Einsatzmöglichkeiten von generativer künstlicher Intelligenz bei der mathematischen Fehleranalyse

NILS BUCHHOLTZ, HAMBURG; FRANZISKA PETERS, HAMBURG

***Zusammenfassung**: Generative künstliche Intelligenz (genKI) besitzt Potenzial für die Unterstützung von Lehrkräften bei Routineaufgaben. Diese Studie untersucht den Einsatz von genKI bei diagnostischen Tätigkeiten anhand von 200 simulierten fehlerhaften Aufgabenbearbeitungen zu schriftlichen Rechenoperationen. Die KI-basierten Fehleranalysen wurden zunächst mit einem Basis-Prompt durchgeführt und in einer zweiten Phase gezielt durch Prompt-Engineering optimiert. Insgesamt zeigen sich erste positive Ansätze, doch die Ergebnisse verdeutlichen, dass das untersuchte GenKI-System ohne spezifisches Training keine verlässliche Fehlerdiagnose leistet.*

***Abstract:** Generative artificial intelligence (genAI) has the potential to support teachers in routine tasks. This study investigates the use of genAI in diagnostic activities using 200 simulated errors of written arithmetic operations. The AI-based error analyses were first carried out once with a basic prompt and in a second phase specifically optimized using prompt engineering. Overall, initial positive approaches are emerging, but the results make it clear that the analyzed genAI system does not provide reliable error analysis without specific training.*

1. Einleitung

Die Bedeutung von Fehlern für das Lernen wird insbesondere in konstruktivistischen Lerntheorien hervorgehoben (Mindnich et al., 2008; Türling, 2014). Fehler der Lernenden zeigen, dass mathematische Inhalte noch nicht vollständig verstanden oder fehlerhaft verinnerlicht wurden. Sie sollten daher von Lehrkräften im Unterricht angemessen adressiert werden (Heinrichs, 2015). Ein konstruktiver Umgang mit einem Fehler durch die Lehrkraft, der diesen als Lernchance begreift, kann dann im weiteren Unterrichtsverlauf zu Lernfortschritten führen, das eigenständige Beheben des Fehlers unterstützen und langfristig u. a. das Entstehen von "negativem Wissen" bei Schülerinnen und Schülern fördern (Fritz-Schubert und Rohde, 2022; Steuer, 2014; Oser et al., 1999). Eine Unterrichtsumgebung, die Schülerinnen und Schüler dazu ermutigt, sich mit ihren Fehlern aktiv auseinanderzusetzen, wird dabei als *positive Fehlerkultur* oder *positives Fehlerklima* bezeichnet (Heinrichs, 2015; Steuer, 2014; Schoy-Lutz, 2005). Es ist jedoch zu beachten, dass nicht alle Fehler automatisch eine Lernchance darstellen, und das Lernen aus Fehlern nicht immer von selbst geschieht (Heinrichs, 2015). Grundlage bleibt deshalb die Diagnose und die Analyse von Fehlern durch die Lehrkraft. Lehrkräfte müssen im Unterrichtsgeschehen vorliegende Fehler erkennen und situativ entscheiden, ob es sich um einen produktiv nutzbaren Fehler handelt, der zum Aufbau von negativem Wissen genutzt werden kann (Oser et al., 1999; Schoy-Lutz, 2005). Daran anknüpfend muss eine Lehrkraft mit lernförderlichen Impulsen mittels ihrer didaktischen und sozialen Kompetenzen auf ihn reagieren (Schoy-Lutz, 2005; Spychiger et al., 1999), wobei fachliches sowie lerntheoretisches und schülerbezogenes diagnostisches Wissen erforderlich sind, da Lernende sehr individuell auf Fehler reagieren (Schoy-Lutz, 2005; Türling, 2014).

Innovative Ansätze zur Unterstützung von Lehrkräften bei der Diagnose und Fehleranalyse ergeben sich seit neuestem durch die Möglichkeit der Nutzung von KI-Technologien, die z. B. in der Form generativer Künstlicher Intelligenz (genKI) zunehmend in den Bildungsbereich integriert werden (Pepin et al., 2025). GenKI-Systeme und Technologien des maschinellen Lernens versprechen insbesondere durch ihre Unterstützung bei der Diagnose und Bewertung mathematischer Leistungen (von vielen Schülerinnen und Schülern gleichzeitig) potenziell eine Entlastung von Lehrkräften (Giannakos et al., 2025; Tan et al., 2025; Weegar & Idestam-Almquist, 2024), so dass Lehrkräfte Zeit gewinnen, ihren Unterricht entsprechend den Bedürfnissen der Schülerinnen und Schüler anzupassen (KMK, 2024; Hankeln et al., 2025). Sie können beispielsweise in intelligenten Tutorsystemen (ITS) Fehleranalysen automatisieren oder im Rahmen von KI-basierten Lernplattformen Lehrkräften detaillierte Lernprozessdaten bereitstellen, die zur formativen Leistungsbewertung herangezogen werden können (Deutscher Ethikrat, 2023; Zeller & Schmid, 2017). Derartige Systeme versprechen u. a. durch das spezifische Training anhand von Daten (Rittle-Johnson et al., 2025; Kim et al., 2026) die Identifizierung konzeptueller Fehlertypen sowie die Echtzeit-Analyse von Schülerlösungen und könnten so langfristig zu einer verbesserten Individualdiagnostik beitragen. Zudem bieten sog. multimodale KI-Modelle bereits jetzt neue Möglichkeiten der

Echtzeit-Lernassistenz, indem sie adaptiv auf Lernausgangslagen und Fehler der Schülerinnen und Schüler reagieren können (vgl. Khan Academy, 2024; Giannakos et al., 2025).

Gleichzeitig ergeben sich aus der Nutzung von KI-Technologien im Bereich der Diagnose neue Herausforderungen. Wie erste Studien in diesem Bereich zeigen, bleibt die Diagnosequalität aufgrund mangelnder Treffsicherheit bislang noch begrenzt (z. B. Bewersdorff et al., 2023; Voßhagen, 2025). Darüber hinaus müssen Lehrkräfte über die notwendige Kompetenz verfügen, um die durch KI-Systeme bereitgestellten Lernprozessdaten kritisch zu hinterfragen und ihre pädagogischen Entscheidungen nicht unreflektiert auf KI-gestützte Fehleranalysen zu stützen (Kasneci et al., 2023; Buchholtz et al., 2024). Der diagnostische Einsatz von KI-Technologie birgt zudem ethische und rechtliche Probleme: Die automatisierte Bewertung von Schülerleistungen kann algorithmische Verzerrungen enthalten (Baker & Hawn, 2021; Luzano, 2024), weshalb derzeit ein breiter Konsens darüber besteht, dass die abschließende Bewertung von Schülerleistungen immer in menschlicher Hand bleiben sollte (U.S. Department of Education, 2023; SWK, 2024; KMK, 2024). Unter anderem stuft aus diesem Grund auch die EU-KI-Verordnung von 2024 KI-Systeme zur Lernergebnisbewertung als hochriskant ein. Neuere Überlegungen der KMK weisen allerdings auf eine generelle Offenheit gegenüber Teilautomatisierungen von Leistungsmessung zur Unterstützung von Lehrkräften hin (KMK, 2024).

Während die Potenziale von KI-Technologien für diagnostische Aufgaben im Bildungsbereich intensiv diskutiert werden, liegen bislang nur wenige empirische Untersuchungen zur Analyse mathematischer Schülerfehler durch genKI-Systeme vor (Voßhagen, 2025; Kim et al., 2026). Insbesondere ist bislang wenig darüber bekannt, wie zuverlässig aktuelle multimodale Sprachmodelle schriftliche Schülerlösungen erkennen, Fehlerphänomene identifizieren und daraus diagnostische Schlussfolgerungen ableiten können. Bisherige Arbeiten fokussieren dabei vor allem auf die mathematische Leistungsfähigkeit generativer KI-Systeme oder ihre Nutzung als Lernunterstützung, während ihr Potenzial zur Analyse mathematischer Schülerfehler bislang deutlich seltener untersucht wurde. Die vorliegende Studie untersucht daher explorativ die Eignung eines exemplarischen genKI-Systems (ChatGPT) für die Analyse schriftlicher Schülerfehler im Bereich der schriftlichen Rechenverfahren. Im Mittelpunkt steht die Frage, inwieweit das System Fehlerphänomene, Fehlermuster, Fehlerursachen und mögliche Lehrkraftinterventionen identifizieren kann sowie welchen Einfluss gezieltes Prompt-Engineering – also die gezielte Manipulation der schriftlichen Befehlseingabe in das genKI-System – auf diese Analysen besitzt. Dazu ist es zunächst erforderlich, innerhalb des theoretischen Hintergrunds den Diagnoseprozess von Fehlern anhand der Betrachtung der fehlerdiagnostischen Kompetenz von Lehrkräften umfassender herauszuarbeiten (Abschnitt 2.1) sowie den aktuellen Diskurs zu allgemeinen Potenzialen und Grenzen der Nutzung von genKI-Systemen im Lehren und Lernen von Mathematik sowie speziell ihrer mathematischen Leistungsfähigkeit nachzuzeichnen (Abschnitt 2.2).

2. Theoretischer Hintergrund

2.1 Die fehlerdiagnostische Kompetenz von Lehrkräften

Schoy-Lutz (2005) sieht in der Analyse von Fehlersituationen durch die Lehrkraft eine wesentliche Grundlage für die Entwicklung einer positiven Fehlerkultur. In einer solchen Kultur bieten Fehler die Möglichkeit, die Denkstrukturen von Schülerinnen und Schülern sichtbar zu machen. Einem wiederholt auftretendem Fehlermuster können verschiedene Auslöser oder Ursachen zugrunde liegen (Fritz, 2021, Heinrichs, 2015). In der mathematikdidaktischen Forschung wird die Identifikation dieser Fehlermuster und Fehlerursachen als Fehleranalyse bezeichnet (Fritz, 2021; Radatz, 1980; Schmassmann, 1992; Jensen & Gasteiger, 2019).

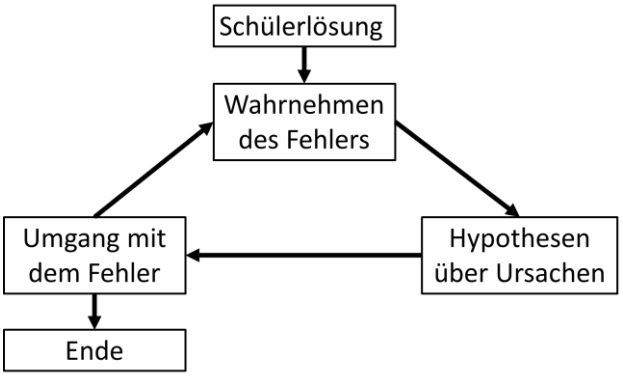


Abb. 1: Prozessmodell der fehlerdiagnostischen Kompetenz (nach Heinrichs, 2015, S. 66)

Heinrichs (2015) beschreibt die Fähigkeit von Lehrkräften, Fehlvorstellungen bei Schülerinnen und Schülern zu erkennen und zu identifizieren, als einen Bestandteil ihrer diagnostischen Kompetenz. Diese Kompetenz stellt eine grundlegende Voraussetzung dar, um Fehler im Unterricht effektiv als Lernchance

adressieren zu können. Ihr in der Forschung breit rezipiertes Prozessmodell stellt insbesondere die Bereiche Fehlerwahrnehmung, das Finden von Ursachenhypothesen und den Umgang mit dem Fehler heraus (vgl. Abb. 1).

Der Fehlerdiagnoseprozess stellt aus verschiedenen Gründen eine erhebliche Herausforderung für Lehrkräfte dar und macht ihn damit zu einer komplexen professionellen Anforderung. In einem Klassenraum ereignen sich zahlreiche Situationen gleichzeitig, an denen individuelle Schülerinnen und Schüler mit unterschiedlichen Bedürfnissen beteiligt sind. Eine Lehrkraft muss in Fehlersituationen all diese Ereignisse parallel beobachten und gleichzeitig in der Lage sein, zeitnah angemessen auf Fehler mit einer geeigneten Lehrkräfteintervention zu reagieren (vgl. Heinrichs, 2015).

