

Spezialisierte KI-Agenten als Werkzeuge im mathematischen Task Design: Eine Interviewstudie zur Interaktion zwischen Lehrkräften und künstlicher Intelligenz

FRANZISKA PETERS, UNIVERSITÄT GIEßEN & SEBASTIAN SCHORCHT, TU DRESDEN,

Zusammenfassung: Die zunehmende Integration von Künstlicher Intelligenz in Bildungskontexte hat ein breites Forschungsinteresse geweckt. Während KI-Technologien im Bildungssektor bereits vielfältig eingesetzt werden, fehlen explorative Untersuchungen zur partnerschaftlichen Zusammenarbeit zwischen Lehrkräften und KI. Die vorliegende Interviewstudie analysiert, wie Lehrkräfte spezialisierte KI-Agenten im mathematischen Task Design nutzen. Die Ergebnisse verdeutlichen, dass der Einsatz von KI-Agenten ein breites Spektrum an Kompetenzen – von analytischen über technische bis hin zu reflexiven Fähigkeiten – erfordert und fördert. Die Lehrkräfte erkennen in den KI-Ausgaben wertvolle Impulse für Differenzierung und Kompetenzorientierung, setzen diese jedoch erst nach kritischer Reflexion und gezielter Anpassung ein.

Abstract: The increasing integration of Artificial Intelligence into educational contexts has sparked broad research interest. While AI technologies are already used in education, exploratory investigations of collaborative partnerships between teachers and AI remain open. This interview study analyses how teachers utilize specialized AI agents in mathematical task design. The findings demonstrate that using AI agents requires and promotes a wide range of competencies – from analytical and technical to reflective skills. While teachers recognize the valuable potential in AI outputs for differentiation and competency orientation, they implement these only after engaging in critical reflection and targeted adaptation.

1. Einleitung

Die Vielfalt an Technologien mit künstlicher Intelligenz (KI) und deren Anwendungsmöglichkeiten hat in den letzten Jahren stark zugenommen, wobei insbesondere Large Language Modelle (LLMs) als Schlüsseltechnologie hervorzuheben sind (Buchholtz et al., 2023; Kasneci et al., 2023). Eine zentrale Innovation im Bereich der LLMs stellen die Transformer-Modelle dar (Vaswani et al., 2017), deren spezielle Architektur die Verarbeitung längerer Textsequenzen und ihrer kontextuellen Zusammenhänge ermöglicht (Wolfram, 2023; Brown et al., 2020; Floridi & Chiriatti, 2020). Auf dieser Technologie basieren leistungsfähige Sprachmodelle wie

GPT-4o (OpenAI, 2024), Gemini 1.5 (Gemini Team Google, 2024), Deep Seek R1 (DeepSeek-AI, 2025) oder Claude 3.5 (Anthropic, 2024). Diese Modelle wurden mit großen Textdatensätzen trainiert und erzeugen komplexe sprachbasierte Ausgaben, die menschenähnlichen Textformen wie Zusammenfassungen, Übersetzungen oder Dialogen ähneln. Die Entwicklungen gehen aktuell über reine Textverarbeitung hinaus: Zunehmend ist auch die benutzerdefinierte Anpassung der Ausgaben mit KI-Agenten möglich. Dabei werden benutzerdefinierte System-Prompts dem persönlichen Prompting durch den Anwender vorgeschaltet und im Hintergrund vom LLM bei der Textgenerierung berücksichtigt. Die Ausgabe ist damit geprägt vom hinzugefügten System-Prompt, der je nach KI-Agent anders ausgestaltet, also benutzerdefiniert, sein kann (Schorcht et al. 2024c). Bildungsplattformen nutzen diese Möglichkeit, um entsprechende KI-Agenten zu bestimmten Themenfeldern anzubieten. Durch diese eingefügten System-Prompts können LLMs sehr spezifische Aufgaben, wie die Unterrichtsplanung oder das Task Design, unterstützen. Die Einführung von LLMs hat damit signifikante Veränderungen im Bildungsbereich bewirkt (Mollick & Mollick, 2023). Lehrkräfte erhalten durch KI-Technologien eine Unterstützung in ihrem pädagogischen Handeln, wie z. B. bei der Korrektur von Prüfungen oder der Erstellung individueller Lernpläne (Luckin et al., 2016). Adaptive Lerntechnologien ermöglichen individualisiertes Lernen, welches die Fähigkeiten, Interessen und Lernfortschritte der Lernenden berücksichtigt (z. B. Ma et al., 2014).

In den aktuellen Handlungsempfehlungen zum Umgang mit KI in schulischen Bildungsprozessen betont die Kultusministerkonferenz (KMK, 2024) das Potenzial von KI-Technologien zur Unterstützung von Lehrkräften bei der Gestaltung personalisierter Lernumgebungen. Gleichzeitig werden jedoch auch notwendige Rahmenbedingungen adressiert, insbesondere die Professionalisierung von Lehrkräften, regulatorische und rechtliche Aspekte sowie die Sicherstellung von Chancengerechtigkeit. Der Deutsche Ethikrat (2023) weist in diesem Zusammenhang auf potenzielle Risiken hin, die sich aus der umfangreichen Sammlung von Lerndaten ergeben, insbesondere hinsichtlich des Schutzes und möglichen Missbrauchs sensibler Informationen. Gleich-

zeitig müssen Lehrkräfte und Lernende auch neue Fähigkeiten entwickeln und eigene Bedenken berücksichtigen (Luckin et al., 2016). Für Vertrauen, Akzeptanz und dann auch den Einsatz von KI-Technologien ist zum einen Transparenz notwendig und zum anderen ein (daraus auch resultierendes) Verständnis für die zugrunde liegenden Entscheidungs- und Berechnungsprozesse der Technologie (Franzoni, 2023). Integration von KI-Technologien im Bildungssystem eröffnet also gleichermaßen Chancen und spezifische Herausforderungen. Sie stellen einerseits leistungsfähige neue Werkzeuge für Lehr- und Lernprozesse bereit, werfen aber andererseits auch wichtige Fragen zur Integration und zum angemessenen Einsatz im Bildungskontext auf. Während sie folglich kreative Möglichkeiten für personalisiertes Lernen, Feedback und Unterrichtsvorbereitung bieten, erfordern sie auch eine kritische Reflexion darüber, wie bewährte pädagogische Praktiken sinnvoll mit diesen neuen Technologien kombiniert werden können. Eine besondere Herausforderung im Zusammenhang mit LLMs stellt dabei die Tendenz zur Halluzination dar, also der Extrapolation und Generierung von fehlerhaften oder erfundenen Informationen (Pal et al., 2023; Zhang et al., 2023). LLMs erzeugen Texte auf Basis statistischer Wahrscheinlichkeiten, die aus großen Mengen von Trainingsdaten abgeleitet wurden. Diese Eigenschaft erlaubt es ihnen, auch kreative oder scheinbar wissensbasierte Antworten zu formulieren. Allerdings verfügen sie nicht über ein echtes Verständnis oder deklaratives Wissen im menschlichen Sinne. Deshalb können sie bei Anfragen mit objektiv richtigen Antworten auch Inhalte generieren, die zwar sprachlich plausibel erscheinen, jedoch faktisch unzutreffend oder irreführend sind. Diese Tendenz zur Halluzination macht eine sorgfältige Überprüfung der KI-Ausgaben zwingend erforderlich. Denn liegen zu einem Thema nur unzureichende oder verzerrte Trainingsdaten vor, extrapoliert das Modell fehlende Informationen und „füllt Lücken“ – ohne eine Möglichkeit der eigenständigen Validierung. Für den schulischen Kontext bedeutet dies, dass Lehrkräfte Outputs erhalten können, die zwar sprachlich korrekt und überzeugend erscheinen, inhaltlich aber fehlerhaft sind.

Während im Managementbereich bereits intensive Forschung zur Interaktion zwischen menschlichen Entscheidungsträgern und KI-Technologien betrieben wird (Hauptmann et al., 2023), und auch in spezifischen Anwendungsfeldern wie Gaming und Luftfahrt entsprechende Studien vorliegen (Liu, Lai

& Tan, 2021; Zhang et al., 2021), besteht im Bildungssektor noch Forschungsbedarf, wenngleich generative KI auch hier in den Fokus rückt. Drei zentrale Perspektiven haben sich im Forschungsbereich herausgebildet: die fachlichen und didaktischen Fähigkeiten von KI (z. B. Schorcht et al., 2024; Sjuts, 2024), ihr Einsatz als Unterstützung für Lehrkräfte bei Planung und Reflexion (z. B. Buchholtz & Huget, 2024; Noster et al., 2024), sowie als Lernhilfe für Schüler:innen (z. B. Dilling et al., 2024). Weitere Forschung zu diesen Perspektiven ist besonders relevant, da KI-Technologien im Bildungsbereich nicht nur verfügbar sind, sondern ihr reflektierter Einsatz auch bildungspolitisch gefordert wird (Kultusministerkonferenz, 2024). Eine explorative Untersuchung der spezifischen Chancen und Herausforderungen der Mensch-KI-Interaktion im Bildungskontext erscheint daher dringend geboten.

In der vorliegenden Interviewstudie wurde deshalb das mathematische Task Design als eine Tätigkeit von Lehrkräften in einer Ko-Konstruktion mit spezialisierten KI-Agenten untersucht. Die Entwicklung mathematischer Aufgaben stellt eine komplexe Anforderung im Unterrichtsalltag dar. Lehrkräfte müssen nicht nur fachlich fundierte, sondern auch differenzierende und curricular anschlussfähige Aufgaben für den eigenen Unterricht entwerfen. Der Bedarf an Unterstützung beim Task Design ist daher sowohl inhaltlich als auch praxisbezogen gegeben. Erfolgreiches Task Design erfordert nicht nur individuelle Kompetenzen, wie flexible Herangehensweisen (Sherin, 2002) und strategische Ressourcennutzung (Choppin, 2011), sondern auch kollaborative Elemente. Die Zusammenarbeit mit Kolleginnen und Kollegen und das Einbringen verschiedener fachlicher Perspektiven (Sensevy et al., 2013) trägt zu einer vertieften inhaltlichen Auseinandersetzung und verstärkten Reflexion der eigenen Praxis bei (Choppin, 2011). Im Bereich des KI-gestützten mathematischen Task Designs sollte daher die Bedeutung der Kollaboration sowie die damit einhergehende Mensch-KI-Interaktion berücksichtigt werden.

Im vorliegenden Beitrag wird zunächst die theoretische Rahmung erörtert, beginnend mit einer Einführung in das mathematische Task Design (Abschnitt 2.1). Anschließend wird das Konzept der KI-Agenten als Werkzeuge in KI-unterstützten kollaborativen Teams in Abschnitt 2.2 näher beleuchtet, mit besonderem Fokus darauf, wie diese Kollaboration zur Entwicklung und Modifikation von mathematischen Aufgaben genutzt werden kann. Dieses Kapitel legt die Basis für die im weiteren Verlauf

vorgestellte empirische Untersuchung, deren Forschungsfokus in Abschnitt 2.3 entwickelt wird. Anschließend wird in der Anlage der Studie (Abschnitt 3) das Forschungsdesign der qualitativen Interviewstudie detailliert beschrieben, einschließlich der Auswahl der Teilnehmerinnen, des Interviewsettings, sowie der Erhebungs- und Auswertungsmethoden. Das darauffolgende Kapitel (Abschnitt 4) widmet sich den Erkenntnissen aus der Interaktion zwischen Lehrkräften und KI-Agenten. In der Diskussion (Abschnitt 5) werden diese Ergebnisse in den Kontext bestehender Forschung gestellt, Potenziale und Herausforderungen der Nutzung von KI-Agenten aufgezeigt und die daraus resultierenden Implikationen für die Praxis erörtert. Abschließend werden in einem Fazit (Abschnitt 6) die zentralen Erkenntnisse zusammengefasst und ein Ausblick auf zukünftige Forschungsperspektiven gegeben.

2. Theoretischer Rahmen

2.1 Mathematisches Task Design

In der Mathematik sind Aufgaben zentraler Bestandteil für die Planung, Durchführung und Reflexion im Unterricht (Watson & Ohtani, 2015). Aufgaben eröffnen Lernmöglichkeiten und können differenzierten Unterricht unterstützen (Bohl et al., 2015). Während im letzten Jahrhundert Aufgaben häufig von Fachleuten (z. B. Lehrbuchautor:innen) entwickelt wurden (Task Design als eigenständiger Prozess), betont die aktuelle Forschung die Bedeutung von Lehrkräften im Prozess des mathematischen Task Design (Jones & Pepin, 2016).

