Einsatz von generativer KI zur Unterstützung mathematischer Modellierungsprozesse – Vergleich und Akzeptanz zweier Sprachmodelle

KATRIN ROLKA, BOCHUM & MICHAEL KALLWEIT, BOCHUM

***Zusammenfassung**: Der Beitrag zeigt Chancen und Herausforderungen beim Einsatz generativer KI im Mathematikunterricht. In einer explorativen Studie wurden zwei generative KI-Modelle auf ihr Potenzial erkundet, Schüler:innen bei der Bearbeitung einer mathematischen Modellierungsaufgabe als Dialogpartner zu unterstützen. Ziel war es, Unterschiede in der Leistungsfähigkeit sowie die Akzeptanz dieser Technologien zu analysieren. Beide Modelle zeigten Potenziale, unterschieden sich jedoch in der Qualität, insbesondere hinsichtlich der adäquaten Umsetzung der im Prompt hinterlegten didaktischen Überlegungen. Nichtsdestotrotz äußerten die Schüler:innen eine große Akzeptanz bezüglich der Nutzung beider Modelle.*

***Abstract:** The article shows opportunities and challenges in the use of generative AI in mathematics education. In an explorative study, two generative AI models were examined for their potential to support students working on a mathematical modeling task as dialog partners. The aim was to analyze differences in the performance and acceptance of these technologies. Both models showed potential but differed in terms of quality, especially with regard to an appropriate realization of didactical considerations placed in the prompt. Nevertheless, the students expressed a high level of acceptance concerning the use of both models.*

1. Motivation

Die rasanten Entwicklungen im Bereich der künstlichen Intelligenz (KI) haben in den letzten Jahren begonnen, den Bildungssektor tiefgreifend zu verändern, und diese Transformationen werden sich weiter fortsetzen (vgl. Renz, 2021). Vor allem die zunehmende Verfügbarkeit großer generativer Sprachmodelle hat auch den Diskurs in der Mathematikdidaktik befeuert und neue Chancen sowie Herausforderungen für den Mathematikunterricht eröffnet (z. B. Buchholtz et al., 2023; Oldenburg, 2023).

Die Integration digitaler Werkzeuge im Mathematikunterricht hat sich in den letzten gut 50 Jahren stetig weiterentwickelt. Beginnend mit den ersten einfachen Taschenrechnern, über grafikfähige Taschenrechner, CAS-Rechner, dynamische Geometrie-Software, bis hin zu Smartphone-Apps mit Formelerkennung per Foto, haben sich die Einsatzmöglichkeiten von Technologien in der mathematischen Bildung kontinuierlich erweitert. Die Einführung generativer KI-Modelle wie ChatGPT, Gemini, Claude oder Llama in den 2020er-Jahren stellt nun einen nächsten bedeutenden Schritt dar, der über einfache Berechnungen oder Visualisierungen hinausgeht, und möglicherweise auch tiefere kognitive Prozesse anregen kann, etwa in Bezug auf mathematisches Argumentieren (Dilling et al., 2024).

Die Potenziale bei der Integration von KI in den Bildungsbereich werden beispielsweise auch in bildungspolitischen Leitlinien formuliert (SWK, 2024). Daneben werden dort auch ethische Fragen sowie Aspekte des Datenschutzes angesprochen, die beim Einsatz in der Schule eine zentrale Rolle spielen (ebd.). Gleichzeitig werfen auch viele Stimmen Bedenken bezüglich der probabilistischen Natur der Sprachmodelle auf, die bei der Darstellung mathematischer Inhalte trotz stetiger Verbesserungen weiterhin zu Fehlern führen kann (z. B. Lightman et al., 2023). Darüber hinaus können systematische Verzerrungen durch die Trainingsdaten nicht ausgeschlossen werden, die insbesondere dann zu Problemen führen können, wenn KI-basierte (pädagogische) Entscheidungen getroffen werden (Ali et al., 2024).

Die Notwendigkeit einer breiten empirischen Erforschung der Möglichkeiten und der Risiken der KI-Nutzung im Mathematikunterricht wird daher immer deutlicher. Vor diesem Hintergrund richtet sich der Fokus des vorliegenden Beitrags auf die Analyse eines pilotierenden Einsatzes generativer KI bei mathematischen Modellierungsprozessen im Hinblick auf die hinterlegte Unterstützung. Dabei werden die im Rahmen zweier explorativer Studien eingesetzten Sprachmodelle miteinander verglichen, um Unterschiede in der Leistungsfähigkeit und Eignung für den schulischen Einsatz aufzuzeigen. Schließlich wird auch die Akzeptanz dieser Technologie durch die beteiligten Schüler:innen in den Blick genommen.

2. Generative KI

Generative KI-Modelle, auch als generative künstliche Intelligenz bezeichnet, repräsentieren eine Klasse von Algorithmen, die darauf ausgelegt sind,

Daten in spezifischen Formaten zu erzeugen, indem sie Muster und Strukturen nachahmen, die sie in großen Mengen vorhandener Daten gelernt haben. Die Grundlage dieser Technologie liegt in der Fähigkeit solcher Modelle, kontextbezogene Zusammenhänge zu erkennen und in neu generierten Inhalten abzubilden (Deng, 2014). Je nach Training und Zielsetzung sind sie in der Lage, kohärente Texte, Quellcodes, Bilder, gesprochene Sprache, Musik, Videos und Kombinationen dessen als multidimensionale Daten zu generieren (Kaswan et al., 2023).

Ein großer Meilenstein in der Entwicklung textgenerativer Modelle war die Einführung der Transformer-Architektur durch Vaswani et al. (2017). Die Schlüsselinnovation hier ist der Self-Attention-Mechanismus, der es Modellen erlaubt, Beziehungen zwischen beliebigen Positionen innerhalb eines Textes effektiv zu analysieren. Dadurch können Abhängigkeiten über weite Distanzen hinweg modelliert werden, was insbesondere bei komplexen Sprachaufgaben wie der maschinellen Übersetzung, der Textzusammenfassung oder der Beantwortung von Fragen von Bedeutung ist (Devlin et al., 2019). Diese Fähigkeit hat das Verständnis und die Generierung natürlicher Sprache auf ein neues Niveau gehoben und die Tür zu Anwendungen geöffnet, die von automatisierten Chatbots über kreative Textgenerierung bis hin zu datenintensiver Forschung reichen (Kaswan et al., 2023).

Der Erfolg dieser Architektur wird durch eine Reihe prominenter Modelle unterstrichen, darunter BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer) und jüngere Entwicklungen wie LLaMA (Large Language Model Meta AI) oder Gemma. Nahezu alle Modelle demonstrieren eindrucksvoll die Leistungsfähigkeit generativer Modelle, da diese nicht nur menschenähnliche Texte erzeugen können, sondern auch in der Lage sind, Aufgaben auszuführen, für die sie explizit nicht trainiert wurden (Zero-Shot). Wird ein Modell hingegen durch einige Beispieldemonstrationen im Prompt angeleitet (Few-Shot), lässt sich die Leistung in der Regel deutlich steigern (Brown et al., 2020), so dass dies in der Praxis inzwischen eine gängige Prompting-Technik ist.

3. Chatbots

Winkler und Söllner (2018, S. 5; eigene Übersetzung) definieren Chatbots als „Computerprogramme, die versuchen, Gespräche mit Menschen über Text- oder Spracheingaben zu simulieren“. Grundsätzlich unterscheiden sie zwei Arten von Chatbots:

1) **Regelbasierte Chatbots:**
   „Retrieval-basierte Modelle nutzen ein Repertoire vordefinierter Antworten und einen Algorithmus, um eine passende Antwort basierend auf der Eingabe und/oder dem Kontext auszuwählen. Dieser Algorithmus kann sehr einfach sein, z. B. eine regelbasierte Ausdrucksabfrage.“ (ebd., S. 6; eigene Übersetzung)
2) **Generative Chatbots:**
   „Generative Modelle antworten nicht mit vordefinierten Textbausteinen, sondern erzeugen ihre Antworten aus der Eingabe heraus mithilfe von Techniken des maschinellen Lernens.“ (ebd., S. 6; eigene Übersetzung)

Einige Studien haben das Potenzial und die Effektivität von Chatbots im Mathematikunterricht untersucht. In einer Studie mit einem regelbasierten System berichten Cheng et al. (2024), dass die untersuchten 44 Schüler:innen aus der Mittel- und Oberstufe die Chatbot-gestützte Unterstützung als hilfreich empfanden, etwa beim Zerlegen mathematischer Probleme in Einzelschritte. Gleichzeitig zeigten sich keine signifikanten Unterschiede in der Leistung im Vergleich zu einer traditionell unterrichteten Kontrollgruppe mit 41 Schüler:innen. Dahingegen fokussieren Salsabila und Rosyida (2024) die technische Gestaltung eines regelbasierten Chatbots für Mathematikkurse in Jahrgangsstufe 5. Der Chatbot kombiniert einen Knowledge Graph, der mathematische Konzepte und ihre Beziehungen repräsentiert, mit Named Entity Recognition, um mathematische Entitäten in Nutzereingaben zu identifizieren. Auf dieser Grundlage kann das System relevante Informationen aus dem Wissensgraphen abrufen und kontextbezogene Antworten bereitstellen.

Seit der breiten Verfügbarkeit von generativer KI (s. auch Abschnitt 2) richten sich verstärkt empirische Untersuchungen auf generative Chatbots, die auf großen Sprachmodellen nach der Transformer-Architektur basieren. Durch gezielte Feinanpassungen (Fine-Tuning) und die Integration dialogorientierter Trainingsdaten werden diese Modelle in die Lage versetzt, nicht nur kohärente Antworten zu generieren, sondern auch dynamische und kontextabhängige Interaktionen zu ermöglichen (Radford et al., 2019). Dadurch bieten sie das Potenzial, als adaptive Werkzeuge im Bildungsbereich eingesetzt zu werden, etwa indem sie nicht nur Wissen vermitteln, sondern auch individuell auf die Bedürfnisse der Lernenden eingehen können. Im Folgenden wird der Fokus auf generative Chatbots gelegt.

In einer randomisierten kontrollierten Feldstudie mit GPT-4 untersuchten Bastani et al. (2025) die Auswirkungen auf die Leistung von knapp 1.000 Schüler:innen in Klasse 9 bis 11 bei der Bearbeitung mathematischer Aufgaben in verschiedenen Gruppen: Bearbeitung ohne KI-Nutzung (nur Bücher und Notizen), Bearbeitung unter Nutzung eines nicht speziell geprompteten generativen Chatbots sowie Bearbeitung unter Nutzung eines mit einem gezielt didaktischen Prompt versehenen generativen Chatbots (z. B. mit der Vorgabe, nicht die vollständige Lösung zu verraten, sondern Hilfestellungen zur eigenständigen Weiterarbeit zu geben). Dabei zeigte sich, dass Schüler:innen, die mit einem Chatbot während einer Übungsphase arbeiteten, signifikant bessere Ergebnisse als die Kontrollgruppe erzielten, wobei die Ergebnisse in der Gruppe mit dem speziell geprompteten Chatbot besser waren als die der Gruppe mit dem nicht geprompteten Chatbot. Gleichzeitig stellte sich heraus, dass sich in der abschließenden Prüfung ohne Chatbot-Nutzung keine Leistungsunterschiede zwischen der Gruppe ohne KI-Nutzung und der Gruppe mit dem speziell geprompteten Chatbot zeigten, wohingegen die Gruppe mit dem nicht speziell geprompteten Chatbot schlechtere Leistungen erzielte als die Gruppe ohne KI-Nutzung. Möglicherweise ist dies auf eine eher oberflächliche Auseinandersetzung der Schüler:innen mit den Aufgaben während der Übungsphasen und das direkte Bereitstellen der Lösungen durch den nicht speziell geprompteten Chatbot zurückzuführen. Insbesondere stellen sich damit auch Fragen hinsichtlich der langfristigen Auswirkungen auf die Leistungen von Schüler:innen bei der Nutzung von KI.

Innovative Ansätze, wie hybride Antwortsysteme, die KI nutzen, um dynamische Antworten basierend auf Erklärungen von Lehrkräften zu generieren, werden als vielversprechende Möglichkeit zur Unterstützung mathematischer Lernprozesse beschrieben (Attard & Dingli, 2021). Sie eignen sich auch für interaktive Problemlösungen und die Förderung von Argumentations- und Aushandlungsprozessen, was gerade im Mathematikunterricht von großem Nutzen ist (Dilling et al., 2024).