2.1.1 Wahrnehmen des Fehlers

Zu Beginn des Prozesses muss der Fehler in Äußerungen oder schriftlichen Ausführungen von Schülerinnen und Schülern zunächst erkannt werden. Dies geschieht, indem die Lehrkraft eine Abweichung von einer Norm oder einem Ziel identifiziert, beziehungsweise das geschriebene Ergebnis als fehlerhaft bewertet (Steuer, 2014; Heinrichs, 2015). Da in der Mathematik der Soll-Zustand eindeutig durch Normen und Ziele festgelegt ist, sollte ein abweichender Ist-Zustand von einer Lehrkraft, die einen Wissensvorsprung und die Deutungshoheit gegenüber Schülerinnen und Schülern besitzt, grundsätzlich erkannt werden (Seifried & Wuttke, 2010).

Nach Radatz (1980) sind Fehler insbesondere bei algorithmischen oder teilautomatisierten Verfahren im Mathematikunterricht – wie etwa im Bereich der vier schriftlichen Rechenverfahren – relativ deutlich zu erkennen (vgl. auch Gerster, 1982). Denn diese Verfahren basieren auf „(…) einer endlichen Folge von eindeutig bestimmten Elementaranweisungen, die den Lösungsweg eines Problems exakt und vollständig beschreiben“ (Ziegenbalg, 2023, S. 344). Diese genannten Elementaranweisungen können jedoch je nach Kontext des betrachteten Problems variieren. In dieser Phase ist daher ein fundiertes Fachwissen der Lehrkräfte über typische Fehlerarten, die in dem jeweiligen Themengebiet auftreten können, von entscheidender Bedeutung, das es ihnen ermöglicht, die unterliegenden Denkprozesse und Argumentationsweisen der Schülerinnen und Schüler schneller nachzuvollziehen (Heinrichs, 2015; Seifried & Wuttke, 2010).

2.1.2 Hypothesen über die Ursache

Den Hauptteil des Diagnoseprozesses bildet das anschließende Aufstellen von Hypothesen über die Fehlerursache (Heinrichs, 2015). Dieser Schritt ermöglicht es, die zugrunde liegenden mentalen Modelle der Schülerinnen und Schüler zu erfassen. Dabei ist zu beachten, dass ein Fehler durch unterschiedliche oder sogar mehrere Ursachen ausgelöst werden kann. Um die Ursachen genauer zu analysieren, kann es hilfreich sein, Fehler anhand verschiedener Klassifikationen einzuordnen (Heinrichs, 2015; Lung, 2020).

In der Literatur werden unterschiedliche *Fehlerklassifikationen* beschrieben, die auf verschiedenen Kategorisierungsbedingungen basieren (Lung 2020). Radatz (1980) nennt u. a. Klassifizierungen nach systematischen und unsystematischen Fehlern, stoffdidaktischen Aspekten, äußeren Erscheinungsbildern sowie Ursachenbeschreibungen. Er argumentiert außerdem, dass eine eindeutige Bestimmung von Fehlerursachen nicht immer möglich ist, da Schülerfehler durch eine Vielzahl von Faktoren ausgelöst werden können. Dazu zählen z. B. die Lehrkraft, der Lehrplan, das schulische Umfeld, die am Unterrichtsprozess beteiligten Variablen sowie das Zusammenspiel all dieser Einflussfaktoren (vgl. auch Türling, 2014). Zudem können die Ursachen nicht strikt voneinander getrennt werden, da sie in einer Wechselbeziehung zueinanderstehen (vgl. Radatz, 1980).

Um Fehlerursachen genauer bestimmen und eingrenzen zu können, können verschiedene Methoden zur Ursachenfindung genutzt werden, die durch direkte oder indirekte Interaktion mit den Schülerinnen und Schülern umgesetzt werden können (vgl. Heinrichs, 2015; Shaughnessy et al., 2021). Radatz (1980) nennt vier zentrale Methoden, mit denen Lehrkräfte, die für eine konstruktive Reaktion auf Fehler notwendigen Informationen gewinnen können: (1) die Analyse schriftlicher Arbeiten und Lösungen, (2) diagnostische Interviews oder informelle Gespräche zwischen Lehrkraft und Schülerin bzw. Schüler, (3) die Verbalisierung der Denkprozesse durch lautes Denken oder (4) die Beobachtung während der Bearbeitung einer Aufgabe. Am effektivsten erscheint jedoch die Kombination dieser Methoden, da sie eine umfassende Informationsgewinnung und präzise Fehleranalyse erlaubt (vgl. ebd.). Im Rahmen dieser Studie werden allerdings nur schriftliche Fehler analysiert.

2.1.3 Umgang mit dem Fehler

Nach Abschluss der Ursachenfindung folgt gemäß Heinrichs (2015) in der Regel eine Reaktion der Lehrkraft (Rach et al., 2012; Gardee & Brodie, 2022). Wurde mit einer passenden Lehrerintervention das Ziel der Einsicht in den Fehler und seine Behebung erreicht schließt das dreischrittige Prozessmodell mit der Phase *Umgang mit dem Fehler* ab, anderenfalls können aus dem Umgang mit dem Fehler und der entsprechenden Schülerreaktion Informationen zur weiteren Wahrnehmung des Fehlerphänomens und zur Interpretation seiner Ursachen geschlossen werden. Ein zielführender Fehlerumgang ermöglicht es Schülerinnen und Schülern, aus ihren Fehlern zu lernen. Türling (2014) beschreibt dafür zentrale Qualitätsmerkmale: Die fachliche Richtigkeit der Lehrkraftäußerungen ist essenziell, um Fehler gezielt aufzugreifen. Ebenso sind eine strukturierte Fehlerbearbeitung sowie kognitive Aktivierung – z. B. durch kognitive Konflikte – entscheidend für den Lernprozess. Zudem sollte die Fehlerverarbeitung adaptiv und vernetzt gestaltet werden, indem sie an das Vorwissen der Lernenden anknüpft und lebensweltliche Bezüge herstellt. Abschließend trägt eine gezielte Sicherung durch Übungen oder Fehleranalysen zur nachhaltigen Korrektur bei.

Wie Lehrkräfte Fehler im Unterricht behandeln, hängt von ihren individuellen Dispositionen ab. Neben fachlichem und fachdidaktischem Wissen der Lehrkraft ist die Berücksichtigung der Lernvoraussetzungen der Schülerinnen und Schüler entscheidend (Spychiger et al., 1999). Darüber hinaus beeinflusst die Fehlerperspektive das Handeln: Während einige Lehrkräfte Fehler als wertvolle Lerngelegenheiten sehen, neigen andere aus der Sorge vor Fehlprägungen zu einer Fehlervermeidungsdidaktik (Türling, 2014). Ein konstruktiver Fehlerumgang erfordert somit sowohl didaktisches Wissen als auch eine reflektierte Fehlerkultur.

2.2 Entwicklung und Potenzial von genKI für das Lehren und Lernen von Mathematik

Im März 2023 veröffentlichte OpenAI die erste Version von GPT-4 und stellte sie im Juli der breiten Öffentlichkeit zur Verfügung (vgl. OpenAI, 2023). Das (im Rahmen dieser Anfang 2024 durchgeführten Studie eingesetzte) Modell GPT-4 zeichnet sich insbesondere durch seine multimodale Fähigkeit aus, neben Texteingaben auch visuelle Daten wie Fotos und Diagramme zu verarbeiten, wodurch sich prinzipiell eine Möglichkeit ergibt, auch schriftliche Darstellungen von Schülerlösungen analysieren zu lassen. Die Potenziale und Grenzen von genKI-gestützten Chatbots wie ChatGPT für den Einsatz im mathematischen Lehren und Lernen werden zunehmend weiter erforscht (Buchholtz et al. 2024; Pepin et al., 2025; Fock & Siller, 2025). Erste Ansätze zeigen, dass genKI-basierte Systeme wie ChatGPT mathematische Konzepte veranschaulichen, personalisierte Lernpfade ermöglichen und adaptive Lernunterstützung bieten können (Buchholtz et al., 2023; Schorcht et al., 2023; Kortenkamp & Dohrmann, 2023). Sind die datenschutzrechtlichen Bestimmungen geklärt und der Einsatz im Unterricht prinzipiell möglich, können Lernende beispielsweise durch Interaktion mit genKI-Systemen ihre kommunikativen und argumentativen Fähigkeiten im mathematischen Diskurs verbessern (Yoon et al., 2024; Dilling & Herrmann, 2024). Zudem erlaubt es KI-generiertes Feedback, eigene Fehler zu erkennen und Lösungswege zu überarbeiten (Zhu et al., 2020). Allerdings birgt der Einsatz von genKI auch Herausforderungen. Ein zentrales Risiko besteht darin, dass Schülerinnen und Schüler sich zu stark auf Chatbots verlassen, wodurch kritisches Denken und Problemlösefähigkeiten eingeschränkt werden könnten (Buchholtz et al., 2024; Kasneci et al., 2023). Auch für Lehrkräfte bietet die Technologie Unterstützung, etwa bei der Erstellung von Unterrichtsplänen oder der Differenzierung von Mathematikaufgaben (Huget & Buchholtz, 2024; Buchholtz & Huget, 2024; Cameron & Mesiti, 2024). GenKI kann dabei helfen, geeignete Lerninhalte auszuwählen, Arbeitsanweisungen zu optimieren und durch gezielte Prompts individuelle Bedürfnisse der Schülerinnen und Schüler zu berücksichtigen – beispielsweise durch die Angabe der jeweiligen Klassenstufe (vgl. Mohr et al., 2023).

2.2.1 Mathematische Fähigkeiten von genKI

Im Kontext dieser Studie steht u. a. die mathematische Leistungsfähigkeit von ChatGPT im Fokus, da eine Diagnose von Schülerfehlern zentral die Bewertung der mathematischen Korrektheit bzw. die Abweichung der Norm beinhaltet. Diese spezifischen Fähigkeiten sind im Rahmen von unterschiedlichen Benchmarking-Untersuchungen bereits Gegenstand von wissenschaftlichen Studien. Diese zeigen durchgängig, dass neuere GPT-Versionen wie GPT-4 mittlerweile zahlreiche (schul-)mathematische Probleme korrekt lösen können, einschließlich komplexer Aufgaben (Schorcht et al., 2023; Schorcht et al., 2024bPepin et al., 2025). Die Verlässlichkeit der Modelle bei komplexeren Aufgaben bleibt jedoch weiterhin begrenzt, da bereits bei einfachen Aufgaben Fehler auftreten können (sog. Datenhalluzinatio-

nen). Ein kritisches Hinterfragen von mathematischen Ergebnissen und die menschliche Überprüfung von KI-Output bleiben daher bis auf Weiteres unerlässlich (Buchholtz et al., 2024).

Um dieses Problem langfristig zu lösen, werden neben allgemeinem Trainingsmaterial aus mathematischen Theorien, Beweisen und Anwendungen in aktuellen Studien zu den mathematischen Fähigkeiten von genKI spezialisierte Datensätze genutzt, um die mathematische Leistungsfähigkeit von genKI-Modellen gezielt durch Feintuning zu verbessern. Diese Datensätze kombinieren natürliche Sprachbeschreibungen mit Python-Code und werden aus mathematischen Webdaten, LaTeX-Quellen, synthetischen Daten und Fachliteratur gespeist (vgl. Liu et al., 2023; Lu et al., 2024; Paster et al., 2023; Yang et al., 2023). Liu et al. (2023) kategorisierten in einem Überblicksartikel über 60 mathematische Datensätze für Training, Benchmarking und spezifische Aufgaben. Besonders Modelle, die auf einer Kombination aus Code und natürlicher Sprache trainiert wurden, zeigen deutliche Fortschritte in mathematischem Schlussfolgern und Präzision und sind in der Lage, mehrstellige Rechenoperationen mit nahezu perfekter Genauigkeit zu bewältigen (Yang et al., 2023).