Eine wichtige Voraussetzung für ein erfolgreiches Task Design ist das fundierte fachliche und didaktische Wissen der Lehrkräfte. Einerseits müssen sie mathematische Inhalte sicher beherrschen, um geeignete Aufgaben erstellen oder auswählen zu können; andererseits sind diagnostische Fähigkeiten gefragt, um den individuellen Wissens- und Kompetenzstand der Lernenden im Task Design einzubeziehen (Kleinknecht & Lankes, 2012). Aufgaben sollten ein angemessenes Maß an kognitiver Komplexität aufweisen, um die Lernenden weder zu über- noch zu unterfordern (Maier et al., 2014). Besonders in heterogenen Lerngruppen ist die sorgfältige Auswahl und Anpassung von Aufgaben entscheidend, um Lernziele zu erreichen und den unterschiedlichen Bedürfnissen gerecht zu werden (Maier et al., 2014; Bohl et al., 2015). Ein besonderes Augenmerk liegt deshalb auf der Gestaltung von Aufgaben, die sowohl an den curricularen Vorgaben

ausgerichtet sind als auch Möglichkeiten zur Differenzierung und Individualisierung bieten. Gut gestaltete, vielseitige Aufgaben können alle Lernenden ansprechen und fördern (Bohl et al., 2015).

Die Forschung beschäftigt sich intensiv mit den komplexen Prozessen, die die Auswahl und Umsetzung mathematischer Aufgaben durch Lehrkräfte betreffen. Dabei werden neben den schon erwähnten kognitiven Anforderungen auch Aspekte wie Überzeugungen der Lehrkräfte, Lehrmaterialien und die Dynamik im Klassenzimmer berücksichtigt. Smith und Stein (1998) heben die zentrale Bedeutung sorgfältig konzipierter Aufgaben für mathematische Lernprozesse hervor. Choppin (2011) analysierte die Entwicklung der professionellen Kompetenz von Lehrkräften im Hinblick auf die Implementation kognitiv aktivierender Aufgaben im Mathematikunterricht. Er stellte fest, dass erfolgreiche Lehrkräfte die Lehrpläne an die Bedürfnisse ihrer Schülerinnen und Schüler anpassen und dabei etwa das Denken der Lernenden berücksichtigen, eigene Beobachtungen im Unterricht mit den Lehrplänen verbinden und empirisch abgeleitetes Wissen entwickeln. Dazu ist ein Verständnis für die durch die Materialien gebotenen Lernmöglichkeiten notwendig, was Drake und Sherin (2009) als „Curriculum Vision“ (Lehrplankompetenz) bezeichnen. Die wachsende Expertise im Umgang mit curricularen Anforderungen führt dazu, dass Lehrkräfte ihre Strategien zur Materialauswahl und -adaptation weiterentwickeln, wodurch sich ihre professionelle Kompetenz stetig erhöht. Sie entwickeln dabei, was Brown (2009) als „Pedagogical Design Capacity“ bezeichnet. Brown's Konzept beschreibt die Fähigkeit einer Lehrkraft, ihre didaktischen Kompetenzen und verfügbaren Ressourcen zielgerichtet für die Unterrichtsgestaltung zu nutzen. Pedagogical Design Capacity wird dabei maßgeblich durch die professionellen Überzeugungen der Lehrkraft sowie ihr fachdidaktisches Wissen geprägt.

Die Untersuchungen von Sherin (2002) verdeutlichen darüber hinaus, dass der Unterricht selbst als Lernumgebung für Lehrkräfte fungiert, indem sie dort ihr Fachwissen mit curricularen Anforderungen und ihrer Wahrnehmung von Lernprozessen der Schülerinnen und Schüler in Beziehung setzen und dadurch neue Erkenntnisse gewinnen. Dabei modifizieren sie ihr bestehendes Wissen und entwickeln neue Strategien. Die Fähigkeit zur adaptiven Unterrichtsgestaltung fördert nicht nur eine flexible und situationsadäquate Vorgehensweise, sondern unterstützt nach Sherin (2002) auch maßgeblich die professionelle Entwicklung der Lehrkräfte. Auch das

Anpassen bestehender Aufgaben trägt zur Entwicklung des mathematischen und didaktischen Wissens von Lehrkräften bei (Pepin, 2015). Es ist also wünschenswert, dass Lehrkräfte im Task Design mitwirken und Aufgaben modifizieren, da sie dabei ihr eigenes Wissen vertiefen und erweitern können.

2.2 KI-Agenten und Lehrkräfte im KI-unterstützten kollaborativen Team

Aufgaben sollen sowohl fachlich korrekt als auch differenziert, sprachsensibel und curricular angebunden sein. Gleichzeitig müssen sie die Heterogenität der Lernenden berücksichtigen und zur aktiven Auseinandersetzung mit dem Lerngegenstand anregen. Doyle (1983) betont, dass Aufgaben die Lernenden durch das Lenken der Aufmerksamkeit auf bestimmte Aspekte beeinflussen sowie durch Hinweise auf die Verarbeitung von Informationen vorgeben. Diese Komplexität macht deutlich, dass die Gestaltung solcher Aufgaben nicht nur Expertise in einem Bereich erfordert, sondern vom Zusammenspiel verschiedener Perspektiven profitiert. An dieser Stelle können KI-unterstützte kollaborative Teams, in diesem Beitrag in Form von spezialisierten KI-Agenten, als potenzielle Unterstützungssysteme angesehen werden. In der Forschung zum professionellen Umgang mit Unterrichtsmaterialien wird betont, dass gerade kollegiale Kooperation zur Reflexion und Qualitätsverbesserung von Aufgaben beiträgt (Choppin, 2011; Sherin, 2002). Hier setzt das Konzept KI-unterstützter kollaborativer Teams an: In solchen Teams bringen Akteure mit unterschiedlichen fachlichen Hintergründen und Verantwortlichkeiten ihre Expertise ein, um gemeinsam an Unterrichtsqualität zu arbeiten (siehe bspw. Widmer-Wolf, 2018). Beispiele hierfür sind Sonderpädagogen, Lehrkräfte für Zweitsprachen sowie Sprachtherapeuten, die im schulischen Alltag konstruktiv mit Lehrkräften zusammenarbeiten und so an den Bildungszielen mitwirken. Dies birgt Potenzial für inklusive Arbeit und Schulentwicklung an Schulen, um den individuellen Anforderungen der Lernenden gerecht zu werden (Werning & Löser, 2010). Die Aufgabenverteilung beruht auf Aushandlungsprozessen, die von den Überzeugungen und Einstellungen der beteiligten Akteure, ihrer jeweiligen beruflichen Sozialisation und Expertise sowie den Strukturen und Praktiken innerhalb der Bildungseinrichtung beeinflusst werden können (Widmer-Wolf, 2018). Da in der Praxis oft weitgehend autonome, unabhängige und voneinander entkoppelte Untereinheiten agieren und die verschiedenen Akteure kaum miteinander in Kontakt

treten (Meyer & Rowan, 1992), sollten Kooperationen, beispielsweise in KI-unterstützten kollaborativen Teams, stärker gefördert und aktiv unterstützt werden.

Das hier vorgestellte Konzept des kollaborativen mathematischen Task Designs integriert empirisch bewährte Ansätze wie das ‚cooperative engineering‘ (Sensevy et al., 2013). Dieser Ansatz basiert auf der systematischen Zusammenarbeit von Lehrkräften und Forschenden, wobei die unterschiedlichen Expertisen der Beteiligten den gemeinsamen Entwicklungsprozess bereichern. Aufgaben werden vorgestellt, konstruktive Hinweise eingebracht, neue Ideen entwickelt und die Aufgaben entsprechend modifiziert. Dieser iterative Prozess aus Entwerfen, Analysieren und Anpassen ermöglicht die Entwicklung von Aufgaben, die auf die Bedürfnisse heterogener Lernenden abgestimmt sind.

Im Hinblick auf den Einsatz von KI-gestützter Werkzeuge beim mathematischen Task Design eröffnet sich die zentrale Forschungsfrage, inwiefern die Interaktion mit spezialisierten KI-Agenten ähnliche Wirkungen entfalten kann. Frühere Studien beforschten bereits den Einsatz von KI-Agenten im Bildungsbereich mit Zielsetzungen wie der Analyse des Verhaltens der Lernenden, dem Vorschlagen geeigneter Kurse und der Verbesserung von Lehrmethoden (Kumar, 2021). Beim Task Design bietet sich ein Ansatz an, bei dem KI-Agenten und Lehrkräfte nicht getrennt voneinander, sondern als KI-unterstütztes kollaboratives Team eng zusammenarbeiten. Die KI-Agenten sind in dem vorliegenden Forschungsansatz nicht als Ersatz für Lehrkräfte zu betrachten, sondern als Werkzeuge in der Zusammenarbeit. Sie bringen verschiedene Blickwinkel ein, wie die Anpassung des Differenzierungsniveaus, der Sprachsensibilität in der Aufgabenstellung oder der Ausrichtung an curriculare Anforderungen.

Küchemann et al. (2023) beforschten bereits, wie angehende Lehrkräfte mit wenig Erfahrung Physikaufgaben mit Hilfe von ChatGPT (GPT-3.5) im Vergleich zur Gestaltung mit Hilfe von Lehrbüchern verändern. Beide Gruppen passten die Aufgaben gut an das geforderte Lernniveau an, jedoch waren Lehrbuchaufgaben klarer formuliert. Auffällig ist, dass von 50 mit ChatGPT erstellten Aufgaben 38 unverändert übernommen wurden, obwohl die Qualität des Outputs als durchschnittlich bewertet wurde. Flexibles KI-gestütztes Task Design im KI-unterstützten kollaborativen Team, bestehend aus Lehrkräften und KI-Agenten, setzt daher voraus,

dass Lehrkräfte den Output der KI-Agenten reflektieren und ggf. anpassen können, wie Küchemann et al. (2023) demonstrierte. Die Anpassung von Outputs kann durch bestimmte Prompt-Techniken der Anwender gesteuert werden. Schorcht et al. (2024b) konnten im Kontext mathematischer Problemlöseaufgaben nachweisen, dass bestimmte Prompt-Techniken wie das Chain-of-Thought-Prompting (Liu et al., 2021; Wei et al., 2023) zwar zu einer detaillierteren Darstellung von Lösungsstrategien durch das LLM führen, die inhaltliche Qualität der mathematischen Lösungen jedoch nicht signifikant verbessert wird.

In Schorcht et al. (2024c) und auch Peters und Schorcht (2024) wurden vier spezialisierte KI-Agenten entwickelt, die als Multi-Agenten-System kollaborativ zur Entwicklung mathematischer Aufgaben eingesetzt wurden. Die Konfigurationen der KI-Agenten integrieren dabei etablierte Prompt-Techniken bereits als System-Prompts. Diese müssen dadurch nicht bei jeder Interaktion neu formuliert werden. Das dadurch entwickelte Multi-Agenten-System arbeitet als KI-unterstütztes kollaboratives Team mit vier spezialisierten KI-Agenten, die aus unterschiedlichen Perspektiven mathematische Aufgaben in einem iterativen, auf Konsens ausgerichteten Prozess analysieren und modifizieren (Schorcht et al., 2024c; Peters & Schorcht, 2024). Die KI-Agenten analysieren Aufgabenstellungen auf Grundlage der aus einer PDF-Datei extrahierten, fachdidaktisch fundierten Informationen. Dabei nutzen sie diese Inhalte zur kontextgesteuerten Textgenerierung, ohne jedoch über eine dauerhafte oder eigenständige Wissensbasis zu verfügen. Zusätzlich liefern sie transparent nachvollziehbare Begründungen für ihre didaktischen Empfehlungen. In den Studien wurde angenommen, dass Lehrkräfte hinsichtlich der Anwendung fachdidaktischen Wissens unterstützt werden können. Die KI-modifizierten Aufgaben wurden anschließend von Lehrkräften im Schuldienst als Experten evaluiert. Die Ergebnisse zeigten, dass ein ausgewogenes Verhältnis zwischen umfangreicher Textgenerierung und präziser Aufgabenformulierung entscheidend ist, was im Output der KI-Agenten nicht immer erreicht wurde. Gleichzeitig konnte das Multi-Agenten-System die Aufgaben durch mögliche Lösungsansätze und spezifische Handlungsaufforderungen bereichern.

2.3 Zusammenfassung und Forschungsfragen

Die differenzierte Analyse der Interaktion zwischen Lehrkräften und KI-Agenten ist grundlegend für die

empirische Fundierung zukünftiger Forschung zum KI-gestützten mathematischen Task Design. Dabei gilt es, den Prozess der Reflexion und Anpassung von KI-generierten Aufgaben nicht nur zu beschreiben, sondern auch die zugrunde liegenden Kompetenzen der Lehrkräfte und potenziellen Hindernisse zu identifizieren. In der bisherigen Forschung liegen Erkenntnisse über die Qualitäten und Mängel der Verwendung von LLMs im Task Design vor (Küchemann et al., 2023), doch spezifische Analysen der Kompetenzen, die Lehrkräfte im Umgang mit solchen Technologien entwickeln oder benötigen, sind bislang selten. Insbesondere ist unklar, inwiefern die Nutzung KI-generierter Ausgaben dazu beitragen kann, Aufgaben zu entwickeln und wie die Kompetenzen von Lehrkräften durch diesen Prozess beeinflusst werden. Dementsprechend sind in dieser Studie folgende Forschungsfrage und die dazugehörigen Unterfragen leitend:

Wie gestaltet sich die Interaktion zwischen Lehrkräften und den verschiedenen KI-Agenten als KI-unterstütztes kollaboratives Team im mathematischen Task Design?