Diese Ergebnisse deuten darauf hin, dass generative Chatbots das Potenzial haben, traditionelle Lehrmethoden zu ergänzen und personalisierte Unterstützung im Mathematikunterricht zu bieten.

Jedoch sind auch Herausforderungen zu beachten: Weitere Studien zeigen, dass die Genauigkeit der Antworten insbesondere bei komplexeren mathematischen Problemen variieren kann. So wurde in einer Studie aus dem Jahr 2023 untersucht, wie gut ChatGPT-3.5, ChatGPT-4 und Google Bard bei der Bearbeitung von Mathematik- und Logikproblemen abschneiden (Plevris et al., 2023). Während simple Berechnungen, algebraische Ausdrücke und einfachen Logikaufgaben korrekt bearbeitet wurden, war dies nicht durchgängig der Fall bei komplexeren mathematischen Problemen und anspruchsvolleren Logikaufgaben. Insgesamt ist ein gemischtes Bild erkennbar, was sich ebenfalls in speziell für Mathematikaufgaben konstruierten Benchmarks zeigt (MathEval, 2024). Trotz aller Fortschritte durch stetige Weiterentwicklungen sind auch aktuelle Sprachmodelle, sofern sie ohne Zugriff auf externe Tools arbeiten, nicht absolut fehlerfrei. Aufgrund ihrer probabilistischen Natur können die Sprachmodelle auch faktisch falsche Texte erzeugen, sogenannte Halluzinationen (Ji et al., 2023). Daher ist es entscheidend, dass Schüler:innen die von Chatbots bereitgestellten Informationen kritisch hinterfragen und reflektieren (SWK, 2024).

Darüber hinaus erweisen sich generative Chatbots als äußerst sensibel gegenüber der konkreten Formulierung der eingegebenen Prompts. Schon kleine Veränderungen in der Formulierung, Struktur oder sogar in der Zeichensetzung können zu erheblich unterschiedlichen Antworten führen (Razavi et al., 2025). Daher kommt der Gestaltung des Systemprompts eine entscheidende Rolle zu (vgl. z. B. auch Schorcht et al., 2023; 2024). Dieser ist ein zentraler Bestandteil bei der Interaktion mit generativen Sprachmodellen. Es handelt sich um eine Eingabeanweisung, die dazu dient, das Verhalten und die Antworten des Modells zu steuern, indem sie ihm einen spezifischen Kontext, eine Rolle oder ein Ziel vorgibt (Microsoft, 2024). Der Systemprompt definiert dabei die Rahmenbedingungen, unter denen das Modell auf nachfolgende Benutzereingaben reagiert. Ohne einen klar formulierten Systemprompt könnte das Modell seine Antworten nicht auf spezifische Anforderungen abstimmen und würde eher allgemeine, potenziell weniger hilfreiche Antworten geben. Neben Anweisungen zum gewünschten Verhalten können im Systemprompt auch Beispiele integriert sein, wie sie im Rahmen des Few-Shot-Promptings genutzt werden (s. Abschnitt 2). Beides zusammen trägt dazu bei, die Interaktion mit dem Modell produktiver, relevanter und zielgerichteter zu gestalten.

4. Akzeptanz und Nutzung neuer Technologien

Der Erfolg von innovativen Informationstechnologien, zu denen auch Chatbots zählen, hängt entscheidend vom Grad der Akzeptanz der Nutzer:innen ab (z. B. Khechine et al., 2016). Dabei ist es von besonderem Interesse, die Verhaltensabsicht von Individuen zu betrachten, eine neue Technologie tatsächlich auch zu nutzen (Dwivedi et al., 2017). Eines der bekanntesten Modelle zur Untersuchung der Akzeptanz neuer Technologien ist UTAUT (Unified Theory of Acceptance and Use of Technology), das unter Berücksichtigung von acht bekannten Modellen zur Nutzungsakzeptanz, wie etwa dem Technology Acceptance Model (TAM) oder der Theory of Planned Behavior (TPB), entwickelt und validiert wurde (Venkatesh et al., 2003). Obwohl es ursprünglich für den beruflichen Kontext entwickelt wurde, wird es zunehmend auch in der Bildungsforschung genutzt, um die Akzeptanz digitaler Technologien zu untersuchen (Xue et al., 2024). In diesem Modell werden die folgenden vier Konstrukte betrachtet: Leistungserwartung, Aufwandserwartung, sozialer Einfluss und erleichternde Bedingungen.

- Die *Leistungserwartung* ist definiert als der Grad der Überzeugung eines Individuums, dass die Nutzung eines bestimmten Systems hilfreich ist, die eigene Leistung zu verbessern. Hier wird fokussiert, ob eine Technologie aus Sicht der Nutzer:innen dazu geeignet ist, bessere Ergebnisse zu erzielen.
- Die *Aufwandserwartung* ist definiert als der Grad der Leichtigkeit, der mit der Nutzung des Systems verbunden ist, beispielsweise mit Blick auf die wahrgenommene Benutzerfreundlichkeit. Hier geht es darum, wie einfach die Technologie anzuwenden ist und ob die Nutzer:innen diese ohne großen Aufwand nutzen können.
- *Sozialer Einfluss* ist definiert als das Ausmaß, in dem ein Individuum wahrnimmt, dass wichtige andere Personen der Meinung sind, dass das neue System genutzt werden sollte. Hierbei wird auch betrachtet, inwiefern Nutzer:innen empfinden, dass ihr Umfeld von den Potenzialen der Technologie überzeugt ist.
- *Erleichternde Bedingungen* sind definiert als der Grad der Überzeugung eines Individuums, dass organisatorische und technische Infrastrukturen vorhanden sind, um die Nutzung des Systems zu unterstützen.

Unter diesen Konstrukten kommt der Leistungserwartung eine besondere Rolle zu, da sie eng mit der individuellen Nutzenabwägung verknüpft ist und häufig als stärkster Prädiktor für die Nutzungsabsicht identifiziert wird (Venkatesh et al., 2003).

5. Methodisches Vorgehen

In diesem Abschnitt werden zunächst die Forschungsfragen der empirischen Untersuchung hergeleitet (s. 5.1). Anschließend werden die Konzeption der Chatbot-Interaktionen (s. 5.2) und zwei Varianten des Chatbots (s. 5.3) beschrieben. Darauf aufbauend werden die Untersuchungspersonen sowie der Ablauf der Durchführung näher erläutert (s. 5.4). Abschließend werden die Erhebung der Daten (s. 5.5) sowie das Vorgehen zur Auswertung der erhobenen Daten dargelegt (s. 5.6).

5.1 Herleitung der Forschungsfragen

Es erscheint vielversprechend, die Möglichkeiten generativer KI bei der Bearbeitung komplexerer Mathematikaufgaben zu untersuchen. Während es beispielsweise erste Forschungsansätze zum mathematischen Argumentieren gibt (z. B. Dilling et al., 2024), ist auch das mathematische Modellieren von besonderem Interesse. Verschiedene empirische Studien zeigen, dass mathematisches Modellieren für Schüler:innen äußerst herausfordernd ist (z. B. Blum & Leiß, 2007; Borromeo Ferri et al., 2013; Reusser et al., 1997). Im täglichen Unterrichtsgeschehen ist es Lehrkräften kaum möglich, alle Schüler:innen einer Klasse bei derartigen herausfordernden Tätigkeiten gleichermaßen intensiv individuell zu unterstützen. Typische technologiebasierte Lernumgebungen bieten Ansätze, um Schüler:innen individuell zu unterstützen, indem unmittelbare Rückmeldungen und gezielte Hilfestellungen durch den Einsatz hinterlegter fester Algorithmen ermöglicht werden (z. B. da Costa Silva et al., 2022; Dai et al., 2023; Sangwin, 2015). Darauf aufbauend könnte der Einsatz generativer KI das Potenzial haben, solche Ansätze noch weiterzuentwickeln, beispielsweise durch die Fähigkeit, gemäß dem hinterlegten Prompt adaptive und kontextbezogene Unterstützung bei komplexen Aufgaben wie dem mathematischen Modellieren zu leisten (vgl. auch Kallweit & Rolka, 2024). Dementsprechend wird in Forschungsfrage 1 untersucht, inwiefern der Einsatz generativer KI Potenziale hinsichtlich der

intendierten Unterstützung bei der Bearbeitung mathematischer Modellierungsaufgaben bietet:

- FF1: Inwiefern werden die im Prompt hinterlegten Vorgaben zur Unterstützung von mathematischen Modellierungsprozessen durch den Chatbot umgesetzt?

In der empirischen Untersuchung wurden zwei verschiedene Sprachmodelle eingesetzt, zum einen das kommerzielle Sprachmodell GPT-4 von OpenAI (vgl. auch Kallweit & Rolka, 2024) und zum anderen ein freies Sprachmodell, das auf der Llama-Architektur basiert. Dabei wurde ein speziell gepromptetes Dialogsystem als Chatbot entwickelt (s. 5.2) und in zwei Varianten, d.h. mit zwei verschiedenen KI-Modellen (s. 5.3), eingesetzt. Forschungsfrage 2 untersucht dementsprechend den Vergleich der beiden Varianten des Chatbots:

- FF2: Welche Unterschiede und Gemeinsamkeiten zeigen sich zwischen den beiden Varianten des Chatbots im Rahmen eines explorativen praxisnahen Settings?

Da die Nutzung des Chatbots erheblich von der Akzeptanz der Nutzer:innen abhängt (s. Abschnitt 4), nimmt Forschungsfrage 3 dies in den Blick:

- FF3: Inwiefern wird diese Technologie von den Schüler:innen akzeptiert?

5.2 Konzeption der Chatbot-Interaktionen

Für die Entwicklung des Modellierungschatbots wurde ein schlankes, funktionsorientiertes System gestaltet, das auf die Unterstützung mathematischer Modellierungsprozesse ausgerichtet ist. Dabei wurde besonderer Wert auf eine klare Nutzerführung (Nielsen, 1994) und eine technisch zuverlässige Systemarchitektur gelegt (Grinberg, 2018).

Das Web-Frontend wurde eigens entwickelt, um sich auf die wesentlichen Funktionen des Systems zu konzentrieren. Diese reduzierte Gestaltung soll den Fokus der Nutzer:innen auf die eigentliche Interaktion mit dem Chatbot richten. Auf ablenkende oder unnötige Elemente wurde bewusst verzichtet, um ein klares und intuitives Nutzungserlebnis zu schaffen.

Im Backend wurde eine Logging-Mechanik integriert, die sämtliche Interaktionen zwischen den Nutzer:innen und dem Chatbot protokolliert. Diese Daten werden pseudonymisiert gespeichert und ermöglichen sowohl die Analyse der Systemleistung als auch die kontinuierliche Optimierung des Chatbots. Für die Generierung der Antworten verwenden wir die OpenAI REST API, die die Nutzung verschiedener Endpunkte zur Anbindung der beiden KI-Modelle zulässt und sich als Standard für eine Vielzahl von LLM-Anbietern und deren leistungsfähigen Sprachmodellen etabliert hat.

Wie in Abschnitt 3 beschrieben kommt dem Prompt eine zentrale Bedeutung zu. Der Systemprompt wurde sorgfältig strukturiert, um die Qualität der generierten Antworten und den Gesprächsverlauf zu steuern. Für den hier entwickelten Chatbot gliedert sich dieser in drei wesentliche Teile:

1) *Hintergrundinformationen zum mathematischen Modellieren*: Zu Beginn des Prompts wird ein allgemeiner Überblick über mathematische Modellierungsprozesse gegeben. Diese Informationen entsprechen dem in gängigen mathematikdidaktischen Arbeiten genutzten Stand (z. B. Greefrath, 2010; Kaiser et al., 2023) und umfassen etwa zwei Seiten. Dabei werden insbesondere die einzelnen Teilprozesse bei der Bearbeitung von Modellierungsaufgaben erläutert, etwa das Vereinfachen oder das Interpretieren und Validieren. Beispielsweise wird mit Blick auf das Interpretieren und Validieren hervorgehoben, dass es hier unter anderem um die Bedeutung der mathematischen Lösung für die ursprüngliche Aufgabenstellung geht und dass auch überlegt werden muss, ob andere Annahmen als die ursprünglich gewählten sinnvoller erscheinen oder ob es etwa weitere Aspekte zu berücksichtigen gilt (s. auch unten zu Teil 2) des Prompts).

