nur wirken, wenn ein Kind weiß, worauf es bei der Aufgabe ankommt: Soll es schnell rechnen, einen Weg begründen oder verschiedene Wege vergleichen? Ein KI-Lernbegleiter sollte dieses Ziel zu Beginn benennen und seine Rückmeldungen erkennbar daran ausrichten (Hattie & Timperley, 2007). Auf jede Rückmeldung muss zudem eine Gelegenheit folgen, sie zu nutzen. Ein Hinweis, nach dem die Aufgabe abgeschlossen wird und die App zur nächsten weitergeht, wirkt nicht. Aufgaben sollten nach einem Hinweis erneut bearbeitbar bleiben, und ein Hinweis sollte eine konkrete nächste Handlung nahelegen.

Song et al. (2026) berichten aus einer Fallstudie in einer Mittelstufenklasse, dass Lernende ein KI-System, dem sie selbst etwas erklären sollten (teachable agent), als Lernbegleiter, Moderator und Mitlösenden wahrnahmen. Ob das Prinzip des Lernens durch Erklären in KI-gestützten Umgebungen auch Lernzuwächse erzeugt, ist damit noch nicht geklärt. Zugleich zeigt sich, dass viele heutige generative KI-Systeme tutorielle Aufgaben wie das gezielte Anregen von Planung, Strategieauswahl und Reflexion ohne spezielle Vorgaben oder Nachtraining nicht zuverlässig erfüllen und eher zu dem bereits beschriebenen preskriptiven Stil neigen (Chudziak & Kostka, 2025; Contel & Cusi, 2025). Solche Systeme müssen also gezielt vorstrukturiert werden, damit sie metakognitive Hilfen von sich aus anbieten und nicht nur auf Fehler reagieren.

Ein zweiter Aspekt der Aktivierung besteht darin, Kinder zur kritischen Prüfung von

KI-Antworten zu befähigen. Er liegt streng genommen quer zum selbstregulierten Lernen, weil er nicht die Steuerung des eigenen Lernprozesses betrifft, sondern den Umgang mit einer fehlbaren Informationsquelle. Für das Lernen mit KI gehört er dennoch hierher, weil beides dieselbe Haltung verlangt: das eigene Vorgehen und das des Systems nicht ungeprüft hinzunehmen. In einer Studie mit 77 Kindern der Jahrgangsstufen 3 bis 5, die mit einem sprachmodellgestützten Roboter Mathematikaufgaben bearbeiteten, konnten die meisten Kinder korrekte von fehlerhaften Antworten unterscheiden (Helal et al., 2024). Diese Fähigkeit hing allerdings deutlich davon ab, ob sie die Aufgabe selbst lösen konnten. Die kritische Prüfung von KI-Antworten setzt fachliches Wissen voraus und ersetzt es nicht. Eine eigene Beobachtung passt dazu: Bei einem Mathetag an einer Grundschule im Januar 2026 forderten Kinder der dritten und vierten Klasse an einer Station einen für sie entwickelten Chatbot mit Aufgaben heraus, die sie selbst sicher lösen konnten, und prüften seine Antworten nach. Dass sich die KI verrechnete, erstaunte die Kinder und führte im Verlauf der Station zu einer prüfenden Haltung. Das ist eine Beobachtung an einem einzelnen Tag, keine Erhebung, und aus ihr folgt keine stabile Fähigkeit. Kritisches Prüfen von KI-Antworten muss deshalb an Inhalten geübt werden, die die Kinder sicher beherrschen, bevor es auf neue Inhalte übertragen wird. Diese Fähigkeit können Kinder bereits in der Primarstufe üben; sie begegnen KI-generierten Inhalten im Alltag zunehmend (Yim & Su, 2025).

Ein KI-Lernbegleiter kann als metakognitiver Partner konzipiert werden, der diese Prozesse anregt. Für die Primarstufe müssen die Fragen an etwas Sichtbarem ansetzen: „Was ist bei diesen drei Aufgaben gleich?“, „Hast du gezählt oder von einer Aufgabe abgeleitet, die du schon kennst?“, „Zeig mir, wo die 8 in deiner Zeichnung steckt.“, „Ist dein Ergebnis größer oder kleiner als 100? Wie kannst du das schnell sehen?“ Offene Reflexionsfragen wie „Was ist dir aufgefallen?“ setzen sprachliche und metakognitive Mittel voraus, über die jüngere Kinder oft noch nicht verfügen; sie funktionieren erst, wenn sie sich auf einen konkreten Gegenstand vor dem Kind beziehen. Das System regt die Reflexion über Fehler an („Welche Idee von vorher könnte hier helfen?“) und kann, wo sinnvoll, Elemente des Zurückerklärens einbauen: Kinder erklären der KI, was sie verstanden haben, und die KI fragt nach und vertieft. Dadurch regt das System eigene Erklärungen, Prüfungen und Reflexionen an, statt selbst als allwissende Instanz aufzutreten.

Für die Primarstufe ist dieses Prinzip zugleich das am wenigsten gesicherte. Die Befunde zum selbstregulierten Lernen mit KI stammen fast ausschließlich aus Sekundarstufe und Hochschule.

Einbettung: Hybride Arrangements und die Rolle der Lehrkraft

Die vierte Gestaltungsebene betrifft den Rahmen, in dem KI-Lernbegleitung stattfindet. Mehrere Studien berichten Vorteile hybrider Arrangements, in denen KI die Lehrkraft nicht ersetzt, sondern entlastet, etwa durch passgenaue Scaffolds im Problemlöseprozess oder Echtzeit-Analysen, damit Lehrkräfte mehr Zeit für pädagogische Interaktion und Beziehungsgestaltung haben (Wezendonk & Veldhuis, 2024; Gonnermann-Müller et al., 2025, Preprint). Eine allgemeine Überlegenheit gegenüber gut gestaltetem rein menschlichem oder rein KI-gestütztem Feedback ist damit nicht belegt.

Ein exploratives Pilotprojekt des Anbieters Eedi mit Google DeepMind in fünf britischen Sekundarschulklassen mit 165 Schülerinnen und Schülern berichtet in einer Pressemitteilung (Eedi & Google DeepMind, 2025), dass ein Hybridarrangement aus KI-Vorschlägen und menschlicher Tutorin gegenüber einem Standardhinweis einen Zuwachs von 10 Prozentpunkten bei der Übertragung auf neue Aufgaben erzielte, gegenüber 4,5 Prozentpunkten bei rein menschlicher Unterstützung. Die Tutorinnen und Tutoren griffen dabei vor allem ein, um das Tempo der KI zu bremsen und ihren Ton abzumildern. Die Ergebnisse sind nicht begutachtet, stammen vom Anbieter selbst und sind vorläufig zu lesen. In der Meta-Analyse von Kaliisa et al. (2026) über 41 Studien mit 4813 überwiegend älteren Lernenden ließ sich zwischen KI-generiertem und menschlichem Feedback kein signifikanter Unterschied in der Lernleistung nachweisen ($g = 0{,}25$, nicht signifikant, bei sehr unterschiedlichen Einzelbefunden). Daraus folgt nicht unmittelbar, dass die Kombination beider Formen besser ist. Plausibel ist die Annahme, weil beide Formen unterschiedliche Stärken haben: KI-Feedback ist unmittelbar verfügbar und niedrigschwellig, Lehrkraft-Feedback ordnet ein und kennt das Kind. Ob die Kombination mehr leistet als jede Form für sich, ist eine offene Forschungsfrage. Dass der Träger der Unterstützung weniger entscheidet als ihre Gestaltung, ist kein neuer Befund: Schon für regelbasierte Tutorsysteme fand VanLehn (2011), dass gut gebaute Systeme in der Größenordnung menschlicher Einzeltutoren wirken ($d = 0{,}76$ gegenüber $d = 0{,}79$) und dass die oft zitierte Überlegenheit menschlichen Tutorings um zwei Standardabweichungen empirisch nicht haltbar ist. Cosentino et al. (2025) sehen in hybriden Feedback-Modellen das Potenzial, die kognitive Belastung zu reduzieren und differenzierte Informationsverarbeitung zu unterstützen.

Ein Gegenbeispiel zur pauschalen Hybrid-Annahme liefern Kestin et al. (2025). In

einem randomisierten Kontrollversuch mit Physikstudierenden einer hochselektiven Hochschule lernten die Teilnehmenden mit einem nach didaktischen Prinzipien gestalteten KI-Tutor in kürzerer Zeit signifikant mehr als in einer Präsenzstunde mit aktivierenden Methoden. Der Befund betrifft eine einzelne Lerneinheit im Hochschulkontext und lässt sich nicht auf den Grundschulunterricht übertragen. Er zeigt aber, dass gut gestaltete KI-Unterstützung nicht grundsätzlich auf menschliche Rahmung angewiesen ist. Ein weiterer Befund betrifft die Unterstützung der Lehrenden selbst. Tutor CoPilot (Wang et al., 2024, Preprint) unterstützt Tutorinnen und Tutoren eines Online-Nachhilfeprogramms in Echtzeit mit Formulierungsvorschlägen. In einem randomisierten Kontrollversuch mit 900 Tutorinnen und Tutoren und 1800 Schülerinnen und Schülern der Klassenstufen K–12 stieg die Wahrscheinlichkeit, ein Thema zu beherrschen, um 4 Prozentpunkte, bei zuvor schwächer bewerteten Tutorinnen und Tutoren um 9 Prozentpunkte. Die Analyse der Nachrichten zeigt, dass sie häufiger leitende Fragen stellten und seltener pauschal lobten.