2.2.2 GenKI und mathematische Fehleranalyse

Generative KI-Modelle wie GPT-4 zeigen zwar im Rahmen diagnostischer Tätigkeiten vielversprechende Ansätze zur Fehleranalyse, weisen jedoch charakteristische Limitationen auf. Ein zentrales Problem ist falsches und fehlerhaftes Feedback. Dies liegt u. a. daran, dass viele Modelle den gesamten Diagnoseprozess in einem einzigen Schritt abwickeln (Daheim et al., 2024). Zudem erkennen genKI-Systeme einfache Rechenfehler oft zuverlässig, sind jedoch bei der Identifikation komplexer Fehlvorstellungen überfordert. Während in Studien für klar definierte Fehlertypen durchaus auch hohe Trefferquoten erzielt wurden, schnitten KI-Modelle bei tiefgehenden konzeptionellen Missverständnissen deutlich schlechter ab (Kim et al., 2026; Voßhagen, 2025; Frieder et al., 2023, Bewersdorff et al., 2023). Die automatische Diagnosequalität hängt daher stark vom Fehlerkontext und der Problemstellung ab (Schorcht et al., 2024a). Vergleicht man die Diagnosequalität von genKI-Systemen mit der menschlicher Lehrkräfte, zeigt sich ein gemischtes Bild. In standardisierten Szenarien mit klar definierten Fehlermustern kann die KI mit Lehrkräften mithalten oder sie sogar übertreffen (Bewersdorff et al., 2023). KI-Modelle stoßen allerdings an Grenzen, wenn Schülerinnen und Schüler Zwischenschritte überspringen oder alternative Notationen verwenden, da dies zu Fehlinterpretationen führen kann.

Um diese Schwächen zu beheben, werden aktuell verschiedene Verbesserungsansätze erforscht. Multimodale KI-Modelle, die neben Text auch handschriftliche Lösungen, Diagramme oder gesprochene Erklärungen analysieren, könnten die Diagnosequalität signifikant steigern. Ebenso zeigen Feintuning-Ansätze mit Bildungsdaten, dass KI-Systeme durch gezielte Trainingsdatensätze in der Fehleranalyse verbessert werden können (Jin et al., 2024; Rittle-Johnson et al., 2025). Schließlich können beispielsweise zweistufige Verifikationsprozesse in KI-Systemen die Genauigkeit der Rückmeldungen verbessern, indem Fehler zuerst systematisch analysiert werden, bevor eine Rückmeldung erfolgt (Daheim et al., 2024).

2.2.3 Prompt-Engineering

Auch durch gezieltes Prompt-Engineering konnte in Studien die Genauigkeit von mathematischem Output gesteigert werden. Frieder et al. (2023) stellten in einer Analyse der Fähigkeiten von ChatGPT mathematische Beweise zu formulieren und komplexe Aufgaben zu lösen fest, dass die Länge eines Prompts keinen signifikanten Einfluss auf dessen Leistung und Ausgabe hatte. Hingegen zeigte sich, dass eine explizite Aufforderung zu einer schrittweisen Vorgehensweise („Step-by-Step"-Handlung) einen positiven Effekt besaß. Dieses spezifische Prompt-Engineering kann die Häufigkeit bestimmter Fehler der KI verringern und somit zu besseren Ergebnissen führen (vgl. Frieder et al., 2023).

Diese Erkenntnis deckt sich mit den Annahmen von Kojima et al. (2023), die beschreiben, dass genKI-Systeme durch *Few-Shot-Learning* erfolgreicher als beim *Zero-Shot-Learning* agieren. Beim Few-Shot-Learning werden die zugrundeliegenden KI-Sprachmodelle mit Beispielaufgaben konfrontiert, um ihr Vorgehen daran auszurichten, während beim Zero-Shot-Learning lediglich die zu erledigende Aufgabe beschrieben wird, ohne dass Beispiele gegeben werden. Allerdings scheint Zero-Shot- und Few-Shot-Learning nicht ausreichend zu sein, wenn für eine Aufgabe mehrstufige Denkprozesse erforderlich sind (vgl. ebd.). In solchen Fällen empfehlen Kojima et al. (2023) das sogenannte *Chain-of-Thought-Prompting*. Dabei wird das KI-Sprachmodell explizit dazu angeleitet, seine Gedankengänge kleinschrittig darzulegen, um dadurch bessere Ergebnisse zu er-

zielen (vgl. ebd.). Diese positive Wirkung wurde anhand arithmetischer Aufgaben nachgewiesen: Zunächst wurden die Aufgaben mit einem Zero-Shot-Prompt bearbeitet, anschließend wurde dem Prompt die Anweisung "*Let's think step by step"* hinzugefügt, um die Ergebnisse zu vergleichen. Es zeigte sich, dass Aufgaben, an denen KI-Sprachmodelle mit Zero-Shot-Learning scheiterten, durch Chain-of-Thought-Prompting erfolgreich gelöst werden konnten (vgl. ebd.). Allerdings führte diese Methode nicht zu einer Verbesserung der Leistung, wenn die Aufgaben allgemeines menschliches Urteilsvermögen erforderten (vgl. ebd.). Ähnliche Ergebnisse erzielten Schorcht et al. (2023) und Schorcht et al. (2024b), die gezielt den Einfluss von Prompt-Engineering auf die mathematischen Ausgaben von ChatGPT im Rahmen von Problemlöseaufgaben untersuchten. In ihrer methodischen Vorgehensweise kombinierten sie die Techniken des Zero-Shot- und Few-Shot-Learnings mit dem Chain-of-Thought-Prompting sowie einer angepassten Variante des Ask-me-Anything-Promptings (vgl. Schorcht et al., 2024b): Durch den Zusatz *„Stelle notwendige Fragen, die du zur Beantwortung der Frage benötigst."* wurde ChatGPT dazu veranlasst, eigenständig Rückfragen zu formulieren.

3. Methodisches Vorgehen

3.1 Forschungsdesign und Forschungsfragen

Durch die vorliegende explorativ angelegte Studie soll vor diesem Hintergrund eine Antwort auf folgende Fragen gefunden werden:

FF1) Wie zuverlässig sind die Fehleranalysen arithmetischer Rechenfehler durch ChatGPT?

und

FF2) Welchen Effekt hat gezieltes Prompt-Engineering auf diese Fehleranalysen?

Um erste wissenschaftliche Erkenntnisse hinsichtlich dieser Fragen zu ermöglichen, wurde der Forschungsprozess in zwei Phasen unterteilt (Abb. 2):

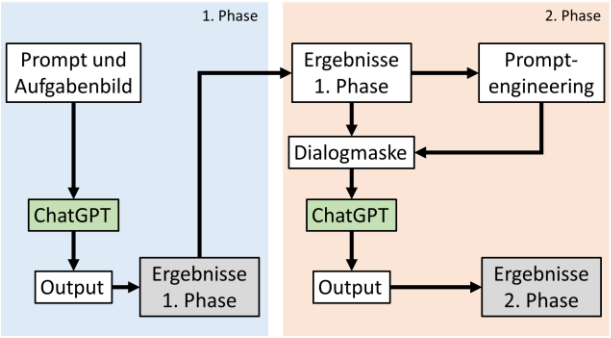


Abb. 2: Forschungsdesign

Zunächst wurde in der ersten Phase (blauer Abschnitt in Abb. 2) die grundlegende Leistungsfähigkeit von ChatGPT in Bezug auf Fehleranalysen im Bereich der vier schriftlichen Rechenarten ohne spezifische Anpassungen des Prompts (d. h. unter den Bedingungen des Zero-Shot Prompting) untersucht. Dabei wurde in ChatGPT pro Rechenart jeweils eine Abbildung mit fünf handschriftlichen Schülerfehlern zusammen mit einem Prompt eingegeben, der eine Aufforderung zur Analyse des Fehlers enthält. Pro Rechenart (Addition, Subtraktion, Multiplikation, Division) wurde dieser Vorgang fünfmal in jeweils neuen Chatverläufen wiederholt. Der generierte Output wurde anschließend qualitativ ausgewertet, um für FF1 erste Aussagen über die Qualität der Fehleranalysen von ChatGPT sowie potenziell unterstützende Faktoren für eine erfolgreiche Analyse zu treffen.

In der zweiten Phase (roter Abschnitt in Abb. 2) sollte in FF2 darauf aufbauend der Effekt von gezieltem Prompt-Engineering analysiert werden. Hierfür wurden aus verschiedenen Studien vorliegende Erkenntnisse zum Prompt-Engineering (vgl. Abschnitt 2.2.3) mit den Ergebnissen aus der ersten Phase kombiniert, um einen initialen Entwurf einer „Dialogmaske" zu entwickeln, darunter verstehen wir eine festgelegte Eingabestruktur für das genKI-System, in der lediglich die zu analysierenden Fehlerbeispiele variabel gehalten werden. Diese Dialogmaske hatte das Ziel, ChatGPT schrittweise durch die Analyse der Schülerfehler zu führen (Schorcht et al., 2024b; Daheim et al., 2024). Mittels mehrerer Testdurchläufe wurde die Dialogmaske optimiert. Mit der finalisierten Dialogmaske wurden alle Schülerfehler erneut durch ChatGPT analysiert, und die generierten Ergebnisse wurden anschließend erneut qualitativ ausgewertet.

3.2 Eingesetzte Fehler und Prompts

Für die Auswahl der Schülerfehler, die von ChatGPT analysiert werden sollten, wurden Untersuchungen, wie die von Frieder et al. (2023) herangezogen, die nahelegen, dass arithmetische Aufgaben für den Chatbot besonders leicht lösbar sind. Die Auswahl der analysierten Rechenfehler wurde dementsprechend auf die vier Grundrechenarten begrenzt und orientierte sich einerseits an klassischen Arbeiten zu Schülerfehlern bei schriftlichen Rechenverfahren (Gerster, 1982), andererseits an neueren mathematikdidaktischen Überlegungen zu Fehlerphänomenen, Fehlermustern und Fehlerursachen (Jensen & Gasteiger, 2019; Nelson & Powell, 2018; Rong & Mononen, 2022). Flüchtigkeitsfehler wurden bewusst

ausgeschlossen, da diese nur etwa 10–30 % der auftretenden Fehler bei Schülerinnen und Schüler ausmachen (vgl. Fritz, 2021). Aus den vier Grundrechenarten wurde schließlich jeweils eines der dort am häufigsten vorkommenden Fehlermuster ausgewählt. Verschiedene Studien nenn hier unterschiedliche Verfahrensfehler, die gehäuft auftreten können (z. B. Jensen & Gasteiger, 2019; Nelson & Powell, 2018). Gerster (1982) beschreibt, dass ein Großteil der Fehler bei der schriftlichen Addition auf Fehler im Einsundeins zurückzuführen ist, die durch das fehlerhafte Zusammenrechnen zweier Ziffern sichtbar werden. Für die Studie wurde daher ein Fehler gewählt, der diese systematische fehlerhafte Berechnung darstellt. Bei der Subtraktion ist ein häufiges Fehlermuster der falsche Umgang mit dem Übertrag, etwa wenn dieser nicht berücksichtigt wird (vgl. ebd.), weshalb dieser Fehler in der Studie verwendet wurde. Bei der schriftlichen Multiplikation zeigen die Ergebnisse von Gerster, dass 30–45 % der Fehler auf Probleme im Einmaleins zurückzuführen sind, wobei die Hälfte dieser Fehler eine inkorrekte Multiplikation mit der Null betrifft (vgl. ebd.). Wir haben daraufhin in den Beispielen der schriftlichen Multiplikation der Studie durchgängig diesen Fehler verwendet. Ein typisches Fehlermuster bei der schriftlichen Division ist ein Fehler beim Berechnen eines Teilprodukts, was oft auf Fehler in der Multiplikation zurückzuführen ist (vgl. ebd.), auch diesen Fehler haben wir ausgewählt und anhand von Beispielberechnungen operationalisiert.

Für die Untersuchung wurden die ausgewählten Fehlermuster jeweils durch fünf Fehlerphänomene in unterschiedlichen Rechenaufgaben repräsentiert, die speziell für diesen Zweck artifiziell erstellt wurden. Die Rechenfehler wurden pro Rechenart GPT-4 jeweils fünfmal hintereinander in Form eines visuellen Bildes präsentiert (siehe Anhang 1), um zufälligen Schwankungen in der Konsistenz der Fehleranalysen vorzubeugen. Auch der Aspekt, dass im Schulalltag Schülerinnen und Schüler unterschiedliche Handschriften haben und verschiedene Schreibgeräte oder Heftarten nutzen, wurde in der Untersuchung berücksichtigt. Zusätzlich wurden die zu analysierenden Rechenfehler in zwei unterschiedlichen Ausführungen präsentiert (Version A: Kugelschreiber auf kariertem Papier; Version B: Bleistift auf einem weißen Blatt), so dass insgesamt 200 simulierte Aufgabenbearbeitungen (4 Rechenarten × 5 Fehlerphänomene × 5 Durchläufe × 2 Versionen) analysiert wurden. Da in der ersten Forschungsphase vollständig auf Prompt-Engineering verzichtet wurde, sollte der in dieser Phase verwendete Prompt ausschließlich eine klare Instruktionsanweisung enthalten, die darauf abzielt, dass GPT-4 einen Fehler im Sinne des Fehlerdiagnoseprozesses nach Heinrichs (2015) analysiert. Unter Berücksichtigung dieser Anforderungen wurde folgende Prompt-Formulierung (P1) für die erste Forschungsphase gewählt:

P1: "Analysiere die auf dem Bild zu sehenden Rechenfehler bezüglich des Fehlermusters, der Fehlerursache und nenne passende Interventionsmöglichkeiten der Lehrkraft."