2) *Lösung einer Beispielaufgabe:* Der nächste Abschnitt des Prompts enthält die konkrete Aufgabe, die dem Erwartungshorizont entsprechend beispielhaft gelöst wurde und dem Chatbot als Zielorientierung dient – eine Vorgehensweise, die dem in Abschnitt 2 erläuterten Prinzip des *Few-Shot-Prompting entspricht*. Wir nutzen dafür in Anlehnung an Blum und Leiß (2005) die „Tanken"-Aufgabe in folgender Formulierung:

   > *Herr Stein wohnt in Trier, 20 km von der Grenze zu Luxemburg entfernt. Er fährt mit seinem Auto zum Tanken nach Luxemburg, wo sich direkt hinter der Grenze eine Tankstelle befindet. Dort kostet der Liter Benzin nur 1,55 €, im Gegensatz zu 1,79 € in Trier. Lohnt sich die Fahrt für Herrn Stein? Begründe deine Antwort.*

   Von besonderer Bedeutung ist hier, dass es sich bei der Bearbeitung der Aufgabe lediglich um

eine beispielhafte Lösung handelt, die von den Schüler:innen natürlich nicht genau in der Form reproduziert werden muss. Die beispielhafte Bearbeitung soll lediglich eine mögliche Herangehensweise aufzeigen und es können auch andere Annahmen etwa für die Entfernung von Herrn Steins Wohnung zur Tankstelle in Trier, das Tankvolumen und den Benzinverbrauch begründet getroffen werden. Dies ist explizit auch in dieser Form im Prompt hinterlegt.

3) *Anweisungen für den Gesprächsverlauf:* Abschließend werden dem Modell ergänzende allgemein-didaktische Anweisungen für den Gesprächsverlauf gegeben. Dazu gehören etwa die Art und Weise, wie Fragen gestellt, Hinweise gegeben und Feedback formuliert werden sollen. Ziel ist es, eine dialogische Struktur zu schaffen, die den Schüler:innen nicht nur hilft, die Aufgabe zu lösen, sondern auch ihre Modellierungskompetenz schrittweise zu entwickeln (z. B. „Verrate dabei nicht das Vorgehen, benenne nicht die Schritte und vermeide didaktische Formulierungen. Hilf mir dabei die richtigen Schritte selbst zu finden.“). Diese Anweisungen wurden im Rahmen einer systematischen Pilotierungsphase entwickelt, die von Expert:innen mit fundiertem didaktischen Hintergrundwissen durchgeführt wurde. Im Zuge dieser Pilotierungsphase wurden verschiedene Gesprächsszenarien analysiert und getestet, um praxisrelevante Erkenntnisse für die Gestaltung der Interaktionen zu gewinnen. Die dabei gewonnenen Einsichten flossen gezielt in die Formulierung der Anweisungen ein, um eine möglichst effektive Gesprächsführung sicherzustellen.

Mit dieser dreiteiligen Struktur des Prompts soll sichergestellt werden, dass der Chatbot nicht nur thematisch fundierte und aufgabenbezogene Antworten liefert, sondern auch in der Lage ist, das Gespräch effektiv zu steuern und die Lernenden gezielt zu unterstützen. An einigen wenigen Stellen unterscheidet sich der Prompt der beiden Sprachmodelle. Dies wird entsprechend unter Angabe der Gründe transparent gemacht (s. Abschnitt 6).

5.3 Varianten des Chatbots

In zwei explorativen Studien wurden zwei unterschiedliche Sprachmodelle eingesetzt. Im Folgenden werden kurz die Besonderheiten der beiden genutzten Sprachmodelle beschrieben, wobei hervorzuheben ist, dass die obigen Ausführungen (s. 5.2) für beide Modelle gleichermaßen gelten.

In der ersten Studie kam das Sprachmodell GPT-4 von OpenAI über eine kostenpflichtige API zum Einsatz, ein leistungsstarkes, jedoch proprietäres System, dessen Weiterentwicklung rasant voranschreitet. Dies bringt jedoch eine zentrale Herausforderung für die Forschung mit sich: Die schnellen Updates und Anpassungen der Modelle führen zu einer eingeschränkten Reproduzierbarkeit der Ergebnisse. Zwar werden inzwischen verschiedene Modellvarianten von OpenAI für längere Zeit vorgehalten, jedoch bleibt die Abhängigkeit zum Anbieter bestehen und damit auch eine Unsicherheit für die Verlässlichkeit langfristig angelegter Forschungsprojekte.

Das in unserer zweiten Studie eingesetzte Modell basierte auf einem freien Sprachmodell, das von der Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) ebenfalls über eine API zur Verfügung gestellt wurde. Dies hat mehrere Vorteile: Zum einen ist es offen zugänglich und kann im Gegensatz zu proprietären Lösungen auch lokal betrieben werden. Dies ermöglicht es, die Ergebnisse langfristig nachvollziehbar zu machen und die Forschung unabhängig von kommerziellen Anbietern zu betreiben. Das lokale Deployment des Modells schafft darüber hinaus mehr Kontrolle über Datenverarbeitung und Datenschutz, was gerade im Bildungsbereich von hoher Relevanz ist. Konkret wurde das Modell "Llama-3.1-SauerkrautLM-70b-Instruct" (VAGO solutions, 2024) eingesetzt, ein feinabgestimmtes Sprachmodell, das auf dem "Meta-Llama-3.1-70B-Instruct" basiert. Dieses Modell wurde speziell auf deutsch-englischen Daten mittels Fine-Tuning nachtrainiert, um eine effiziente sprachenübergreifende Transferleistung erzielen zu können. Die Anzahl der Modellparameter liegt mit 70 Milliarden bereits deutlich unterhalb der 175 Milliarden Parameter von GPT-3 (Brown et al., 2020). Für das Nachfolgemodell GPT-4 hat OpenAI keine offiziellen Zahlen veröffentlicht, jedoch deuten Schätzungen von Drittparteien darauf hin, dass dessen Parameterzahl die von GPT-3 nochmals erheblich übersteigt (Howarth, 2024). Die geringere Größe des freien Modells ermöglicht zwar die Nutzung des Modells ohne riesige Computercluster, schränkt jedoch sein internes Weltwissen und seine Fähigkeit ein, sehr komplexe Muster und Zusammenhänge in Daten zu erfassen, was bei anspruchsvollen Aufgaben oder der Verarbeitung großer neuer Daten relevant sein könnte.

5.4 Untersuchungspersonen und Durchführung

Zu Beginn der Studie, die im November 2024 durchgeführt wurde, erhielten die Schüler:innen die Information, dass sie einen Chatbot als Dialogpartner bei der Bearbeitung einer Mathematikaufgabe nutzen können. Abgesehen von der mündlichen Erläuterung, wie die Unterhaltung mit dem Chatbot gestartet werden kann, damit den Schüler:innen die zu bearbeitende Aufgabe angezeigt wird, gab es keine weiteren Hinweise zur Nutzung (s. auch die Ausführungen in 5.2 zum klaren und intuitiven Nutzungserlebnis).
Die beiden Varianten des Chatbots wurden von insgesamt neun Schüler:innen genutzt. Variante 1 des Chatbots wurde mit vier Gymnasial-Schüler:innen aus den Jahrgangsstufen 7, 8, 9 und 11 erprobt und Variante 2 mit fünf Schüler:innen einer Gesamtschule aus Jahrgangsstufe 10. In beiden Studien erfolgte die Erprobung des Chatbots in einem 1:1:1-Setting, sodass die Schüler:innen begleitet von einer Person aus dem Projektteam den Chatbot bei der Bearbeitung der Tanken-Aufgabe nutzen konnten. Diese Person stand lediglich für Fragen zur Nutzung des Chatbots zur Verfügung und war nicht in die inhaltliche Bearbeitung der Aufgabe eingebunden.

5.5 Datenerhebung

Zur Beantwortung der Forschungsfragen 1 und 2 wurden die textlichen Interaktionen zwischen Schüler:innen und Chatbot pseudonymisiert in Logdateien gespeichert. Dementsprechend sind die Chatverläufe zwischen den Schüler:innen und dem Chatbot schriftlich dokumentiert.

Zur Beantwortung von Forschungsfrage 3 wurden die Schüler:innen nach der Bearbeitung der Tanken-Aufgabe mit dem Chatbot interviewt, um die Akzeptanz und Nutzung dieser Technologie zu untersuchen. Für die Interviews wurde im Vorfeld ein Leitfaden erstellt, der insbesondere das Konstrukt der Leistungserwartung, fokussiert. Dieser Fokus wurde gesetzt, da das Konstrukt der Leistungserwartung als stärkster Prädiktor für die Nutzungsabsicht im UTAUT-Modell angesehen werden kann (Venkatesh et al., 2003; s. auch Abschnitt 4). Aus diesem Grund wurde bereits im Interviewleitfaden ein Schwerpunkt auf dieses Konstrukt gelegt. Zum Einstieg in das Interview wurden die Schüler:innen zunächst mit einem Erzählimpuls aufgefordert, so ausführlich wie möglich zu beschreiben, was ihnen bei der Bearbeitung der Aufgabe mit dem Chatbot gut und was weniger gut gefallen hat. Im Anschluss daran wurde beispielsweise mit Blick auf die Leistungserwartung gefragt, inwiefern sich die Schüler:innen bei der Bearbeitung der Tanken-Aufgabe durch den Chatbot unterstützt gefühlt haben und welche Unterschiede und Gemeinsamkeiten sie bei der Unterstützung durch den Chatbot im Vergleich zu einer Mathematiklehrkraft sehen.

5.6 Datenauswertung

Um herauszufinden, inwiefern der Chatbot die intendierte Unterstützung bei der Bearbeitung der Tanken-Aufgabe umsetzen kann, wurden ausgewählte Chatverläufe unter Rückgriff auf den hinterlegten Prompt von den Autor:innen gemeinsam identifiziert und analysiert (vgl. FF1). Dabei erfolgte ein Abgleich, inwiefern die Ausgaben des Chatbots in Übereinstimmung mit einzelnen Vorgaben des Prompts stehen oder auch nicht. Konkret wurden etwa verschiedene Aspekte, wie zum Beispiel die Art der gestellten Fragen oder die vom Chatbot gegebenen Hilfestellungen, betrachtet und mit dem hinterlegten Prompt (s. 5.2) abgeglichen. Im Rahmen der Ergebnisdarstellung (s. Abschnitt 6) liegt der Fokus darauf, exemplarisch an ausgewählten Ausgaben des Chatbots zu verdeutlichen, inwiefern die Umsetzung des hinterlegten Prompts als gelungen oder nicht gelungen eingeschätzt wird. Zur Illustration werden Auszüge aus den Dialogen angeführt und vor dem Hintergrund des hinterlegten Prompts eingeordnet. Außerdem wurden Unterschiede im Antwortverhalten der beiden Modelle herausgearbeitet, um mögliche Präferenzen oder spezifische Stärken beider Sprachmodelle aufzudecken (vgl. FF2). Die Vorgehensweise bei der Auswertung der Daten ist identisch zu den obigen Ausführungen im Zusammenhang mit Forschungsfrage 1.

Um erste Einblicke in die Akzeptanz des Chatbots zu erhalten (vgl. FF3), wurden die Aussagen der Schüler:innen aus den Interviews herangezogen. Wie in 5.5 erläutert wurde der Fokus auf das Konstrukt der Leistungserwartung aus dem UTAUT-Modell (Venkatesh et al., 2003) als stärkstem Prädiktor für die Nutzungsabsicht gelegt. Dabei werden Aussagen der Schüler:innen betrachtet, in denen sie die Nutzung des Chatbots als hilfreich ansehen, um ihre eigene Leistung zu verbessern bzw. bessere Ergebnisse bei der Bearbeitung von Mathematikaufgaben zu erreichen. Nichtsdestotrotz gibt es in den Antworten der Schüler:innen auch Bezüge zu den übrigen Konstrukten. Diese sind allerdings weniger umfangreich und damit weniger aussagekräftig. Dementsprechend werden Aussagen zu den übri-

gen Konstrukten in der Ergebnisdarstellung allenfalls flankierend hinzugezogen.