Aus der Perspektive der hybriden Einbettung ergeben sich mehrere Konsequenzen für das Systemdesign. Die Generierung von Feedback muss auf fachdidaktisch fundierten Informationen beruhen. Lehrkräfte benötigen Einblick in den Interaktionsverlauf und Einstellungsmöglichkeiten. Die Forschung zu Übersichtsanzeigen für Lehrkräfte (Dashboards) zeigt, dass deren Nutzen von der Deutungsarbeit der Lehrkraft abhängt. In einer Beobachtungsstudie in 38 Mathematikstunden der Grundschule konsultierten Lehrkräfte die Anzeige im Mittel 8,3-mal pro Stunde und leiteten daraus vor allem aufgaben- und prozessbezogene Rückmeldungen ab (Molenaar & Knoop-van Campen, 2019). Die angezeigten Daten wirken also nicht von selbst, sondern über das, was die Lehrkraft mit ihnen anfängt. Das legt nahe, Anzeigen zu priorisieren, statt Verlaufsdaten zu häufen: Welches Kind braucht jetzt die Lehrkraft, und warum? In einer Studie mit Mittelstufenklassen, in der ein Werkzeug den Lehrkräften anzeigte, welche Kinder im Tutorsystem gerade Hilfe brauchten, lernten die Kinder mehr als ohne diese Anzeige (Holstein et al., 2018). Ein direkter Vergleich beider Anzeigeformen liegt nicht vor. Eine Übersichtsanzeige für die Grundschule sollte deshalb weniger Verlaufsdaten und mehr handlungsleitende Hinweise enthalten und im laufenden Unterricht in Sekunden erfassbar sein. Die KI kann Vorschläge für Aufgaben, Hilfen oder erste Diagnosen machen; die Entscheidung bleibt beim Menschen. Aussichtsreich ist vor allem die Unterstützung auf Aufgabenebene beim Problemlösen. In einer heterogenen Klasse erreichen Lehrkräfte dort oft nicht alle Kinder rechtzeitig. Im systematischen Review von Eti et al. (2026) werden vor allem solche Systeme als

erfolgreich beschrieben, die Fehlvorstellungen erkennen, passende Hinweise geben und Lernende schrittweise zu selbstständigem Problemlösen führen; ein Wirksamkeitsvergleich zwischen diesen und anderen Gestaltungen liegt der Übersicht nicht zugrunde.

Die Einbindung von Lehrkräften in Entwicklung und Einsatz von KI-Tutorsystemen bleibt bisher oft unzureichend mitgedacht. Guerino et al. (2023) sowie Wezendonk und Veldhuis (2024) betonen, dass lehrkraftzentrierte Gestaltungsansätze und Programme zur KI-Kompetenz von Lehrkräften notwendig sind, um die Integration in den Unterricht und die Akzeptanz zu sichern. Professionalisierung und Fortbildung zur Einbindung und Prüfung von KI sind Voraussetzung für einen verantwortlichen Einsatz (Holmes et al., 2019; KMK, 2024; Wang & Nie, 2023) und gehören in die Lehrkräfteaus- und -weiterbildung. Fortbildung sollte technische Grundlagen sowie pädagogische Chancen, Risiken und Prüfkriterien umfassen. Grundlage bleibt das fachdidaktische Wissen, mit dem Lehrkräfte die KI als Assistenz im Lernarrangement prüfen und einordnen.

KI als Anstoß für mathematische Aktivitäten mit Material

Die Gestaltungsprinzipien beschreiben, wie KI-Lernbegleitung in der Interaktion zwischen System und Kind funktionieren sollte. Ebenso wichtig ist, wie sich diese Interaktion in das Gesamtarrangement mathematischen Lernens einfügt. KI soll nicht dazu führen, dass Kinder nur noch auf Bildschirme schauen. Forschung zu gegenständlichen Bedienoberflächen (Tangible Interfaces) und sozialen Robotern zeigt, dass KI auch Interaktion in der physischen Welt anregen kann (Ligthart et al., 2023). KI sollte als Anstoß für mathematische Aktivitäten dienen und dabei den Wechsel zwischen den Darstellungsebenen unterstützen: zwischen Handlung am Material (enaktiv), Bild und Skizze (ikonisch) sowie Zahl, Term und Sprache (symbolisch; Bruner, 1966). Lernwirksam ist weniger die einzelne Ebene als der Übersetzungsprozess zwischen ihnen, in der deutschen Didaktik als intermodaler Transfer bezeichnet (Wartha & Schulz, 2012). Der produktivste Impuls eines KI-Systems ist deshalb häufig keine Erklärung, sondern eine Aufforderung zum Darstellungswechsel: „Kannst du das mit den Plättchen legen?“, „Zeichne, was du gerechnet hast.“, „Wo steckt die 7 in deiner Zeichnung?“ Solche Impulse verlagern die Denkarbeit zurück zum Kind und verlassen zugleich den Bildschirm.

Die KI kann Handlungen mit Material sowie Skizzen, Notizen oder Fotos einbeziehen

und dazu Fragen stellen, beispielsweise beim Modellieren von Sachaufgaben in der App Rechengeschichten. Dort können neben der Spracheingabe auch Skizzen, Notizen und Fotos der Modellierung mit dem KI-Lernbegleiter besprochen werden. Gerade Sachaufgaben zeigen, wie fein die Grenze zwischen Unterstützung und Übernahme verläuft. Kinder scheitern hier selten am Rechnen, häufiger am sinnentnehmenden Lesen, an unbekannten Wörtern, an Signalwortstrategien wie „zusammen heißt plus“ und daran, das Ergebnis auf seine Sinnhaftigkeit zu prüfen. Ein KI-Lernbegleiter kann sprachlich entlasten, indem er vorliest, ein Wort erklärt, die Situation nachfragt oder bei Kindern mit Deutsch als Zweitsprache in der Erstsprache nachfragt, ohne die mathematische Struktur vorwegzunehmen. Formuliert er die Aufgabe dagegen so um, dass die Rechenoperation bereits erkennbar ist, entfällt genau die Leistung, um die es geht. Eine Rückfrage in der Erstsprache setzt zudem zweierlei voraus: dass sie als Einstiegshilfe und nicht als Dauerlösung gedacht ist, weil der Aufbau der deutschen Fachsprache selbst zum Lernziel gehört, und dass die Lehrkraft den Verlauf nachvollziehen kann, etwa über eine mitgeführte deutsche Fassung. Für die Praxis ergibt sich daraus eine einfache Prüffrage: Nimmt die KI dem Kind das Lesen ab oder das Modellieren?

Die Spracheingabe ist dabei ein empfindlicher Punkt. Kinderstimmen werden von automatischer Spracherkennung deutlich schlechter erkannt als Erwachsenenstimmen; gängige Modelle erreichen bei Kindersprache Wortfehlerraten, die deutlich über denen bei Erwachsenen liegen, untersucht bislang vor allem für englischsprachige Kinder (Attia et al., 2024), verstärkt durch Dialekt, Mehrsprachigkeit und Geräusche im Klassenraum. Ein Erkennungsfehler bleibt für das nachgeschaltete Sprachmodell unsichtbar. Es deutet das Missverstandene als inhaltliche Äußerung und im ungünstigen Fall als Fehlvorstellung. Systeme für die Primarstufe sollten deshalb Unsicherheit sichtbar machen und rückfragen, statt zu deuten: Sie zeigen oder lesen vor, was verstanden wurde, und lassen das Kind korrigieren, bevor das Sprachmodell darauf antwortet.

Für den Primarbereich liegen deutlich weniger Studien vor als für Sekundarstufe und Hochschule. Die vorhandenen Ergebnisse stimmen vorsichtig optimistisch: Die Meta-Analyse von Hwang (2022) findet für regelbasierte KI-Systeme in der Grundschulmathematik einen kleinen positiven Effekt ($g = 0{,}35$), und die wenigen Studien zu dialogischen Systemen deuten darauf hin, dass Kinder profitieren, wenn die Lernumgebung sinnvoll gestaltet ist und digitale Begleitung mit analogem Arbeiten verknüpft wird (Kuo et al., 2026; Liu et al., 2025; Ruan et al., 2020, dort mit menschengesteuertem Chatbot). Eine explorative Fallstudie mit zwei

sechsjährigen Kindern (Rumbelow & Coles, 2024) beschreibt, wie eine App zur Objekterkennung von Cuisenaire-Stäben bei einem der beiden Kinder innerhalb weniger Minuten einen Wechsel von der Deutung der Stäbe als Zähleinheiten hin zur Deutung als Längen anstieß; beim zweiten Kind trat dieser Wechsel nicht ein. Solche Ergebnisse sind Hinweise auf ein Gestaltungspotenzial, keine Wirksamkeitsbelege. Übertragen auf gängiges Material im deutschen Anfangsunterricht hieße das: Ein Kind legt eine Aufgabe mit Wendeplättchen am Zwanzigerfeld, fotografiert die Legung, und das System fragt nach, statt zu bewerten: „Welche Aufgabe hast du hier gelegt?“, „Warum hast du fünf und fünf zusammengeschoben?“ Der Gewinn liegt nicht in der Erkennung selbst, sondern darin, dass das Kind sein Handeln versprachlichen muss. Das Foto soll die Handlung dokumentieren, nicht ersetzen. Wie verlässlich die Erkennung gelingt, hängt vom Material ab: Klar unterscheidbares strukturiertes Material wird gut erkannt, während handgezeichnete Skizzen, Kinderhandschrift und vor allem das genaue Zählen von Objekten für heutige Bildmodelle noch fehleranfällig sind.

KI-gestütztes Üben kann die Automatisierung von Rechensätzen unterstützen; eine kleinere, in einer wenig verbreiteten Zeitschrift veröffentlichte Studie mit achtjährigen Kindern über sechs Wochen berichtet größere Flüssigkeitsgewinne als beim reinen Auswendiglernen (Samuelsson, 2023). Für die Primarstufe ist dabei die Reihenfolge entscheidend: Automatisierung setzt Verständnis und ableitende Strategien voraus. Bei Kindern, die noch zählend rechnen, ist Übung unter Zeitdruck nicht nur wenig wirksam, sondern kann die zählende Strategie verfestigen und Vermeidungsverhalten verstärken (Gaidoschik, 2010; Wartha & Schulz, 2012). Ein KI-Lernbegleiter müsste daher unterscheiden können, ob ein Kind eine Aufgabe zählend oder ableitend löst, und im ersten Fall nicht Tempo, sondern Strategieaufbau adressieren, etwa über die Kraft der Fünf, das Verdoppeln oder das Zerlegen am Zwanzigerfeld. Adaptive Systeme für Kinder mit Rechenschwierigkeiten werden in Prototypenstudien erprobt (Hocine et al., 2023). Spielerische Ansätze sind ebenfalls untersucht: Sayed et al. (2022) berichten Verbesserungen besonders bei leistungsschwächeren Schülerinnen und Schülern durch adaptive, spielerische Inhalte.