- F1 Welche Kompetenzen zeigen Lehrkräfte beim Nutzen KI-generierter Outputs im Prozess des Task Designs?
- F2 Welche Schwierigkeiten treten bei der Nutzung der KI-Agenten auf und wie werden sie von den Lehrkräften adressiert?
- F3 Welche Arten von Prompt-Techniken verwenden Lehrkräfte in ihrer Interaktion mit großen Sprachmodellen im mathematischen Task Design?

Zur Beantwortung dieser Forschungsfragen wurde eine Interviewstudie mit fünf Lehrkräften durchgeführt, die vier speziell trainierte KI-Agenten zum mathematischen Task Design einsetzten. Aufgrund des explorativen Charakters der Studie wurde ein induktives Vorgehen gewählt und während des gesamten Forschungsprozesses eine Offenheit gegenüber neuen Erkenntnissen bewahrt, die sich über bekannte Kategorien hinaus ergeben könnten.

3. Anlage der Studie

Da bislang weder qualitative noch quantitative Erkenntnisse zum Einsatz von KI-Agenten im mathematischen Task Design vorliegen, ist eine a-priori-Hypothesenbildung im Kontext des vorliegenden Untersuchungsgegenstandes nicht zweckmäßig. Daher erfordert die Untersuchung einen explorativen, hypothesengenerierenden Ansatz, der ein

qualitatives Vorgehen rechtfertigt. Hypothesen können erst durch die Analyse der empirischen Daten entwickelt werden (Flick, 2016). Zur Bewertung einer bisher unerforschten Maßnahme – wie dem Einsatz von KI-Agenten im mathematischen Task Design – empfiehlt sich der Einsatz „qualitativer Fallstudien, an denen sich die Wirksamkeit einer Methode besser erkennen lasse“ (Bortz & Döring, 2006, S. 110). Aus diesem Grund wird in der vorliegenden Studie das neue Forschungsfeld explorativ beschrieben und somit eine deskriptive Bestandsaufnahme erstellt. Aufbauend auf dieser Bestandsaufnahme kann dann an einzelnen empirischen Daten konkreter nachgezeichnet werden, wie der Einsatz der KI-Agenten durch die Lehrkräfte realisiert wird und welchen Beitrag dieser Einsatz innerhalb der Lehrkraft-KI-Interaktion leistet.

3.1 Datenerhebung

Um die Lehrkraft-KI-Interaktion qualitativ zu untersuchen, wurden im Herbst 2024 Interviews mit fünf ausgebildeten Grundschullehrkräften geführt, die die KI-Agenten aus Schorcht et al. (2024c) zur Unterstützung des mathematischen Task Design nutzen sollten. Um die Kollaboration nicht nur zu ermöglichen, sondern auch besonders ertragreich zu machen, wurde das LLM instruiert Self-Explanations (Huang et al., 2023) zu generieren. Es sollten beispielsweise verschiedene Schritte der Aufgabenanalyse und -modifikation befolgt und jeweils dargestellt sowie Feedback zur Aufgabe gegeben werden. Indem die KI-Agenten ihren Output mit Erklärungen und Begründungen versehen, kann bei den Nutzenden ein Gefühl von Transparenz entstehen (Huang et al., 2023), auch wenn unklar bleibt, inwieweit diese Erklärungen die tatsächlichen Entscheidungsprozesse abbilden. Die empfundene Transparenz geht mit einer wahrgenommenen Nachvollziehbarkeit der Entscheidungen einher und kann zur Vertrauensbildung sowie Akzeptanz auf Seiten der Lehrkräfte beitragen (Franzoni, 2023). Durch diese Kooperation können die Stärken der KI-Agenten – wie schnelle Analysen und kreative Vorschläge – mit dem didaktischen Wissen und den Erfahrungswerten der Lehrkräfte kombiniert werden.

Die Untersuchungsgruppe umfasste fünf Lehrkräfte mit verschiedenen beruflichen Hintergründen. Mittels vorab durchgeführter Fragebögen wurden ihre professionellen und KI-bezogenen Erfahrungen erfasst, um ihre Einschätzungen zum KI-gestützten mathematischen Task Design im Kontext ihrer individuellen Ausgangsbedingungen analysieren zu

können. Die gewonnenen Vorabinformationen erlauben es, den Einfluss verschiedener beruflicher Hintergründe auf die KI-Nutzung zu analysieren und Empfehlungen für die Kompetenzförderung abzuleiten. Die unterschiedlichen Hintergründe der Lehrkräfte ermöglichen zudem eine breite Bestandsaufnahme möglicher Nutzungsweisen der KI-Agenten. Die teilnehmenden Lehrkräfte weisen folgende Charakteristika auf:

Die Teilnehmenden zeigen eine ausgewogene Verteilung hinsichtlich ihrer Tätigkeitsfelder. Vier Lehrkräfte (Lehrkraft 1 bis 4) unterrichten an Grundschulen, eine (Lehrkraft 5) an einer Förderschule, alle in den Klassenstufen 1 bis 4. Mit Berufserfahrungen zwischen 5 und 11 Jahren nehmen sie verschiedene zusätzliche Funktionen wie Fachleitung, IT- und Medienbeauftragte, Personalvertretung und Mentoring wahr, einige sind auch universitär tätig. Die KI-Erfahrung der Gruppe variiert deutlich. Während zwei Teilnehmende noch nie bewusst ein großes Sprachmodell verwendet haben, nutzen drei Probanden LLMs bereits für Übersetzungen, Fragebögen und administrative Aufgaben – sowohl privat als auch im pädagogischen Kontext. Diese heterogene Zusammensetzung ermöglicht vielfältige Perspektiven auf das KI-gestützte mathematische Task Design.

Im Interviewsetting arbeiteten die Lehrkräfte an Computerstationen, die ihnen parallel den Zugriff auf vier verschiedene KI-Agenten in getrennten Fenstern ermöglichten. Dabei handelte es sich um GPT-Modelle, die über unterschiedliche System-Prompts jeweils spezifische Expertisen innerhalb eines KI-unterstützten kollaborativen Teams repräsentierten; eine ausführliche Beschreibung findet sich in Schorcht et al. (2024).

- Der inhaltsorientierte KI-Agent überprüft den mathematischen Kern der Aufgabe.
- Der kompetenzorientierte KI-Agent unterstützt Lehrkräfte dabei, mathematische Kompetenzen zu identifizieren und die Aufgabe dahingehend weiterzuentwickeln. Der KI-Agent orientiert sich an den fünf inhaltsbezogenen Kompetenzen gemäß UNESCO (2020) sowie an mathematikspezifischen Tätigkeiten bzw. prozessbezogenen Kompetenzen (hier der Prozessstandards der NCTM (o. D.)).
- Der differenzierungsorientierte KI-Agent untersucht Aufgaben in Hinblick auf natürliche Differenzierung (Wittmann, 2010) und berücksichtigt die sechs Merkmale natürlich differenzier-

ter Lerngelegenheiten und Aufgaben: Offenheit, Komplexität, niedrige Einstiegsschwelle, „Rampen“ für leistungsstarke Lernende, Diskussionsbedarf und ein hohes Potenzial zur kognitiven Aktivierung (Krauthausen & Scherer, 2022).

- Der sprachensible KI-Agent untersucht Aufgaben auf sprachliche Barrieren (Abshagen, 2015) und gibt Vorschläge zur sprachlichen Vereinfachung oder Unterstützung.

Die Lehrkräfte erhielten im Setting klare Anweisungen zur Verwendung der KI-Agenten in der Umgebung ChatGPT (Version GPT-4) sowie zum Arbeitsauftrag der KI-gestützten Aufgabenmodifikation und wurden dazu angehalten, die generierten Outputs im Sinne der introspektiven Erhebungsmethode des Think-Aloud-Protocol (Sasaki, 2008; Konrad, 2010) zu reflektieren. Sie wurden außerdem dazu angeregt, den generierten Output bei Bedarf anzupassen und Feedback zu Qualität und Nützlichkeit zu geben (Tabelle 1).

Einstieg	Beschreiben Sie diese Klasse und warum Sie die Aufgabe für diese Lerngruppe ausgewählt haben.
	Welche KI-Agenten könnten bei der Anpassung der Aufgabe an die Bedürfnisse Ihrer Lerngruppe hilfreich sein?
Task Design Prozess	Wählen Sie einen der Chatbots aus, um mit der Modifizierung der Aufgabe zu beginnen.
	Fügen Sie die Aufgabe in das Chat-Feld ein.
	Überprüfen und kommentieren Sie dann die Ausgabe.
Reflexion	Arbeiten Sie gerne mit einem oder mehreren KI-Agenten weiter.
	Überprüfen und kommentieren Sie das Endergebnis.
	Reflektieren Sie die KI-gestützte Aufgabenentwicklung und Ihre eigene Arbeitsweise.

Tab. 1: Kurzfassung des Interviewleitfadens

Nach detaillierter Erklärung des Arbeitsauftrags zur KI-gestützten Aufgabenmodifikation wurden die Lehrkräfte gebeten, eine Aufgabe mit Bezug zu einer ihrer aktuellen oder kürzlich unterrichteten Klassen auszuwählen und diese Wahl zu begründen. Dies berücksichtigt, dass Lehrkräfte mathematisches Task Design stets – bewusst oder unbewusst – im Kontext konkreter Anwendungsfälle vorneh-

men. Aus einem Pool von sechs verfügbaren Aufgaben (a bis f) konnten die Lehrkräfte dafür eine Aufgabe für die weitere Bearbeitung selbstständig auswählen. Drei von diesen Aufgaben entsprachen den einfachen Aufgaben für Grund- und Sekundarstufe aus Schorcht et al. (2024c), die andere Hälfte beinhaltete die entsprechenden anspruchsvolleren Problemlöseaufgaben. Die ersten drei Aufgaben deckten die mathematischen Bereiche Arithmetik, Muster und Strukturen sowie rationale Zahlen ab. Die drei Problemlöseaufgaben befassten sich mit dem Lösen linearer Gleichungssysteme:

- a) $12 + 3$
- b) $1 + 3 + 5 + 7 + 9 + \dots =$
- c) $1/2 \cdot 1/4$
- d) Wie kann man genau sechs Liter Wasser aus dem Fluss holen, wenn man nur zwei Gefäße zum Abmessen hat, einen Vier-Quart-Eimer und einen Neun-Quart-Eimer? (Polya, 1957, S. 226)
- e) Ein Auto fährt mit einer Geschwindigkeit von 10 km/h. Vier Stunden später fährt ein zweites Auto los, um das erste Auto zu überholen. Es benutzt dieselbe Strecke und fährt 30 km/h. In wie vielen Stunden wird das zweite Auto das erste Auto überholen? (Cooper & Sweller, 1987, S. 361)
- f) Ein Mann betrat einen Obstgarten durch 7 Tore und nahm dort eine bestimmte Anzahl von Äpfeln mit. Als er den Obstgarten verließ, gab er dem ersten Wächter die Hälfte der Äpfel, die er hatte, und 1 Apfel mehr. Dem zweiten Wächter gab er die Hälfte seiner verbliebenen Äpfel und 1 Apfel mehr. Dasselbe tat er bei den übrigen fünf Wächtern und verließ den Obstgarten mit 1 Apfel. Wie viele Äpfel hat er in der Obstplantage gesammelt? (Leonardo de Pisa, 1202)

Anschließend sollten die Teilnehmenden die Nützlichkeit der KI-Agenten erproben und einen der Agenten als Start-Tool auswählen. Die aufgezeichneten Daten beinhalten zum einen die Bildschirmaufnahmen des gesamten Arbeitsprozesses der Lehrkräfte sowie Tonaufnahmen ihrer begleitenden Äußerungen. Außerdem wurden die Chatverläufe exportiert und Notizen der Lehrkräfte gesichert. Für die ausführliche Analyse erfolgte die Datenfixierung in Form von erweiterten Transkripten. Dazu wurde zunächst das Audiomaterial transkribiert, sodass das gesamte Interview in schriftlicher Form vorlag. Anschließend wurden die Transkripte durch Hinzu-

nahme der jeweiligen Auszüge aus den exportierten Chatverläufen angereichert, sodass jederzeit verständlich ist, auf welche KI-Ausgabe sich die entsprechende Lehrkraft bezieht.

3.2 Datenauswertung

Insgesamt wurden etwa sechs Stunden Audiomaterial aufgezeichnet; zusammen mit Auszügen aus den Chatverläufen ergaben sich daraus rund 80 Seiten Transkript für die Analyse. Die Auswertung erfolgte nach dem Ansatz der qualitativen Inhaltsanalyse nach Mayring (2015) in der Software MAXQDA (2022). Die Annotation der Daten wurde von einer der beiden Autor:innen sowie einer geschulten wissenschaftlichen Hilfskraft vorgenommen. Es wurden induktive Kategorien gebildet, um zentrale Muster und Kompetenzen zu identifizieren, die im Umgang mit den KI-Ausgaben eine Rolle spielen. Fokus war die Analyse der Reflexionen und Adaptionen der Lehrkräfte. Da die Äußerungen der Lehrkräfte die zentralen Perspektiven für die Forschungsfrage liefern, konzentrierte sich die Analyse ausschließlich auf diese Äußerungen. Die Aussagen des Interviewers wurden nicht als primäre Datenquelle verwendet, sondern dienten nur als Kontext für die Antworten der Lehrkräfte.