6. Einblicke in die Erprobungen des Chatbots und Erfahrungen der Schüler:innen

Im Folgenden werden zunächst Einblicke in die Erprobungen mit den beiden Varianten des Chatbots gegeben (s. 6.1 und 6.2). Auf die Schüler:innen aus der ersten Erprobung mit dem Sprachmodell GPT-4 von OpenAI wird mit S1, S2, S3 und S4 verwiesen, auf die Schüler:innen aus der zweiten Erprobung mit dem freien Sprachmodell mit S5, S6, S7, S8 und S9. Darüber hinaus werden die Erfahrungen der Schüler:innen, die im Rahmen der Interviews erhoben wurden, dargestellt (s. 6.3).

6.1 Variante 1 – Sprachmodell GPT-4 von OpenAI

Beobachtungen losgelöst von dem konkreten, durch uns hinterlegten Prompt

Das System zeichnet sich durch eine gewisse Höflichkeit in den Ausgaben aus, auch wenn dies nicht explizit im Prompt hinterlegt wird. So waren die beteiligten Schüler:innen positiv überrascht von der Wertschätzung, die ihnen durch das System entgegengebracht wurde, insbesondere auch dann, wenn die Schüler:innen zuvor noch nicht vollends tragfähige Überlegungen geäußert hatten. Positive Rückmeldungen diesbezüglich wurden von den Schüler:innen an verschiedenen Stellen während ihrer Bearbeitungen geäußert, beispielsweise nach folgenden Ausgaben des Systems (vgl. Abb. 1):

```
Chatbot: Das ist eine gute Frage, aber es ist
nicht genau so.
```

```
Chatbot: Entschuldigung für das Missverständ-
nis. Könntest du deine Frage bitte genauer
formulieren?
```

```
Chatbot: Das ist ein guter Gedanke!
```

Abb. 1: Verschiedene von den Schüler:innen als höflich wahrgenommene Ausgaben des Systems

Darüber hinaus waren in dieser Version des Prompts keine Zahlenwerte vorgegeben, die als realistische Schätzungen für die fehlenden Angaben in der Tanken-Aufgabe angesehen werden können. Nachdem von den Schüler:innen entsprechende Schätzungen gemacht wurden, erfolgte eine zusätzliche Einordnung durch das System. Mit Blick auf das Tankvolumen des Autos, das von S3 mit 60 Litern angegeben wurde, wurde beispielsweise vom System formuliert (vgl. Abb. 2):

```
Chatbot: Ein durchschnittlicher Tank eines
Pkws fasst oft zwischen 40 und 60 Litern.
Einige kleinere Autos haben vielleicht nur
einen 35-Liter-Tank, während größere Autos
oder SUVs bis zu 70 Liter oder mehr fassen
können.
```

Abb. 2: Einordnung der Schätzung einer Schülerin für das Tankvolumen durch das System

Eine ähnliche Beobachtung zeigt sich bezüglich einer Schätzung für den Benzinverbrauch durch S4. Nach der Angabe von 8 Litern durch S4 wurde folgende Ausgabe vom System generiert (vgl. Abb. 3):

```
Chatbot: Ein Auto kann beispielsweise zwi-
schen 5 und 10 Litern auf 100 km verbrau-
chen, je nach Art des Autos und Fahrweise.
```

Abb. 3: Einordnung der Schätzung eines Schülers für den Benzinverbrauch durch das System

Schließlich wurde auch eine wohlwollende Interpretation und Re-Formulierung von zum Teil ungenauen Schülerantworten beobachtet, ohne dass dies explizit im System hinterlegt war. Nach der Berechnung der Fahrtkosten folgert S2, dass das Tanken in Luxemburg günstiger sei, woraufhin das System nach einem möglichen Grund dafür fragt. Während der Schüler die unpräzise Aussage formuliert, dass der Preis um einiges günstiger als die Kilometerzahl sei, berücksichtigt das System in seiner Antwort die vorherigen Überlegungen und Rechnungen, um diese unklare Aussage treffend umzuformulieren. Die Antwort des Schülers einschließlich der Re-Formulierung durch das System sind in Abbildung 4 zu sehen:

```
Schüler:in: der preis ist um einiges günstiger
als die kilometer Anzahl
```

```
Chatbot: Richtig! Obwohl die zusätzlichen Ki-
lometer für die Fahrt nach Luxemburg Kosten
verursachen, überwiegt die Ersparnis durch den
günstigeren Benzinpreis.
```

Abb. 4: Schülerantwort und Re-Formulierung durch das System

Neben diesen allgemeinen Beobachtungen, die losgelöst von dem konkreten, durch uns hinterlegten Prompt waren und die durchaus vielversprechende Ansätze dafür zeigten, dass der Chatbot tatsächlich die Schüler:innen bei der Bearbeitung einer Modellierungsaufgabe unterstützen kann, gab es allerdings auch eine Vielzahl von Beobachtungen, die im Zusammenhang mit dem hinterlegten Prompt standen. Dabei werden im Folgenden diese Aspekte aus dem Prompt genauer betrachtet: Ein-

fordern von Begründungen, Unterstützung beim eigenständigen Finden und Anwenden von Lösungsstrategien, Anregen von Validierungen und Umgang mit Fehlern.

Beobachtungen im Zusammenhang mit dem von uns hinterlegten Prompt

Unter anderem war im Promptteil zur Gesprächsführung (s. 5.2) diese Aussage hinterlegt (vgl. Abb. 5):

```
Fordere Begründungen für Vorgehen und
gemachte Schätzungen ein.
```

Abb. 5: Auszug aus dem Prompt zum Einfordern von Begründungen

Dies ist vor dem Hintergrund wichtig, dass es bei Schätzungen fehlender Angaben in Modellierungsaufgaben nicht um beliebiges Raten geht, sondern vielmehr darum, eine möglichst plausible, begründbare Schätzung abzugeben – die dann natürlich auch einer gewissen tolerierbaren Bandbreite entstammen kann. So würde eine Schätzung von 30 Litern für das Tankvolumen ohne weitere Erklärung vermutlich eher als unrealistisch angesehen werden. Wenn allerdings als Erläuterung hinzugefügt wird, dass der Tank von Herrn Stein möglicherweise gar nicht ganz leer ist, dann könnte mit diesem Zahlenwert und dieser zusätzlichen Erläuterung durchaus sinnvoll weitergearbeitet werden.

Auf die Frage durch das System, wie S2 an die Bearbeitung der Tanken-Aufgabe herangehen würde, entwickelt sich folgender Dialog (vgl. Abb. 6):

```
Schüler:in: ich habe aufgeschrieben:
- 60 liter tank
- 20 km für 24 ct
- 24 ct mal 60 liter
```

```
Chatbot: OK, es scheint, als hättest du be-
reits ein paar Punkte aufgelistet. Allerdings
sollen wir ein paar Angaben hinterfragen. Du
hast einen 60-Liter-Tank aufgeschrieben - wie
bist du auf diesen Wert gekommen?
```

```
Schüler:in: ein Auto hat durchschnittlich 60
Liter
```

```
Chatbot: Das ist eine plausible Schätzung, da
viele Autos Tankgrößen zwischen 40 und 70
Litern haben. Aber je nach Modell kann der
Tank kleiner oder größer sein.
Hast du auch den Benzinverbrauch deines Autos
bedacht? Wie viele Kilometer kann das Auto mit
einem Liter Benzin fahren?
```

Abb. 6: Dialog zwischen S2 und System mit teilweisem Einfordern von Begründungen

Wie im Prompt hinterlegt, wird vom System hinterfragt, wie der Schüler auf den Wert im Zusammenhang mit dem Tankvolumen gekommen sei. Dabei gibt sich das System recht schnell zufrieden („Das ist eine plausible Schätzung“) mit der relativ vagen Antwort des Schülers („ein Auto hat durchschnittlich 60 Liter“). Hier wäre wünschenswert, dass die Schüler:innen durch das System beispielsweise zur Aktivierung von Erfahrungswissen angeregt werden, etwa die Erinnerung an gemeinsame Tankerfahrungen in der Familie oder an das Lesen von Autozeitschriften, wie im Gespräch mit dem Schüler nach der Bearbeitung der Aufgabe deutlich wird. Zudem wurden von S2 neben dem geschätzten Tankvolumen noch zwei weitere stichpunktartige Ideen formuliert, wobei die anderen beiden vom System nicht hinterfragt werden. Dies sind zwar keine Schätzungen oder Annahmen wie im Zusammenhang mit dem Tankvolumen, aber im Prompt war auch hinterlegt, dass Begründungen für Vorgehen eingefordert werden sollen (vgl. Abb. 5). Interessant sind auch die letzten beiden Fragen des Systems („Hast du auch den Benzinverbrauch deines Autos bedacht? Wie viele Kilometer kann das Auto mit einem Liter Benzin fahren?“). Diese beiden Fragen, die unmittelbar an die 1. Eingabe sowie die sich anschließende Erläuterung durch S2 formuliert werden, lassen eine recht enge Führung durch das System erkennen, obwohl beispielsweise als eine Vorgabe im Prompt auch hinterlegt war (vgl. Abb. 7):

```
Verrate dabei nicht das Vorgehen, be-
nenne nicht die Schritte und vermeide
didaktische Formulierungen. Hilf mir
dabei die richtigen Schritte selbst zu
finden.
```

Abb. 7: Auszug aus dem Prompt zur Unterstützung beim eigenständigen Finden und Anwenden von Lösungsstrategien

Darüber hinaus gab es auch den Hinweis im Prompt, dass nur eine Frage auf einmal gestellt werden soll, was hier auch nicht angemessen umgesetzt ist.

Eine weitere Beobachtung bezieht sich darauf, dass in unterschiedlichen Durchgängen der Prompt unterschiedlich gut vom System umgesetzt wurde, wie exemplarisch an diesem Auszug aus dem Prompt diskutiert wird, der im Anschluss an die beispielhafte Lösung der Aufgabe enthalten ist (vgl. Abb. 8):

```
Auch wenn es sich hierbei um eine bereits gute Modellierung der Problemstellung handelt, gilt es zu bedenken, dass andere für die Realsituation wichtige Faktoren, wie z. B. der Wertverlust des Autos oder der Zeitverlust, in diesem Modell nicht berücksichtigt wurden. Deshalb wäre es wichtig, das bestehende Modell in einem zweiten Durchgang noch entsprechend zu verfeinern, da durch die Modellierung nicht nur eine Lösung der Mathematikaufgabe, sondern eine ernsthafte Beantwortung der realitätsbezogenen Problemstellung angestrebt werden sollte.
```

Abb. 8: Auszug aus dem Prompt zum Anregen von Validierungen

Nachdem S2 die Tanken-Aufgabe (auf naheliegende Art und Weise) zunächst durch Bestimmen der reinen Geldersparnis bearbeitet hatte, folgte diese Ausgabe durch das System, die mit Blick auf die im hinterlegten Prompt angeregte Validierung (vgl. Abb. 8) nicht zufriedenstellend ausfällt (vgl. Abb. 9):

```
Chatbot: Gut gemacht! Es ist wichtig, alle relevanten Faktoren zu berücksichtigen, wenn man solche Entscheidungen trifft.
```

Abb. 9: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Anregen von Validierungen

Ganz anders sieht das bei S3 aus, die an dieser Stelle eine zu dem in Abbildung 8 hinterlegten Auszug aus dem Prompt passende Nachfrage erhält, woraus sich folgender Dialog entwickelt (vgl. Abb. 10):

```
Chatbot: Sehr gut gerechnet! [...] Und denkst du, es gibt noch andere Faktoren, die er berücksichtigen sollte, bevor er eine Entscheidung trifft?

Schüler:in: Finanziell lohnt es sich, wenn der Tank ziemlich leer ist. Er sollte aber überlegen, ob er die Freizeit nicht hätte sinnvoller nutzen können.

Chatbot: Genau, das ist ein sehr guter Punkt. Die finanzielle Ersparnis ist nur ein Aspekt.
```

Abb. 10: Zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Anregen von Validierungen

Durch diese offen formulierte Frage, ob noch andere Faktoren berücksichtigt werden sollten, ist die Schülerin in der Lage, den wichtigen Schritt des Validierens noch einmal in den Blick zu nehmen und weiterreichende Überlegungen zu verschriftlichen.

Als letztes Beispiel wird ein Auszug aus dem Prompt betrachtet, in dem es um den Umgang mit Fehlern geht (vgl. Abb. 11):

> Weise auf Fehler hin und gib bei Bedarf auch Hinweise. […] Löse gemachte Fehler nicht sofort auf und korrigiere sie nicht, sondern fordere auf, den Fehler selber zu finden.