Kinder sollen mathematische Aussagen, auch und gerade von KI, kritisch prüfen, begründen und miteinander diskutieren (Kortenkamp, 2024; Aufenanger et al., 2023). KI-Lernbegleitung in der Grundschule sollte daher vor allem als Anstoß für reichhaltige mathematische Aktivitäten dienen, die digitale und analoge Handlungen miteinander verbinden und an fachdidaktischen Überlegungen

ausgerichtet sind.

Wann auf KI-Lernbegleitung verzichtet werden sollte

Aus den vier Prinzipien folgt spiegelbildlich, in welchen Situationen ein Einsatz derzeit nicht sinnvoll ist. Vor jeder Frage nach der Gestaltung steht die Frage, ob KI hier etwas leistet, das Material, Gespräch und eigenes Probieren nicht leisten. Ein Kind, das Zahlen zerlegt, braucht Plättchen und eine Frage, keinen Chatbot.

Beim erstmaligen Aufbau von Grundvorstellungen mit Material tritt der Bildschirm in Konkurrenz zur Handlung und zum Gespräch. Hier ist die Lehrkraft die passendere Instanz; KI kann allenfalls im Anschluss zur Versprachlichung des Gehandelten beitragen.

Bei Aufgaben, deren Lernwert gerade im eigenständigen Ringen liegt, etwa offene Problemstellungen, verkürzt jede sofort verfügbare Hilfe genau den Prozess, um den es geht. Hier ist die Lehrkraft, die Zeit gibt und im richtigen Moment nachfragt, dem System überlegen.

Bei diagnostischen Einschätzungen mit Folgen, etwa Förderentscheidungen, dem Verdacht auf Rechenschwierigkeiten oder der Leistungsbewertung, ist eine KI-gestützte Einschätzung bestenfalls ein Hinweis, nie eine Grundlage. Die Verantwortung liegt bei der Lehrkraft.

Wenn die Lehrkraft keinen Zugriff auf den Interaktionsverlauf hat, entfällt das vierte Prinzip. Eine unbeaufsichtigte KI-Lernbegleitung in der Primarstufe, deren Ausgaben niemand prüft, ist nach dem gegenwärtigen Kenntnisstand nicht zu verantworten.

Schließlich entstehen zusätzliche Hürden statt Unterstützung, wenn die sprachlichen Voraussetzungen der Kinder eine dialogische Interaktion nicht zulassen, etwa bei geringer Lesekompetenz ohne verlässliche Sprachausgabe oder bei unzuverlässiger Spracherkennung. Die Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (SWK, 2024) empfiehlt für die Grundschule eine weitgehende Abstinenz von Sprachmodellen und stattdessen den systematischen Aufbau basaler Lese- und Schreibkompetenzen; eine reguläre Nutzung sieht sie erst ab Klasse 8 vor. Die hier beschriebenen Prinzipien beziehen sich nicht auf den freien Zugang zu einem Chatbot, den die SWK im Blick hat, sondern auf eng

vorstrukturierte, in eine Lernanwendung eingebettete Systeme. Ob eine solche Einbettung die Bedenken der SWK ausräumt, ist eine offene Frage; die vorliegenden Befunde sprechen dafür, dass es auf genau diese Vorstrukturierung ankommt.

Was du als Lehrkraft prüfen kannst

Die vier Prinzipien lassen sich in Fragen übersetzen, mit denen du einen KI-Lernbegleiter im Schulalltag beurteilen kannst. Ein einfaches Verfahren: Bearbeite die App selbst und gib dabei bewusst die Fehler ein, die du aus deiner Klasse kennst.

- **Fundierung.** Gib eine typische Fehlvorstellung ein, etwa $42 - 8 = 46$ (das Kind zieht ziffernweise die kleinere von der größeren Ziffer ab), oder antworte auf die Frage, wie viele Zehner in 240 stecken, mit „vier“ statt „24“. Erkennt das System, welche Vorstellung dahintersteckt, oder meldet es nur „falsch, versuch es noch einmal“? Schlägt es bei einer Aufgabe wie $8 + 6$ vor, weiterzuzählen, statt eine ableitende Strategie anzuregen? Dann ist es für Kinder mit Rechenschwierigkeiten nicht geeignet. Bei $8 + 1$ ist Weiterzählen dagegen unbedenklich; entscheidend ist, ob das System dort zählen lässt, wo eine Strategie zur Verfügung stünde.
- **Dosierung.** Fordere fünfmal hintereinander einen Hinweis an, ohne selbst etwas zu rechnen. Gibt das System irgendwann die Lösung heraus? Fragt es nach einer eigenen Idee? Ein System, das auf Knopfdruck beliebig oft Lösungen liefert, wird in der Klasse genau so genutzt werden. Prüfe auch den umgekehrten Fall: Was passiert, wenn du dreimal denselben Fehler machst, ohne Hilfe anzufordern? Und lass dich nicht davon beeindrucken, dass Kinder in einer App sichtbar viel richtig machen; Übungserfolg mit Hilfe ist kein Beleg für Lernen.
- **Aktivierung.** Gib eine richtige Antwort mit falscher Begründung ein. Reagiert das System auf die Begründung oder nur auf das Ergebnis? Lobt es alles? Widerspricht es, wenn du eine falsche Behauptung überzeugt vorträgst?
- **Einbettung.** Siehst du, welche Hilfen ein Kind bekommen hat? Kannst du Funktionen abschalten? Kannst du die Rückmeldungen einer Stunde in den Tagen danach nachlesen, bevor du mit dem Kind sprichst? Nimmt die KI dem Kind bei Sachaufgaben das Lesen ab oder das Modellieren?

Für den Unterricht folgt daraus: Beim Automatisieren von Rechenfertigkeiten ist unmittelbare, knappe Rückmeldung richtig; hier darf die KI sofort sagen, ob etwas

stimmt. Das gilt allerdings erst, wenn ein Kind die Aufgaben verstanden hat und nicht mehr zählend löst. Kinder, die noch zählen, brauchen Strategieaufbau, kein schnelleres Feedback. Beim Aufbau von Verständnis und beim Problemlösen sollte sie erst nach einem eigenen Versuch eingreifen und dann mit einer Frage beginnen. Bei einem neuen Verfahren, etwa einem Rechenweg über den Zehner, ist der Einstieg über ein gemeinsam betrachtetes Lösungsbeispiel sinnvoller als der eigene Versuch; für den erstmaligen Aufbau von Grundvorstellungen gilt das nicht, dort gehen Material und Gespräch vor. In der Übersicht ist nicht nur auffällig, wer viele Hinweise anfordert, sondern auch, wer nie welche anfordert und trotzdem Fehler macht. Und Kinder trauen sich einen eigenen Versuch nur zu, wenn Fehler in der Klasse als Arbeitsmaterial und nicht als Makel behandelt werden. Diesen Rahmen kann kein System herstellen. Eine ausführlichere Prüfliste mit fünf Fragen bietet der SKILL-Check im Beitrag zum SKILL-Modell.

Zentrale Belege im Überblick

Die folgende Tabelle ordnet die wichtigsten empirischen Belege dieses Beitrags nach Stichprobe, Design und Übertragbarkeit auf die Primarstufe. Sie macht sichtbar, dass die große Mehrheit der Wirkungsbefunde aus höheren Bildungsstufen stammt.

Quelle	Stichprobe	Design	Befund	Übertragbarkeit auf die Primarstufe
Bastani et al. (2025)	rund 1000 Schülerinnen und Schüler, Klassen 9–11, Türkei	Feldexperiment	Freier Chatbot-Zugang: mehr Erfolg beim Üben, 17 % schlechter im Test ohne KI; Variante mit Hinweisen: Test auf Kontrollniveau	Mechanismus plausibel, für Grundschule ungeprüft
Gerlich (2025)	666 Erwachsene	Querschnittsbefragung	$r = -0{,}68$ zwischen KI-Nutzung und kritischem Denken	Nur Hypothese; keine Kinderdaten
Kumar et al. (2025)	1818 Erwachsene online	Zwei präregistrierte Experimente	Erklärungen wirksamer als Lösungen; Vorteil am größten nach eigenem Versuch; fehlerhafte Erklärungen weniger schädlich als erwartet	Prinzip plausibel, Übertragung offen
Makransky et al. (2025)	175 Studierende; 234 Sekundarstufe II	Präregistriertes Experiment und Replikation	Didaktisch vorstrukturierter Tutor besser als generisches Modell im Sofort-Test; verzögerter Test nur in Studie 2 signifikant	Übertragung offen