Die Anpassungen der eingesetzten Prompts in der zweiten Phase schildern wir aufgrund der besseren Nachvollziehbarkeit im Ergebnisteil (Abschnitt 5.2).

3.3 Datenauswertung

Eine qualitative Inhaltsanalyse ermöglicht es, die vorliegenden Texte (d. h. die Chatverläufe) regelgeleitet und nachvollziehbar zu untersuchen. Die Methode eignet sich insbesondere für die Untersuchung sprachlicher Daten, bei denen vorab definierte Analysekategorien mit offen auftretenden Phänomenen im Material in Beziehung gesetzt werden sollen (Mayring, 2010). In dieser Studie wird die Anwendung der zusammenfassenden und strukturierenden Inhaltsanalyse kombiniert (vgl. Mayring 2010). Ziel ist es, generelle Aussagen über alle gesammelten Daten treffen zu können und gleichzeitig Zusammenhänge zwischen verschiedenen Kategorien aufzudecken. Die Auswertung erfolgte mithilfe der Software MAXQDA. Das Kategoriensystem wurde zunächst deduktiv aus dem theoretischen Rahmen entwickelt und im Auswertungsprozess bei Bedarf präzisiert. Zur Absicherung der Nachvollziehbarkeit und intersubjektiven Verständigung wurde ein Teil des Datenmaterials zusätzlich von einer zweiten Person unabhängig kodiert. Anschließend wurden die Kodierungen verglichen und auftretende Unterschiede diskutiert, bis ein gemeinsames Verständnis der Kategorien erzielt wurde. Die Ergebnisse dieser Diskussion flossen in die weitere Anwendung des Kategoriensystems ein.

Die Kodierung der ChatGPT-Fehleranalysen erfolgte nicht ausschließlich auf Grundlage formaler Kriterien, sondern unter Berücksichtigung der fachlichen Angemessenheit der jeweiligen Aussage im Kontext des zugrunde liegenden Fehlerphänomens. Das folgende Beispiel (5.-Addition-Version A, Pos. 11-13) verdeutlicht den Kodierungsprozess:

ChatGPT: „*5. **Fehlermuster**: Im fünften Beispiel wurde 9 + 13 als 23 statt als 22 gelöst.*

***Fehlerursache**: Ähnlich wie im ersten Beispiel könnte dies ein Zählfehler sein. Das Kind könnte Probleme mit dem Konzept des Übertrags von den Einern zu den Zehnern haben.*

***Intervention**: Die Lehrkraft könnte mit dem Kind Übungen durchführen, bei denen der Fokus auf dem Übertrag von den Einern zu den Zehnern liegt, und das Kind auffordern, laut zu zählen, um den Prozess zu verinnerlichen."*

Die Ausgabe wurde codiert als: Rechenaufgabe korrekt wiederholt; Fehlerphänomen erkannt; Korrekte Lösung bekannt; mögliches Fehlermuster erkannt; Ursachenhypothese nimmt Bezug auf Fehlermuster; Ursachen anhand stoffdidaktischer Aspekte; äußerliche Ursache – syntaktisch; Lehrkräfteintervention nimmt Bezug auf Ursachenhypothese; informell, konstruktiv, hohe Schüleraktivität.

Abgrenzend von einer vollständigen qualitativen Inhaltsanalyse, die alle Informationen in den Texten umfassend betrachtet und kodiert, wurde festgelegt, dass der Kodierungsprozess an bestimmten Stellen der ChatGPT-Fehleranalysen abgebrochen werden soll. Aus der Beschreibung des Analyseprozesses als einem aufeinander aufbauenden Vorgang ergab sich, dass es nicht zielführend erschien, Ursachenhypothesen zu betrachten, wenn das zugrunde liegende Fehlerphänomen nicht korrekt erkannt wurde. Entsprechend wurden folgende Abbruchkriterien für den Kodierungsprozess definiert:

- Die Rechenoperation in der Aufgabenstellung wurde nicht korrekt erkannt.
- Das Fehlerphänomen wurde nicht korrekt erkannt.
- Das Fehlermuster passt nicht zum Fehlerphänomen.
- Die Ursachenhypothese steht in keinem Bezug zum Fehlermuster.
- Die Interventionen stehen nicht im Zusammenhang mit der Ursachenhypothese

Das Forschungsdesign der Studie umfasste somit in beiden Forschungsphasen, einen eigenen künstlichen Datensatz. Diese künstlichen Datensätze wurden unabhängig voneinander erhoben und jeweils separat ausgewertet. Da der generierte Output im Kontext der Forschungsfrage auf seine Qualität hin untersucht werden sollte, wurde das in Tabelle 1 dargestellte Kategoriensystem genutzt, um die Güte der Fehleranalysen einzuordnen.

Rechenoperation	Rechenoperation erkannt		
	Rechenoperation nicht erkannt*		
Fehlerphänomen	Fehlerphänomen erkannt		
	Fehlerphänomen nicht erkannt*		
Fehlermuster	Mögliches Fehlermuster erkannt	Korrekte Lösung bekannt	Wird erkenntlich
	Mögliches Fehlermuster nicht erkannt*		Wird nicht erkenntlich
Ursachen-hypothese	Nimmt Bezug auf Fehlermuster	Allgemeine Ursachen - Ursachen anhand des äußeren Erscheinungsbildes (semantisch/syntaktisch) - Ursachen anhand stoffdidaktischer Aspekte - Unsystematische Fehler/ Flüchtigkeitsfehler - Systematische Fehler	
	Nimmt keinen Bezug auf Fehlermuster*		
Lehrkraft-intervention	Nimmt Bezug auf Ursachenhypothese	Wertschätzung durch Lehrkraft - konkrete Hilfe - hohe Schüler:innenaktivität	Instruktiv - konstruktiv
	Nimmt keinen Bezug auf Ursachenhypothese*		Formell - informell
*Wenn festgestellt: Abbruch Kodierung			

Tab. 1: Schwerpunkte des Kategoriensystems (eigene Darstellung)

4. Ergebnisse

Für die Auswertung der Ergebnisse aus der ersten und zweiten Phase wurden die festgelegten Kategorien schrittweise analysiert. Dabei lag der Fokus insbesondere auf der Identifikation von Regelmäßigkeiten sowie auf hervorstechenden Aspekten innerhalb der Daten. Nachfolgend werden die Ergebnisse getrennt für die erste (Abschnitt 4.1) und für die zweite Phase (Abschnitt 4.3) dargestellt.

4.1 Ergebnisse der Fehleranalysen in Phase 1

Die Kategorie *Rechenoperation erkannt* untersuchte, ob ChatGPT die vorgelegten Abbildungen zunächst korrekt wiedergeben konnte. Dabei wurden lediglich in 47 % der Fälle (94) die Aufgaben und Schülerlösungen korrekt erkannt, in 46,5 % (93) traten beim Wiedergeben Fehler auf, und in 6,5 % (13) der Fälle war keine Bewertung in dieser Kategorie möglich, weil die Aufgaben selbst unter der Promptformulierung P1 gar nicht durch ChatGPT wiedergegeben wurden. Besonders auffällig war die unzuverlässige Erkennung bei Multiplikationsaufgaben: In keinem einzigen Fall wurde das Multiplikationszeichen korrekt interpretiert; stattdessen wurde es durchgehend als Divisionssymbol gedeutet. Dies führte zu inhaltlich falschen Aussagen zur Aufgabenstellung und trug erheblich zur nachfolgenden fehlerhaften Analyse bei. Auch bei Additions- und Subtraktionsaufgaben kam es teils zu Verwechslungen einzelner Ziffern (z. B. wurde der Übertrag als Teil des Ergebnisses gedeutet), was auf Schwächen in der Interpretation der handschriftlichen Notation hinweist. Bei der Version B traten zudem wiederholt

Fehler auf, bei denen ChatGPT fälschlicherweise Brüche identifizierte, obwohl diese in der Aufgabenstellung nicht vorkamen – ein Effekt, der vermutlich auf die spezifische grafische Gestaltung der Abbildungen zurückzuführen ist. Insgesamt deutet das Ergebnis auf eine inkonsistente visuelle Verarbeitung der Bilder durch das multimodale Modell hin (vgl. auch Schorcht et al., 2024a). Für die zweite Phase wurde daher beschlossen, die Rechenaufgaben neu und besonders leserlich aufzuschreiben und einzuscannen sowie im initialen Prompt eine Erklärung der jeweiligen Rechenverfahren zu integrieren, um Erkennungsfehler durch unklare Handschrift und Verfahrensverwechslungen (speziell Multiplikation und Division) zu vermeiden.

Ein zentrales Anliegen war die Analyse der *Fehlerphänomenerkennung* im nächsten Schritt. Von den insgesamt noch 107 der 200 analysierten Aufgabenbearbeitungen, bei denen die Darstellung korrekt erkannt oder nicht wiederholt wurde (d.h. ausschließlich Addition, Subtraktion und Division), konnte ChatGPT nur in 24,3 % (26) der Fälle ein *Fehlerphänomen* erkennen. Dabei zeigten sich interessanterweise deutliche Unterschiede zwischen den Rechenoperationen: Bei den übrig gebliebenen 46 von den ursprünglich 50 Subtraktionsaufgaben lag die Erkennungsrate bei 48 % (22 erkannt, 24 nicht erkannt), bei den übrig gebliebenen Additionen (14 %) und Divisionen lag die Erkennungsrate deutlich darunter (2,5 %). Dies lässt auf eine bessere Modellvertrautheit mit dem Fehlertyp der Subtraktion (fehlender Übertrag) schließen. Auffällig war, dass ChatGPT zwar teilweise korrekte Ergebnisse nannte, dabei jedoch fehlerhafte oder widersprüchliche Beschreibungen des Lösungswegs lieferte. So wurden beispielsweise korrekte Rechenschritte als falsch bezeichnet oder falsche Rechenschritte als richtig. Besonders bei Divisionen kam es zu fehlerhaften Aussagen, in denen ChatGPT Zahlen erfand oder falsche Operationen einsetzte. Die inkonsistente Fehlerbeschreibung lässt vermuten, dass das Modell GPT-4 als erstes multimodales KI-Sprachmodell noch keine robuste Repräsentation des schriftlichen Rechenverfahrens besitzt. Ein Beispiel illustriert dies deutlich (vgl. Abb. 3):

Abb. 3: Divisionsaufgabe aus Phase 1 (Version A)

ChatGPT: *„Beim ersten Schritt subtrahiert der Schüler korrekt 45 von 45 und bekommt 0. Danach zieht er 315 von 8 ab, was nicht möglich ist, und notiert 15, was ein Fehler ist. Dann zieht er 15 von 0 ab und erhält wieder 0, was mathematisch inkorrekt ist.“*

Für die zweite Phase wurde daher auch aus diesen Ergebnissen gefolgert, dass der initiale Prompt eine Erklärung der abgebildeten Rechenverfahren enthalten sollte.

Die Fähigkeit zur Benennung *möglicher Fehlermuster* wurde ebenfalls analysiert. Insgesamt wurden von den verbleibenden 26 Fällen, in denen ein Fehlerphänomen korrekt erkannt wurde, nur in 17 Fällen tatsächliche Fehlermuster korrekt benannt. In 11 von diesen 17 Fällen wurde das Muster durch die Beschreibung der Rechnung erkennbar gemacht, und nur in sechs Fällen explizit mit fachdidaktischen Begriffen benannt (z. B. „Übertragsfehler“). Die Analyseform scheint dabei eine Rolle zu spielen: Bei gebündelten Analysen, in denen ChatGPT mehrere Aufgaben gemeinsam analysierte statt einzelne, wurden häufiger konkrete und passende Fehlermuster genannt. Dies legt nahe, dass die Analyseleistung von ChatGPT auch hier möglicherweise von der Struktur des Prompts beeinflusst wird. Für die zweite Phase wurde daher vorgesehen, die Aufgaben in gebündelter Form analysieren zu lassen und Fehlermuster explizit als Kategorie im Prompt zu definieren.