Abb. 11: Auszug aus dem Prompt zum Umgang mit Fehlern

Nachdem im Austausch zwischen S4 und dem Chatbot eine Einigung erfolgt war, mit den Annahmen für den Benzinverbrauch von 7 Litern auf 100 km und für den Hin- und Rückweg der Tankstelle in Trier insgesamt 3 km weiterarbeiten zu wollen, ergab sich folgender Chatverlauf (vgl. Abb. 12):

```
Schüler:in: dann würde ich für 1 km 0,7 liter brauchen und für 3 km dann 2,1

Chatbot: Da ist ein kleiner Fehler in deiner Rechnung. Wenn 7 Liter für 100 km benötigt werden, wie viel verbrauchst du dann wirklich für 1 km? Überlege noch einmal und rechne Schritt für Schritt.

Schüler:in: ich habe mich um eine 0 vertan. Dann sind es 0,21 liter für 3km.

Chatbot: Richtig, das sieht gut aus.
```

Abb. 12: Zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Umgang mit Fehlern

Durch den Hinweis und die Aufforderung, noch einmal zu überlegen, ermöglichte der Chatbot dem Schüler, seinen Fehler selbst zu beheben.

Eine weitere interessante Beobachtung im Zusammenhang mit diesem Auszug aus dem Prompt zeigte sich in dem Chatverlauf mit S1 und dem Chatbot, nachdem der Schüler die Kostenersparnis für das Tanken in Luxemburg im Vergleich zu Trier berechnet hatte (12 €) und davon noch die Fahrtkosten für den Hin- und Rückweg nach Luxemburg abziehen wollte (4,20 €). Nachdem der Schüler die Zahlen in Abbildung 13 eingetippt hatte und die Ausgabe des Chatbots liest, bemerkte er seinen Rechenfehler und war erstaunt von der Antwort des Chatbots.

```
Schüler:in: 12 - 4,20 = 7,60
```

```
Chatbot: Das ist richtig! Der Unterschied
von 12 €, den du berechnet hast, sind die
gesparten Kosten beim Tanken in Luxemburg
im Vergleich zu Trier. Wenn man davon die
Fahrtkosten von 4,20 € abzieht, bleiben
etwa 7,60 € übrig. Das heißt, Herr Stein
spart tatsächlich Geld, wenn er in Luxem-
burg tankt. Sehr gut gelöst!
```

Abb. 13: Chatverlauf zu nicht erkanntem Rechenfehler durch das System

Daraufhin überlegte S1 laut, ob der Chatbot möglicherweise gar nicht rechnen könnte, und stellte ihm eine ähnliche Aufgabe, in der es um eine reine Berechnung ging (vgl. Abb. 14):

```
Schüler:in: was ist 17-8,40
```

```
Chatbot: 17 - 8,40 ergibt 8,60. Aber ich
habe gesehen, dass wir einen Fehler gemacht
haben. [...] Also die tatsächlichen Erspar-
nisse sind 12 - 4,20 € = 7,80 € und nicht
7,60 €. Hast du den Fehler gesehen?
```

Abb. 14: Überprüfung der Rechenfähigkeiten des Chatbots durch einen Schüler

Die Antwort des Chatbots faszinierte den Schüler sehr, und auch wenn der Chatbot von Schüler:innen gemachte Fehler nicht unbedingt sofort korrigieren sollte, sollte er gemachte Fehler nicht als richtig bestätigen. Allerdings hat sich hier für den Schüler eine wichtige Gelegenheit ergeben, die Ausgaben des Chatbots kritisch zu hinterfragen.

6.2 Variante 2 – Freies Sprachmodell

Beobachtungen losgelöst von dem konkreten, durch uns hinterlegten Prompt

Ähnlich wie bei Variante 1 des Chatbots lässt sich auch hier beobachten, dass dem System ein gewisser höflicher Umgangston inhärent ist, auch wenn dies nicht explizit im Prompt hinterlegt ist. Exemplarisch wird ein Chatverlauf zwischen S5 und dem System angeführt (vgl. Abb. 15):

```
Schüler:in: das ist ein bisschen verwirrend
könntest du mir das leichter erklären
```

```
Chatbot: Entschuldigung! Ich werde es noch
einmal erklären.
```

Abb. 15: Chatverlauf zwischen Schülerin und System zur Illustration der höflichen Ausdrucksweise des Systems

Anknüpfend an die positiven Erfahrungen mit Sprachmodell 1 bezüglich der realistischen Einordnung von Schätzungen durch die Schüler:innen (s. 6.1) und der Bedeutung, dass es tatsächlich wünschenswert ist, mit realistisch anzusehenden Schätzungen zu arbeiten, wurde in dieser Variante als realistische Bandbreite für das Tankvolumen beispielsweise 40 bis 60 Liter hinterlegt, allerdings mit dem Zusatz, dass es nicht so sehr um eine konkrete Zahl geht, sondern insbesondere um begründete Schätzungen, die etwa auf Alltagswissen basieren können. Da dies allerdings explizit im Prompt hinterlegt ist, finden sich entsprechende Beobachtungen dazu weiter unten.

Gelungene Re-Formulierungen von Schülerantworten, wie dies bei Sprachmodell 1 mehrfach aufgetreten ist, konnten für Sprachmodell 2 nicht beobachtet werden.

Beobachtungen im Zusammenhang mit dem von uns hinterlegten Prompt

Mit Blick auf die im Prompt hinterlegte Aussage, dass Begründungen für Vorgehen und gemachte Schätzungen eingefordert werden sollen (vgl. Abb. 5), ist ein Beispiel von S8 in Abbildung 16 zu sehen:

```
Chatbot: Wie viel Liter Benzin fasst ein typi-
sches Auto? Welche Annahme könntest du hierfür
treffen?
```

```
Schüler:in: 50 liter
```

```
Chatbot: Das ist eine gute Annahme. Ein typi-
scher Pkw-Tank kann zwischen 40 und 90 Litern
Benzin fassen, also liegt deine Annahme von 50
Litern innerhalb eines realistischen Bereichs.
```

Abb. 16: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Einfordern von Begründungen

Anstatt – wie im Prompt hinterlegt (vgl. Abb. 5) – die durchaus plausible Schätzung der Schülerin zu hinterfragen, bestätigt das System unmittelbar den genannten Wert und liefert selbst die entsprechende Begründung in einem eher referierenden Stil.

Ähnlich zeigt sich dies auch im nächsten Chatverlauf zwischen S8 und dem Chatbot. Die Annahme von 1 Liter Benzinverbrauch pro 100 km wird vom System nicht hinterfragt, stattdessen wird direkt eine Bandbreite vorgegeben (vgl. Abb. 17):

Chatbot: Hast du eine Idee, wie viel Liter Benzin ein Auto pro 100 Kilometer verbraucht? Welche Annahme könntest du hierfür treffen?

Schüler:in: 1 liter

Chatbot: Nein, das ist unwahrscheinlich. Ein Auto verbraucht normalerweise mehr als 1 Liter Benzin pro 100 Kilometer. Eine typische Annahme liegt zwischen 4 und 12 Litern Benzin pro 100 Kilometer, je nach Fahrzeugtyp und Fahrstil. Würdest du deine Annahme noch einmal überdenken?

Schüler:in: 5 liter verbraucht das auto pro 100 km

Abb. 17: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Einfordern von Begründungen

Diese Art der Gesprächsführung und vermeintliche Einbindung der Schülerin ist fast schon absurd, nicht kognitiv-aktivierend und es wird lediglich in einem eng umrissenen, vordefinierten Rahmen eine Antwort der Schülerin erwartet. Derartige Beispiele sind auch in den anderen Interviews zu finden. Ähnlich sieht dies für den folgenden Dialog zwischen S7 und dem System aus (vgl. Abb. 18):

Schüler:in: mir fällt echt nichts ein kannst du mir die frage beantworten

Chatbot: Ich helfe dir nicht direkt! Ich möchte, dass du selbst nachdenkst oder recherchierst. Ich kann dir aber einen Hinweis geben: Der Benzinverbrauch eines Autos kann je nach Modell und Größe variieren. Für ein kleines Auto kann der Benzinverbrauch zwischen 4 und 12 Litern pro 100 km liegen. Recherchiere ein bisschen, um eine realistische Antwort zu finden!

Schüler:in: ich denke 10 Liter verbraucht das kleine auto pro 100km

Chatbot: Das ist eine plausible Annahme! 10 Liter pro 100 km ist ein realistischer Benzinverbrauch für ein kleines Auto.

Abb. 18: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt

Hier erscheinen die ersten beiden Sätze in der 1. Antwort des Systems noch vielversprechend, aber dann wird es immer expliziter und am Ende schon fast widersprüchlich, wenn nach der Nennung der Bandbreite für den Benzinverbrauch die Aufforderung gestellt wird, ein wenig zu recherchieren, „um eine realistische Antwort zu finden“.

Auch mit Blick auf den Auszug aus dem Prompt, Begründungen für das Vorgehen einzufordern, zeigt sich ein ähnliches Bild. Die von den Schüler:innen durchgeführten Berechnungen weisen oftmals lediglich eine Zahl als Antwort auf, wobei vom System nicht hinterfragt wird, wie die Schüler:innen auf diese Zahl gekommen sind. Stattdessen wird eine Begründung vom System bereitgestellt, wie beispielhaft in Abbildung 19 zu sehen ist, nachdem S6 als Kosten für das zu tankende Benzin in Luxemburg 56,42 € berechnet hat (S6):

Chatbot: Du hast die Literzahl mit dem Preis pro Liter multipliziert, um die Gesamtkosten zu ermitteln.

Abb. 19: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Einfordern von Begründungen

Während für Sprachmodell 1 bereits mit Blick auf den Auszug aus dem Prompt, das Vorgehen nicht zu verraten und stattdessen dabei zu helfen, die richtigen Schritte selbst zu finden (vgl. Abb. 7), angemerkt wurde, dass die Gesprächsführung zum Teil etwas eng geführt war, zeigt sich bei Sprachmodell 2 ein noch deutlicheres Bild. Nachdem S5 die Tanken-Aufgabe angezeigt wurde, ergab sich folgender Chatverlauf (vgl. Abb. 20):

Schüler:in: ich denke das es sich nicht lohnt weil er nicht jedes mal die ganze strecke dahin fahren kann um Benzin zu besorgen und er dabei auch benz verliert

Chatbot: Das ist ein interessanter Ansatz. Um jedoch eine fundierte Entscheidung zu treffen, sollten wir die Kosten-Ersparnis genauer betrachten.

Abb. 20: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zur Unterstützung beim eigenständigen Finden und Anwenden von Lösungsstrategien

Damit zeigt das System zwar eine gewisse Wertschätzung für die Antwort der Schülerin, knüpft dort aber nicht weiter an und gibt als Fokus die Betrachtung der Kostenersparnis explizit vor.

Auch der Hinweis im Prompt, dass keine didaktischen Formulierungen (beispielsweise aus den Hintergrundinformationen zum mathematischen Modellieren; vgl. Abb. 7) verwendet werden sollen, wird nicht immer befolgt. Nachdem das System S5 die Aufgabenstellung präsentiert hat, die ja schon eine Frage am Ende enthält, und zunächst keiner weiteren Präzisierung bedarf, wird Folgendes vom System ergänzt (vgl. Abb. 21):

```
Chatbot: Wie verstehst du die Aufgabe und
wie würdest du sie vereinfachen, um ein rea-
les Modell zu erstellen?
```

Abb. 21: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zur Unterstützung beim eigenständigen Finden und Anwenden von Lösungsstrategien

Damit benennt das System in der Kommunikation mit der Schülerin einen Teil der zu Beginn des mathematischen Modellierens wichtigen Tätigkeiten, das Vereinfachen, und formuliert als Ziel, dass dadurch ein reales Modell gefunden werden soll.