Quelle	Stichprobe	Design	Befund	Übertragbarkeit auf die Primarstufe
Abdulsalam & Aroyehun (2025, Preprint)	Nachhilfedialoge; Altersangaben fehlen	Dialoganalyse	Höflichkeit hängt negativ mit bewerteter Qualität zusammen; keine Lernergebnisse	Hinweis auf Systemverhalten, nicht auf Lernen
Kestin et al. (2025)	194 Physikstudierende	Randomisierter Kontrollversuch	KI-Tutor lernwirksamer als aktivierende Präsenzstunde	Nicht übertragbar; zeigt, dass gute Gestaltung nicht zwingend menschliche Rahmung braucht
Kaliisa et al. (2026)	41 Studien, 4813 Lernende, überwiegend Hochschule	Meta-Analyse	Kein signifikanter Unterschied KI- vs. menschliches Feedback (g = 0,25)	Keine Aussage für Grundschule
Eedi & Google DeepMind (2025, Pressemitteilung)	165 Schülerinnen und Schüler, Sekundarstufe, Großbritannien	Exploratives Pilotprojekt	Gegenüber einem Standardhinweis: Mensch + KI +10 Prozentpunkte, Mensch allein +4,5	Vorläufig; Anbieterquelle
Wang et al. (2024, Preprint)	900 Tutorinnen und Tutoren, 1800 Lernende K–12	Randomisierter Kontrollversuch	Formulierungshilfe für Tutorinnen und Tutoren: +4 Prozentpunkte Themenbeherrschung	Teilweise, Nachhilfekontext
Kuo et al. (2026)	Fünftklässlerinnen und Fünftklässler, Taiwan	Interventionsstudie	Leistungsschwächere Kinder profitieren am stärksten von adaptiver Unterstützung	Direkt relevant
Liu et al. (2025)	Grundschulkinder, Textaufgaben	Interventionsstudie	Leistungsgewinne; wahrgenommene Qualität der Unterstützung entscheidend	Direkt relevant
Ruan et al. (2020)	72 Kinder, Klassen 3–5, USA	Experiment; Chatbot im Wizard-of-Oz-Aufbau	Narrativ steigert Engagement; tutorieller Chatbot verbessert Lernergebnisse	Grundschulkinder, aber kein KI-System im Einsatz
Helal et al. (2024)	77 Kinder, Jahrgang 3–5	Nutzungsstudie mit Experiment	Fehlererkennung hängt vom eigenen Können ab	Direkt relevant
Rumbelow & Coles (2024)	2 Sechsjährige	Explorative Fallstudie	Objekterkennung stieß bei einem Kind einen Darstellungswechsel an	Hinweis, kein Beleg
Steenbergen-Hu & Cooper (2013)	34 Stichproben, Klassen 1–12	Meta-Analyse	Intelligente Tutorsysteme gegenüber regulärem Unterricht: g = 0,01 bis 0,09	Teilweise; dämpft Erwartungen
Liu et al. (2026)	22 Studien, 46 Stichproben, 5232 Teilnehmende	Meta-Analyse	Generative KI in der Mathematik: g = 0,53; größere Effekte bei kleinen Stichproben	Kaum Grundschulstudien enthalten; Publikationsverzerrung möglich

Quelle	Stichprobe	Design	Befund	Übertragbarkeit auf die Primarstufe
Hwang (2022)	21 Studien, 30 Stichproben, Grundschulmathematik	Meta-Analyse	Regelbasierte KI-Systeme: g = 0,35	Direkt relevant, aber Systeme vor der Verbreitung generativer KI
Létourneau et al. (2025)	28 Studien, 4597 Lernende, Klassen 1–12	Systematische Übersicht	Überwiegend positive Effekte; nur 4 Studien (14 %) mit Grundschulkindern	Belegt die Forschungslücke selbst
Wisniewski et al. (2020)	435 Studien, 994 Effektstärken, altersübergreifend	Meta-Analyse	Feedback d = 0,48; informationsreiches Feedback d = 0,99; Verstärkung und Sanktionierung d = 0,24	Gut, altersübergreifend
Kluger & DeNisi (1996)	überwiegend Erwachsene, auch Arbeitskontexte	Meta-Analyse	Über ein Drittel der Feedbackinterventionen verschlechtert die Leistung, vor allem bei Fokus auf die Person	Mechanismus plausibel; Stichproben nicht aus der Schule

Konsequenzen für die technische Umsetzung

Dieser Abschnitt richtet sich an Entwicklerinnen und Entwickler; für den Unterrichtseinsatz ist er nicht erforderlich. Die vier Prinzipien stellen Anforderungen, an denen Anwendungen bei der Umsetzung scheitern können, und aus dem Forschungsstand folgen einige Konsequenzen, die über die didaktische Argumentation hinausgehen. Die folgenden Aussagen beruhen auf dem technischen Stand vom September 2026 und auf Erfahrungen aus der Entwicklung im Projekt PRIMA-KI. Sie sind Konstruktionsregeln, keine empirisch geprüften Wirkaussagen, und sie veralten schneller als der übrige Beitrag.

Der Lernstand gehört in die Anwendung, nicht in das Sprachmodell. Ein Sprachmodell verfügt über kein Modell des Lernenden, sondern nur über den Text, den es im aktuellen Gespräch sieht. Je länger dieses Gespräch wird, desto unzuverlässiger befolgt es seine ursprünglichen Vorgaben. Der Lernstand muss deshalb von der App geführt werden: bearbeitete Aufgaben, Fehlerarten, Anzahl angeforderter Hilfen, Bearbeitungsdauer, wiederkehrende Muster. Diese Information wird dem Modell in verdichteter, strukturierter Form übergeben, nicht als vollständiger Gesprächsverlauf. Das Sprachmodell deutet dann eine einzelne Äußerung im Licht dieses Lernstands; die Fortschreibung des Lernstands bleibt Aufgabe der Anwendung. Diese Trennung ist zugleich die Voraussetzung dafür, dass Lehrkräfte den Lernstand einsehen können und dass er über eine Sitzung hinaus

Bestand hat.

Prüfbare Entscheidungen gehören in den Programmcode. Der wirksamste Schutz gegen Halluzinationen, also frei erfundene, aber plausibel klingende Ausgaben, ist kein KI-Verfahren, sondern eine Zuständigkeitsgrenze. Ob eine Antwort richtig ist, welche Aufgabe als Nächstes kommt, wie viele Hilfen ein Kind schon bekommen hat: Das prüft und steuert die Anwendung, nicht das Sprachmodell. Fading und Hilfebegrenzung sind beides Anwendungslogik, aber mit verschiedenen Steuergrößen. Fading richtet sich nach dem geschätzten Kompetenzstand aus dem Lernendenmodell: Die höchste angebotene Hilfestufe sinkt, wenn ein Kind einen Aufgabentyp zunehmend allein bewältigt. Die Begrenzung bei Missbrauch richtet sich nach dem Nutzungsmuster innerhalb einer Aufgabe: Wer in wenigen Sekunden dreimal den nächsten Hinweis anfordert, bekommt eine Rückfrage statt der nächsten Stufe. Ein Hilfebudget setzt nur das Zweite um; wer damit auch das Erste abbilden will, verknappt die Hilfe für die Kinder, die sie am nötigsten brauchen. Ein Sprachmodell hält eine solche Zurückhaltung im laufenden Gespräch von sich aus nicht durch. Vorstrukturierung über einen Systemprompt steuert das Modell nur wahrscheinlichkeitsbasiert, nicht verbindlich, und gerade Verbote der Art „gib keine vollständige Lösung heraus“ lassen sich durch einfache Nachfragen unterlaufen. Hinzu kommt, dass Kinder ein solches System aktiv testen; Eingaben, die die Vorgaben außer Kraft setzen sollen, sind in einer Klasse binnen weniger Tage in Umlauf. Verlässlich wird die Zurückhaltung von Lösungen erst, wenn die Anwendung entscheidet, welche Hinweisstufe überhaupt an das Modell übergeben wird, und Ausgaben vor der Anzeige prüft. Was das Modell nicht kennt, kann es nicht ausplaudern; bei einfachen Rechenaufgaben, deren Lösung das Modell selbst findet, muss die Prüfung der Ausgabe diese Aufgabe übernehmen. Das steht in Spannung zur Kontingenzforderung aus dem Abschnitt zur Dosierung. Aufgelöst wird sie so: Adaptiv bleiben Diagnose und Formulierung; welche Hilfestufe freigegeben wird, entscheidet die Anwendung nach Regeln, die den geschätzten Kompetenzstand einbeziehen.

Diese Grenze lässt sich allerdings nicht überall ziehen. Ob ein Rechenergebnis stimmt, kann die Anwendung entscheiden; ob eine gesprochene Begründung, eine Skizze oder ein selbst gefundener Rechenweg mathematisch stichhaltig ist, kann sie nicht entscheiden. Wo die Anwendung nichts prüfen kann, sollte das System nicht bewerten, sondern nachfragen und die Beurteilung dem Kind oder der Lehrkraft überlassen.

Fachdidaktisches Wissen an die Aufgaben binden. Die im Abschnitt zur Fundierung genannten Wissensgrundlagen lassen sich auf mehreren Wegen bereitstellen: über den Systemprompt, über strukturiertes Wissen, das an die Aufgaben selbst gebunden ist und situationsabhängig zugespielt wird, oder über ein spezialisiertes Nachtraining des Modells. Der zweite Weg ist in der Praxis der ergiebigste, weil er dasselbe Wissen für die Anwendung, für die Aufgabenauswahl und für die Rückmeldung an die Lehrkraft nutzbar macht. Fehlertaxonomien gehören deshalb als maschinenlesbare Markierungen an Aufgaben und typische Falschantworten, nicht als Fließtext in eine Anweisung. Diese Markierungen sind zugleich die Voraussetzung dafür, den Lernstand fortschreiben zu können: Ein Verfahren wie Knowledge Tracing braucht Aufgaben, denen die geforderten Fähigkeiten zugeordnet sind. Ein umfassendes eigenes Nachtraining scheitert für den deutschsprachigen Primarbereich derzeit meist am fehlenden Datenmaterial; für eng umgrenzte Teilaufgaben, etwa das Kategorisieren typischer Fehler, kann eine Anpassung mit vergleichsweise wenigen, fachdidaktisch kuratierten Beispielen genügen. Ein einzelner, immer längerer Regelkatalog im Systemprompt ist nicht die Lösung: Je mehr Vorgaben gleichzeitig gelten, desto unzuverlässiger werden sie befolgt. Sinnvoller sind kurze, auf die jeweilige Lernsituation zugeschnittene Vorgaben.

