
ChatGPT als Mathematik-Tutor:

Wie unterstützen ChatGPT 3.5, 4.0 und 4.0 mit Wolfram Schüler und Studierende beim Lernprozess?

Bachelorarbeit zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)
im Studiengang Allgemeine Informatik
an der Fakultät für Informatik und Ingenieurwissenschaften
der Technischen Hochschule Köln
Campus Gummersbach

vorgelegt von: Sarah Langrehr
Matrikel-Nr.: 111 388 13
Adresse: Heckenhofstraße 28a
53819 Neunkirchen-Seelscheid
sarah.langrehr@smail.th-koeln.de

eingereicht bei: Prof. Dr. Wolfgang Konen
Technische Hochschule Köln

Zweitgutachter: Prof. Dr. Daniel Gaida
Technische Hochschule Köln

Gummersbach, 12.09.2024

Kurzfassung/*Abstract*

The present study investigates the efficiency differences between various versions of ChatGPT as a mathematics tutor for students at the high school and university levels. The versions examined include ChatGPT 3.5, ChatGPT 4.0, and ChatGPT 4.0 integrated with Wolfram. Through an analysis of 225 datasets consisting of explanations generated by these AI models for mathematical problems at the upper secondary and university levels, the fundamental suitability of these AIs for generating mathematical explanations was assessed. The perspective of the target group was incorporated through a quantitative survey involving 38 university students and 13 tutoring recipients.

The results of the data collection indicate varying degrees of suitability among the different versions of ChatGPT for use as a mathematics tutor. It was found that the generated responses were not necessarily correct across all versions, although the proportion of correctly answered questions increased significantly from version 3.5 to 4.0. Combined with the survey's finding of the respondents' limited ability to identify these incorrect answers, this highlights the need for enhanced training in critically questioning AI-generated responses. Furthermore, the importance of verifying and scrutinizing even seemingly plausible answers is emphasized.

Schlüsselwörter: ChatGPT, adaptive Lerntechnologien, Künstliche Intelligenz, Bildung.

Danksagung

Diese Arbeit wäre ohne die Unterstützung und Begleitung vieler wertvoller Menschen nicht möglich gewesen, denen ich an dieser Stelle meinen tiefen Dank aussprechen möchte. Die Unterstützung dieser Menschen durch ihre Zeit, Geduld und ihre Erfahrung war von unschätzbarem Wert und hat wesentlich zur erfolgreichen Fertigstellung dieser Arbeit beigetragen. Für die zahlreichen hilfreichen Hinweise und Anmerkungen bin ich äußerst dankbar.

Besonderer Dank gilt meinem Betreuer, Prof. Dr. Wolfgang Konen, der mich mit seinem fundierten Wissen, wertvollen Ratschlägen und stets konstruktivem Feedback während der gesamten Arbeit begleitet hat. Die regelmäßigen Gespräche und Rückmeldungen auf meine Fragen, auch während der vorlesungsfreien Zeit, waren für mich eine außerordentlich große Unterstützung und haben dazu beigetragen, diese Arbeit in geregelte Bahnen zu lenken. Auch möchte ich Prof. Dr. Daniel Gaida für die wertvollen Hinweise zu Beginn meiner Arbeit danken, die mir geholfen haben, einen guten Einstieg in das Thema zu finden.

Ebenso möchte ich Frau Flory, der Studienkreisleitung der Filialen Seelscheid und Rösrath, meinen Dank für die Möglichkeit, die Umfrage unter den Nachhilfe Nehmenden durchführen zu dürfen und die Ermöglichung der äußerst flexiblen Arbeitszeiten aussprechen. Sowie allen Nachhilfe Nehmenden und Studierenden, die sich bereit erklärt haben, an der Umfrage teilzunehmen und somit zur Datengrundlage meiner Forschung beigetragen haben. Die zahlreichen, unerwarteten und wertvollen Anregungen und Ratschläge, die ich per E-Mail von einigen diesen Teilnehmenden erhalten habe, waren informativ und haben die Perspektiven meiner Untersuchung erweitert.