Durch die induktive Entwicklung eines Kategoriensystems, das auf die Forschungsfragen nach den Kompetenzen, Potentialen und Schwierigkeiten abgestimmt ist, lassen sich die relevanten Aspekte der Lehrkraft-KI-Interaktion identifizieren. Dieses Kategoriensystem fungiert als eine Art "Inventar" wichtiger Aspekte des Untersuchungsgegenstandes (Bortz & Döring, 2006, S. 381) und ermöglicht einen ersten Überblick in die empirischen Daten. In der vorliegenden Arbeit dient das Kategoriensystem somit als Übersicht der zentralen Aspekte in der Interaktion zwischen Lehrkräften und KI-Agenten im mathematischen Task Design. Anschließend können die Interpretationen der Ergebnisse in Richtung der Hauptfragestellung erfolgen.

Das entwickelte Kategoriensystem und das kodierte Material ermöglichen nicht nur die gezielte Auswahl von Passagen für eine qualitative Detailanalyse, sondern auch die Untersuchung absoluter Häufigkeiten. Die Häufigkeitsverteilungen der Kategorien ermöglichen Rückschlüsse auf zentrale Aspekte des Untersuchungsgegenstands, da häufig auftretende Kategorien besondere Relevanz aufweisen (Mayring, 2012). Die Betrachtung quantitativer Aspekte soll dabei Einblicke in die Nutzung der KI-Agenten durch die Lehrkräfte liefern und eine Basis

für systematische quantitative Untersuchungen schaffen. Im vorliegenden Beitrag wird der Fokus auf die Analyse der Prompt-Techniken und Eingabevarianten, die im Hinblick auf die Lehrkraft-KI-Interaktion besonders spannend erscheinen, gelegt.

4. Ergebnisse

4.1 Beobachtete Kompetenzen und Schwierigkeiten der Lehrkräfte im KI-gestützten mathematischen Task Design

Für die Datenanalyse wurde ein methodischer Rahmen mit Kategoriendefinitionen, Abstraktionsniveau und Analyseeinheiten aus den Beobachtungen entwickelt. Kodiert wurden Äußerungen der Lehrkräfte über ihre Wahrnehmung, Einschätzung und Nutzung von KI in unterrichtlichen Kontexten. Die Kategorien beziehen sich dabei auf die Perspektive der Lehrkraft – also darauf, wie sie die KI-Technologie erlebt, bewertet und in ihre pädagogische Praxis einordnet. Es werden sowohl explizite Aussagen (z. B. "Ich muss das kritisch hinterfragen") als auch implizite Einstellungen (z. B. Zustimmung zu KI-generierten Ausgaben ohne Prüfung) erfasst, sofern sie aus dem Kontext erschließbar sind. Zwei Personen kodierten das Material unabhängig voneinander und zwischen den Kodierphasen fanden Reflexionssitzungen zur Überprüfung und Überarbeitung der induktiv entwickelten Kategorien statt. Insgesamt wurden so 21 Kategorien entwickelt, die mit Definitionen und exemplarischen Ankerbeispielen dokumentiert wurden. In einem nächsten Schritt wurden diese zu sechs Hauptkategorien subsummiert, sodass ein Kategoriensystem entstand. Im Hinblick auf die Forschungsfragen wird zunächst der Teil des Kategoriensystems dargestellt, der die Frage beantwortet, welche Kompetenzen Lehrkräfte zeigen, um KI-generierte Outputs beim mathematischen Task Design zu nutzen (F1).

Die induktiv entwickelten Kategorien bieten eine strukturierte Übersicht über die Kompetenzen, die Lehrkräfte im Umgang mit den KI-Agenten während des mathematischen Task Designs zeigten und sind in fünf Hauptkategorien subsummiert: analytische Kompetenzen, technische Kompetenzen, adaptive Kompetenzen, didaktische Kompetenzen und Reflexionskompetenzen. Anschließend an die Kompetenzen, die für die Nutzung von KI-Agenten identifiziert werden konnten, sind zudem Schwierigkeiten aufgetaucht, die Lehrkräfte bei der Nutzung von KI-Agenten im mathematischen Task Design erleben.

Hauptkategorien	Subkategorien
Analytische Kompetenzen	Identifikation von bereichernden Ausgaben
	Validierung der fachlichen Korrektheit (Fehlerdiagnose)
	Kritische Bewertung der didaktischen Qualität
	Integration von Erfahrungswerten
Technische Kompetenzen	Kenntnis der Funktionsmechanismen von KI
	Kenntnisse der Möglichkeiten der KI-Agenten
	Gezieltes Prompting
Adaptive Kompetenzen	Fehlerkorrektur
	Anpassung: spezifische Kontexte
	Anpassung: spezifische Lernziele
	Weiterentwicklung der Ausgabe
Didaktische Kompetenzen	Berücksichtigung von Lernzielen
	Orientierung an Kompetenzentwicklung
	Entwicklung und Anwendung von Differenzierungsstrategien
Reflexionskompetenzen	Offenheit für neue Perspektiven
	Reflexion hinsichtlich der Aufgabe
	Reflexion der eigenen Rolle
Inhaltliche Schwierigkeiten/ Aufwand	Hoher Leseaufwand/ Umfangreiche Textmenge
	Erfordernis einer gezielten und detaillierten Anleitung
	Notwendigkeit, alle Inhalte kritisch zu hinterfragen
Technische Schwierigkeiten	Bildgenerierung

Tab 2: Kompetenzen und Schwierigkeiten der KI-gestützten mathematischen Task Design (reduzierte Darstellung des induktiv gewonnenen Kategoriensystems)

Die induktiv gewonnenen Kategorien umfassen sowohl inhaltliche als auch technische Schwierigkeiten und geben Einblicke in die Herausforderungen, die mit der Integration von KI-Agenten beim mathematischen Task Design verbunden sind. Im Folgenden werden die Hauptkategorien und ihre Subkategorien ausführlich beschrieben und mit Beispielen (d. h. konkreten Textpassagen) illustriert.

4.1.1 Analytische Kompetenzen

Die qualitative Inhaltsanalyse zeigt, dass Lehrkräfte analytische Kompetenzen einsetzen, um KI-generierte Ausgaben kritisch zu bewerten und an ihren eigenen Kontext anzupassen. So wurde die Subkategorie *Identifikation von bereichernden Ausgaben* codiert, wenn Lehrkräfte die KI-Ausgaben

filterten und priorisierten, um diese gezielt für das mathematische Task Design zu nutzen. Lehrkräfte versuchten dabei, zwischen nützlichen und irrelevanten Vorschlägen zu unterscheiden. Lehrkraft 3 beispielsweise bemerkte im Hinblick auf die Ausgabe des differenzierungsorientierten KI-Agenten unter Verwendung von Aufgabe b) folgendes:

Aber zum Beispiel mit geraden Zeilen finde ich schon auch eine plausible Differenzierung, dass die einfacher ist. Oder halt auch zu sagen, man begrenzt das auf eine bestimmte Anzahl an Zahlen. Es steht ja erst in zehn geraden Zahlen. Ja, das finde ich gut. (Lehrkraft 3)

Dies zeigt, wie Lehrkräfte ein Verständnis bezüglich der ausgegebenen Aufgabe und eine Einschätzung zur Qualität der Ausgabe anwenden.

Zudem zeigt sich die *Validierung der fachlichen Korrektheit* der Vorschläge des großen Sprachmodells als zentraler Bestandteil im Umgang mit KI-Agenten. Lehrkräfte erkannten und korrigierten potenzielle Fehler während der Verwendung der KI-Agenten (Fehlerdiagnose).

Genau, mit dem Kombinieren von Mengen bin ich jetzt nicht so ganz einverstanden. Da sind wir ja in der Kombinatorik und nicht in der Mengenlehre. Man kombiniert auch nicht Mengen, sondern man kann zwei Elemente von einer Menge miteinander in Relation setzen. (Lehrkraft 5)

Als weiterer Aspekt wurde in diesem Zusammenhang die Subkategorie zugeordnet, wenn die Überprüfung der mathematischen Tiefe im Fokus stand:

Also ich stolper darüber, dass die mathematische Tiefe darin nicht gesehen wird. (Lehrkraft 5)

Diese Aussage verdeutlicht, dass das Fachwissen eine wichtige Rolle spielt. Insbesondere, weil KI-Ausgaben häufig generisch oder ungenau ausfallen, konnte die Subkategorie gebildet werden.

Neben der inhaltlichen Richtigkeit scheint zudem die *kritische Bewertung der didaktischen Qualität* ein Bestandteil des KI-gestützten mathematischen Task Designs zu sein. In Abgrenzung zur Oberkategorie *Didaktische Kompetenzen (allgemein)* verstehen wir hierunter die gezielte Bewertung der Qualität einer didaktischen Maßnahme oder eines Prompts. Lehrkräfte prüften während der Verwendung der KI-Agenten, ob die generierten Aufgaben die intendierten Lernziele adressierten und Lernende aktivieren könnten. In Bezug auf eine KI-Ausgabe zur Aufgabe a) analysiert Lehrkraft 4 beispielsweise folgendes:

Ich würde das mit den Kontexten auch anders einordnen. Dass ja die Addition quasi erst mit Blick auf die

Grundvorstellung aufgebaut und vorbereitet wird und dass man dafür Kontexte braucht. Und dass sich daraus dann langsam das Verständnis von der Operation, also von der Addition ergibt. (Lehrkraft 4)

Diese Subkategorie spiegelt somit die Bedeutung einer didaktischen Reflexion wider, um die Lernwirksamkeit der KI-generierten Aufgabe sicherzustellen.

Lehrkräfte bezogen zudem ihre eigenen Erfahrungen ein, um die Vorschläge der KI-Agenten zu bewerten und weiterzuentwickeln. Die Subkategorie *Integration von Erfahrungswerten* wurde daher identifiziert, wenn der Umgang mit KI-Agenten nicht nur von den Ausgaben abhängt, sondern auch stark durch die individuelle Perspektive der Lehrkräfte geprägt war.

4.1.2 Technische Kompetenzen

Technische Kompetenzen umfassen die Fähigkeit, die Funktionsweise der KI-Agenten zu verstehen und optimal zu nutzen.

Ein grundlegendes Verständnis der KI-Technologie zeigte sich als hilfreich, um die Potenziale und Grenzen von KI-Agenten einschätzen zu können. Die Subkategorie *Kenntnis der Funktionsmechanismen von KI* verweist auf Aspekte in der Lehrkraft-KI-Interaktion, die auf entsprechendes Wissen hindeuten. Diese Kompetenz schien die Interaktion mit KI-Agenten zu erleichtern und potenzielle Fehlanwendungen zu reduzieren.

Wenn man zum Beispiel so im Bereich Sachrechnung, wo man viele Daten braucht oder so was, wo es darum geht, kannst du mir irgendwie eine Statistik zu irgendwas erstellen, damit kann es vielleicht eine große Hilfe sein. (Lehrkraft 4)

Die Nutzung der KI-Agenten durch Lehrkräfte zeigte, dass Kenntnisse über deren spezifische Fähigkeiten (z. B. Differenzierung, Sprachsensibilität oder Bildgenerierung) dazu beitragen können, sie als Experten-Perspektive innerhalb eines KI-unterstützten kollaborativen Teams zu sehen und zielgerichtet einsetzen zu können. Lehrkraft 2 verwies beispielsweise auf die Möglichkeiten eines KI-Agenten:

Ich kann den noch mal fragen, wie der die sprachlich vereinfacht darstellen würde. (Lehrkraft 2)

Die Subkategorie *Kenntnisse der Möglichkeiten der KI-Agenten* wurde daher verwendet, wenn Aussagen der Lehrkräfte auf Wissen über spezifische Funktionen und Einsatzpotenziale von KI-Agenten verweisen.

Durch die Subkategorie *gezieltes Prompting* wird deutlich, dass Lehrkräfte überlegte Eingaben formulierten, um den Zusammenhang zwischen der Qualität der Prompts und der Qualität der Ausgaben zu nutzen. Lehrkraft 5 verwies beispielsweise darauf, dass dem KI-Agenten gezielte Prompts formuliert werden müssen:

Also das macht er jetzt nicht von sich, sondern ich müsste ihm jetzt ja eigentlich sagen. (Lehrkraft 5)

Diese Fähigkeit scheint eine Kompetenz im Umgang mit großen Sprachmodellen zu sein, da sie die Grundlage für zweckdienliche Ergebnisse bildet.

4.1.3 Adaptive Kompetenzen

Adaptive Kompetenzen umfassen die Fähigkeit, die Ausgaben der KI-Agenten flexibel anzupassen und weiterzuentwickeln. Während der Verwendung der KI-Agenten zeigten Lehrkräfte, dass sie nicht nur Fehler in den KI-Ausgaben erkennen, sondern diese auch gezielt beheben können. Lehrkraft 2 beispielsweise bemerkte in einer Ausgabe Inhalte, die nicht zur genannten Klassenstufe passen:

Das ist nicht Klasse 4. Ich habe aber Klasse 4 geschrieben. ... [tippt] Und dann schreibe ich dem auch mal noch Klasse 4 dazu. (Lehrkraft 2)

An dieser Stelle müsste durch gezieltes Prompting die Ausgabe nachgesteuert werden. Die Subkategorie *Fehlerkorrektur* wurde daher kodiert, wenn Lehrkräfte die Outputs der KI nicht nur passiv übernehmen, sondern aktiv überarbeiten und anpassen.