Kleine Modelle bringen einen Zielkonflikt mit sich. Für eng umgrenzte Aufgaben kann ein kleineres, spezialisiertes Modell ausreichen, das mit weniger Rechenleistung auskommt und sich lokal auf dem Gerät oder auf eigenen beziehungsweise vertraglich gebundenen Servern betreiben lässt, sodass die Eingaben der Kinder den eigenen Verantwortungsbereich nicht verlassen. Das entschärft Datenschutz- und Energiefragen; auf Schul-Tablets ist die Grenze meist der Arbeitsspeicher. Kleine Modelle sind derzeit aber gerade dort am schwächsten, wo dieser Beitrag die höchsten Anforderungen stellt: beim Deuten offener Kinderäußerungen und beim Erkennen der Vorstellung hinter einem Fehler. Sie folgen komplexen Anweisungen weniger zuverlässig, und die geforderte Zurückhaltung bei Lösungen ist eine solche Anweisung. Realistisch ist heute eine Arbeitsteilung: klar entscheidbare Aufgaben wie Richtig-falsch-Prüfung, die Kategorisierung erwartbarer Fehler und die Aufgabenauswahl lokal und regelbasiert, das offene Gespräch über ein leistungsfähigeres Modell, und nur dort, wo es didaktisch gebraucht wird.

Prüfverfahren mit unabhängigem Signal. Verfahren, bei denen mehrere Modellinstanzen arbeitsteilig zusammenwirken und eine von ihnen die Ausgaben

der anderen prüft (Multi-Agent-Ansätze), oder bei denen ein zweites Sprachmodell die Ausgaben des ersten bewertet (LLM-as-Judge), verbessern in mehreren Arbeiten die bewertete Qualität von Scaffolds und senken die Rate offensichtlich falscher Rückmeldungen (Gonnermann-Müller et al., 2025, Preprint; Qian et al., 2026). Diese Ergebnisse beziehen sich auf Qualitätsbewertungen, nicht auf gemessenes Lernen. Wo ein Sprachmodell als Prüfinstanz eingesetzt wird, ist zuvor zu zeigen, dass seine Urteile mit denen fachkundiger Personen übereinstimmen; Cohn et al. (2026) berichten für ihre Prüfmodelle eine hohe Übereinstimmung mit menschlicher Konsenskodierung. Ein Sprachmodell, das ein Sprachmodell prüft, teilt zudem dessen systematische Fehler, besonders wenn beide aus derselben Modellfamilie stammen, und zeigt bekannte Verzerrungen: eine Bevorzugung längerer Antworten und eine Bevorzugung der eigenen Ausgaben (Zheng, Chiang, et al., 2023). Verlässlich sind solche Prüfungen vor allem, wenn die prüfende Instanz eine unabhängige Information nutzt: die bekannte Lösung, eine Regel, eine hinterlegte Wissensquelle. Im Mathematikbereich lässt sich rechnerische Korrektheit eindeutig nachrechnen, bevor eine Ausgabe das Kind erreicht, verlässlich allerdings nur dort, wo die Anwendung die Rechnung selbst kennt und nicht erst aus freiem Text herauslesen muss.

Antwortzeit, Infrastruktur und Kosten. Jede Prüfstufe kostet Antwortzeit. Ein Gespräch mit einem Kind der zweiten Klasse gelingt nur, wenn die Antwort in wenigen Sekunden kommt. Der Zielkonflikt lässt sich entschärfen, wenn die Anwendung nicht auf das Modell wartet: Was sie selbst weiß, zeigt sie sofort, etwa die Rückmeldung zum Ergebnis oder die erste hinterlegte Hinweisstufe, während die frei formulierte Rückmeldung im Hintergrund entsteht und geprüft wird. Eine ganze Klasse arbeitet gleichzeitig über eine Schulinfrastruktur, die zeitweise nicht verfügbar ist; eine Anwendung muss deshalb auch ohne KI-Anbindung sinnvoll weiterlaufen. Und die laufenden Kosten pro Kind entscheiden darüber, ob ein Konzept über den Modellversuch hinaus Bestand hat. Daraus folgt eine sparsame Grundhaltung: Das Sprachmodell wird an den wenigen Stellen eingesetzt, an denen offene Sprache gebraucht wird, und nicht dort, wo eine hinterlegte Rückmeldung dasselbe leistet.

Prüfen, bevor Kinder damit arbeiten. Ein KI-Lernbegleiter muss systematisch und wiederholt geprüft werden. Sinnvoll ist eine feste Sammlung realistischer Kinderäußerungen und typischer Fehler mit dem jeweils erwarteten Systemverhalten, gegen die nach jeder Änderung an Vorgaben oder Modellversion erneut geprüft wird, denn Modellwechsel verändern das Verhalten oft unbemerkt

und lassen sich nicht immer aufschieben, weil Anbieter einzelne Modellversionen nach einiger Zeit abstellen. Geprüft wird dabei nicht der Wortlaut der Antwort, sondern ob sie bestimmte Eigenschaften hat: keine vollständige Lösung, ein Bezug auf den genannten Rechenweg, eine Rückfrage statt einer Bewertung. Da dasselbe Modell auf dieselbe Eingabe unterschiedlich antwortet, muss jeder Fall mehrfach durchlaufen werden; das Ergebnis ist eine Fehlerquote, keine Ja-nein-Aussage, und sie wird an einer vorab festgelegten Schwelle gemessen. Ein Lernbegleiter, der in einem von zwanzig Fällen die Lösung ausgibt, fällt in einer Klasse mit 25 Kindern täglich auf. Die Prüfliste für Lehrkräfte im Abschnitt oben lässt sich dafür direkt in Prüffälle übersetzen. Dazu gehören eine gezielte Prüfung auf unerwünschte Ausgaben, die Frage, ob das System einem Kind widerspricht, das eine falsche Begründung überzeugt vorträgt, und die Bewertung einer Stichprobe durch fachdidaktisch geschulte Personen. Benchmarks für pädagogische Qualität liegen inzwischen vor (Macina et al., 2025); sie ersetzen die Prüfung am eigenen Anwendungsfall nicht, geben ihr aber einen Rahmen. Die Verlässlichkeit muss für Modell, Version, Aufgabe und Lerngruppe jeweils geprüft werden.

Ethische Anforderungen an KI-Lernbegleitung

Die Gestaltungsprinzipien zielen auf die lernwirksame Ausgestaltung von KI-Lernbegleitung. Lernwirksamkeit allein ist aber kein hinreichendes Kriterium, besonders nicht, wenn Kinder die Lernenden sind. Ethische Fragen stellen sich beim Lernen mit generativer KI unabhängig von der Lernwirksamkeit, und sie gehen über den häufig im Vordergrund stehenden Datenschutz hinaus. Holmes et al. (2022) fordern ein gemeinschaftlich entwickeltes Ethik-Rahmenwerk, das Fairness, Transparenz, den Handlungsspielraum der Lernenden (Agency) und pädagogische Verantwortung umfasst. Die Kultusministerkonferenz (KMK, 2024) empfiehlt für Grund- und Förderschulen einen vorsichtigen, forschungsbasierten Einsatz von KI mit Fokus auf Basiskompetenzen, Inklusion, Chancengerechtigkeit und datenschutzkonforme, altersangemessene Lösungen.

Dass die Forschung hier Lücken aufweist, zeigt eine Übersichtsarbeit zu KI und dem Wohlergehen von Lernenden (Human Flourishing) in der Sekundarstufe: Die Forschungslage ist stark leistungsorientiert, während ethische, metakognitive und lehrkraftbezogene Perspektiven unterbelichtet bleiben (Fock & Siller, 2025). Almheiri et al. (2025) sowie Cárdenas et al. (2025) identifizieren ethische Herausforderungen und Skalierungsprobleme als Haupthindernisse für den breiten Einsatz von KI-Tutorsystemen. Arbeiten zum psychologischen Profiling mit großen

Sprachmodellen (Rosenfelder et al., 2025) zeigen, wie treffgenau Modelle aus Texten Persönlichkeits- und Wertemuster ableiten können. Ein intransparentes System könnte solche Profile ohne Wissen der Kinder und ihrer Eltern erstellen. Gulz und Haake (2021) betonen zudem die Notwendigkeit, Adaptivität mit inklusiver Pädagogik und Barrierefreiheit zu verbinden, ohne Lernende mit besonderen Bedürfnissen zu stigmatisieren. Dazu gehört ein Punkt, der im Klassenzimmer unmittelbar sichtbar wird: Kinder sehen die Bildschirme ihrer Nachbarn. Differenzierung, die als Abstufung erkennbar wird, wirkt anders als Differenzierung, die sich in der Art der Rückfrage verbirgt.

Aus diesen Befunden und Forderungen lassen sich konkrete ethische Anforderungen an einen verantwortlichen KI-Lernbegleiter ableiten. Er muss datensparsam arbeiten und psychologische Profilbildung vermeiden. Das ist anspruchsvoller, als es klingt, denn offene Eingaben von Kindern sind nicht steuerbar: Kinder schreiben und sprechen Namen, Orte und Persönliches in ein Eingabefeld. Datensparsamkeit muss deshalb an der Speicherung ansetzen und daran, wie wenig überhaupt an einen externen Dienst übermittelt wird. Praktisch heißt das: pseudonyme Kennungen statt Namen, dauerhafte Speicherung nur des fachlichen Lernstands und getrennte Datenbestände. Gesprächsverläufe bleiben nur so lange verfügbar, wie die Lehrkraft sie für die Nachbereitung braucht, etwa wenige Tage, und werden danach automatisch gelöscht. Bei externen Diensten braucht es einen Auftragsverarbeitungsvertrag, der die Verarbeitung im europäischen Raum und den Ausschluss der Weiterverwendung für das Training vorsieht. Verarbeitung auf dem Gerät entschärft viele dieser Fragen, ist aber in der Leistungsfähigkeit begrenzt. Die Qualitätsprüfung durch Fachleute braucht Dialoge; sie sollte auf eine ausdrücklich dafür erhobene, pseudonymisierte Stichprobe gestützt werden und nicht auf den laufenden Datenbestand. Für Forschungsprojekte kommt hinzu, dass Gesprächsverläufe personenbezogene Daten von Kindern sind und eine eigene Rechtsgrundlage sowie die Einwilligung der Erziehungsberechtigten erfordern. Rechtlich zu beachten ist außerdem die europäische KI-Verordnung (Europäische Union, 2024): KI-Systeme, die Lernleistungen bewerten oder den Zugang zu Bildung steuern, gelten als Hochrisikosysteme mit entsprechenden Pflichten, und Transparenzpflichten gegenüber den Nutzenden gelten unabhängig davon.