In der Kategorie *Ursachenhypothesen* wurde untersucht, ob ChatGPT Erklärungen für beobachtete Fehler anbot. Fast immer bezogen sich genannte Ursachen direkt auf das zuvor identifizierte Fehlermuster, was auf eine gewisse diagnostische Kohärenz des Modells hindeutet. Inhaltlich dominierten Ursachen mit Bezug zu stoffdidaktischen Aspekten (z. B. Probleme mit dem Stellenwertverständnis, Übertrag, Bündelung). Flüchtigkeitsfehler und systematische Fehlkonzepte wurden seltener erwähnt. Die Analyseform spielte auch hier eine Rolle: In gebündelten Analysen wurden meist mehrere Hypothesen formuliert, während bei sequenzieller Aufgabenbearbeitung häufig nur eine Ursache genannt wurde. Eine ausgewogene Verteilung zwischen semantischen (z. B. „*Fehlerursache: Dies könnte auf ein grundlegendes Missverständnis der Subtraktion hinweisen oder darauf, dass der Schüler versehentlich addiert hat*.“ 4.-Subtraktion-Version A, Pos. 26) und syntaktischen Ursachen (z. B. „*Der Schüler könnte auch Schwierigkeiten haben, die Subtraktionsregel "wenn der Minuend kleiner als der Subtrahend ist, muss man borgen*" (1.-Subtraktion-Version B,

Pos. 16) anzuwenden.) (14 vs. 15) zeigte zudem, dass ChatGPT unterschiedliche Erklärungstypen adressiert, dabei anscheinend jedoch keine klare Systematik verfolgt. Für die zweite Phase wurde deshalb entschieden, auch die Ursachenanalyse über eine präzisere Promptstruktur mit zu fördern.

Schließlich wurde die Kategorie *Lehrkraftinterventionen* ausgewertet. Insgesamt konnten fallübergreifend 56 Interventionsvorschläge von ChatGPT kodiert werden, von denen der Großteil konkrete Hilfestellungen oder Hinweise zur Aktivierung der Lernenden enthielt. 36 Vorschläge förderten eine aktive Rolle der Schülerinnen und Schüler, 29 enthielten didaktisch konkrete Unterstützungsmaßnahmen (z. B. Visualisierungen, Strategievermittlung). Deutlich unterrepräsentiert waren hingegen wertschätzende Hinweise wie etwa „*Es ist wichtig, dass die Lehrkraft ermutigend vorgeht und dem Schüler hilft, ein positives Verhältnis zur Mathematik und zum Lernen aus Fehlern zu entwickeln*." (1.-Subtraktion-Version A, Pos. 19) (nur 5 von 56 Segmenten), die ein entscheidendes Element innerhalb einer positiven Fehlerkultur darstellen (Steuer, 2014; Schoy-Lutz, 2005). Auch hier zeigte sich: Gebündelte Analysen führten häufiger zu elaborierten, mehrdimensionalen Interventionen. Das Verhältnis zwischen instruktiven und konstruktiven Maßnahmen war insgesamt ausgeglichen, wobei ChatGPT sich in der Regel auf den einzelnen Schüler bzw. eine einzelne Schülerin bezog – auch ohne explizite Angabe im Prompt. Diese Tendenz wurde für die zweite Phase aufgenommen, indem der Prompt spezifizierte, dass es sich um Fehler eines/einer einzelnen Lernenden handelt.

4.2 Entwicklung der Dialogmaske für Phase 2

In der zweiten Phase des Forschungsprozesses sollte durch den gezielten Einsatz von Prompt-Engineering die Qualität der Fehleranalysen von ChatGPT verbessert werden. Da es sich bei Fehleranalysen um einen mehrschrittigen Arbeitsprozess handelt, sollte die Analyse der fehlerhaften Aufgabenbearbeitungen durch ChatGPT in Form eines Dialogs gestaltet werden (vgl. Schorcht et al., 2023; 2024b). Diese Technik wurde berücksichtigt, indem die Analyse in mehrere, klar voneinander abgegrenzte Prompts unterteilt wurde, sodass ChatGPT nicht alle Aufgabenstellungen auf einmal bearbeiten musste (vgl. Daheim et al., 2023). Dabei wurde jeweils auch ein Wartegebot formuliert (Buchholtz & Huget, 2024). Der erste Prompt (P2), der zu Beginn der Analyse in ChatGPT eingegeben wurde, enthielt eine Erklärung des Rechenverfahrens, das den zu analysierenden Schülerfehlern zugrunde liegt (siehe Anhang 2). Im Prompt wurde ChatGPT zusätzlich darüber informiert, welche Rolle es einzunehmen hat, welche Aufgabe es zu erfüllen gilt und welches Verhalten vermieden werden soll, um unerwünschte Reaktionen zu verhindern (vgl. Mollick & Mollick, 2023). Darüber hinaus wurde explizit angegeben, welches Rechenverfahren analysiert werden soll.

P2: "Du sollst eine Lehrkraft dabei unterstützen, Schülerfehler bei schriftlichen [Rechenarts]-Aufgaben zu analysieren, indem die Fehlerphänomene beschrieben werden und mögliche Fehlermuster, -ursachen und Lehrkraftinterventionen genannt werden. Wir gehen Schritt für Schritt vor. Fahre erst mit dem nächsten Analyseschritt fort, wenn du dazu aufgefordert wirst und antworte lediglich auf die gestellten Fragen und Anweisungen. Zuerst erkläre ich dir anhand eines korrekten Beispiels, das auf dem Bild zu sehen ist, wie das schriftliche [Rechenart]-verfahren funktioniert. Warte im Anschluss auf die nächste Anweisung." (Erklärung des Rechenverfahrens)

In den beiden nachfolgenden Prompts P3 und P4 wurde ChatGPT zunächst für die Erkennung des Fehlerphänomens angewiesen, eigene Lösungen für die gezeigten Rechenaufgaben zu erstellen. Diese Aufgaben wurden dem Prompt in Form von maschinell geschriebenen Darstellungen hinzugefügt, um sicherzustellen, dass ChatGPT die gezeigten Inhalte korrekt interpretieren und analysieren konnte.

P3: "Auf diesem Bild sind 5 fehlerhafte schriftliche [Rechenarts]-Aufgaben eines Schülers zu sehen, die nach dem obigen Verfahren berechnet wurden. Beschreibe mir zunächst nur, wie die korrekte schriftliche [Rechenart] der Aufgaben aussehen muss und warte im Anschluss auf weitere Anweisungen."

Im nächsten Schritt wurde ChatGPT aufgefordert, die zuvor angefertigten eigenen Lösungen Schritt für Schritt mit den gezeigten fehlerhaften Aufgabenbearbeitungen zu vergleichen.

P4: „Vergleiche nun deine eigenen Rechnungen Schritt für Schritt und deine Lösungen Stellenwert für Stellenwert mit den fehlerhaften schriftlichen [Rechenart] der Aufgaben, die auf dem Bild zu sehen sind. Das kann beispielsweise so aussehen: [Beispiel für Beschreibung]. Warte im Anschluss auf weitere Anweisungen."

Diese Zweiteilung wurde im Optimierungsprozess entwickelt, da ChatGPT bei einer direkten Aufforderung zur Beschreibung der dargestellten Rechnun-

gen oft entweder mit Musterlösungen zu den Aufgaben antwortete oder lediglich das zugrunde liegende Rechenverfahren beschrieb. Um die Anzahl der korrekt beschriebenen Rechnungen zu erhöhen, wurden zusätzliche Anweisungen integriert. ChatGPT sollte die eigenen Rechnungen und Lösungen Schritt für Schritt mit den Schülerlösungen vergleichen und erhielt ein Beispiel für eine mögliche Beschreibung.

Im nächsten Schritt des Dialogs wurde ChatGPT aufgefordert, mögliche Fehlermuster zu den zuvor erkannten Fehlerphänomenen aufzulisten. Die Ergebnisse der ersten Phase hatten gezeigt, dass ChatGPT keine klare Definition des Begriffs „Fehlermuster" besaß und stattdessen teilweise lediglich die dargestellten Rechnungen beschrieb. Um diesem Problem entgegenzuwirken, wurde der entsprechende Prompt um eine angepasste Begriffsdefinition von Fehlermustern ergänzt. Der Prompt wurde weiter konkretisiert, indem die spezifischen Fehlerphänomene, auf die sich ChatGPT beziehen sollte, ausdrücklich erwähnt wurden. Zusätzlich wurde die explizite Aufforderung hinzugefügt, eine Liste möglicher Fehlermuster zu erstellen sowie das Ask-me-anything-Prompting angepasst, um ChatGPT dazu zu bewegen, durch gezielte Rückfragen sicherzustellen, dass die eigene Beschreibung korrekt ist (vgl. Mollick & Mollick, 2023, S. 4).

P5: „Im nächsten Schritt der Analyse sollst du mir auf Grundlage der erkannten Fehlerphänomene bei den Aufgaben (…) mit den erkannten Fehlern (…) eine Liste von möglichen Fehlermustern nennen, die auf diese Rechnungen zutreffen. Fehlermuster lassen sich identifizieren, wenn bestimmte Fehlerphänomene über die Aufgaben hinweg gehäuft auftreten, oder wenn hinter den Fehlerphänomenen eine gewisse innere Logik zu erkennen ist (wenn also bei strukturell gleichen Aufgaben auch strukturell gleiche Fehlerphänomene auftreten). Frage mich im Anschluss, welches erkannte Fehlermuster richtig ist und warte auf weitere Anweisungen."

Da es nicht zielführend erschien, im weiteren Verlauf der Analyse falsch benannte Fehlermuster einzubeziehen, wurde entschieden, ChatGPT an dieser Stelle im Dialog explizit mitzuteilen, auf welches Fehlermuster es sich für den restlichen Analyseprozess beziehen soll.

P6: „Ein erkanntes Fehlermuster ist korrekt. [Nennung des Fehlermusters] Warte auf weitere Anweisungen."

Im darauffolgenden Abschnitt des Dialogs wurde ChatGPT aufgefordert, mögliche Fehlerursachen zu den zuvor identifizierten Fehlermustern zu benennen. Um relativ konsistente Ergebnisse des genKI-Systems zu gewährleisten, wurde – analog zum Fehlermuster-Prompt – eine Begriffsdefinition von Fehlerursachen in den Prompt integriert.

P7: *„Im nächsten Schritt der Analyse sollst du mir mögliche Fehlerursachen für das erkannte Fehlermuster nennen. Hinter Fehlerphänomenen können verschiedene tiefer liegende Fehlerursachen stehen, die sich auf den eigentlichen Fehlerprozess beziehen. Hinter einem Fehlerphänomen können verschiedene Fehlerursachen stehen. Die Fehlerursachen erschließen sich häufig erst im Gespräch mit Schülerinnen und Schülern (und auch dann nicht immer). Während Fehlermuster angeben, wie Schülerinnen und Schüler häufig falsch vorgehen, erklären Fehlerursachen, warum sie dies tun."*

Dabei wurden im Sinne einer positiven Fehlerkultur gezielt die genannten Eigenschaften – Schüleraktivierung, wertschätzender Umgang und konkrete Unterstützung – eingefordert.

P8: „Als letzten Schritt der Analyse sollst du mir mögliche Interventionen der Lehrkraft auf diesen Schülerfehler nennen. Lehrkraftinterventionen sollten möglichst wertschätzend gegenüber Schülerinnen und Schülern sein, Schülerinnen und Schüler sollten möglichst aktiv sein, Interventionen sollten konkrete Hilfen für Schülerinnen und Schüler darstellen und einen Bezug zu den möglichen Fehlerursachen haben."