Die Anpassung der Aufgaben an den jeweiligen Unterrichtskontext bzw. die jeweilige Lerngruppe scheint sich zudem in der Verwendung mit KI-Agenten zu zeigen. Die Subkategorie *Anpassung an spezifische Kontexte* wurde daher gebildet, um die Intention zur Anpassung der Ausgaben, die Lehrkräfte bei der Nutzung von generischen Ausgaben zeigten, zu erfassen.

Aufgaben wurden von den untersuchten Lehrkräften nicht nur kontextuell, sondern auch zielgerichtet angepasst, damit Ausgaben zu den angestrebten Kompetenzen der Lernenden passen. Die Subkategorie *Anpassung an spezifische Lernziele* markiert diese Aussagen der Lehrkräfte zur fachdidaktischen Zielorientierung.

Lehrkräfte nutzten die Ausgaben zusätzlich als Grundlage für kreative Erweiterungen und Anpassungen.

Das heißt, jetzt wäre die Frage, ob man daraus was basteln kann. (Lehrkraft 5)

Die Subkategorie *Weiterentwicklung der Ausgabe* zeigt damit die Rolle von Lehrkräften als aktive Gestalter im Prozess mathematischen Task Designs.

4.1.4 Didaktische Kompetenzen

Die qualitative Inhaltsanalyse verdeutlicht, dass Lehrkräfte didaktische Kompetenzen einsetzen, um KI-Ausgaben in lernförderliche Aufgaben zu überführen. Lehrkräfte berücksichtigten beispielsweise die curricularen Anforderungen der Ausgaben und erfragten ggf. die Lernziele, die mit einer Aufgabe erzielt werden sollen. Sie forderten in ihren Prompts explizit die Berücksichtigung ihrer eigenen Lernziele ein. Die Subkategorie *Berücksichtigung von Lernzielen* zeigt daher die Verbindung zwischen bewusst formulierten Lernzielen und den Vorschlägen der KI-Agenten.

Auch verfolgten die Lehrkräfte das Ziel, dass KI-generierte Aufgaben nicht nur der Wissensvermittlung dienen, sondern auch prozessbezogene Kompetenzen fördern. Während der KI-gestützten Aufgabenmodifikation b) überlegt Lehrkraft 1:

Oder vielleicht gibt es auch eine Aufgabe, wo sie noch mal eine eigene Aufgabe entwickeln müssen, in der sie dann begründen, dass die ähnlich ist wie die, damit sie dieses Muster verstehen. Das heißt, ich versuche jetzt nochmal die Aufgabe hier reinzubringen und sage, ich ... möchte, dass diese Aufgabe nicht nur ausgerechnet wird, sondern dass die Kinder, dass die Lernenden das Muster verstehen und begründen. (Lehrkraft 1)

Die Subkategorie *Orientierung an Kompetenzentwicklung* verweist auf diese und ähnliche Aussagen, die eine langfristige Förderung prozessbezogener Kompetenzen in den Blick nehmen.

Lehrkräfte nutzten die KI-Agenten auch, um Aufgaben für heterogene Gruppen zu gestalten. Diese Fähigkeit wurde von Lehrkräften vor allem im Zusammenhang der Ausarbeitung von Aufgaben für inklusive oder leistungsheterogene Lerngruppen thematisiert. Die Subkategorie *Entwicklung und Anwendung von Differenzierungsstrategien* nimmt genau diese Anwendungsfälle in den Blick.

4.1.5 Reflexionskompetenz

Selbstreflexive Kompetenzen beziehen sich auf die Reflexion über die eigene Rolle und den Umgang mit den KI-generierten Ausgaben. Die Lehrkräfte zeigten sich größtenteils aufgeschlossen gegenüber den KI-Agenten:

Ich finde es schon insgesamt ganz interessant, wie man mit ihm so in einen Austausch treten kann. (Lehrkraft 4)

Diese Offenheit schien den Lehrkräften zu ermöglichen, die Perspektive der KI-Agenten als wertvolle Ergänzung wahrzunehmen und das mathematische Task Design als kollaborativen Prozess zu verstehen (Choppin, 2011). Die Subkategorie *Offenheit für neue Perspektiven* markiert diese Fälle.

Sie reflektierten zudem, wie ihr eigenes Wissen über die Aufgabe durch die KI-Ausgabe beeinflusst wurde. Dabei gingen die Lehrkräfte teilweise offen mit eigenen ‚Denklücken‘ um:

Es ist auch spannend, dass jetzt wirklich auch mal die vielen Merkmale von natürlicher Differenzierung aufgeführt werden. Also, dass man so ins Denken kommt, okay, wie, wenn ich differenzieren will. ... Aber ich finde es ziemlich gut, weil es sind ja alles Dinge, die man schon mal gehört hat, aber jetzt vielleicht nicht in der Planung konkret drauf kommen würde. (Lehrkraft 3)

Die Subkategorie *Selbstreflexion hinsichtlich der Aufgabe* wurde daher kodiert, wenn die Arbeit mit KI-Agenten die Lehrkräfte zur Reflexion bezüglich der eigenen Konzeption von Unterrichtsmaterialien anregte.

Die Lehrkräfte schienen zudem herausgefordert, ihre didaktische und methodische Rolle bei der Unterrichtsplanung und beim mathematischen Task Design neu zu definieren. Lehrkraft 1 beispielsweise nahm die Perspektive der KI-Agenten offen an, erkannte dadurch Lücken im eigenen Task Design noch und sah gleichzeitig ihre eigene Verantwortung, auf dieser Grundlage nun die Aufgabe weiterzuentwickeln und zu optimieren:

Genau so, also das hat mir jetzt schon mal geholfen, also Zahlen und Operation ist mir klar, das weiß ich, aber dass jetzt quasi noch, dass er jetzt noch sagt, okay, prozessbezogene Kompetenz argumentieren und begründen, ist für mich jetzt noch mal ein guter Hinweis, dass ich weiß, okay, ich versuche jetzt noch mal die Aufgabe irgendwie so zu verändern. (Lehrkraft 1)

Die damit kodierte Subkategorie *Reflexion der eigenen Rolle* verdeutlicht somit die Dynamik, die durch den Einsatz von KI-Agenten zum Ausdruck kommt.

4.1.6 Inhaltliche Schwierigkeiten und Aufwand

Diese Hauptkategorie beschreibt die inhaltlichen Herausforderungen, die mit der Analyse, Anpassung und Anwendung der KI-Ausgaben verbunden sind. Dabei ist zu beachten, dass ‚Schwierigkeiten‘ auch durch fehlende Kompetenzen entstehen können – entsprechend wurden bestimmte Aspekte teilweise im mehreren Codierungen auftauchen. Dies wird an entsprechender Stelle erläutert.

Ein Teil der Lehrkräfte bemerkte, dass die Textmenge der generierten Outputs oft umfangreich ist und einen *hohen Leseaufwand* erfordert.

Also, ich glaube, mir wäre das, würde das mehr Aufwand machen, das mir dann durchzulesen, was er da sagt. (Lehrkraft 4)

Die ausführlichen KI-Ausgaben scheinen eine effiziente Nutzung zu erschweren, da Lehrkräfte – besonders unter Zeitdruck – relevante Inhalte aus umfangreichen Texten herausfiltern mussten.

Die Qualität der Ausgaben hängt bekannterweise stark von den Prompts ab (Liu et al., 2021). Auch Lehrkräfte beschrieben, dass eine präzise Anleitung notwendig ist, um brauchbare Ergebnisse zu erhalten.

Jetzt kann ich mir ja mal überlegen, worauf ich hinaus wollte. Vielleicht muss ich ein bisschen mehr anleiten. Aber man merkt schon, man braucht ja ganz schön viel Wissen eigentlich, damit man das benutzen kann. (Lehrkraft 5)

Diese Schwierigkeit verdeutlicht, dass ein fundiertes Verständnis von Prompting für den Umgang mit KI-Agenten hilfreich sein kann. Die Subkategorie *Erfordernis einer gezielten und detaillierten Anleitung* wird daher codiert, wenn Lehrkräfte gezielt Fragen formulierten und klare Anweisungen gaben. Dies deutet auf erhöhte Anforderungen an ihre technologischen Kompetenzen hin. Dabei zeigen sich zwei Seiten derselben Medaille: Einerseits werden gezielte Anleitung und technologische Kompetenz als Voraussetzung für den erfolgreichen Einsatz sichtbar – andererseits treten genau dort inhaltliche Schwierigkeiten auf, wenn diese Kompetenzen (noch) fehlen. Diese Verbindung führt wie eingangs erwähnt dazu, dass bestimmte Aspekte bzw. Äußerungen sowohl unter ‚Schwierigkeiten‘ als auch unter ‚Kompetenzen‘ doppelt codiert wurden.

Lehrkräfte schienen sich zudem gezwungen zu sehen, die von den KI-Agenten erzeugten Inhalte umfassend zu überprüfen, da Fehler, Unklarheiten oder unpassende Vorschläge häufig vorkamen. Die Subkategorie *Notwendigkeit, alle Inhalte kritisch zu hinterfragen* verweist auf die Rolle der Lehrkräfte bei der aktiven Prüfung und Validierung der KI-Ausgaben und meint den allgemeinen Anspruch an einen reflektierten Umgang mit Inhalten – auch wenn kein konkreter Fehler vorlag (Abgrenzung zur Kategorie *Fehlerdiagnose*, in der konkret korrektive Hinweise auf spezifische fachliche Fehler gegeben werden). In den Beobachtungen zeigt sich, dass dies mit einem erhöhten kognitiven und zeitli-

chen Aufwand verbunden war. Beispielsweise beschreibt Lehrkraft 4:

Und das dann immer zu hinterfragen, als mich einfach hinzusetzen und mir selber was zu überlegen, was ich mit [der Aufgabe] oder sowas machen will. (Lehrkraft 4)

4.1.7 Technische Schwierigkeiten

Technische Schwierigkeiten umfassen die Herausforderungen im Umgang mit der technologischen Funktionalität der KI-Agenten und weisen auf technische Barrieren hin. Einige Lehrkräfte berichteten beispielweise von Schwierigkeiten bei der Generierung und Anpassung von visuellen Materialien durch die KI-Agenten:

Ja, witzig. Das ist ein perfektes Bild. Die Aufgabe ist drin. Aber er checkt nicht, dass die [Autos] gleich einen Unfall bauen. ... Es gibt ein Bild, aber es hilft nicht weiter, weil da andere Angaben stehen. (Lehrkraft 2)

Die Ausgaben schienen oft nicht den gewünschten Anforderungen zu entsprechen oder technisch nicht umsetzbar zu sein. Die Fähigkeit der KI-Agenten zur Bildgenerierung wurde von den Lehrkräften als begrenzt wahrgenommen. Die Subkategorie *Bildgenerierung* weist damit auf das generelle Problem der Generierung fachdidaktisch durchdachter Abbildungen hin.

4.2 Verwendete Prompt-Techniken

Für die Analyse der absoluten Häufigkeiten der Prompt-Typen und Eingabevarianten wurde zunächst (ebenfalls durch induktive Kategorienbildung) herausgearbeitet, welche Arten von Anfragen die Lehrkräfte an die KI-Agenten stellen. Folgende Prompt-Typen konnten im Material identifiziert werden (Tabelle 3).

Die verschiedenen Prompt-Typen lassen sich in zwei Hauptkategorien einteilen: *Interaktionsleitende Prompts* steuern den Dialog mit dem KI-Agenten, während *didaktische Prompts* die inhaltliche und pädagogische Gestaltung der Aufgaben fokussieren. Die Kategorien beschreiben zunächst zwei unterschiedliche Aspekte der Lehrkraft-KI-Interaktion.

Hauptkategorien	Subkategorien	H
Interaktionsleitende Prompts:	Weiterführender Prompt	18
	Ausschließlich Aufgabenstellung als Initialprompt	12
	Konkretisierung des vorherigen Prompts	11
	Präziser Initialprompt	3
	Referenzierender Prompt	1
Didaktische Prompts	Visuelle Darstellung	10
	Anforderung zusätzlicher Aufgaben	8
	Förderung prozessbezogene Kompetenzen	5
	Anpassung der Differenzierung/Differenzierungsvorschläge	5
	Sprachliche Vereinfachung/ Unterstützung	5
	Ergänzung von Lernzielen	2
	Abfrage mathematischer Grundlagen	1
	Anpassung der Aufgabenschwierigkeit	1

Tab. 3: Absolute Häufigkeiten der verschiedenen Arten von Prompt-Techniken in der Lehrkraft-KI-Interaktion während des Task Designs

4.2.1 Interaktionsleitende Prompts

Interaktionsleitende Prompts wurden eingesetzt, um die Kommunikation zwischen Lehrkraft und KI-Agenten effizient zu gestalten. Sie steuerten die Interaktion durch klare Anweisungen und gezielte Eingaben.