Der Lernbegleiter muss barrierearm sein und multimodale Interaktion (Sprache, Text, Bild, Handlung) für unterschiedliche Lernbedürfnisse nutzen (Hocine et al., 2023), wobei benachteiligte Lernende bei der Gestaltung gezielt in den Blick genommen werden. Seine Funktionsweise muss in Grundzügen erklärbar und

nachvollziehbar sein. Ein Lernendenmodell kann dem Kind auch zugänglich gemacht werden: Offene Lernendenmodelle machen sichtbar, was das System über den Lernstand annimmt, und geben dem Kind die Möglichkeit, dieser Einschätzung zu widersprechen (Bull & Kay, 2016). Untersucht sind sie bislang fast nur mit älteren Lernenden; ob Grundschulkinder eine sichtbare Zustandsanzeige als Hilfe zur Selbsteinschätzung oder als Bewertung erleben, ist offen. Der Lernbegleiter muss Kinder zur kritischen Prüfung von KI-Antworten anregen und dazu beitragen, dass kritisches Denken gestärkt wird statt geschwächt.

Hinzu kommt eine Anforderung, die sich aus dem Alter der Lernenden ergibt. Kinder im Grundschulalter reagieren auf ein geduldiges, freundliches und jederzeit verfügbares Gegenüber mit Beziehungsaufbau. Sie duzen es, erzählen ihm Privates und fragen es, ob es sie mag. Ein KI-Lernbegleiter sollte deshalb weder eine Persönlichkeit noch eine Freundschaft simulieren, seinen Status als Programm auf Nachfrage klar benennen und Gespräche, die den Lerngegenstand verlassen, an Menschen zurückverweisen. Die Beziehung zum Kind ist Aufgabe der Lehrkraft und der Eltern; das System hat sie weder zu ersetzen noch zu imitieren. Der übergeordnete Grundsatz lautet, dass der Lernbegleiter Selbsttätigkeit und Entscheidungsspielräume der Lernenden stärken soll.

Das SKILL-Modell als Orientierungsrahmen

Für das Projekt PRIMA-KI wurde das SKILL-Modell als Ordnungsrahmen für den Einsatz generativer KI beim Lernen entwickelt (Urff, 2026). Seine Grundregel lautet: Je geringer die Kompetenz einer Person, KI in einer konkreten Sache lernwirksam zu nutzen, desto mehr didaktische Lenkung braucht der Einsatz, aus dem Werkzeug oder durch die Lehrkraft. Lenkung wird über drei Hebel beschrieben (Scaffolding, Leitplanken, Kontextualisierung), und drei Stufen des Einsatzes (Eingebettet, Begleitet, Offen) markieren, wie viel davon eine Person braucht. Maßgeblich ist die schwächste von vier Teilkompetenzen: Wissen über die Sache; Fehler in KI-Antworten erkennen; das eigene Vorgehen steuern und den Griff zur Lösung zurückstellen; Fragen formulieren und Antworten verarbeiten. Für die Grundschule bedeutet das in der Regel Stufe 1: eine eng vorstrukturierte, in eine Lernanwendung eingebettete KI, deren Grenzen außerhalb des Dialogs verankert sind, begleitet von der Lehrkraft.

Die in diesem Beitrag entwickelten Gestaltungsprinzipien beschreiben, wie ein Werkzeug auf dieser Stufe gebaut sein muss. Fundierung und Dosierung

entsprechen den Hebeln Kontextualisierung und Scaffolding, die Hilfebegrenzung den Leitplanken, die Einbettung der Begleitung durch die Lehrkraft. Der SKILL-Check des Modells übersetzt diese Anforderungen in fünf Prüffragen für ein konkretes Werkzeug. Das SKILL-Modell ist ein Ordnungsrahmen, kein empirisch geprüftes Wirkmodell; es macht die Annahmen dieses Beitrags planbar, belegt sie aber nicht.

Bilanz und offene Fragen

Die vier Gestaltungsprinzipien lassen sich in einem Satz zusammenfassen: Ein KI-Lernbegleiter für die Grundschule ist dann hilfreich, wenn er weiß, wie Kinder Mathematik lernen, wenn er weniger sagt, als er könnte, wenn er zum Nachdenken und Nachfragen auffordert, statt zu erklären, und wenn eine Lehrkraft sieht und steuert, was er tut. Fehlt eine dieser Bedingungen, ist der Verzicht auf den Einsatz die bessere Entscheidung. Für die Primarstufe bleiben diese Prinzipien forschungsbasierte Gestaltungsannahmen. Die große Mehrheit der Wirkungsbefunde stammt aus Sekundarstufe, Hochschule oder Erwachsenenstichproben; die Meta-Analyse zur Grundschule betrifft regelbasierte Systeme, und die wenigen Grundschulstudien zu dialogischen Systemen sind klein oder explorativ.

Daraus ergeben sich offene Forschungsfragen. Es fehlen Langzeitstudien, die KI-gestützte Lernbegleitung über einzelne Interventionen hinaus untersuchen, insbesondere zur Frage, ob adaptives Scaffolding zu dauerhaftem Kompetenzaufbau führt oder ob sich Nachteile erst mit zeitlicher Verzögerung zeigen. Es fehlen Studien, die die Bedingungen der Primarstufe systematisch berücksichtigen: geringere Lesekompetenz, andere Interaktionsmuster, das Zusammenspiel mit konkretem Material und die Frage, wie lange ein Kind ohne Hilfe suchen sollte. Ungeklärt ist, wie Lehrkräfte KI-Lernbegleitung im Alltag einer heterogenen Grundschulklasse mit begrenzter Infrastruktur integrieren, wer wann welche Aufmerksamkeit erhält und wie hoch die zusätzliche Belastung der Lehrkraft dabei ist. Hinzu kommt ein methodisches Problem, das die Bewertung aller genannten Systeme betrifft: Die Qualität von KI-Lernbegleitern wird derzeit überwiegend an Ersatzgrößen gemessen, an Expertenurteilen über Dialoge, an Benchmark-Ergebnissen oder an Bewertungen durch andere Sprachmodelle. Ob ein System, das nach diesen Maßstäben gut abschneidet, zu Lernzuwachs führt, ist bisher kaum untersucht. Solange dieser Zusammenhang offen ist, sind gute Benchmark-Werte kein Wirksamkeitsnachweis.

Im Projekt PRIMA-KI entstehen auf Basis der hier formulierten Prinzipien appintegrierte KI-Lernbegleitungen, die in Zyklen aus Entwicklung, Erprobung und Überarbeitung (Design-Based Research) erforscht und weiterentwickelt werden. Die Ergebnisse dieser Zyklen werden auf urff.app dokumentiert und fließen in die nächste Fassung dieses Beitrags ein.

Literatur

Abdulsalam, R. O., & Aroyehun, S. (2025). Large language models approach expert pedagogical quality in math tutoring but differ in instructional and linguistic profiles [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.20780

Alemdag, E. (2025). The effect of chatbots on learning: A meta-analysis of empirical research. Journal of Research on Technology in Education, 57(2), 459–481. https://doi.org/10.1080/15391523.2023.2255698

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. International Journal of Artificial Intelligence in Education, 26(1), 205–223. https://doi.org/10.1007/s40593-015-0089-1

Almheiri, A. S. B., Albastaki, H., & Alrashdan, H. (2025). AI-based tutoring systems in education. In B. I. Edwards, H. Abuhassna, D. Olugbade, O. A. Ojo & W. A. Jaafar Wan Yahaya (Hrsg.), Generators, bots, and tutors (S. 185–210). IGI Global. https://doi.org/10.4018/979-8-3373-0847-0.ch007

Attia, A. A., Liu, J., Ai, W., Demszky, D., & Espy-Wilson, C. (2024). Kid-Whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7, 74–80. https://doi.org/10.1609/aies.v7i1.31618

Aufenanger, S., Herzig, B., & Schiefner-Rohs, M. (2023). Künstliche Intelligenz und Schule. Aufgaben für Unterricht und die Organisation (von) Schule. In C. de Witt, C. Gloerfeld & S. E. Wrede (Hrsg.), Künstliche Intelligenz in der Bildung (S. 199–218). Springer Fachmedien. https://doi.org/10.1007/978-3-658-40079-8_10

Awang, L. A., Yusop, F. D., & Danaee, M. (2025). Current practices and future direction of artificial intelligence in mathematics education: A systematic review. International Electronic Journal of Mathematics Education, 20(2), em0823.

https://doi.org/10.29333/iejme/16006

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122 (Korrektur: 122(34), e2518204122. https://doi.org/10.1073/pnas.2518204122)

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough & J. R. Pomerantz (Hrsg.), Psychology and the real world: Essays illustrating fundamental contributions to society (S. 56–64). Worth Publishers.

Bruner, J. S. (1966). Toward a theory of instruction. Belknap Press of Harvard University Press.