4.3 Ergebnisse der Fehleranalysen in Phase 2

Die zweite Phase der Untersuchung hatte das Ziel, aufbauend auf den Schwächen der ersten Phase, gezielt die Analysefähigkeit von ChatGPT durch Prompt-Engineering zu verbessern. Die erklärten Rechenverfahren wurden in dieser Phase unter Anwendung von P2 inklusive der Erklärung des Verfahrens zunächst in 72,5 % der Fälle (29 von 40 Durchläufen) formal korrekt durch ChatGPT wiedergegeben. Allerdings traten erneut bei der Multiplikation Bilderkennungsschwierigkeiten auf, obwohl die Erklärung des Verfahrens nicht handschriftlich eingegeben wurde.

Die Rechenoperationen in den einzelnen Schülerfehlern wurden unter Anwendung von P3 und P4 in 169 von 200 Fällen korrekt identifiziert (84,5 %). In 26 Fällen wurde die Rechenoperation allerdings nicht korrekt wiedergegeben und in fünf Fällen gar nicht wiedergegeben. Interessanterweise erfolgte in Phase zwei eine nahezu durchgängig korrekte Erkennung der Multiplikation (96 %), was möglicherweise

auf die Erklärung des Rechenverfahrens zurückzuführen ist. Lediglich die Division wurde diesmal nur in 58 % der Fälle erkannt. So behauptete ChatGPT bei der dritten Divisionsaufgabe des 3. Durchlaufs der Version B, dass anstelle der eigentlichen Schülerlösung 58 (vgl. Abb. 4) der Schüler oder die Schülerin wie folgt gerechnet habe:

ChatGPT: "Der Fehler passiert hier: Er teilt 33 durch 4 und schreibt als Ergebnis 5, obwohl es 8 sein sollte. Er multipliziert 5 mit 4, um 20 zu schreiben, statt 32. Das lässt einen Rest von 13 statt 1 übrig, den er nicht beachtet. Das falsche Schülerergebnis ist 55"

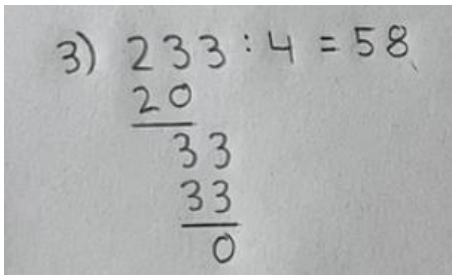


Abb. 4: Divisionsaufgabe aus Phase 2 (Version B)

Diese und ähnliche Fehler deuten darauf hin, dass die Verbesserung der Aufgabenwahrnehmung zwar signifikant war, aber die Zahlenerkennung auch mit dem Prompt-Engineering eine zentrale Schwachstelle des genKI-Systems bleibt, vermutlich infolge der unscharfen Interpretation von Handschrifterkennung und Texteinbettung durch das Modell GPT-4.

Ein zentrales Anliegen war erneut die Erkennung von Fehlerphänomenen, also die Frage, ob ChatGPT in der Lage ist, das eigentliche Problem in der Schülerlösung zu identifizieren. Von den 174 von 200 korrekt oder nicht wiedergegebenen betrachteten Aufgaben in Phase zwei konnte ChatGPT in 58 Fällen (33,3 %) ein Fehlerphänomen korrekt erkennen. Im Vergleich zur ersten Phase (24,3 %) ist dies als eine Verbesserung zu werten. Allerdings war die Verteilung nicht gleichmäßig: Während bei den Subtraktionsaufgaben mit 50 % die Erkennungsrate erneut am höchsten lag, wurden nur 26 % der Multiplikationsfehler und kein einziger Divisionsfehler korrekt erkannt. Zudem wurden bei den Aufgaben von Version B über 70 % der Fehlerphänomene nicht erkannt. Diese Diskrepanzen deuten darauf hin, dass die Verbesserung durch Promptgestaltung nicht in allen Fällen wirksam war – möglicherweise beeinflusst durch Unterschiede in der Schreibweise, Struktur oder der visuellen Lesbarkeit der Aufgabenformate. Inhaltlich zeigte sich zudem, dass ChatGPT zwar oft korrekte Ergebnisse und Rechenwege präsentieren konnte, aber erneut widersprüchliche Aussagen produzierte – etwa indem es zunächst eine korrekte Musterlösung präsentierte, später aber die Schülerrechnung falsch erklärte oder umgekehrt. In mehreren Fällen behauptete ChatGPT, ein Ergebnis sei falsch, obwohl es korrekt war, oder beschrieb korrekte Teilschritte als fehlerhaft.

Deutlich verbessert zeigte sich hingegen die Fehlermusternennung. Während in der ersten Phase Fehlermuster meist nur implizit beschrieben wurden (etwa durch Paraphrasierung der Schülerrechnung), konnte ChatGPT in der zweiten Phase – unterstützt durch eine im Prompt definierte Beschreibung von Fehlermustern – in vielen Fällen spezifische, korrekte Bezeichnungen liefern. Insgesamt wurden 64 zutreffende Fehlermuster (Addition, Subtraktion und Multiplikation) identifiziert. Dabei lag die Erkennungsrate bei Subtraktionen (78 %) erneut deutlich höher als bei Multiplikationen (33,3 %) und Additionen (45,2 %). Gleichzeitig wurden jedoch auch 50 unzutreffende oder nicht plausible Fehlermuster benannt – teilweise durch Vermengung von Fehlerursache und -symptom oder durch Zuschreibungen, die dem tatsächlich beobachteten Schülerfehler widersprachen. Positiv hervorzuheben ist außerdem, dass in fast allen zutreffenden Fällen eine explizite Benennung des Musters (z. B. „Zählfehler" oder „Übertragsfehler") erfolgte und nicht lediglich eine Beschreibung.

Ein weiteres zentrales Diagnoseelement war die Ursachenhypothesenfindung. ChatGPT zeigte sich hier deutlich differenzierter als in Phase eins. In nahezu allen Fällen, in denen ein Fehlermuster korrekt benannt wurde, konnte auch eine plausible Ursache angegeben werden. Von 157 genannten Ursachen, die einen Bezug zum Fehlermuster herstellten, konnten 34,4 % semantischen, 28,7 % syntaktischen, 44,6 % allgemeinen, 55,4 % stoffdidaktischen, 9,5 % unsystematischen und 10,2 % systematischen Ursachen zugeordnet werden. Am häufigsten nannte das Modell stoffdidaktische Ursachen – etwa fehlendes Verständnis für den Stellenwert, Schwierigkeiten mit dem Übertrag oder Probleme bei der Entbündelung. Auffällig war, dass bei gebündelten Analysen häufiger mehrere Ursachen genannt wurden – im Gegensatz zu sequenziellen Analysen, bei denen meist nur eine Hypothese formuliert wurde. Daraus ergibt sich erneut die Bedeutung der Analyseform (gebündelt vs. einzeln) für die Tiefe der Diagnostik.

In der zweiten Phase wurden 214 Lehrkraftinterventionen analysiert, die größtenteils Bezug auf vorher genannte Ursachenhypothesen nahmen. Durch die

Prompt-Anpassungen konnte der Anteil wertschätzender Interventionen leicht von 8,9 % in Phase eins auf 16,8 % gesteigert werden, blieb jedoch im Vergleich zu konkreten Hilfen (53,7 %) und hoher Schüleraktivität (57 %) erneut deutlich geringer. Die meisten Interventionen waren informell (70,1 %) und instruktiv (56,5 %), wobei sich erneut eine thematische Häufung zum Fehlerbild „Übertrag" zeigte. Insgesamt deutet sich an, dass ChatGPT im Bereich der Lehrerintervention auf ein standardisiertes Repertoire zurückgreift, das inhaltlich an die jeweilige Diagnose angepasst wird.

5. Zusammenfassung

Die erste Phase der Studie offenbarte vor allem zentrale Schwächen der ChatGPT-basierten Fehlerdiagnose: Während einfache Aufgaben zwar halbwegs korrekt wiedergegeben und grundlegende Fehlermuster erkannt wurden, zeigte sich eine erhebliche Variabilität in der Aufgabenwahrnehmung, Beschreibung von Fehlerphänomenen und Konsistenz der Analysen. Besonders der Einfluss der Analyseform (gebündelt vs. sequenziell) erwies sich als bedeutend – gebündelte Analysen führten zu systematischeren und differenzierteren Aussagen. Insgesamt blieb die Zuverlässigkeit der Fehleranalysen deutlich hinter den Erwartungen zurück. Diese Beobachtungen flossen direkt in die Gestaltung der zweiten Untersuchungsphase ein, insbesondere in die Optimierung der Prompts (z. B. explizite Definitionen, strukturiertere Aufgabenbilder, Vorgaben zur Diagnoseabfolge). Die Ergebnisse zu FF1 verdeutlichen, dass genKI-Systeme wie ChatGPT für Fehleranalysen zwar Potenzial aufweisen, gleichzeitig aber klare Steuerungsimpulse benötigen, um im didaktisch-diagnostischen Kontext sinnvoll eingesetzt werden zu können.

Zusammenfassend zu FF2 lässt sich sagen, dass durch die gezielte Promptgestaltung in der zweiten Phase in bestimmten Teilen der Fehleranalysen eine deutliche Steigerung der Diagnosequalität von ChatGPT erreicht wurde. Vor allem die Nennung korrekter Fehlermuster und die systematischeren Ursachenanalysen zeigen, dass das genKI-System durch textuelle Strukturierung und Definitionen zu einem gewissen Maß gezielt steuerbar ist. Dennoch bestehen weiterhin Limitierungen. Die Konsistenz in der Analyse, die Zuverlässigkeit bei der Zahlenwahrnehmung und die Fähigkeit zur reflexiven Kohärenzprüfung blieben weiter eingeschränkt. Die Ergebnisse unterstreichen die Notwendigkeit, genKI-Systeme nicht als autonomes Diagnoseinstrument einzusetzen, sondern sie durch pädagogisch gestaltete Prompts, kontrollierte Aufgabenformate und menschliche Supervision didaktisch zu rahmen. Die zweite Phase machte deutlich, dass durch gezielte Gestaltung zwar große Fortschritte möglich sind – aber auch, dass eine vollständige Automatisierung der Fehlerdiagnose derzeit (noch) nicht realistisch ist.

6. Fazit und Ausblick

Die vorliegende Studie untersuchte das Potenzial generativer KI-Systeme – insbesondere ChatGPT – zur Unterstützung von Mathematiklehrkräften bei diagnostischen Tätigkeiten. Im Zentrum stand die Frage, inwieweit genKI in der Lage ist, arithmetische Schülerfehler zu analysieren, Ursachenhypothesen zu formulieren und geeignete Interventionen vorzuschlagen sowie welchen Einfluss gezieltes Prompt-Engineering auf diese Leistungen hat. Die Ergebnisse zeigten, dass die Fehleranalysen ohne strukturierende Prompts stark variieren. Sowohl die Erkennung von Zahlen als auch die Analyse von Fehlerphänomenen und Fehlermustern erwies sich als inkonsistent und damit in der ersten Phase kaum verlässlich. Erst durch gezielte Steuerungsimpulse – etwa klare Definitionsvorgaben, bildliche Strukturierung und explizite Aufforderungen zur Diagnoseabfolge – konnte in der zweiten Phase eine Verbesserung erreicht werden. Dazu zählte eine präzisere Wiedergabe von Zahlenarten und Rechenoperationen, eine geringere Zahl falsch interpretierter Ziffern sowie eine differenziertere Beschreibung von Fehlermustern, die nicht mehr mit bloßen Rechenschrittwiederholungen verwechselt wurden.

Besonders bei Subtraktionsaufgaben zeigte ChatGPT eine vergleichsweise hohe Diagnosequalität – vermutlich, weil entsprechende Fehlerbilder häufiger im Trainingsmaterial vertreten waren. Auch die Anzahl genannter Ursachenhypothesen und Lehrkraftinterventionen nahm zu, wobei letztere weiterhin seltener wertschätzende Elemente enthielten. Dennoch blieben zentrale Herausforderungen bestehen: Die Bild- und Zahlenwahrnehmung ist störanfällig, die Modellantworten enthalten zum Teil Widersprüche, und die Fähigkeit zur kohärenten Überprüfung eigener Aussagen ist nicht ausgeprägt. Ein auffälliges Ergebnis der Studie ist daher die deutlich variierende Erkennungsleistung je nach Rechenoperation. Bei Multiplikations- und Divisionsaufgaben traten systematisch mehr Fehler auf. Diese Unterschiede lassen sich nicht allein durch technische Bildverarbeitung erklären, sondern könnten auf kulturelle Einflüsse in den Trainingsdaten hinweisen. Viele genKI-Modelle,

darunter auch GPT-4, wurden überwiegend mit englischsprachigen und international geprägten Quellen trainiert. Die dort gebräuchlichen schriftlichen Rechenverfahren unterscheiden sich teilweise deutlich von den in deutschsprachigen Klassenzimmern üblichen Verfahren – insbesondere bei der schriftlichen Multiplikation und Division. Die empirischen Hinweise aus dieser Studie legen nahe, dass ChatGPT mit diesen Verfahren weniger vertraut ist, was sich negativ auf die Analyseleistung auswirkt.