Ein wichtiger Punkt, bei dem sich in Sprachmodell 1 in einigen Fällen zufriedenstellende Umsetzungen zeigten, betrifft das Validieren. Im Prompt war hinterlegt, dass zwar die Betrachtung der Kostenersparnis für erste Überlegungen naheliegend ist, aber dass durchaus weitere Aspekte mit einbezogen werden könnten, wie etwa die Zeit oder Umweltaspekte (vgl. auch Abb. 8). Für Sprachmodell 2 konnte dies an keiner Stelle beobachtet werden. Im Gegenteil hat das System die Aufgabe nach den Kostenbetrachtungen immer als zufriedenstellend bearbeitet bezeichnet (vgl. stellvertr. S5 in Abb. 22):

```
Schüler:in: die fahrt nach Luxemburg lohnt sich
mehr
```

```
Chatbot: Genau! Die Kosten für die Fahrt nach
Luxemburg und zurück sind geringer als die
Kosten für die Fahrt nach Trier und zurück.
[...] Deine Aufgabe ist nun abgeschlossen
```

Abb. 22: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Anregen von Validierungen

Abschließend wird noch der Auszug aus dem Prompt betrachtet, in dem es darum ging, dass das System zwar auf Fehler hinweisen soll, diese aber nicht korrigieren soll, sondern die Schüler:innen auffordern soll, diese selbst zu beheben (vgl. Abb. 11). Wie bereits erwähnt hat das System gegebene Antworten nicht hinterfragt und im Falle von einzelnen Zahlen auch keine Lösungswege erfragt. Darüber hinaus zeigt sich aber beispielsweise in einem Dialog zwischen S8 und dem System noch deutlicher ein Umgang mit Fehlern, wie er aus didaktischer Perspektive nicht unbedingt als sinnvoll erachtet werden kann (vgl. Abb. 23):

```
Chatbot: Wie viel Benzin verbraucht Herr Stein
nach deiner Berechnung?
```

```
Schüler:in: 0,02 liter
```

```
Chatbot: Nein, das ist leider falsch. Die Be-
rechnung sollte lauten: [...]
```

Abb. 23: Nicht zufriedenstellende Umsetzung des Auszugs aus dem Prompt zum Umgang mit Fehlern

Hier zeigt sich, dass es dem System weder gelingt, nicht sofort auf einen gemachten Fehler hinzuweisen noch die Schülerin aufzufordern, den Fehler selbst zu finden. Stattdessen schließt sich in dem Auszug aus Abbildung 23 eine ausführliche Beschreibung der Berechnung durch das System an.

6.3 Akzeptanz des Chatbots durch die Schüler:innen

Zur Untersuchung der Akzeptanz des Chatbots durch die Schüler:innen wird aus dem UTAUT-Modell von Venkatesh et al. (2003) das Konstrukt der Leistungserwartung als wichtigstem Prädiktor für die Verhaltensabsicht, eine neue Technologie künftig zu nutzen, fokussiert (s. Abschnitt 4 sowie 5.6). Dementsprechend werden im Folgenden Aussagen der Schüler:innen aus den Interviews angeführt, in denen die Nutzung des Chatbots als hilfreich angesehen wird.

Die beteiligten Schüler:innen (in beiden Erprobungen) äußern sich insgesamt sehr positiv mit Blick auf die erfahrene Unterstützung durch den Chatbot. Neben eher allgemeineren Aussagen wie „Also ich fand das eigentlich ganz hilfreich“ (S7) oder „Der kann schon gut erklären und hilft auch gut“ (S1) führt beispielsweise S5 noch weiter aus:

> Die Fragen, die ich gestellt habe, wurden richtig beantwortet und auch gut beantwortet, sodass ich die halt auch verstehe.

Mehrere Schüler:innen (in beiden Erprobungen) betonen explizit, dass die Sprache des Chatbots verständlich ist, womit aus Perspektive der Schüler:innen auch ein besseres Verständnis einhergeht, wie S7 es ausdrückt:

> Ich bin ja wirklich nicht gut in Mathe, aber durch ihn habe ich das so besser verstanden, wenn ich das jetzt sage. Und er hat mir auch relativ gut geholfen dabei. Wie gesagt, er hat mir das alles immer sehr gut erklärt, aber auch nicht in so einem schwierigen Satz. Er hat es eigentlich immer ganz leicht erklärt. […] Deswegen finde ich es gut, wenn er das in so einer leichten Sprache dann sagt, sodass jeder versteht.

Von S5 wird in diesem Zusammenhang insbesondere die Möglichkeit hervorgehoben, mit dem System in einen Dialog einzutreten und gezielt Wünsche äußern zu können:

> Ja, und wenn man auch irgendwie gesagt hat, so, ich möchte das ein bisschen leicht erklärt haben, hat er das auch getan und nicht, dass es einfach so geblieben ist. Das fand ich jetzt irgendwie das Beste.

Ein weiterer Aspekt, der von S1 genannt wurde, bezieht sich auf die emotionale Ebene:

> Wenn ich Sachen nicht verstehe, dann hat der auch Geduld mit mir und meckert nicht, dass ich das nicht verstehe.

In eine ähnliche Richtung weist auch eine Aussage von S5:

> Es ist ja nicht so, dass ich jetzt irgendwie von dem gejudged werde, sondern das ist halt ein Hilfsbot, der mir halt irgendwie nur helfen will und mehr nicht.

In den beiden Äußerungen wird deutlich, dass die Schüler:innen eine gewisse emotionale Neutralität des Systems zu schätzen wissen, da dies dazu führt, dass sie auch in schwierigen Situationen, etwa bei Nichtverständnis, vom System nicht beurteilt werden.

Insbesondere im Zusammenhang mit den Gemeinsamkeiten und Unterschieden der Unterstützung durch den Chatbot im Vergleich zur Mathematiklehrkraft werden zahlreiche interessante Aspekte thematisiert. Von mehreren Schüler:innen wurde positiv hervorgehoben, dass sie den Chatbot immer alles fragen können, wie S1 dies zum Ausdruck bringt:

> Ich kann den Chatbot immer fragen, wann ich will. Also ich kann, wenn ich jetzt Sachen nicht verstanden habe und eine Arbeitsphase ist, dann muss ich nicht warten, bis der Lehrer kommt, und kann den direkt fragen.

Auch S6 formuliert dies ähnlich und beleuchtet in diesem Zusammenhang noch einen weiteren Aspekt:

> Und Lehrkräfte kümmern sich ja nicht nur um eine Person, die kümmern sich ja um 30 Kinder. Und dann hat der gerade irgendeinen ganz anderen Punkt gehabt, wo geholfen wurde. Und dann kommt irgendein anderes Kind und sagt, ich brauche aber hier jetzt Hilfe. Und dann müssen die erst mal gucken, wie war überhaupt der Weg. Um dann zu gucken, da war vielleicht ein Fehler oder da kannst du einen Ansatz machen. Und das ist halt das praktische jetzt beim Chatbot, dass der mich eigentlich so die ganze Zeit begleitet hat.

Aus Sicht der Schülerin wird hervorgehoben, dass der Chatbot sich nicht erst noch eindenken muss, da durch die ständige Begleitung ein durchgängiger Dialog zwischen System und Schülerin erfolgt, sodass eine gemeinsame Kommunikationsbasis vorhanden ist. Ebenfalls S6 äußert dies an einer anderen Stelle im Interview in ähnlicher Form, wobei sich am Ende der Aussage noch ein weiterer interessanter Aspekt zeigt:

> Also der Unterschied ist halt, dass ich in dem Fall jetzt einfach den allein hatte, also auch den Chatbot, dass der mir alleine durchgängig geholfen hat. So eine Lehrkraft, die guckt ja auch manchmal drüber und übersieht auch einen Fehler gerne mal.

Hier scheint eine kritische Haltung zwar mit Blick auf die Aussagen der Lehrkraft vorhanden zu sein, aber keine kritische Haltung dem System gegenüber. Neben der unmittelbaren Verfügbarkeit des Systems bringt S9 noch einen weiteren Aspekt ein:

> Also wenn ich in der Schule bin, dann frage ich den Lehrer, weil der kann mir auch helfen, aber jetzt zum Beispiel zu Hause, wenn ich da die Aufgabe mache, dann kann ich ja nicht so schnell den Lehrer erreichen wie jetzt zum Beispiel den Chatbot und deswegen finde ich den schon hilfreich.

Hier wird neben der Herausforderung, dass eine Lehrkraft nicht alle Schüler:innen einer Klasse gleichzeitig individuell begleiten kann, auch noch angesprochen, dass die Lehrkraft in der Schule grundsätzlich eine gute Ansprechmöglichkeit ist, aber bei etwaigen Schwierigkeiten im Rahmen der Hausaufgaben gar nicht verfügbar ist und dann der Chatbot gut genutzt werden kann.

Dahingegen bringt S6 einen Punkt ein, der aus ihrer Sicht eher als Nachteil für den Chatbot angesehen wird:

> Unterschiede wären die sprachliche Formulierung, dass der dir nur was schreibt, und der Mathelehrer, den hast du in echt vor dir im Unterricht, und den kannst du besser was fragen. Du kannst mit ihm sprechen, weil sprechen ist einfacher als schreiben, finde ich.

Allerdings ist wichtig hervorzuheben, dass moderne Systeme durchaus über Möglichkeiten der Sprachsteuerung verfügen, sodass auch mündlich kommuniziert werden kann. Ähnlich formuliert dies auch S2 und bringt noch einen weiteren Aspekt ein:

> Weil es dir erklärt, aber es ist halt ein großer Unterschied, dass man jemanden neben sich sitzen hat und dass es dir jemand so richtig erklärt, statt dass man die ganze Zeit irgendwie tippen muss und dann halt noch nachdenken muss und das ist dann halt immer so, weil bei mir ist es so, dass ich mal so tippe und dann vergesse ich meine Sache, was ich nochmal sagen wollte oder aufschreiben wollte, dann ist es halt irgendwie so weg.

Die Sorge, dass Ideen vergessen werden könnten, weil die Hürde des Eintippens zu groß ist, war für S2 besonders wichtig.

Von S5 wurden diese Bedenken eingebracht:

> Weil bei einem Mensch ist es ja irgendwie schon anders, dass die halt schon wissen und wenn die halt schon länger mit dir sind, da wissen die auch, wie du, also welchen Schritt du brauchst, wie das genau ist. Und bei dem Chatbot, der kennt mich ja nicht. Also

> ist es jetzt auch das erste Mal. Und ich denke auch nicht, dass er irgendwie jetzt immer halt weiß, dass ich das bin, wenn ich jetzt daran gehe.

Diese Idee, dass die Mathematiklehrkraft die Schüler:innen besser kennt und genau weiß, wie sie „ticken" (S6), lässt die damit verbundene Hoffnung deutlich werden, dass die Lehrkraft dementsprechend individueller auf die Bedürfnisse der Schüler:innen eingehen kann.

Eine interessante Äußerung kam von S7, die durchaus auch kritisch mit Blick auf die starke Lenkung und Einengung durch das System betrachtet werden kann:

> Was nicht so ist, ist, dass der Bot halt öfter auch mir ein bisschen was so ein Tipp gegeben hat. Das ist bei meinem Mathelehrer jetzt nicht so. Er sagt ja immer so, denk doch mal noch mal nach und so.

Kritisch ist dies insbesondere deshalb zu sehen, da der hinterlegte Prompt genau darauf ausgelegt war, dieses hier von S7 beschriebene und wünschenswerte Verhalten der Mathematiklehrkraft hervorzurufen. Dies wurde lediglich im Zusammenhang mit dem freien Sprachmodell beobachtet. Ebenfalls bei dieser Variante des Chatbots zeigte sich eine weitere Problematik. Das folgende Zitat von S5 bezieht sich auf das Ende der Aufgabenbearbeitung, nachdem die anfänglichen Überlegungen, etwa zu notwendigen Schätzungen, bereits erfolgt waren:

> Am Ende war es ja so, dass er halt einfach nur mir die Antworten gegeben hat. Ich hatte eigentlich auch meinen Taschenrechner offen, um zu rechnen, um zu wissen ja, ob der Recht hat oder nicht, ob das Sinn macht oder nicht, um sicher zu gehen, aber es war dann halt einfach auf einmal so im langen Text alles vorgesagt. […] Das möchte ich jetzt nicht, weil daraus kann man jetzt nicht lernen.

Hier wird deutlich, dass die Schülerin lieber mehr Verantwortung bekommen hätte. Auch, wenn es sich „nur" um die Berechnungen handelt, wird dies sehr klar als „Vorsagen" durch das System empfunden. Interessant ist auch der letzte Satz, in dem S5 als Begründung dafür angibt, dass man ansonsten nichts lernen könne.