Ein *präziser Initialprompt* fungierte als Ausgangspunkt der Interaktion und definierte, was die KI-Agenten generieren sollen. In Anlehnung an Liu et al. (2021) wurde in dieser Studie ebenfalls deutlich, dass präzise Formulierungen dazu beitragen können, Missverständnisse zu verringern und die Generierung zielgerichteter zu gestalten.

Ich zeige dir gleich eine Aufgabe. Zeig mir bitte welche Kompetenzen und Lernziele durch diese Aufgabe bei den Lernenden gefördert werden. (Lehrkraft 1)

In der Analyse zeigt sich, dass präzise formulierte Initialprompts mit Ausgaben einhergehen, die von den Lehrkräften als relevant eingeschätzt wurden. Dies wurde von einigen Lehrkräften auch mit einer effizienteren Arbeitsweise in Verbindung gebracht.

Die *ausschließliche Aufgabenstellung als Initialprompt* beschränkt sich auf die mathematische Aufgabe ohne Kontextinformationen und überlässt es den KI-Agenten, relevante Bezüge herzustellen. In einigen Fällen wurden dabei kontextuell weniger passende oder allgemeinere Ausgaben beobachtet.

Durch die *Konkretisierung des vorherigen Prompts* passten Lehrkräfte zudem ihre Eingaben an, um die Ausgaben der KI-Agenten gezielt zu steuern. Die

Konkretisierung verweist auf iterative Prozesse, in denen Lehrkräfte generierte Ergebnisse reflektieren und ihre Prompts entsprechend anpassen.

Die Lehrkräfte formulierten außerdem *weiterführende Prompts*, die inhaltliche Erweiterung durch neue Impulse erzielen, indem sie die KI-Agenten zu vertiefenden Antworten aufforderten, etwa durch neue Argumente, zusätzliche Perspektiven oder weiterführende Fragestellungen. Lehrkraft 1 begründet die Eingabe eines weiterführenden Prompts folgendermaßen:

...damit ich noch mal weiß, okay, was könnte ich noch zusätzlich zu der Aufgabe hinzufügen, damit [die Schüler:innen] nochmal argumentieren und begründen müssen. (Lehrkraft 1)

Solche Prompts machen eine kontinuierliche Interaktion sichtbar, bei der Lehrkräfte die KI-Ausgabe bewusst steuern und gezielt anreichern.

Im Unterschied dazu beziehen sich *Referenzierende Prompts* lediglich auf bereits generierte Inhalte oder Beispiele, ohne diese inhaltlich weiterzuentwickeln. Sie dienen primär der Wiederholung, Anwendung oder Strukturierung von Bekanntem (Abgrenzung Subkategorie *Weiterführender Prompt*). Ein Beispiel liefert Lehrkraft 4, die im Prozess des mathematischen Task Design den differenzierungsorientierten KI-Agenten auf ein zuvor besprochenes Beispiel verweist:

Das kann ich nochmal versuchen, ob ich ihn dahin bringen kann. Ich möchte, kann ich ihn jetzt erinnern an das, was er vorhin gemacht hat? [Interviewer: Ja] Ich möchte, wie in dem Beispiel mit den Quadratzahlen, gerne eine natürlich differenzierende Aufgabe, differenzierende Aufgabe haben, die die Kinder zum Untersuchen von Rechteckszahlen und dem Zusammenhang mit Additionsaufgaben veranlasst. (Lehrkraft 4)

Hier steht die Verknüpfung mit vorherigen Inhalten im Vordergrund, nicht deren Weiterentwicklung. Obwohl Lehrkraft 4 noch nie bewusst ein großes Sprachmodell verwendet hat, wurde hier außerdem intuitiv die Prompt-Technik des Few-Shots (Liu et al., 2021), d. h. die Eingabe mit Minimalbeispiel, verwendet.

4.2.2 Didaktische Prompts

Didaktische Prompts fokussieren sich auf die elementarmathematische und pädagogische Gestaltung der generierten Aufgaben. Sie beziehen sich auf konkrete Anforderungen, die die Aufgaben für den Unterricht relevant machen.

Lehrkräfte forderten beispielsweise die KI-Agenten auf, alternative oder ergänzende Aufgaben zu erstellen. Diese *Anforderung zusätzlicher Aufgaben* in Prompts weist darauf hin, dass Lehrkräfte die KI-Agenten auch zur Generierung vielfältiger Aufgabenformate einsetzten und zusätzliche Anregungen für die Unterrichtsgestaltung einholten.

In Prompts, die auf die *Ergänzung von Lernzielen* abzielten, forderten die Lehrkräfte die KI-Agenten dazu auf, Aufgaben mit bestimmten curricularen Zielsetzungen zu verknüpfen.

Oder vielleicht gibt es auch eine Aufgabe, wo sie noch mal eine eigene Aufgabe entwickeln müssen, in der sie dann begründen, dass die ähnlich ist wie die, damit sie dieses Muster verstehen. (Lehrkraft 1)

Die Lehrkräfte forderten zudem spezifische *Differenzierungsvorschläge* durch Prompts zur *Anpassung der Differenzierung*, um Aufgaben zu gestalten, die auf verschiedene Leistungsniveaus zugeschnitten sind. Lehrkraft 2 erfragte beispielsweise im Kontext der Aufgabe b) nach Hilfen für Lernende, die Schwierigkeiten haben das Muster der Zahlenfolge zu erkennen:

Welche Hilfe könnte ich Kindern geben, um einen Einstieg in die Aufgabe zu kriegen. (Lehrkraft 2)

Dieser Typ des Promptings stand im Zusammenhang mit Lehrkräften, die im Kontext heterogener Lerngruppen Aufgaben generieren wollten. Lehrkräfte setzten diesen KI-Agenten ein, um insbesondere individualisierte Aufgabenformate zu entwickeln. Die Subkategorie verdeutlicht damit eine Strukturierungsmöglichkeit der Aufgabe und nicht die mathematisch, inhaltliche Ausgestaltung, wie es in Abgrenzung zur Subkategorie *Anpassung der Aufgabenschwierigkeit* der Fall ist.

Prompts, die die *Sprachliche Vereinfachung/Unterstützung* fokussierten, forderten die Anpassung der Sprache an bestimmte Zielgruppen. Diese Prompts verwiesen im Task Design darauf, dass Lehrkräfte sprachliche Anforderungen berücksichtigten und Inhalte so gestalteten, dass sie für Lernende mit unterschiedlichen sprachlichen Voraussetzungen verständlich sind.

In Prompts zur *visuellen Darstellung* baten Lehrkräfte um die Erstellung oder Integration visueller Elemente, die die Aufgaben unterstützen sollten. Die Integration von Visualisierungen zeigte, dass Lehrkräfte multimodale Ansätze beim mathematischen Task Design berücksichtigten.

Mit Prompts zur *Förderung prozessbezogener Kompetenzen* zielten die Lehrkräfte darauf ab, Aufgaben

zu erstellen, die z. B. mathematisches Problemlösen oder mathematisches Argumentieren (Kultusministerkonferenz, 2022) fördern. Diese Prompts weisen somit darauf hin, dass Lehrkräfte Aufgaben erarbeiteten, die neben der Wissensabfrage auch prozessbezogene Kompetenzen adressieren.

Die Lehrkräfte forderten mit Prompts zur *Abfrage mathematischer Grundlagen* von den KI-Agenten die Erklärung oder Bereitstellung mathematischen Wissens zu einer Aufgabe. Diese Prompts deuten darauf hin, dass Lehrkräfte den Ausgaben der KI mit Zurückhaltung begegneten und Wert auf Transparenz sowie die Möglichkeit zur Überprüfung der Inhalte legten u. a. auch in Hinblick auf die kontextbasierten Informationen, die in den Generierungsprozess einfließen – ein Vorgehen, das sich mit der bereits diskutierten Logik der Self-Explanation (Huang et al., 2023) zusammenbringen lässt.

Die Prompts zur *Anpassung der Aufgabenschwierigkeit* zielten darauf ab, Aufgaben einfacher oder schwieriger zu gestalten. Sie sind also in Abgrenzung zur Kategorie *Anpassung der Differenzierung / Differenzierungsvorschläge* inhaltsorientiert.

Hast du ähnliche Aufgaben mit unterschiedlichen Schwierigkeitsniveaus? (Lehrkraft 3)

In diesem Zusammenhang wurde sichtbar, dass Lehrkräfte beim KI-gestützten mathematischen Task Design auf die individuellen Lernvoraussetzungen ihrer Schüler:innen eingehen.

Anschließend an die Identifikation der Prompt-Typen und Eingabevarianten wurden die Kodehäufigkeiten, d. h. die absoluten Häufigkeiten der verschiedenen Prompts analysiert. Insgesamt wurden 82 Prompts kodiert. Tabelle 3 bietet eine Übersicht über die Verteilung der verschiedenen Prompt-Typen, die von Lehrkräften bei der Nutzung der KI-Agenten verwendet wurden, welche im folgenden Abschnitt diskutiert werden.

5. Diskussion

Die Ergebnisse der Interviewstudie machen deutlich, dass Lehrkräfte vielfältige Kompetenzen zeigten, um KI-generierte Ausgaben im Prozess des Task Designs zu nutzen (F1). Die KI-Agenten lieferten Anregungen für das mathematische Task Design, die die Lehrkräfte adaptierten und häufig überarbeiteten oder weiterentwickelten. Insbesondere in den Bereichen Differenzierung und Kompetenzorientierung wurde die Unterstützung durch die KI-Agenten von den Lehrkräften besonders hervorgehoben. Gleichzeitig wurde der Bedarf an didaktisch-methodischen Kompetenzen sichtbar, um

die Ausgaben kritisch bewerten und anpassen zu können.

Dafür muss man halt sein Ziel kennen. ... Bei der bildlichen Darstellung und bei der symbolischen und der Alltagsebene würde mir jetzt auch, wenn ich didaktisch draufgucke, das Konzeptionelle fehlen (Lehrkraft 5)

Während der induktiven Kategorienbildung war das Auftreten der Kategorie *Reflexionskompetenz* ein unerwarteter Befund. Ursprünglich wurde das Material lediglich im Hinblick auf fachdidaktische Kompetenzen sowie auftretende Schwierigkeiten kodiert. Die empirischen Daten zeigen jedoch, dass die Interaktion mit den KI-Agenten von einigen Lehrkräften auch als Impuls für die Reflexion über Prozesse des Task Designs beschrieben wurde – ähnlich dem Perspektivwechsel in einem multiprofessionellen Team (Widmer-Wolf, 2018). Diese Reflexionskompetenz, die sich induktiv als Kategorie aus dem Datenmaterial ergab, erwies sich als ein bedeutendes und vielversprechendes Phänomen. Daraus ergibt sich ein wertvolles Potenzial für weitere Forschung, da die identifizierten Reflexionsmuster nicht nur als Potenzial von KI-Agenten untersucht werden können, sondern auch die professionelle Weiterentwicklung von Lehrkräften unterstützen können.

Hinsichtlich der Schwierigkeiten, die bei der Nutzung von KI-Agenten auftreten, zeigten Lehrkräfte nur wenige Möglichkeiten, wie diese adressiert werden können (F2). Die technischen Schwierigkeiten bei der Bildgenerierung deuten darauf hin, dass technologische Entwicklungen im Bereich der Visualisierung weiter verbessert werden müssen, um den spezifischen Bedürfnissen von Lehrkräften gerecht zu werden. Allerdings kann dieser Schwierigkeit von den Lehrkräften alleine nicht begegnet werden. Hier ist jedoch davon auszugehen, dass zukünftige Versionen großer LLMs Verbesserungen zeigen könnten. In diesem Zusammenhang wurde auch der erhöhte Aufwand durch die Anwendung bestimmter Kompetenzen im mathematischen Task Design bemerkt – zum einen in Bezug auf das kritische Hinterfragen und Bewerten der Ausgaben und zum anderen in Bezug auf den Leseaufwand aufgrund der generierten Textmenge. Dabei wird deutlich, dass Lehrkräfte längere Textausgaben der KI teilweise als zu ausführlich empfinden und sich kurze, prägnante Formulierungen wünschen.

Die Verteilung der Prompt-Typen verdeutlicht hingegen, welche Arten von Prompts von Lehrkräften eingesetzt wurden (F3). Bemerkenswert scheint, dass interaktionsleitende Prompts ebenso häufig

(sogar minimal häufiger) eingesetzt wurden wie didaktische Prompts. 45 der 82 Prompts waren interaktionsleitender Natur, während 37 der Prompts didaktische Ziele spezifizierten. Die am häufigsten verwendeten Eingaben bezogen sich auf das Weiterführen generierter Ausgaben (18) auf interaktionsleitender Ebene und auf die Anforderung visueller Darstellungen (10) sowie zusätzlicher Aufgaben (8) auf didaktischer Ebene. Im Folgenden soll die Verteilung der Prompt-Typen im Detail betrachtet werden.