Damit stellt sich nicht nur die Frage nach der didaktischen Justierbarkeit solcher Modelle, sondern auch nach der Repräsentation kulturell unterschiedlicher Bildungspraktiken im Training. Eine sinnvolle Weiterentwicklung von genKI im Bildungsbereich müsste daher sowohl technisch (etwa durch multimodale und kontextbewusste Modellarchitekturen) als auch inhaltlich (durch ausgewogene, domänenspezifische Trainingsdaten) erfolgen. Nur so kann gewährleistet werden, dass solche Systeme langfristig für verschiedene Bildungskontexte adaptierbar und valide einsetzbar sind.

Die Schwächen in der Bilderkennung sind andererseits zum Teil auf die Architektur aktueller Sprachmodelle zurückzuführen. ChatGPT basiert – auch in multimodalen Varianten wie GPT-4 – auf einem Zusammenspiel zwischen Sprachtransformern und visuellen Encodern (z. B. Vision-Transformer oder CLIP-ähnlichen Modulen). Diese Modelle wandeln Bildinhalte in semantische Vektorrepräsentationen um, die dem Sprachmodell zur weiteren Verarbeitung bereitgestellt werden. Zwar erlaubt dies die Analyse komplexer Eingaben wie handschriftlicher Schülerlösungen, doch zeigen sich bei mathematischen Aufgabenstellungen mit strukturellen oder positionssensitiven Anforderungen (z. B. Überträge) weiterhin deutliche Schwächen. Die semantische Interpretation genügt daher nicht, um präzise formale Abläufe nachzuvollziehen. Zukünftige Entwicklungen gehen daher über rein textbasierte Verfahren hinaus: Multimodale Modelle, die Text, Bild und mathematische Repräsentationen (wie LaTeX oder Code) integriert verarbeiten, werden derzeit intensiv erforscht.

Für den schulischen Einsatz bedeutet das: ChatGPT kann – trotz Limitierungen – als Ideengeber im Kontext didaktischer Reflexion dienen. Lehrkräfte sollten jedoch keine automatisierte Diagnose erwarten, sondern die KI als Werkzeug zur Hypothesenbildung und Anregung von Interventionen verstehen. Besonders dort, wo strukturierte Prompts verwendet werden und die Analyse auf typische Fehlerbilder abzielt, kann genKI sinnvoll unterstützen, ersetzt aber nicht das professionelle Urteilsvermögen. Vielmehr ist eine symbiotische Nutzung denkbar: Menschliche Expertise und maschinelle Vorschläge greifen ineinander, um diagnostische Routinen effizienter und vielfältiger zu gestalten.

Danksagung

Wir danken Robin Weber für die Durchführung der Studie und die getätigten Analysen.

Literatur

Baker, R. & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E. & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters, *Computers and Education: Artificial Intelligence, 5,* 100177. https://doi.org/10.1016/j.caeai.2023.100177

Buchholtz, N. & Huget, J. (2024). Möglichkeiten für die Anwendung von ChatGPT für angehende Mathematiklehrkräfte – Von guten und schlechten Prompts. *Mitteilungen der mathematischen Gesellschaft Hamburg, 44*, 11–31.

Buchholtz, N., Schorcht, S., Baumanns, L., Huget, J., Noster, N., Rott, B., Siller, H.-S. & Sommerhoff, D. (2024). Damit rechnet niemand! Sechs Leitgedanken zu Implikationen und Forschungsbedarfen zu KI-Technologien im Mathematikunterricht. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 117*, 15–24. https://ojs.didaktik-der-mathematik.de/index.php/mgdm/article/view/1249/1403

Cameron, S. & Mesiti, C. (2024). What kind of mathematics teacher is ChatGPT? Identifying the pedagogical practices preferenced by generative AI tools when preparing lesson plans. In J. Visnovska, E. Ross & S. Getenet (Hrsg.), *Proceedings of the 46th annual conference of the Mathematics Education Research Group of Australasia* (S. 135–142). MERGA.

Daheim, N., Macina, J., Kapur, M., Gurevych, I. & Sachan, M. (2024). Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen (Hrsg.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, (S. 8386–8411). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.478/

Deutscher Ethikrat (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz: Stellungnahme*. www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

Dilling, F. & Herrmann, M. (2024). Using large language models to support pre-service teachers mathematical reasoning—an exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence, 7*:1460337. https://doi.org/10.3389/frai.2024.1460337

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P.C. & Berner, J. (2023). *Mathematical Capabilities of ChatGPT* (arXiv Preprint No. 2301.13867). arXiv. https://doi.org/10.48550/ARXIV.2301.13867

Fritz, J. (2021). *Schülerfehler im Problemlöseunterricht - Empirische Erkundungen zum Umgang der Lehrperson mit Schülerfehlern im mathematischen Problemlöseunterricht.* Technische Universität Carolo-Wilhelmina zu Braunschweig. https://leopard.tu-braunschweig.de/receive/dbbs_mods_00069975

Fritz-Schubert, E. & Rohde, T. (2022). *Holpern, stolpern, weiterkommen: für eine konstruktive Fehlerkultur in der Schule*. 1. Auflage. Pädagogik. Beltz.

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

Gardee, A. & Brodie, K. (2022). Relationships between teachers' interactions with learner errors and learners' mathematical identities. *International Journal of Science and Mathematics Education, 20*, 193–214. https://doi.org/10.1007/s10763-020-10142-1

Gerster, H.-D. (1982). *Schülerfehler bei schriftlichen Rechenverfahren: Diagnose und Therapie. Studienbücher Mathematik Didaktik.* 102. Herder.

Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leo, D., … Rienties, B. (2025). The promise and challenges of generative AI in education. *Behaviour & Information Technology, 44*(11), 2518–2544. https://doi.org/10.1080/0144929X.2024.2394886

Hankeln, C., Kroehne, U., Voss, L., Gross, S. & Prediger, S. (2025). Developing digital formative assessment for deep conceptual learning goals: Which topic-specific research gaps need to be closed?. *Educational Technology Research and Development, 73,* 1953–1973. https://doi.org/10.1007/s11423-025-10486-x

Heinrichs, H. (2015). *Diagnostische Kompetenz von Mathematik-Lehramtsstudierenden: Messung und Förderung. Perspektiven der Mathematikdidaktik.* Springer Fachmedien Wiesbaden GmbH.

Huget, J. & Buchholtz, N. (2024). ChatGPT als Reflexionsinstrument zur Förderung von Unterrichtsplanungskompetenzen von Lehramtsstudierenden. In P. Ebers, F. Rösken, B. Barzel, A. Büchter, F. Schacht & P. Scherer (Hrsg.), *Beiträge zum Mathematikunterricht 2024. 57. Jahrestagung der Gesellschaft für Didaktik der Mathematik* (S. 925-928). WTM. http://dx.doi.org/10.17877/DE290R-24623

Jensen, S. & Gasteiger, H. (2019). „Ergänzen mit Erweitern" und „Abziehen mit Entbündeln" – Eine explorative vergleichende Studie zu spezifischen Fehlern und zum Verständnis des Algorithmus. *Journal für Mathematikdidaktik 40*(2), 135–167. https://doi.org/10.1007/s13138-018-00139-3

Jin, H., Kim, Y., Su Park, Y., Tilekbay, B., Son, J. & Kim, J. (2024). Using Large Language Models To Diagnose Math Problem-solving Skills At Scale. In Association for Computing Machinery, Inc. (ACM) (Hrsg.), Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), July 18–20, 2024, Atlanta, GA, USA. ACM, New York, NY, USA, 5 Seiten. https://doi.org/10.1145/3657604.3664697

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . & 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

Khan Academy (2024, 13. Mai). GPT-4o (Omni) mathtutoring demo on Khan Academy [Video]. youtu.be/IvXZCocyU_M

[KMK] Ständige Konferenz der Kultusminister (2024). *Handlungsempfehlung für die Bildungsverwaltung zum Umgang mit Künstlicher Intelligenz in schulischen Bildungsprozessen. Themenspezifische Handlungsempfehlung* (Beschluss der Bildungsministerkonferenz vom 10.10.2024).

Kim, Y., Jin, H., Doh, H., Kim, E., Jung, D., Kim, S., Choi, K., Son, J. & Kim, J. (2026). Benchmarking large language models for diagnosing students' cognitive skills from handwritten math work (arXiv Preprint No. 2504.00843). arXiv. https://arxiv.org/abs/2504.00843

Kojima, T., Shane Gu, S., Reid, M., Matsuo, Y. & Iwasawa, Y. (2022). *Large Language Models are Zero-Shot Reasoners* (arXiv Preprint No. 2205.11916). arXiv. https://doi.org/10.48550/ARXIV.2205.11916

Kortenkamp, U. & Dohrmann, C. (2023). Pre-service teacher training with ai: Using chatgpt discussions to practice teacher-student discourse. In M. Ayalon, B. Koichu, R. Leikin, L. Rubel & M. Tabach (Hrsg.), *Proceedings of PME 46* (S. 187–194, Vol. 3).

Liu, W., Hu, H., Zhou, J., Ding, Y., Li, J., Zeng, J., He, M., Chen, Q., Jiang, B., Zhou, A. & He, L. (2023). *Mathematical language models: A survey* (arXiv Preprint No. 2312.07622). arXiv. https://arxiv.org/abs/2312.07622

Lu, Z., Zhou, A., Wang, K., Ren, H., Shi, W., Pan, J., Zhan, M. & Li, H. (2024). *MathCoder2: Better math reasoning from continued pretraining on model-translated mathematical code* (arXiv Preprint No. 2410.08196). arXiv. https://arxiv.org/abs/2410.08196

Lung, J. (2020). *Schulcurriculares Fachwissen von Mathematiklehramtsstudierenden. Struktur, Entwicklung und Einfluss auf den Studienerfolg*. Wiesbaden.

Luzano, J. (2024). Assessment in mathematics education in the sphere of artificial intelligence: A systematic review on its threats and opportunities. *International Journal of Academic Multidisciplinary Research, 8*(2), 100–104.

Mayring, P. (2010). Qualitative Inhaltsanalyse. In G. Mey und K. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (S. 601–613). VS Verlag für Sozialwissenschaften.

Mindnich, A., Wuttke, E. & Seifried, J. (2008). Aus Fehlern wird man klug? Eine Pilotstudie zur Typisierung von Fehlern und Fehlersituationen. In E.-M. Lankes (Hrsg.), *Pädagogische Professionalität als Gegenstand empirischer Forschung* (S. 153–164). Münster: Waxmann.

Mohr, G., Reinmann, G., Blüthmann, N., Lübcke, E. & Kreinsen, M., (2023). *Übersicht zu ChatGPT im Kontext Hochschullehre. Hamburger Zentrum für Universitäts- Lehren und Lernen.* Zugegriffen am 25.11.2024. https://www.hul.uni-hamburg.de/selbstlernmaterialien/dokumente/hul-chatgpt-im-kontext-lehre-2023-01-20.pdf

Mollick, E. R. & Mollick, L. (2023). *Assigning AI: Seven Approaches for Students, with Prompts* (arXiv Preprint No. 2306.10052). arXiv. https://arxiv.org/abs/2306.10052

Nelson, G. & Powell, S. R. (2018). Computation error analysis: Students with mathematics difficulty compared to typically

achieving students. *Assessment for Effective Intervention, 43*(3), 144–156.

OpenAI (2023) *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses.* https://openai.com/index/gpt-4/

Oser, F., Hascher, T. & Spychiger, M. (1999). Lernen aus Fehlern. Zur Psychologie des ‚negativen' Wissens. In F. Oser & W. Althof (Hrsg.), *Fehlerwelten: vom Fehlermachen und Lernen aus Fehlern. Beiträge und Nachträge zu einem interdisziplinären Symposium aus Anlass des 60. Geburtstages von Fritz Oser,* (S. 11–42). Leske + Budrich.