Bull, S., & Kay, J. (2016). SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. International Journal of Artificial Intelligence in Education, 26(1), 293–331. https://doi.org/10.1007/s40593-015-0090-8

Canonigo, A. M. (2024). Levering AI to enhance students‘ conceptual understanding and confidence in mathematics. Journal of Computer Assisted Learning, 40(6), 3215–3229. https://doi.org/10.1111/jcal.13065

Cárdenas, R., Vásquez, H. G. E., Gamboa, D. A. P., Arteaga-Arcentales, E., & Carrera, J. E. M. (2025). Exploring AI-powered adaptive learning systems and their implementation in educational settings: A systematic literature review. International Journal of Innovative Research and Scientific Studies, 8(4), 832–842. https://doi.org/10.53894/ijirss.v8i4.7961

Chudziak, J. A., & Kostka, A. (2025). AI-powered math tutoring: Platform for personalized and adaptive education. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos & S. Isotani (Hrsg.), Artificial intelligence in education (Lecture Notes in Computer Science, Bd. 15882, S. 462–469). Springer. https://doi.org/10.1007/978-3-031-98465-5_58

Cohn, C., Rayala, S., Srivastava, N., Fonteles, J., Jain, S., Luo, X., Mereddy, D., Mohammed, N., & Biswas, G. (2026). A theory of adaptive scaffolding for LLM-based

pedagogical agents. Proceedings of the AAAI Conference on Artificial Intelligence, 40(3), 1757–1765. https://doi.org/10.1609/aaai.v40i3.37154

Contel, F., & Cusi, A. (2025). Investigating the role of ChatGPT in supporting metacognitive processes during problem-solving activities. Digital Experiences in Mathematics Education, 11(1), 167–191. https://doi.org/10.1007/s40751-024-00164-7

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4(4), 253–278. https://doi.org/10.1007/BF01099821

Cosentino, G., Anton, J., Sharma, K., Gelsomini, M., Giannakos, M. N., & Abrahamson, D. (2025). Generative AI and multimodal data for educational feedback: Insights from embodied math learning. British Journal of Educational Technology, 56(5), 1686–1709. https://doi.org/10.1111/bjet.13587

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 227, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Dinsmore, D. L., & Fryer, L. K. (2026). What does current genAI actually mean for student learning? Learning and Individual Differences, 125, 102834. https://doi.org/10.1016/j.lindif.2025.102834

Eedi & Google DeepMind. (2025, 11. November). New exploratory research from Eedi and Google DeepMind reveals human-in-the-loop AI tutoring outperforms human-only support [Pressemitteilung]. https://www.eedi.com/news/new-exploratory-research-from-eedi-and-google-deepmind-reveals-human-in-the-loop-ai-tutoring-outperforms-human-only-support

Eti, N., Mosia, M., & Egara, F. O. (2026). The role of AI-driven personalised learning in enhancing mathematics problem-solving skills: A systematic review. Frontiers in Computer Science, 8. https://doi.org/10.3389/fcomp.2026.1813431

Europäische Union. (2024). Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (KI-Verordnung). Amtsblatt der Europäischen

Union, L, 12.7.2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Fock, A., & Siller, H.-S. (2025). Generative artificial intelligence in secondary STEM education in the light of human flourishing: A scoping literature review. International Journal of STEM Education, 12(1), 67. https://doi.org/10.1186/s40594-025-00589-5

Gaidoschik, M. (2010). Wie Kinder rechnen lernen – oder auch nicht: Theoretische und empirische Grundlagen einer entwicklungspsychologisch orientierten Rechendidaktik. Peter Lang. https://doi.org/10.3726/978-3-653-01218-7

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), 6. https://doi.org/10.3390/soc15010006 (Korrektur: 15(9), 252. https://doi.org/10.3390/soc15090252)

Gonnermann-Müller, J., Haase, J., Fackeldey, K., & Pokutta, S. (2025). FACET: Teacher-centred LLM-based multi-agent systems – Towards personalized educational worksheets [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.11401

Guerino, G., Challco, G. C., Veloso, T. E., Oliveira, L., Penha, R. S. D., Melo, R. F., Vieira, T., Marinho, M. L. M., Macario, V., Bittencourt, I. I., Isotani, S., & Dermeval, D. (2023). Teacher-centered intelligent tutoring systems: Design considerations from Brazilian public school teachers. Anais do XXXIV Simpósio Brasileiro de Informática na Educação (SBIE 2023) (S. 1419–1430). Sociedade Brasileira de Computação. https://doi.org/10.5753/sbie.2023.235159

Gulz, A., & Haake, M. (2021). No child left behind, nor singled out: Is it possible to combine adaptive instruction and inclusive pedagogy in early math software? SN Social Sciences, 1, 205. https://doi.org/10.1007/s43545-021-00205-7

Guo, J., Ma, Y., Jang, H., Li, T., Wu, J., Huang, D., Han, F., Noetel, M., Liao, K., Tang, X., & Kui, X. (2025). The impact of artificial intelligence on primary school students‘ motivation and engagement: A systematic review [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ecspn_v1

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Helal, M., Holthaus, P., Wood, L., Velmurugan, V., Lakatos, G., Moros, S., & Amirabdollahian, F. (2024). When the robotic maths tutor is wrong – Can children

identify mistakes generated by ChatGPT? In 2024 5th International Conference on Artificial Intelligence, Robotics and Control (AIRC) (S. 83–90). IEEE. https://doi.org/10.1109/AIRC61399.2024.10672220

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). Educational Psychologist, 42(2), 99–107. https://doi.org/10.1080/00461520701263368

Hocine, N., Moussa, M. B. O., & Ali, S. A. (2023). PosiCalculia: An adaptive virtual environment for children with learning difficulties. In 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA) (S. 1–6). IEEE. https://doi.org/10.1109/inista59065.2023.10310592

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. European Journal of Education, 57(4), 542–570. https://doi.org/10.1111/ejed.12533

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promise and implications for teaching and learning. Center for Curriculum Redesign.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. International Journal of Artificial Intelligence in Education, 32(3), 504–526. https://doi.org/10.1007/s40593-021-00239-1

Holstein, K., McLaren, B. M., & Aleven, V. (2018). Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In C. Penstein Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren & B. du Boulay (Hrsg.), Artificial intelligence in education (Lecture Notes in Computer Science, Bd. 10947, S. 154–168). Springer. https://doi.org/10.1007/978-3-319-93843-1_12

Hwang, S. (2022). Examining the effects of artificial intelligence on elementary students‘ mathematics achievement: A meta-analysis. Sustainability, 14(20), 13185. https://doi.org/10.3390/su142013185

Jančařík, A., Michal, J., & Novotná, J. (2023). Using AI chatbot for math tutoring. Journal of Education Culture and Society, 14(2), 285–296.

https://doi.org/10.15503/jecs2023.2.285.296

Kaliisa, R., Misiejuk, K., López-Pernas, S., & Saqr, M. (2026). How does artificial intelligence compare to human feedback? A meta-analysis of performance, feedback perception, and learning dispositions. Educational Psychology, 46(1), 80–111. https://doi.org/10.1080/01443410.2025.2553639

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. Educational Psychology Review, 19(4), 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educational Psychologist, 51(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458. https://doi.org/10.1038/s41598-025-97652-6

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin, 119(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. Educational Psychology Review, 19(3), 239–264. https://doi.org/10.1007/s10648-007-9049-0

Kortenkamp, U. (2024). Wieviel Mathe braucht der Mensch? Mathematische Kernkompetenzen im Angesicht von KI. In A. S. Steinweg (Hrsg.), Schule im Wandel

– Mathematikunterricht im Wandel: Tagungsband des AK Grundschule in der GDM 2024 (S. 57–72). University of Bamberg Press. https://doi.org/10.20378/irb-104036

Kultusministerkonferenz [KMK]. (2024). Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen [Beschluss vom 10.10.2024]. Kultusministerkonferenz. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf

Kumar, H., Rothschild, D. M., Goldstein, D. G., & Hofman, J. M. (2025). Math education with large language models: Peril or promise? In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos & S. Isotani (Hrsg.), Artificial intelligence in education (Lecture Notes in Computer Science, Bd. 15880, S. 60–75). Springer. https://doi.org/10.1007/978-3-031-98459-4_5

Kuo, B.-C., Bai, Z.-E., & Lin, C.-H. (2026). Developing an AI learning companion for mathematics problem solving in elementary schools. Computers & Education, 240, 105463. https://doi.org/10.1016/j.compedu.2025.105463

Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. npj Science of Learning, 10(1), 29. https://doi.org/10.1038/s41539-025-00320-7

Ligthart, M. E. U., de Droog, S. M., Bossema, M., Elloumi, L., Hoogland, K., Smakman, M. H. J., Hindriks, K. V., & Ben Allouch, S. (2023). Design specifications for a social robot math tutor. In G. Castellano, L. Riek, M. Cakmak & J. Leite (Hrsg.), Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (S. 321–330). ACM/IEEE. https://doi.org/10.1145/3568162.3576957

Listyaningrum, P., Retnawati, H., Harun, H., & Ibda, H. (2024). Digital learning using ChatGPT in elementary school mathematics learning: A systematic literature review. Indonesian Journal of Electrical Engineering and Computer Science, 36(3), 1701–1710. https://doi.org/10.11591/ijeecs.v36.i3.pp1701-1710

Liu, B., Zhang, W., & Wang, F. (2026). Can generative artificial intelligence effectively enhance students‘ mathematics learning outcomes? A meta-analysis of empirical studies from 2023 to 2025. Education Sciences, 16(1), 140. https://doi.org/10.3390/educsci16010140

Liu, J., Sun, D., Sun, J., Wang, J., & Yu, P. L. H. (2025). Designing a generative AI enabled learning environment for mathematics word problem solving in primary schools: Learning performance, attitudes and interaction. Computers and Education: Artificial Intelligence, 9, 100438. https://doi.org/10.1016/j.caeai.2025.100438

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901–918. https://doi.org/10.1037/a0037123

Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., & Sachan, M. (2025). MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (S. 204–221). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.11

Makransky, G., Shiwalia, B. M., Herlau, T., & Blurton, S. (2025). Beyond the “wow” factor: Using generative AI for increasing generative sense-making. Educational Psychology Review, 37(3), 60. https://doi.org/10.1007/s10648-025-10039-x

Mason, J., Burton, L., & Stacey, K. (1982). Thinking mathematically. Addison-Wesley.

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners‘ self-regulated learning. Computers and Education: Artificial Intelligence, 3, 100070. https://doi.org/10.1016/j.caeai.2022.100070

Molenaar, I., & Knoop-van Campen, C. A. N. (2019). How teachers make dashboard information actionable. IEEE Transactions on Learning Technologies, 12(3), 347–355. https://doi.org/10.1109/TLT.2018.2851585

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children’s motivation and performance. Journal of Personality and Social Psychology, 75(1), 33–52. https://doi.org/10.1037/0022-3514.75.1.33

Opesemowo, O. A. G., & Adewuyi, H. O. (2024). A systematic review of artificial intelligence in mathematics education: The emergence of 4IR. Eurasia Journal of Mathematics, Science and Technology Education, 20(7), em2478. https://doi.org/10.29333/ejmste/14762

Padberg, F., & Benz, C. (2021). Didaktik der Arithmetik: Fundiert, vielseitig, praxisnah (5. Aufl.). Springer Spektrum.