Die interaktionsleitenden Prompts steuern die Interaktion mit den KI-Agenten und bilden die Grundlage für die generierten Ausgaben. In Hinblick auf alle Initialprompts lässt sich nur ein geringer Anteil präziser Initialprompts feststellen (3 von 15 Initialprompts). In 12 von 15 Fällen nutzten die Lehrkräfte minimalistische Eingaben, verwendeten ausschließlich die Aufgabenstellung als Initialprompt und verzichteten auf weiterführende Eingaben. Dies deutet darauf hin, dass die Nutzung der KI-Agenten häufig explorativ beginnt. In vielen Fällen wurden Erwartungen an die KI-Technologie erst im Verlauf der Interaktion präzisiert, anstatt sie bereits zu Beginn klar zu formulieren. Dieser Ansatz kann generische Ausgaben zur Folge haben, da die KI-Agenten ohne Kontext arbeiten. Dies schien auch die Lehrkräfte zu erkennen. Begannen sie mit einem anderen KI-Agenten zu arbeiten, wurde oft entweder ein präziser Initialprompt oder ein weiterführender Prompt verwendet. Im Verlauf der Interviews zeigte sich, dass Lehrkräfte zunehmend klar definierte Aufgabenstellungen formulierten. In der durch den Initialprompt initiierten Interaktion folgen als häufige Schritte die Konkretisierung des vorherigen Prompts (11 von 45 interaktionsleitenden Prompts) sowie die eben erwähnten weiterführenden Prompts (18 von 45 interaktionsleitenden Prompts). Die hohe Anzahl an Konkretisierungen deutet darauf hin, dass Lehrkräfte iterativ arbeiten und ihre Eingaben verfeinern, um kontinuierlich relevantere und präzisere Ergebnisse zu erhalten. Dieser iterative Prozess ist zentral für die Optimierung der Ausgaben. Auch die hohe Anzahl an weiterführenden Prompts weist darauf hin, dass Lehrkräfte die KI-Agenten nutzen, um auf vorherigen Ausgaben aufzubauen und weiterführende Ideen zu generieren. Dies unterstreicht das Potenzial für ein gemeinsames, aufeinander aufbauendes mathematisches Task Design.

Didaktische Prompts stellen konkrete pädagogische Anforderungen dar und zeigen, wie Lehrkräfte die KI-Agenten für ihre didaktischen Ziele einsetzen. Als

häufigste Prompt-Typen konnten die Anforderung visueller Darstellung und zusätzlicher Aufgaben (8 von 37 didaktischen Prompts) kodiert werden. Die häufige Anfrage nach visuellen Darstellungsmöglichkeiten (10 von 37) deutet auf die Bedeutung hin, die Lehrkräfte visuellen Unterstützungen für das Verständnis von Aufgaben zuschreiben. Prompts mit Darstellungswechseln deuten darauf hin, dass Lehrkräfte Darstellungsnetzwerke nutzen möchten und verschiedene Zugänge zu Inhalten erproben. Die häufigen Anfragen nach zusätzlichen Aufgaben zeigen, dass KI-Agenten zur Erweiterung und Variation bestehender Materialien herangezogen werden. Drei Prompt-Typen wurden mit jeweils 5 von 37 didaktischen Prompts kodiert: Die Anforderung von spezifischen Differenzierungsvorschlägen, die Anforderung von sprachlicher Vereinfachung oder Unterstützung sowie die Anforderung von Hinweisen zur Förderung prozessbezogener Kompetenzen. Die Prompts zu Differenzierungsvorschlägen und individualisierten Aufgaben deuten darauf hin, dass Lehrkräfte KI-Agenten einsetzen, um Aufgaben für heterogene Lerngruppen zu variieren. Prompts zur sprachlichen Anpassung beziehen sich häufig auf Verständlichkeit und sprachliche Zugänglichkeit für Lernende mit unterschiedlichen Sprachkompetenzen. In Prompts, die mathematisches Argumentieren, Problemlösen oder andere prozessbezogene Kompetenzen (Kultusministerkonferenz, 2022) thematisieren, wird sichtbar, dass Lehrkräfte Aufgaben entwickeln, die über reine Inhaltsvermittlung hinausgehen. Nur bei 2 von 37 didaktischen Prompts wurden Lernziele explizit in den Prompts thematisiert und ihre Ergänzung eingefordert. Diese geringe Anzahl könnte darauf hinweisen, dass die Lehrkräfte die Lernziele implizit mitdenken, worauf auch die Äußerung von Lehrkraft 1 hindeutet.

Die Frage wäre halt auch, ob, also wie ich jetzt starte, weißt du, weil ich weiß ja jetzt schon eigentlich, was ich mit so einer Aufgabe zum Ziel habe, die Frage wäre jetzt, ob ich noch mal gucke, was mir der kompetenzorientierte Chatbot sagt, welche Kompetenzen überhaupt durch so eine Aufgabe gefördert werden. (Lehrkraft 1)

Das Lernziel wird als Ausgangspunkt und so grundlegend für das Task Design gesehen, dass es nicht direkt in die Interaktion eingebracht oder abgefragt werden muss. Noch seltener werden mathematische Grundlagen abgefragt (1 von 37 didaktischen Prompts). Der geringe Anteil suggeriert, dass Lehrkräfte ihr fachspezifisches Grundlagenwissen als ausreichend einschätzen oder die detaillierten Aus-

führungen des LLMs als implizite mathematische Auffrischung nutzen. Lehrkraft 3 bestätigt dies:

Also es ist ja dann nochmal ein Auffrischen von bestimmten Inhalten. (Lehrkraft 3)

Dies deutet darauf hin, dass der Dialog mit KI-Agenten als Gelegenheit zur Auffrischung mathematischer Grundlagen wahrgenommen wird. Auch die Anpassung der Aufgabenschwierigkeit wurde nur einmal angefordert. Obwohl es für die KI-Agenten ein leichtes wäre, Aufgaben in verschiedenen Schwierigkeitsniveaus darzustellen, verlangen die Lehrkräfte nicht einfach nach einer leichteren Aufgabe, sondern nach Möglichkeiten, den Verstehensprozesse zu unterstützen, ohne das Niveau der Aufgabe zu verändern.

Abschließend ist einzuräumen, dass die Annotation der Daten ausschließlich durch eine der beiden Autor:innen sowie eine geschulte wissenschaftliche Hilfskraft vorgenommen und im Rahmen einer doppelten Codierung überprüft wurde. Dadurch kann nicht ausgeschlossen werden, dass subjektive Einschätzungen oder eine eingeschränkte Perspektivenvielfalt die Objektivität und Konsistenz der Annotation beeinflusst haben.

6. Schlussfolgerung

Die induktiv entwickelten Kategorien systematisieren die erforderlichen Kompetenzen und auftretenden Schwierigkeiten beim KI-gestützten mathematischen Task Design. Sie zeigen die vielfältigen Anforderungen an Lehrkräfte, die sich in der Integration von KI-Agenten und ihren Experten-Perspektiven in die Unterrichtsplanung zeigen. In den Daten wird deutlich, dass bei der Nutzung von KI-Agenten verschiedene Kompetenzen zum Einsatz kommen (darunter technologische, analytische, adaptive, fachdidaktische und selbstreflexive Fähigkeiten), um das Potenzial der Outputs beim Task Design auszuschöpfen. Diese Erkenntnisse sind sowohl für die Lehrkräfteausbildung als auch für die Weiterentwicklung von KI-Agenten von zentraler Bedeutung.

Lehrkräfte zeigten selten Lösungsansätze für Schwierigkeiten beim Einsatz von KI-Agenten. Technische Probleme bei der Bildgenerierung erfordern Verbesserungen zukünftiger Sprachmodelle, die Lehrkräfte nicht selbst beheben können. Zusätzlich empfanden die Lehrkräfte die Arbeit mit den KI-Agenten durch die notwendige kritische Bewertung der KI-Ausgaben auch als erhöhten Arbeitsaufwand. Zu ausführliche Textausgaben trugen zu dieser Wahrnehmung der Lehrkräfte bei, die sich

wiederum kurze und prägnante Formulierungen durch die KI-Agenten wünschten.

Die Analyse der Prompt-Typen zeigte zwei wesentliche Funktionen innerhalb der Interaktion zwischen Lehrkräften und KI-Agenten:

- *Interaktionssteuerung:* Lehrkräfte nutzen Prompts, um die Kommunikation mit den KI-Agenten zu leiten und relevante Ausgaben zu generieren. Diese Prompts (z. B. präziser Initialprompt, Konkretisierung des vorherigen Prompts) dienen der Steuerung der Interaktion zwischen Lehrkraft und KI-Agenten und legen den Rahmen für die KI-generierten Aufgaben fest. Die iterative Anpassung der Prompts macht deutlich, dass Lehrkräfte den Arbeitsprozess mit KI-Agenten häufig als reflektiert und schrittweise angelegt gestalten.
- *Didaktische Zielorientierung:* Die didaktischen Prompts zeigen, dass Lehrkräfte KI-Agenten nutzten, um Aufgaben zu gestalten, die sowohl auf curriculare Vorgaben als auch auf individuelle Lernbedürfnisse Bezug nehmen. Prompts wie die Anpassung der Differenzierung oder die Ergänzung von Lernzielen folgten dabei einer klaren fachdidaktischen Ausrichtung und ermöglichen es, die KI-Ausgaben präzise auf spezifische Unterrichtserfordernisse abzustimmen.

Die einmalige Kodierung des Prompts zur Anpassung der Aufgabenschwierigkeit lässt darauf schließen, dass Lehrkräfte KI-Agenten potenziell als Unterstützung bei der Individualisierung des Unterrichts in Betracht zogen. Gleichzeitig weist sie auf ein mögliches Einsatzfeld hin, in dem KI-Agenten zur Differenzierung im Unterricht beitragen könnten. Es sollte nicht (nur) darum gehen, Aufgaben auf verschiedenen Niveaustufen anzubieten, sondern vielmehr darum, eine aktive Interaktion mit dem LLM zu ermöglichen und konkrete Vorschläge zu diskutieren, wie mit diesen Differenzierungsvorschlägen gearbeitet werden kann.

Die Ergebnisse geben weiterhin Anlass zu folgenden praxisbezogenen Überlegungen und Handlungsempfehlungen, die sich unmittelbar aus den beobachteten Kompetenzen und Vorgehensweisen der Lehrkräfte ableiten:

- *Unterstützung bei der Anpassung generierter Inhalte:* Die Lehrkräfte zeigten, dass sie KI-Ausgaben kritisch prüfen und gezielt anpassen können. Tools sollten daher stärker darauf ausgerichtet sein, flexible Bearbeitungen und schnelle Modifikationen zu ermöglichen, um

diese Kompetenz effektiv zu unterstützen. So könnten Lehrkräfte beispielsweise selbstständiger und intuitiver den Leseaufwand reduzieren oder Anpassungen in der Bildgenerierung vornehmen, was eine effizientere Integration der KI-Technologien ins Task Design ermöglichen würde.

- *Förderung reflexiver Nutzungskompetenzen:* Die iterative Arbeit mit Prompts verdeutlicht die Notwendigkeit, den Einsatz von KI-Agenten als einen dynamischen und reflexiven Prozess und Teil der didaktischen Arbeit zu begreifen. Lehrkräfte sollten in Weiterbildungen darin unterstützt werden, kontinuierlich die generierten Ergebnisse zu analysieren und anzupassen, um den größtmöglichen Nutzen für die Unterrichtsentwicklung zu erzielen.
- *Ausbau von Prompt-Kompetenzen:* Erste Anwendungen von Prompt-Techniken verdeutlichen, dass Lehrkräfte Ansätze zur Steuerung von KI bereits erproben, diese aber noch nicht systematisch einsetzen. Hieraus ergibt sich die Empfehlung, entsprechende Fortbildungsangebote zu entwickeln, die den gezielten Einsatz von Prompt-Techniken schulen. Diese Maßnahme könnte Lehrkräfte dabei unterstützen, die identifizierten Herausforderungen, wie das Erfordernis präziser Anleitungen und die Notwendigkeit der kritischen Validierung von Inhalten, erfolgreich zu bewältigen.

Weiterhin konnten in Bezug auf die theoretische Rahmung Erkenntnisse in zwei zentralen Perspektiven erzielt werden: Erkenntnisse zur Lehrkraft-KI-Interaktion sowie Erkenntnisse zum mathematischen Task Design.

Im Bereich der *Lehrkraft-KI-Interaktion* zeigte sich, dass der Einsatz KI-gestützter Werkzeuge zur mathematischen Aufgabenerstellung eine inhaltliche Nutzung der KI-Ausgaben ermöglichen kann. Die Stärke kollaborativer Prozesse wird primär in der Integration multipler Perspektiven deutlich (Choppin, 2011), die innerhalb eines KI-unterstützten kollaborativen Teams zusammenwirken (vgl. Widmer, 2018). Dies fördert sowohl die vertiefte inhaltliche Auseinandersetzung als auch die Entwicklung reflexiver Praktiken. In der vorliegenden Studie zeigte sich diesbezüglich neben der Reflexion über mathematische Aufgaben auch die Reflexion der eigenen Rolle im Prozess des mathematischen Task Designs, die durch diese Form der Zusammenarbeit angeregt wurde. Zusätzlich wurde deutlich, dass KI-gestütztes mathematisches Task Design dazu bei-

tragen kann, Lücken in den Aufgabenvorschlägen oder im eigenen fachdidaktischen oder fachwissenschaftlichen Wissen zu identifizieren, die ohne eine Zusammenarbeit mit KI-Agenten möglicherweise unentdeckt geblieben wären.