Oser, F. & Spychiger, M. (2005). *Lernen ist schmerzhaft: zur Theorie des negativen Wissens und zur Praxis der Fehlerkultur.* Beltz-Pädagogik. Beltz.

Paster, K., Dos Santos, M., Azerbayev, Z. & Ba, J. (2023). Openwebmath: An open dataset of high-quality mathematical web text (arXiv Preprint No. 2310.06786). arXiv. https://arxiv.org/abs/2310.06786

Pepin, B., Buchholtz, N. & Salinas-Hernández, U. (2025). A Scoping Survey of ChatGPT in Mathematics Education. *Digital Experiences in Mathematics Education,* 11, 9–41. https://doi.org/10.1007/s40751-025-00172-1

Rach, S., Ufer,S. & Heinze, A. (2012). Lernen aus Fehlern im Mathematikunterricht - kognitive und affektive Effekte zweier Interventionsmaßnahmen. *Unterrichtswissenschaft, 40*, 213–234.

Rittle-Johnson, B., Adler, R., Durkin, K., Burleigh, L., King, J. & Crossley, S. (2025). Detecting math misconceptions: An AI benchmark dataset. In J. Wilson, C. Ormerod & M. Beiting Parrish (Hrsg.), *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Vol. 2: Works in Progress* (S. 20–24). National Council on Measurement in Education. NCME.

Radatz, H. (1980). *Fehleranalysen im Mathematikunterricht.* Vieweg.

Rong, L. & Mononen, R. (2022). Error analysis of students with mathematics learning difficulties in Tibet. *Asian Journal for Mathematics Education, 1*(1), 52-65.

Schmassmann, M. (1992). Fehleranalyse – ein Weg zum Verstehen. In D. Jost, J. Erni & M. Schmassmann (Hrsg.), *Mit Fehlern muss gerechnet werden: Mathematischer Lernprozess, Fehleranalyse, Beispiele und Übungen* (S. 35–42). Zürich: Sabe.

Schorcht, S., Baumanns, L., Huget, J., Buchholtz, N., Peters, F. & Pohl, M. (2023). Ask Smart to Get Smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. *Mitteilungen der Gesellschaft für Didaktik der Mathematik 115*, 12–23.

Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F. & Pohl, M. (2024a). Lernt die KI nun Sehen und Zeichnen? Herausforderungen der Bildgenerierung und Bildinterpretation mit ChatGPT in der Mathematikdidaktik. *Mitteilungen der Gesellschaft für Didaktik der Mathematik, 116,* 22–29. https://ojs.didaktik-der-mathematik.de/index.php/mgdm/article/view/1223/1381

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,* 1386075. https://doi.org/10.3389/feduc.2024.1386075

Schoy-Lutz, M. (2005). *Fehlerkultur im Mathematikunterricht: theoretische Grundlegung und evaluierte unterrichtspraktische Erprobung anhand der Unterrichtseinheit „Einführung in die Satzgruppe des Pythagoras". Texte zur mathematischen Forschung und Lehre, 39*. Franzbecker.

Seifried, J. & Wuttke, E. (2010). Student errors: how teachers diagnose them and how they respond to them. *Empirical Research in Vocational Education and Training, 2*(2), 147–162.

Shaughnessy, M., DeFino, R., Pfaff, E. & Blunk, M. (2021). I think I made a mistake: how do prospective teachers elicit the thinking of a student who has made a mistake? *Journal of Mathematics Teacher Education, 24*, 335–359. https://doi.org/10.1007/s10857-020-09461-5

Spychiger, M, Oser, F., Hascher, T. & Mahler, F. (1999). Entwicklung einer Fehlerkultur in der Schule. In W. Althof (Hrsg.), *Fehlerwelten* (S. 43–70). VS Verlag für Sozialwissenschaften.

Steuer, G. (2014). *Fehlerklima in der Klasse: zum Umgang mit Fehlern im Mathematikunterricht.* Research. Springer VS.

[SWK] Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (2024). *Large Language Models und ihre Potenziale im Bildungssystem: Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz.* www.swk-bildung.org/content/uploads/2024/02/SWK-2024-Impulspapier_LargeLanguage-Models.pdf

Tan, L. Y., Hu, S., Yeo, D. J. & Cheong, K. H. (2025). A Comprehensive Review on Automated Grading Systems in STEM Using AI Techniques. *Mathematics, 13*(17), 2828. https://doi.org/10.3390/math13172828

Türling, J. M. (2014). Die professionelle Fehlerkompetenz von (angehenden) Lehrkräften. Eine empirische Untersuchung im Rechnungswesenunterricht. Springer VS. https://link.springer.com/book/10.1007/978-3-658-04931-7

U.S. Department of Education, Office of Educational Technology (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations*. Washington, DC.

Voßhagen, J. (2025). Fehleranalyse neu gedacht: generative KI als Werkzeug zur individuellen Förderung im Mathematikunterricht? In L. Schick, M. Platz & A. Lambert (Hrsg.), *Beiträge zum Mathematikunterricht 2025*, 1484. WTM.

Weegar, R., Idestam-Almquist, P. (2024). Reducing Workload in Short Answer Grading Using Machine Learning. International Journal of Artificial Intelligence in Education, 34, 247–273. https://doi.org/10.1007/s40593-022-00322-1

Yang, Z., Ding, M., Lv, Q., Jiang, Z., He, Z., Guo, Y., Bai, J. & Tang, J. (2023). *GPT can solve mathematical problems without a calculator* (arXiv Preprint No. 2309.03241). arXiv. https://doi.org/10.48550/arXiv.2309.03241

Yoon, H., Hwang, J., Lee, K., Roh, K. H., and Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education*, 56, 1531–1551. https://doi.org/10.1007/s11858-024-01629-0

Zeller, C. & Schmid, U. (2017). *Automatic generation of analogous problems to help resolving misconceptions in an intelligent tutor system for written subtraction*. Vortrag auf der 24th International Conference on Case Based Reasoning, Atlanta, GA, 31th October – 2nd November 2016.

Zhu, M., Liu, O, L. & Lee, H.-S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers &*

Education, 143, 103668. https://doi.org/10.1016/j.compedu.2019.103668

Ziegenbalg, J. (2023). Algorithmisches Arbeiten. In R. Bruder, A. Büchter, H. Gasteiger, B. Schmidt-Thieme, H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 341–368). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-66604-3_11

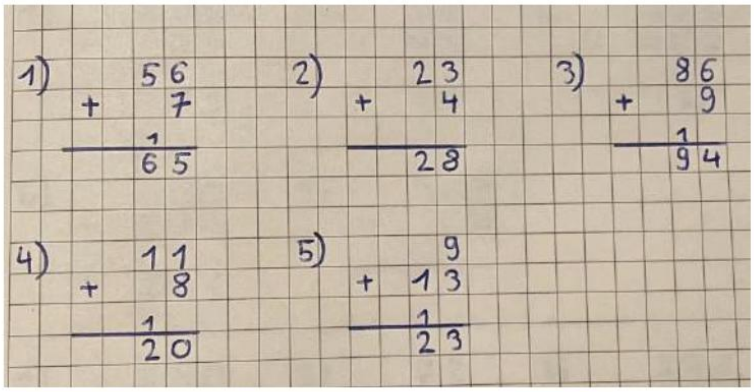
Anschrift der Verfasser:innen

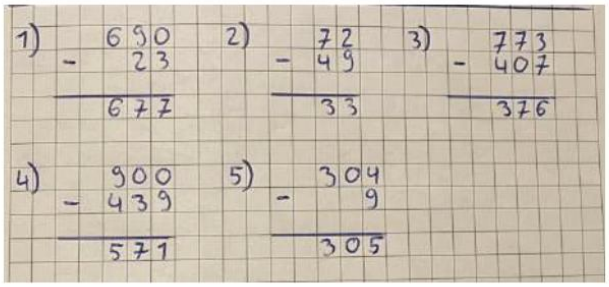
Nils Buchholtz
Universität Hamburg
Fakultät für Erziehungswissenschaft
Didaktik der Mathematik - Sekundarstufe
Von-Melle-Park 8,
20146 Hamburg
nils.buchholtz@uni-hamburg.de

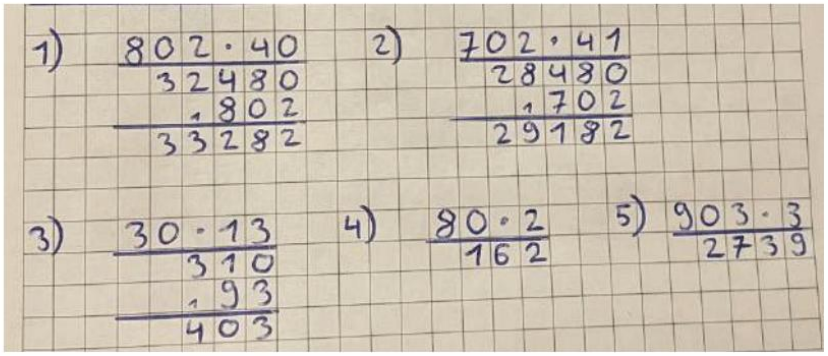
Franziska Peters
Universität Hamburg
Fakultät für Erziehungswissenschaft
Didaktik der Mathematik - Primarstufe
Von-Melle-Park 8,
20146 Hamburg
Franziska.peters-2@uni-hamburg.de

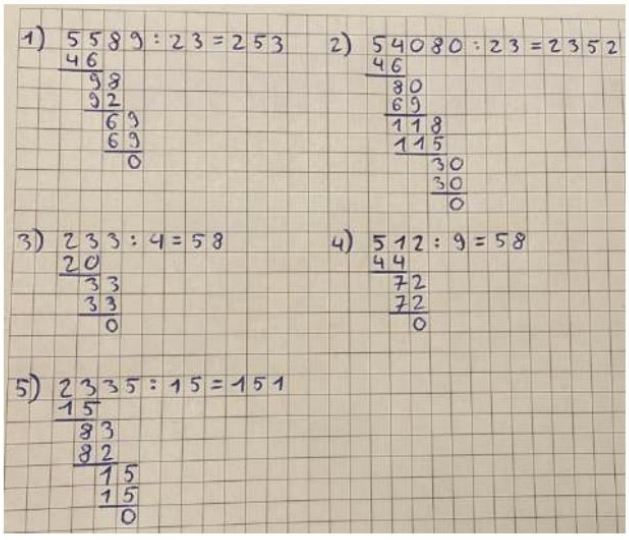
Anhang

Anhang 1: Aufgabenbilder erste Phase Schüler A









Anhang 2: Spezifischer Teil Dialogmaske Addition

Prompt:

Du sollst eine Lehrkraft dabei unterstützen, Schülerfehler bei schriftlichen Additionsaufgaben zu analysieren, indem die Fehlerphänomene beschrieben werden und mögliche Fehlermuster, -ursachen und Lehrkraftinterventionen genannt werden. Wir gehen Schritt für Schritt vor. Fahre erst mit dem nächsten Analyseschritt fort, wenn du dazu aufgefordert wirst und antworte lediglich auf die gestellten Fragen und Anweisungen. Zuerst erkläre ich dir anhand eines korrekten Beispiels, das auf dem Bild zu sehen ist, wie das schriftliche Additionsverfahren funktioniert. Warte im Anschluss auf die nächste Anweisung.

Bei der schriftlichen Addition werden beide Summanden stellengerecht untereinander notiert. Die untereinanderstehenden Ziffern innerhalb eines Stellenwerts werden jeweils separat addiert, wobei mit dem kleinsten Stellenwert begonnen wird (E + E, Z + Z, H + H). Die Summe der Teilrechnungen wird stellengerecht jeweils in der entsprechenden Spalte unter dem Strich notiert. Ist das Teilergebnis zweistellig (im Beispiel 13 Zehner in der Zehnerspalte), wird nur der Wert des Einers im Ergebnis notiert und die übrigen 10 Einheiten des Teilergebnisses (im Beispiel 10 Zehner) werden gebündelt als Übertrag beim nächstgrößeren Stellenwert vermerkt (im Beispiel eine kleine 1 für 1 Hunderter). Dieser Übertrag wird im darauffolgenden Rechenschritt zu dem nächsten Teilergebnis addiert. Das Gesamtergebnis ist nach der letzten Teilrechnung direkt unter dem Strich ablesbar.