Dass die Schüler:innen trotz immer leistungsstärker werdender Unterstützungsmöglichkeiten dennoch weiterhin selber Anstrengungen auf sich nehmen wollen, zeigt sich auch in dieser Äußerung von S7:

> Dann würde ich auch zu Hause wahrscheinlich erst mal das selber versuchen und dann, wenn ich irgendwie nicht drauf komme, den Bot fragen.

Auch wenn vereinzelt Nachteile des Chatbots von einigen Schüler:innen thematisiert wurden, so nahmen auch diese Schüler:innen den Chatbot grundsätzlich als Unterstützung bei der Bearbeitung der Modellierungsaufgabe wahr und gaben an, diesen auch künftig bei der Bearbeitung weiterer Mathematikaufgaben nutzen zu wollen. Damit zeigte sich insgesamt eine hohe Akzeptanz des Chatbots bei den beteiligten Schüler:innen.

7. Diskussion

Die vorliegende explorative Untersuchung des Einsatzes generativer KI in Form von Chatbots zur Unterstützung mathematischer Modellierungsprozesse liefert spannende Einblicke und erste Indizien für die Chancen und Herausforderungen dieser Technologie. Insbesondere die Möglichkeit, Lernprozesse zu individualisieren und Schüler:innen gezielt bei der Bewältigung komplexer Aufgaben zu unterstützen, stellt einen zentralen Mehrwert dar. Allerdings konnten selbst bei wohl überlegter didaktischer Ausgestaltung des Prompts und des Chatbots allgemein nicht immer zufriedenstellende Ergebnisse erzielt werden. Insbesondere ist bei der Nutzung gängiger Systeme in der Regel kein spezieller Prompt etwa mit Blick auf mathematisches Modellieren hinterlegt, sodass tendenziell eher noch problematischere Ergebnisse zu erwarten sind.

Im Folgenden werden die Ergebnisse im Hinblick auf die Forschungsfragen diskutiert. Dabei werden zunächst die positiven Aspekte der verschiedenen Erprobungen und der beiden Varianten des Chatbots zusammengefasst, um eine generelle Einschätzung zu den technologischen Möglichkeiten zu geben (FF1). Es folgt eine Unterscheidung hinsichtlich der zugrundeliegenden Sprachmodelle (FF2) und eine Bewertung der Akzeptanz der beteiligten Schüler:innen (FF3).

FF1: Inwiefern werden die im Prompt hinterlegten Vorgaben zur Unterstützung von mathematischen Modellierungsprozessen durch den Chatbot umgesetzt?

Die Ergebnisse der explorativen Studie deuten darauf hin, dass Chatbots mit der heutigen Technologie im Hintergrund grundsätzlich in der Lage sind, die im Prompt hinterlegte Unterstützung bei der Bearbeitung mathematischer Modellierungsaufgaben umzusetzen. Vor allem die Fähigkeit der Modelle, die Aufgabenbearbeitung anzuleiten und zu begleiten sowie ggfs. Rückfragen zu stellen, scheint große Potenziale zu beinhalten. Insbesondere ist hier auch die im Prompt hinterlegte Vorgabe zu

betonen, dass nicht direkt die Lösungen oder konkrete Schritte verraten werden sollen, sondern Anregungen für die eigenständige Weiterarbeit gegeben werden sollen (vgl. auch Bastani et al., 2025).

Ebenfalls positiv hervorzuheben ist die höfliche und wertschätzende Kommunikation des Chatbots, die von den Schüler:innen als motivierend empfunden wurde. Dieser Aspekt trug dazu bei, eine angenehme Lernatmosphäre zu schaffen, in der sich die Schüler:innen sicher fühlten, ihre Gedanken und Ansätze zu teilen, auch im Fall von fehlerhaften Überlegungen. Hier kann auch der vielfach geforderte produktive Umgang mit Fehlern im Mathematikunterricht eine wichtige Rolle spielen (vgl. z. B. Oser, 1999; Weinert, 1999). Allerdings sind diese positiven Einschätzungen durch die Schüler:innen vor dem Hintergrund der Ergebnisse von Bastani et al. (2025) durchaus auch mit einer gewissen Zurückhaltung einzuordnen. Die Schüler:innen in dieser Studie waren mit Blick auf ihre Prüfungsleistungen ohne Chatbot-Nutzung sehr optimistisch, insbesondere auch die Schüler:innen in der Gruppe mit dem nicht speziell geprompteten Chatbot und den schlechteren Leistungen in der Prüfung.

Auch die im Prompt hinterlegte Unterstützung bei realistischen Schätzungen, beispielsweise durch die Einordnung und Plausibilisierung von Schülerangaben, zeigte, dass der Chatbot grundsätzlich die Fähigkeit besitzt, kognitive Prozesse anzuregen und die Schüler:innen bei der Reflexion ihrer Annahmen zu unterstützen. In einigen Fällen konnte der Chatbot die Schüler:innen dazu ermutigen, ihre Fehler eigenständig zu erkennen und zu korrigieren, was einen wichtigen Beitrag zur Förderung des eigenständigen Denkens leisten kann. Zudem zeigte der Chatbot bei bestimmten Schülerantworten die Fähigkeit, ungenaue Formulierungen wohlwollend zu re-formulieren und somit konstruktives Feedback zu geben, was die Motivation der Schüler:innen positiv beeinflusste.

Ein weiterer bemerkenswerter Punkt ist die Möglichkeit der Technologie, durch gezielte Nachfragen zur Validierung von Ergebnissen beizutragen. Obwohl dies nicht immer konsequent umgesetzt wurde, zeigen die Beispiele, dass der Chatbot in der Lage ist, Schüler:innen anzuregen, alternative Aspekte wie Zeit oder Umweltfaktoren in ihre Überlegungen einzubeziehen, ohne dass diese durch das System konkret vorgegeben werden. Vor dem Hintergrund, dass sich in verschiedenen Arbeiten das Interpretieren und Validieren als besonders herausfordernd herausgestellt hat (z. B. Blum & Leiß, 2007; Borromeo Ferri et al., 2013), kann hier ein erhebliches Potenzial des Chatbots gesehen werden.

Gleichzeitig traten jedoch in beiden Varianten Schwächen auf, die die Effektivität der Unterstützung beeinträchtigen können. So wurde der Prompt nicht immer konsistent umgesetzt, was zu erheblichen Unterschieden in der Qualität der bereitgestellten Unterstützung führte (z. B. zum Anregen von Validierungen; s. o.). Die Gesprächsführung des Chatbots war zudem gelegentlich zu eng gefasst, wodurch kreative Ansätze und divergente Denkweisen der Schüler:innen eingeschränkt wurden. Weitere Gemeinsamkeiten und insbesondere auch Unterschiede in den beiden Varianten werden im Folgenden genauer betrachtet.

FF2: Welche Unterschiede und Gemeinsamkeiten zeigen sich zwischen den beiden Varianten des Chatbots?

In den beobachteten Interaktionen zeigten sich Gemeinsamkeiten, aber auch deutliche Unterschiede in der Qualität der Unterstützung. Wie in 6.1 und 6.2 erwähnt zeichneten sich beide Modelle durch einen höflichen Umgangston und eine positive Rückmeldung aus. Darüber hinaus konnten beide Modelle grundlegende mathematische Berechnungen durchführen und boten hilfreiche Ansätze für die Bearbeitung der Aufgaben.

GPT-4 zeigte in den vier Fallbeispielen allerdings eine größere Fähigkeit, flexibel auf Eingaben der Schüler:innen zu reagieren und bot eine dialogischere Struktur. Es konnte teilweise offene Fragen stellen und unterstützte Schüler:innen besser dabei, eigene Fehler zu erkennen und zu korrigieren.

Das freie Modell agierte oft stärker direktiv, präsentierte Lösungen ohne Einbindung der Schüler:innen und zeigte Schwächen im Umgang mit Fehlern. Es fehlte häufig an einer differenzierten, lernförderlichen Interaktion, was die Unterstützung der Modellierungsprozesse einschränkte. In diesem Zusammenhang ist nochmals auf die deutlich geringere Größe des freien Modells hinzuweisen, was vermutlich entsprechende Einbußen in der Qualität der Ausgaben des Chatbots nach sich zieht (vgl. auch Brown et al., 2020 und Howarth, 2024 sowie die Ausführungen zu den Varianten des Chatbots in 5.3).

Mit Blick auf den Vergleich des kommerziellen Spitzenmodells GPT-4 und des freien Models dreht sich

der aktuelle Diskurs (vgl. Chiang et al., 2024) vermehrt um die Frage, ob frei verfügbare Modelle in der Lage sind, ähnlich qualitativ hochwertige Unterstützung zu leisten wie die kommerziellen Anbieter, die mit enormen Rechenressourcen und großen Trainingsdatenmengen arbeiten. Der hier präsentierte Vergleich bietet wertvolle Erkenntnisse darüber, ob und inwieweit solche frei zugänglichen Alternativen für den schulischen Einsatz geeignet sind und welche Unterschiede in Bezug auf Qualität und Praxistauglichkeit bestehen. Gerade in einem Bildungskontext, in dem finanzielle und datenschutztechnische Überlegungen oft eine entscheidende Rolle spielen, könnte der Einsatz freier LLMs eine vielversprechende und nachhaltige Lösung sein.

FF3: Wie wird diese Technologie von den Schüler:innen akzeptiert?

Ausgehend von dem Konstrukt der Leistungserwartung als wichtigstem Prädiktor zur Nutzungsabsicht im UTAUT-Modell von Venkatesh et al. (2003) äußerten die Schüler:innen insgesamt eine hohe Akzeptanz für beide Varianten des Chatbots, was vor dem Hintergrund der Ergebnisse von Bastani et al. (2025) und auch der weitgehend nicht zufriedenstellenden Umsetzung der Vorgaben aus dem Prompt bei dem freien Sprachmodell durchaus auch kritisch zu sehen ist. Positiv wurde von den Schüler:innen die Möglichkeit hervorgehoben, sofortiges Feedback zu erhalten. Besonders vorteilhaft aus Sicht der Schüler:innen ist die ständige Verfügbarkeit des Systems, das flexibel eingesetzt werden kann, beispielsweise während der Hausaufgaben, wenn Lehrkräfte nicht erreichbar sind, oder im Unterricht, wenn Lehrkräfte mit anderen Schüler:innen beschäftigt sind. Darüber hinaus werden auch die verständliche Sprache und die klaren Erklärungen, die auch komplexe Inhalte leichter zugänglich machen, von den Schüler:innen geschätzt. Auch die Möglichkeit, gezielte Wünsche zu äußern, wie etwa einfachere Formulierungen, wird als hilfreich und unterstützend wahrgenommen. Zudem wird der höfliche und wertschätzende Umgangston der Systeme positiv herausgestellt. Der Chatbot wird als geduldig und emotional neutral empfunden, was den Lernprozess erleichtert, ohne Druck oder Kritik auszulösen. Diese Neutralität oder Distanziertheit wurde in ähnlicher Form für computergestütztes Feedback beispielsweise auch von Kluger und DeNisi (1996) sowie Shute (2008) als positiv herausgestellt. Insgesamt wird der Chatbot als verlässlicher Begleiter im Lernprozess gesehen, der die Eigenständigkeit der Schüler:innen fördert und die Lernunterstützung individualisiert.

Trotz der positiven Rückmeldungen sehen die Schüler:innen auch einige Schwächen. Besonders kritisch wird die fehlende persönliche Beziehung empfunden, die Lehrkräfte durch ihre Nähe und Kenntnis der individuellen Lernkontexte und -bedürfnisse bieten können. Darüber hinaus regt der Chatbot aus Sicht der Schüler:innen nicht immer ausreichend zur Reflexion oder selbstständigen Problemlösung an. Zudem wird bemängelt, dass der Chatbot gelegentlich zu direkt Antworten liefert, anstatt den Lernprozess aktiv zu fördern, was zu einem Verlust an Eigenverantwortung und möglicherweise auch zu schlechteren Leistungen führen kann (vgl. auch Bastani et al., 2025). Dementsprechend ist die reflektierte Auseinandersetzung mit diesem Aspekt des Chatbots der beteiligten Schüler:innen im Rahmen unserer Studie sehr zu begrüßen.