In Bezug auf das *mathematische Task Design* wurde deutlich, dass die Pedagogical Design Capacity (Brown, 2009) durch die Arbeit mit den KI-Agenten gefordert und gleichzeitig gefördert werden kann. Die Lehrkräfte reflektieren und modifizieren ihre Aufgaben aktiv, was zu einer kontinuierlichen Weiterentwicklung des mathematischen und fachdidaktischen Wissens beiträgt (Pepin, 2015). Das Nachdenken über die Anpassung von Aufgaben erfordert und stärkt insbesondere analytische und adaptive Kompetenzen, die eine flexible und anpassungsfähige Herangehensweise ermöglichen und so zur professionellen Weiterentwicklung der Lehrkräfte beitragen (Sherin, 2002). Zudem wird deutlich, dass Kompetenzen zur strategischen Nutzung von Ressourcen auf individueller Ebene notwendig sind (Choppin, 2011). Dies umfasst – wie das induktiv entwickelte Kategoriensystem zeigt – technische, analytische, adaptive, didaktische und reflexive Kompetenzen, die im Umgang mit den KI-Agenten gleichermaßen gefordert und gefördert werden können. Zusammengefasst verdeutlichen die Ergebnisse, dass die Nutzung von KI-Agenten sowohl kollaborative und reflektive Praktiken in der Lehrkraft-KI-Interaktion als auch Kompetenzentwicklung und Weiterentwicklung mathematischer Aufgaben auf vielfältige Weise unterstützen kann.

Transkriptionslegende

Für eine gute Lesbarkeit sind die Transkripte so weit wie möglich der Schriftsprache angenähert. Es wurde folgende Formatierung genutzt:

- *kursiv*: Vorlesen der KI-Ausgabe
- (x Sek.): Pause von x Sekunden
- /: Wort- oder Satzabbruch
- [Handlung]: Beschreibung von Mimik, Gestik und Handeln

Danksagung

Wir danken den Lehrkräften, die sich die Zeit genommen haben, an unserer Studie teilzunehmen sowie den begutachtenden Personen für die hilfreichen Anmerkungen.

Literatur

- Abshagen, M. (2015). *Praxishandbuch Sprachbildung Mathematik. Sprachsensibel unterrichten – Sprache fördern*. Klett.
- Anthropic. (2024, June). *Claude 3 model card addendum*. Anthropic. <https://www.anthropic.com/news/claude-3-model-card-addendum>

- Bohl, T., Drücke-Noe, C., Hoppe, H., Kleinknecht, M., Maier, U. & Metz, K. (2015). Was bringt diese Aufgabe? Lehrerinnen und Lehrer analysieren das kognitive Potenzial von Aufgaben. *Pädagogik* 5(15), 28–31.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer.
- Brown, M. W. (2009). The teacher-tool relationship: Theorizing the design and use of curriculum materials. In J. T. Remillard, B. A. Herbel-Eisenmann & G. M. Lloyd (Hrsg.), *Mathematics teachers at work: Connecting curriculum materials and classroom instruction* (S. 17–36). Routledge.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Pohl, M. & Schorcht, S. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative KI-Sprachmodelle. *Mitteilungen der GDM*, 114, 19–26
- Buchholtz, N. & Huget, J. (2024). ChatGPT as a reflection tool to promote the lesson planning competencies of pre-service teachers. In E. Faggiano, A. Clark-Wilson, M. Tabach & H.-G. Weigand (Eds.), *Proceedings of the 17th ERME Topic Conference MEDA 4* (S. 129–136). University of Bari.
- Choppin, J. (2011). Learned adaptations: Teachers' understanding and use of curriculum resources. *Journal of Mathematics Teacher Education*, 14, 331–353. <https://doi.org/10.1007/s10857-011-9170-3>
- Cooper, G. & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79(4), 347–362. <https://doi.org/10.1037/0022-0663.79.4.347>
- DeepSeek-AI (2025). *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. arXiv. <https://arxiv.org/abs/2501.12948>
- Dilling, F., Herrmann, M., Müller, J., Pielsticker, F. & Witzke, I. (2024b). Initiating interaction with and about ChatGPT – an exploratory study on the angle sum in triangles. In E. Faggiano, A. Clark-Wilson, M. Tabach & H.-G. Weigand (Eds.), *Proceedings of the 17th ERME Topic Conference MEDA 4* (S. 145–152). University of Bari Aldo Moro.
- Drake, C. & Sherin, M. G. (2009). Developing curriculum vision and trust: Changes in teachers' curriculum strategies. In J. T. Remillard, B. Herbel-Eisenmann & G. Lloyd (Hrsg.), *Mathematics teachers at work: Connecting curriculum materials and classroom instruction* (S. 321–337). New York: Routledge.
- Doyle, W. (1983). Academic work. *Review of Educational Research*, 53(2), 159–199.
- Franzoni, V. (2023). From Black Box to Glass Box: Advancing Transparency in Artificial Intelligence Systems for Ethical and Trustworthy AI. In O. Gervasi et al. (Hrsg.), *Computa-*

- tional Science and Its Applications – ICCSA 2023 Workshops* (S. 118–130). Springer. https://doi.org/10.1007/978-3-031-37114-1_9
- Floridi, L. & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines*, 30, 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Gemini Team, Google (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. arXiv <https://arxiv.org/abs/2403.05530>
- Hauptman, A. I.; Schelble, B. G.; McNeese, N.J.; Madathil, K. C. (2023). Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming, *Computers in Human Behavior*, 138. <https://doi.org/10.1016/j.chb.2022.107451>
- Huang, S., Mamidanna, S., Jangam, S., Zhou, Y. & Gilpin, L. H. (2023). *Can large language models explain themselves? study of LLM-generated self-explanations*. arXiv. <https://arxiv.org/abs/2310.11207>
- Jones, K. & Pepin, B. (2016). Research on mathematics teachers as partners in task design. *Journal of Mathematics Teacher Education*, 19, 105–121. <https://doi.org/10.1007/s10857-016-9345-z>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. & Kasneci, G. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and individual differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kleinknecht, M. & Lankes, E.-M. (2012). Kompetenzvermittlung im Unterricht: Eine neue Lern- und Aufgabenkultur an der Schule etablieren. *Schulleitung und Schulentwicklung*, 57(2), 1–16.
- Konrad, Klaus (2010). Lautes Denken. In G. May & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (476–490). Springer.
- Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E. & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics tasks development? *Physical Review Physics Education Research*, 19(2), 1–14. <https://doi.org/10.1103/PhysRevPhysEducRes.19.020128>
- Kultusministerkonferenz (2024). *Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen*. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2024/2024_10_10-Handlungsempfehlung-KI.pdf
- Kumar, P. M. (2021). Special issue on Artificial Intelligence in Engineering Education. *Computer Applications in Engineering Education* 29 (2), 311–311. <https://doi.org/10.1002/cae.22398>
- Krauthausen, G. & Scherer, P. (2022). *Natürliche Differenzierung im Mathematikunterricht*. Kallmeyer.
- Leonardo de Pisa (1202). *Liber Abaci*.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2021). *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing*. arXiv. <https://arxiv.org/abs/2107.13586>
- Luckin, R., Holmes, W., Griffiths, M. & Forcier, L. B. (2016). *Intelligence Unleashed. An argument for AI in Education*. London: Pearson.
- 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>
- Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz.
- Maier, U., Bohl, T., Drücke-Noe, C., Hoppe, H., Kleinknecht, M. & Metz, K. (2014). Das kognitive Anforderungsniveau von Aufgaben analysieren und modifizieren können: Eine wichtige Fähigkeit von Lehrkräften bei der Planung eines kompetenzorientierten Unterrichts. *Beiträge zur Lehrerinnen- und Lehrerbildung*. 32(3), 340–358.
- MAXQDA (2022). *Software für qualitative Datenanalyse*. Sozialforschung GmbH.
- Meyer, J. W. & Rowan, B. (1992). The Structure of Educational Organizations. In J. W. Meyer & R. Scott (Hrsg.), *Organizational Environments: Ritual and Rationality* (S. 71–97). Sage.
- Mollick, Ethan R. & Mollick, L. (2023). Assigning AI: Seven Approaches for Students, with Prompts. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4475995>
- NCTM (n.d.) *Principles, Standards, and Expectations*. <https://www.nctm.org/Standards-and-Positions/Principles-and-Standards/Principles-Standards-and-Expectations/>
- Noster, N., Gerber, S. & Siller, H.-S. (2024). Pre-Service Teachers’ Approaches in Solving Mathematics Tasks with ChatGPT. *Digital Experiences in Mathematics Education* 10, 543–567. <https://doi.org/10.1007/s40751-024-00155-8>
- OpenAI. (2024, Oktober). *GPT-4o System Card*. arXiv. <https://arxiv.org/abs/2410.21276>
- Pal, A., Umapathi, L. K. & Sankarasubbu, M. (2023). *Med-HALT: Medical Domain Hallucination Test for Large Language Models*. arXiv. <https://arxiv.org/abs/2307.15343>
- Pepin, B. (2015). *Enhancing mathematics/STEM education: A ‘resourceful’ approach*. Technische Universiteit Eindhoven.
- Peters, F. & Schorcht, S. (2024). AI-supported Mathematical Task Design with a GPT Agent Network. In E. Faggiano, A. Clark-Wilson, M. Tabach, H.-G. Weigand. (Hrsg.), *Proceedings of the 17th ERME Topic Conference MEDA4*. University of Bari Aldo Moro.
- Polya, G. (1957). *How to solve it: A new aspect of mathematical method*. Princeton University Press.
- Sasaki, T. (2008). Concurrent think-aloud protocol as a socially situated construct. *International Review of Applied Linguistics in Language Teaching*, 46 (4), 349–374.
- Schorcht, S., Buchholtz, N. & Baumanns, L. (2024b). Prompt the Problem – Investigating the Mathematics Educational Quality of AI-Supported Problem Solving by Comparing Prompt Techniques. *Frontiers in Education*, 9. Article 1386075. <https://doi.org/10.3389/feduc.2024.1386075>
- Schorcht, S., Peters, F. & Kriegel, J. (2024c). Communicative AI Agents in Mathematical Task Design: A Qualitative Study of GPT Network Acting as a Multi-professional Team. *Digital*

Experience in Mathematics Education. Springer Nature.
<https://doi.org/10.1007/s40751-024-00161-w>

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2022). *Bildungsstandards für das Fach Mathematik Primarbereich*.
https://www.kmk.org/fileadmin/veroeffentlichungen_besc_hluesse/2022/2022_06_23-Bista-Primarbereich-Mathe.pdf

Sensevy, G., Forest, D., Quilio, S. & Morales, G. (2013). Cooperative engineering as a specific design-based research. *ZDM – Mathematics Education*, 45(7), 1031–1043.

Sherin, M. G. (2002). When Teaching Becomes Learning. *Cognition and Instruction*, 20(2), 119–150.
https://doi.org/10.1207/S1532690XCi2002_1

Sjuts, J. (2024). Sinn oder Unsinn? ChatGPT beim Lösen von Aufgaben und Problemen in Mathematik. *MNU-Journal* 77(2), 128–136

UNESCO/USAID (2020). *Global Proficiency Framework for Mathematics: Grades 1–9*. UNESCO Institute for Statistics & USAID.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Watson, A. & Ohtani, M. (2015). *Task design in mathematics education: An ICMI study* 22. Springer.
<https://doi.org/10.1007/978-3-319-09629-2>

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. & Zhou, D. (2022). *Self-consistency improves chain of thought reasoning in language models*. arXiv.
<https://arxiv.org/abs/2203.11171>

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. & Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv.
<https://arxiv.org/abs/2201.11903>

Werning, R. & Löser, J. M. (2010). Inklusion: aktuelle Diskussionslinien, Widersprüche und Perspektiven. *Die Deutsche Schule* 102(2), 103–114.

Widmer-Wolf, P. (2018). Kooperation in multiprofessionellen Teams an inklusiven Schulen. In T. Sturm & M. Wagner-Willi (Hrsg.), *Handbuch schulische Inklusion* (S. 298–313). Barbara Budrich.

Wittmann, E. (2010). Natürliche Differenzierung im Mathematikunterricht der Grundschule – vom Fach aus. In P. Hanke, G. Möwes-Butschko, A. Hein, D. Berntzen & A. Thielges (Hrsg.), *Anspruchsvolles Fördern in der Grundschule* (S. 63–78). Zentrum für Lehrerbildung.

Wolfram, S. (2023). *What is ChatGPT doing ... and why does it work?* Wolfram Media, Inc.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F. & Shi, S. (2023). *Siren's song in the AI ocean: A survey on hallucination in large language models*. arXiv.
<https://arxiv.org/abs/2309.01219>

Anschrift der Verfasser:innen

Franziska Peters
 Justus-Liebig-Universität-Gießen
 Institut für Didaktik der Mathematik
 Karl-Glöckner-Straße 21 C
 35394 Gießen
franziska.peters@uni-giessen.de

Sebastian Schorcht
 Technische Universität Dresden
 Fakultät für Erziehungswissenschaft
 Grundschulpädagogik/Mathematik
 Weberplatz 5
 01062 Dresden
sebastian.schorcht@tu-dresden.de