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? Educational Psychologist, 40(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1

Qian, K., Liu, S., Li, T., Raković, M., Li, X., Guan, R., Molenaar, I., Nawaz, S., Swiecki, Z., Yan, L., & Gašević, D. (2026). Towards reliable generative AI-driven scaffolding: Reducing hallucinations and enhancing quality in self-regulated learning support. Computers & Education, 240, 105448. https://doi.org/10.1016/j.compedu.2025.105448

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. Cognitive Science, 38(1), 1–37. https://doi.org/10.1111/cogs.12086

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. Educational Psychologist, 38(1), 15–22. https://doi.org/10.1207/S15326985EP3801_3

Rosenfelder, A., Levitin, M. D., & Gilead, M. (2025). Towards social superintelligence? AI infers diverse psychological traits from text without specific training, outperforming human judges. Computers in Human Behavior: Artificial Humans, 6, 100228. https://doi.org/10.1016/j.chbah.2025.100228

Ruan, S., He, J., Ying, R., Burkle, J., Hakim, D., Wang, A., Yin, Y., Zhou, L., Xu, Q., AbuHashem, A. A., Dietz, G., Murnane, E. L., Brunskill, E., & Landay, J. A. (2020). Supporting children’s math learning with feedback-augmented narrative technology. In Proceedings of the Interaction Design and Children Conference (S. 567–580). ACM. https://doi.org/10.1145/3392063.3394400

Rumbelow, M., & Coles, A. (2024). The promise of AI object-recognition in learning mathematics: An explorative study of 6-year-old children’s interactions with Cuisenaire rods and the Blockplay.ai app. Education Sciences, 14(6), 591. https://doi.org/10.3390/educsci14060591

Salden, R. J. C. M., Koedinger, K. R., Renkl, A., Aleven, V., & McLaren, B. M. (2010). Accounting for beneficial effects of worked examples in tutored problem solving. Educational Psychology Review, 22(4), 379–392. https://doi.org/10.1007/s10648-010-9143-6

Samuelsson, J. (2023). Arithmetic fact fluency supported by artificial intelligence. Frontiers in Education Technology (Scholink), 6(1), 13. https://doi.org/10.22158/fet.v6n1p13

Sayed, W. S., Noeman, A., Abdellatif, A., Abdelrazek, M., Badawy, M. G., Hamed, A. E. A., & El-Tantawy, S. (2022). AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging e-learning platform. Multimedia Tools and Applications, 82(3), 3303–3333. https://doi.org/10.1007/s11042-022-13076-8

Schneider, R. J. (o. J.). Der Einsatz von KI zur Unterstützung bei der Unterrichtsvorbereitung: Wie lassen sich KI-generierte Übungsaufgaben für den Mathematikunterricht der Grundschule fachdidaktisch bewerten? [Unveröffentlichtes Manuskript].

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In The Twelfth International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=tvhaxkMKAn

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795

Son, T. (2024). Artificial intelligence in mathematics education: A systematic literature review on intelligent tutoring systems. Journal of Educational Research in Mathematics, 34(2), 187–208. https://doi.org/10.29275/jerm.2024.34.2.187

Song, Y., Kim, J., Liu, Z., Li, C., & Xing, W. (2026). Students‘ perceived roles, opportunities, and challenges of a generative AI-powered teachable agent: A case of middle school math class. Journal of Research on Technology in Education, 58(3), 535–554. https://doi.org/10.1080/15391523.2024.2447727

Ständige Wissenschaftliche Kommission der Kultusministerkonferenz [SWK]. (2024). Large Language Models und ihre Potenziale im Bildungssystem [Impulspapier]. https://doi.org/10.25656/01:28303

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of

intelligent tutoring systems on K–12 students‘ mathematical learning. Journal of Educational Psychology, 105(4), 970–987. https://doi.org/10.1037/a0032447

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students‘ academic learning. Journal of Educational Psychology, 106(2), 331–347. https://doi.org/10.1037/a0034752

Sweller, J. (2020). Cognitive load theory and educational technology. Educational Technology Research and Development, 68(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tabak, I. (2004). Synergy: A complement to emerging patterns of distributed scaffolding. Journal of the Learning Sciences, 13(3), 305–335. https://doi.org/10.1207/s15327809jls1303_3

Urff, C. (2026). Das SKILL-Modell: Wie viel Lenkung braucht Lernen mit KI? urff.app. https://urff.app/skill-modell-ki-unterricht-grundschule/

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. Educational Psychology Review, 22(3), 271–296. https://doi.org/10.1007/s10648-010-9127-6

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner & E. Souberman, Hrsg.). Harvard University Press.

Wang, D., Shan, D., Ju, R., Kao, B., Zhang, C., & Chen, G. (2025). Investigating dialogic interaction in K12 online one-on-one mathematics tutoring using AI and sequence mining techniques. Education and Information Technologies, 30(7), 9215–9240. https://doi.org/10.1007/s10639-024-13195-9

Wang, J., & Fan, W. (2025). The effect of ChatGPT on students‘ learning

performance, learning perception, and higher-order thinking: Insights from a meta-analysis. Humanities and Social Sciences Communications, 12(1), 1–21. https://doi.org/10.1057/s41599-025-04787-y

Wang, L., & Nie, Z. (2023). Research on adaptive learning in K-12 education in the perspective of teachers' artificial intelligence literacy: Development, technology, improvement strategies. In 2023 5th International Conference on Computer Science and Technologies in Education (CSTE) (S. 1–8). IEEE. https://doi.org/10.1109/cste59648.2023.00059

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2410.03017

Wartha, S., & Schulz, A. (2012). Rechenproblemen vorbeugen. Cornelsen Scriptor.

Weidlich, J., Gašević, D., Drachsler, H., & Kirschner, P. A. (2025). ChatGPT in education: An effect in search of a cause. Journal of Computer Assisted Learning, 41(5), e70105. https://doi.org/10.1111/jcal.70105

Wezendonk, A., & Veldhuis, M. (2024). Adaptieve leersystemen en de didactische besluitvorming van basisschoolleerkrachten bij rekenen-wiskunde. Tijdschrift OnderwijsPraktijk Studies. https://doi.org/10.54657/tops.13844

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. British Journal of Educational Technology, 55(1), 10–33. https://doi.org/10.1111/bjet.13334

Wu, X.-Y., Radloff, J., Yeter, I., Wang, L., & Chiu, T. K. F. (2026). Designing artificial intelligence chatbots for self-regulated learning from a systematic review based on Habermas's three interests. Interactive Learning Environments, 34(5), 3141–3164. https://doi.org/10.1080/10494820.2025.2563086

Yim, I. H. Y., & Su, J. (2025). Artificial intelligence literacy education in primary schools: A review. International Journal of Technology and Design Education, 35(5), 2175–2204. https://doi.org/10.1007/s10798-025-09979-w

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023, Datasets and Benchmarks Track). arXiv-Fassung: https://doi.org/10.48550/arXiv.2306.05685

Zheng, L., Niu, J., Zhong, L., & Gyasi, J. F. (2023). The effectiveness of artificial intelligence on learning achievement and learning perception: A meta-analysis. Interactive Learning Environments, 31(9), 5650–5664. https://doi.org/10.1080/10494820.2021.2015693

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Weiterführende Literatur

Die folgenden Titel wurden bei der Erarbeitung des Beitrags herangezogen, werden im Text aber nicht direkt zitiert.

Aru, J., & Laak, K.-J. (2025). Developing an AI-based general personal tutor for education. Trends in Cognitive Sciences, 29(11), 957–960. https://doi.org/10.1016/j.tics.2025.09.010

Bach, K. M., Reinhold, F., & Hofer, S. (2025). Unlocking math potential in students from lower SES backgrounds – using instructional scaffolds to improve performance. npj Science of Learning, 10(1), 66. https://doi.org/10.1038/s41539-025-00358-7

Buchholtz, N., Schorcht, S., Baumanns, L., Huget, J., Noster, N., Rott, B., Siller, H.-S., & Sommerhoff, D. (2024). Damit rechnet niemand! Sechs Leitgedanken zu Implikationen und Forschungsbedarfen zu KI-Technologien im Mathematikunterricht. Mitteilungen der Gesellschaft für Didaktik der Mathematik, 117.

Gisiger, M. (2025, 17. April). Die Rolle von Künstlicher Intelligenz im Lernen – Chancen und Risiken [Blogbeitrag]. https://text.tchncs.de/gisiger/die-rolle-von-kunstlicher-intelligenz-im-lernen-chancen

-und-risiken

Harahap, R. (2024). The role of ChatGPT in enhancing mathematics education: A systematic review. Annals of the Vietnam Academy of Science and Technology, 28(2s), 511–524. https://doi.org/10.52783/anvi.v28.2753

Li, M. (2024). Integrating artificial intelligence in primary mathematics education: Investigating internal and external influences on teacher adoption. International Journal of Science and Mathematics Education. https://doi.org/10.1007/s10763-024-10515-w

Mott, B., Gupta, A., Glazewski, K., Ottenbreit-Leftwich, A., Hmelo-Silver, C., Scribner, A., Lee, S., & Lester, J. (2023). Fostering upper elementary AI education: Iteratively refining a use-modify-create scaffolding progression for AI planning. In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 2 (S. 647). ACM. https://doi.org/10.1145/3587103.3594170

Ninaus, M., & Sailer, M. (2022). Closing the loop – The human role in artificial intelligence for education. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.956798

Topkaya, Y., Doğan, Y., Batdı, V., & Aydın, S. (2025). Artificial intelligence applications in primary education: A quantitatively-supported mixed-meta method study [Preprint]. Preprints. https://doi.org/10.20944/preprints202501.2263.v1

Vitale, A., & Dello Iacono, U. (2024). Using social robots as inclusive educational technology for mathematics learning through storytelling. European Public & Social Innovation Review, 9, 1–17. https://doi.org/10.31637/epsir-2024-672