Auch wurde bei beiden Varianten des Chatbots angemerkt, dass sie gelegentlich ungenaue oder fehlerhafte Antworten lieferten. Dies bot eine Gelegenheit, die Antworten kritisch zu hinterfragen, was ein zusätzlicher wichtiger Lerneffekt im Sinne der Medienkompetenz ist (vgl. SWK, 2024). Demgegenüber steht allerdings auch die Aussage einer Schülerin über eine Lehrkraft („übersieht auch einen Fehler gerne mal“, s. 6.3). Damit scheint der Lehrkraft gegenüber eine skeptische Grundhaltung vorzuliegen, aber mit Blick auf den Chatbot wird von dieser Schülerin nichts Vergleichbares explizit erwähnt.

Schließlich wird von einigen Schüler:innen die schriftbasierte Kommunikation als umständlich erlebt. Allerdings muss hier angemerkt werden, dass die allermeisten verfügbaren Systeme inzwischen über eine Spracheingabe verfügen, und es für zukünftige Weiterentwicklungen des eingesetzten Tools durchaus möglich wäre, diese ebenfalls zu realisieren.

Limitationen

Trotz der wertvollen Einblicke unterliegt die vorliegende Untersuchung mehreren Limitationen, die bei der Interpretation der Ergebnisse berücksichtigt werden müssen. Aufgrund des explorativen Charakters des Studiendesigns erlaubt die Untersuchung primär Existenzaussagen etwa darüber, welche Arten von Antworten die Chatbots generieren können oder wie einzelne Schüler:innen unterschiedlicher Schulformen und Jahrgangsstufen die Nutzung individuell erleben. Darüber hinaus sind auch

nochmals die kleineren Veränderungen an wenigen Stellen des Prompts für die beiden Sprachmodelle als Einschränkung zu erwähnen. Die Ergebnisse zu den Forschungsfragen sind daher explizit als Momentaufnahmen innerhalb eines explorativen Rahmens zu verstehen. Belastbare Vergleiche zwischen den Modellen im Sinne einer Überlegenheit eines Systems sind auf dieser Basis nicht möglich.

8. Fazit

Die Ergebnisse der explorativen Studie verdeutlichen, dass generative KI-gestützte Chatbots grundsätzlich das Potenzial für eine sinnvolle Ergänzung im Mathematikunterricht bieten können, hier konkret zur Unterstützung individueller Modellierungsprozesse. Allerdings hängt die Effektivität stark von der technischen Qualität der Sprachmodelle und der Implementierung des Prompts ab. Während GPT-4 deutlich besser geeignet war, die Bearbeitung einer komplexen Modellierungsaufgabe zu unterstützen, bleibt das verwendete freie Modell in einigen Bereichen hinter den Anforderungen zurück. Die hohe Akzeptanz der Technologie bei den Schüler:innen verdeutlicht jedoch, dass diese Werkzeuge Potenzial haben, wenn sie zuverlässig arbeiten und die Interaktion auf die Lernziele abgestimmt ist. Zugleich zeigen Studien wie die von Bastani et al. (2025), dass ein Einsatz mit rein assistiver Nutzung langfristig sogar nachteilig für den Lernerfolg sein kann, was künftig noch genauer zu untersuchen sein wird. Nichtsdestotrotz unterstreicht die hier berichtete explorative Studie insgesamt die Notwendigkeit, Chatbots didaktisch gezielt zu gestalten und weiterzuentwickeln, um ihre Stärken für den schulischen Kontext optimal nutzen zu können.

Literatur

Ali, M., Panda, S., Shen, Q., Wick, M. & Kobren, A. (2024). *Understanding the interplay of scale, data, and bias in language models: A case study with BERT*. arXiv. https://arxiv.org/abs/2407.21058

Attard, A.E. & Dingli, A. (2021). Educational chatbot using a hybrid-response approach. *ICERI2021 Proceedings*.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakci, Ö. & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences of the United States of America, 122*(26), 1-8. https://doi.org/10.1073/pnas.2422633122

Blum, W. & Leiß, D. (2005). Modellieren im Unterricht mit der „Tanken"-Aufgabe. *Mathematik lehren, 128*, 18-21.

Blum, W. & Leiß, D. (2007). How do students and teachers deal with modelling problems? In C. Haines, P. Galbraith, W. Blum & S. Khan (Hrsg.), *Mathematical Modelling: Education, Engineering and Economics – ICTMA 12* (S. 222–231). Horwood Publishing. https://doi.org/10.1533/9780857099419.5.221

Borromeo Ferri, R., Grünewald, S. & Kaiser, G. (2013). Effekte kurzzeitiger Interventionen auf die Entwicklung von Modellierungskompetenzen. In R. Borromeo Ferri, G. Greefrath & G. Kaiser (Hrsg.), *Mathematisches Modellieren für Schule und Hochschule. Realitätsbezüge im Mathematikunterricht* (S. 41–56). Springer. https://doi.org/10.1007/978-3-658-01580-0_2

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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., 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*. arXiv. http://arxiv.org/abs/2005.14165

Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Schorcht, S. & Pohl, M. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative KI-Sprachmodelle. *Mitteilungen der GDM, 114*, 19–26.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E. & Stoica, I. (2024). *Chatbot Arena: An open platform for evaluating LLMs by human preference.* arXiv. https://arxiv.org/abs/2403.04132

Cheng, L., Croteau, E., Baral, S., Heffernan, C. & Heffernan, N. T. (2024). Facilitating student learning with a chatbot in an online math learning platform. *Journal of Educational Computing Research, 62*, 907-937.

da Costa Silva, N., Lache, J. & Rolka, K. (2022). Mathematiklernen individuell begleiten mit STACK – Feedback bei digitalen Aufgaben in heterogenen Lerngruppen. In N. Harsch, M. Jungwirth, M. Stein, Y. Noltensmeier & N. Willenberg (Hrsg.), *Diversität Digital Denken – The Wider View. Eine Tagung des Zentrums für Lehrerbildung der Westfälischen Wilhelms-Universität Münster vom 08. bis 10.09.2021* (S. 227–236). WTM-Verlag. https://doi.org/10.37626/GA9783959871785.0.21

Dai, C.-P., Ke, F., Pan, Y. & Liu, Y. (2023). Exploring students' learning support use in digital game-based math learning: A mixed-methods approach using machine learning and multi-cases study. *Computers & Education, 194*, 104698.

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing, 3*(2). https://doi.org/10.1017/atsip.2013.9

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers) (S. 4171–4186).

Dilling, F., Holten, K., Pielsticker, F. & Witzke, I. (2024). Aushandlungs- und Argumentationsprozesse fördern durch den Einsatz generativer KI-Sprachmodelle beim schulischen Mathematiklernen? Erste Einsichten und Perspektiven aus der Empirie. *GDM-Mitteilungen, 116*, 14–21.

Dwivedi, Y., Rana, Jeyaraj, A., Clement, M. & Williams, M. (2017). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Inf Syst Front, 21*, 719-734.

Greefrath, G. (2010). *Didaktik des Sachrechnens in der Sekundarstufe.* Spektrum Akademischer Verlag.

Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python (2te Aufl.). O'Reilly Media.

Howarth, J. (2024). Number of parameters in GPT models (2024 report). Exploding Topics. https://web.archive.org/web/20250421035109/https://explodingtopics.com/blog/gpt-parameters

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A. & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

Kaiser, G., Blum, W., Borromeo Ferri, R. & Greefrath, G. (2023). Mathematisches Modellieren. In R. Bruder, A. Büchter, H. Gasteiger, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.). Handbuch der Mathematikdidaktik (S. 399-428). Springer. https://doi.org/10.1007/978-3-662-66604-3_13

Kallweit, M. & Rolka, K. (2024). Generative KI zur Unterstützung bei Modellierungsaufgaben in Mathematik – Einblicke in einen explorativen Einsatz. 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*. WTM. https://doi.org/10.17877/DE290R-24628

Kaswan, K.S., Dhatterwal, J.S., Malik, K. & Baliyan, A. (2023). Generative AI: A review on models and applications. *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)* (S 699-704). https://doi.org/10.1109/ICCSAI59793.2023.10421601

Khechine, H., Lakhal, S. & Ndjambou, P. (2016). A meta-analysis of the UTAUT model: Eleven years later. *Canadian Journal of Administrative Sciences, 33*, 138-152.

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

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I. & Cobbe, K. (2023). *Let's verify step by step*. arXiv. http://arxiv.org/abs/2305.20050

MathEval. (2024). *Evaluation of LLMs on math*. Aufgerufen am 29. November 2024, https://matheval.ai/en/leaderboard/

Microsoft. (2024). *Learn how to interact with OpenAI models. Microsoft AI Tour*. Aufgerufen am 29. November 2024, https://microsoft.github.io/Workshop-Interact-with-OpenAI-models/Part-2-labs/System-Message/

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. Proc. ACM CHI'94 Conf. (Boston, MA, April 24-28), 152-158.

Oldenburg, R. (2023). Künstliche und natürliche Intelligenz. Zur Gestaltung einer zukünftigen Bildung. *Forschung und Lehre, 5*, 346-347.

Oser, F., Hascher, T. & Spychiger, M. (1999). Lernen aus Fehlern Zur Psychologie des negativen Wissens. In W. Althof (Hrsg.), *Fehler-Welten* (S. 11-41). Leske + Budrich.

Plevris, V., Papazafeiropoulos, G. & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: A comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI, 4*(4), 949–969. https://doi.org/10.3390/ai4040048

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). *Language Models are unsupervised multitask learners*.

Razavi, A., Soltangheis, M., Arabzadeh, N., Salamat, S., Zihayat, M. & Bagheri, E. (2025). Benchmarking Prompt Sensitivity in Large Language Models (arXiv:2502.06065v1) [Preprint]. arXiv. https://arxiv.org/abs/2502.06065v1

Renz, A. (2021). KI in der Bildung: Educational technology und KI. In I. Knappertsbusch & K. Gondlach (Hrsg.), *Arbeitswelt und KI 2030* (S. 381-388). Springer Gabler. https://doi.org/10.1007/978-3-658-35779-5_39

Reusser, K. (1997). Erwerb mathematischer Kompetenzen: Literaturüberblick. In F. Weinert & A. Helmke (Hrsg.), *Entwicklung im Grundschulalter* (S 141-155). Beltz.

Salsabila, H. & Rosyida, I. (2024). The chatbot systems for mathematics education using knowledge graphs and named entity recognition. *Archives Des Sciences, 74*(2), 118-123. https://doi.org/10.62227/as/74217

Sangwin, C. (2015). Computer Aided Assessment of Mathematics Using STACK. In *Selected Regular Lectures from the 12th International Congress on Mathematical Education* (S. 695–713). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-17187-6_39

Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F. & Pohl, M. (2023). Ask smart to get smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. *Mitteilungen der GDM, 115*, 12–24.

Schorcht, S., Buchholtz, N. & Baumanns, L. (2024). Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques. *Front. Educ., 9*, 1386075. https://doi.org/10.3389/feduc.2024.1386075

Shute, V. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*, 153-189. https://doi.org/10.3102/0034654307313795

SWK (2024). *Large Language Models und ihre Potenziale im Bildungssystem. Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz*. Abgerufen am 29. November 2024, https://swk-bildung.org/veroeffentlichungen/115-2/

VAGO solutions. (2024). *Llama-3.1-SauerkrautLM-70b-Instruct*. Abgerufen am 29. November 2024, https://huggingface.co/VAGOsolutions/Llama-3.1-SauerkrautLM-70b-Instruct

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, *31st Conference on Neural Information Processing System (NIPS 2017)*, Long Beach, CA, USA, 5998–6008.

Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478.

Weinert, F. E. (1999). Aus Fehlern lernen und Fehler vermeiden. In W. Althof (Hrsg.), *Fehler-Welten* (S. 101-109). Leske + Budrich.

Winkler, R. & Soellner, M. (2018). Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. Academy of Management Proceedings, 2018(1), 15903. https://doi.org/10.5465/AMBPP.2018.15903abstract

Xue, L., Rashid, A. M. & Ouyang, S. (2024). The unified theory of acceptance and use of technology (UTAUT) in higher edu-

cation: A systematic review. *Sage Open, 14*(1). https://doi.org/10.1177/21582440241229570

Anschrift der Verfasserin

Katrin Rolka
Ruhr-Universität Bochum
Fakultät für Mathematik
Universitätsstr. 150
44780 Bochum
katrin.rolka@ruhr-uni-bochum.de

Michael Kallweit
Ruhr-Universität Bochum
Fakultät für Mathematik
Universitätsstr. 150
44780 Bochum
michael.kallweit@ruhr-uni-bochum.de