Abschließend möchte ich ganz herzlich meiner Familie danken, deren unermüdliche Unterstützung und Liebe die Grundpfeiler meines Erfolgs sind. Ihr habt mir Kraft und Zuversicht geschenkt und wart immer da, wenn ich euch brauchte. Eure Rolle in diesem Kapitel meines Lebens ist unermesslich und ich bin euch zutiefst dankbar. Ebenso danke ich von Herzen meinen Freunden, die mich auf diesem Weg stets ermutigt und begleitet haben. Ich bin unendlich dankbar für jeden einzelnen von euch, der mich auf diesem Weg mit Rat und Tat unterstützt hat.

Sarah Langrehr
Gummersbach, 06.09.2024

Inhaltsverzeichnis

Tabellenverzeichnis	V
Abbildungsverzeichnis	VII
Abkürzungsverzeichnis	VIII
1 Einleitung	1
1.1 Ziel und Fragestellung	2
1.2 Methodik und Vorgehen	4
1.3 Motivation	5
1.4 Aufbau der Arbeit	5
2 Theoretischer Rahmen	7
2.1 Künstliche Intelligenz	7
2.1.1 Begriffserklärung und Geschichte	7
2.1.2 Vergleich künstlicher und menschlicher Intelligenz	12
2.1.3 Grundlagen des maschinellen Lernens	13
2.1.4 Künstliche neuronale Netzwerke	14
2.2 ChatGPT	19
2.2.1 Entstehung	20
2.2.2 Unterschiede ChatGPT 3.5 und 4.0	21
2.2.3 Wolfram Alpha	22
2.2.4 Verwendung und Prompten	25
2.2.5 Probleme und Kritik	26
2.3 ChatGPT in der Lehre	28
2.3.1 Konstruktivismus	29
2.3.2 Forschungsstand und Ausblick	30
3 Methodik	33
3.1 Versuchsaufbau und Durchführung	34
3.2 Bewertungskriterien	37
3.2.1 Korrektheit	38
3.2.2 Vollständigkeit	38
3.2.3 Sprache und Terminologie	39

3.2.4	Strukturierung	39
3.2.5	Reproduzierbarkeit	40
4	Forschungsergebnisse	41
4.1	Ergebnisse der Datenauswertung	41
4.1.1	Korrektheit	42
4.1.2	Vollständigkeit	48
4.1.3	Sprache und Terminologie	50
4.1.4	Strukturierung	53
4.1.5	Zusammenfassung der Ergebnisse	55
4.2	Ergebnisse der Umfrage	56
4.2.1	Korrektheit	58
4.2.2	Vollständigkeit	61
4.2.3	Sprache und Terminologie	63
4.2.4	Verständnis	64
4.2.5	Zusammenfassung der Ergebnisse	67
5	Diskussion	69
5.1	ChatGPT 3.5	69
5.2	ChatGPT 4.0	74
5.3	ChatGPT 4.0 mit Wolfram Alpha	78
5.4	Einschränkungen der Datenerhebung	82
6	Fazit	84
	Literatur	90
	Anhang	96

Tabellenverzeichnis

2.1	Definition künstlicher Intelligenz in vier Kategorien.	11
4.1	Korrekte Antworten aller Fragen von ChatGPT 3.5	46
4.2	Korrekte Antworten aller Fragen von ChatGPT 4.0	47
4.3	Korrekte Antworten aller Fragen von ChatGPT 4.0 mit Wolfram Alpha	48
4.4	Konfusionsmatrix: Bewertung der ChatGPT-3.5-Antworten durch Studierende	59
4.5	Konfusionsmatrix: Bewertung der ChatGPT-3.5-Antworten durch Nachhilfe Nehmende	59
4.6	Konfusionsmatrix: Bewertung der ChatGPT-4.0-Antworten durch Studierende	60
4.7	Konfusionsmatrix: Bewertung der ChatGPT-4.0-Antworten durch Nachhilfe Nehmende	60
4.8	Konfusionsmatrix: Bewertung der Wolfram-Antworten durch Studierende	61
4.9	Konfusionsmatrix: Bewertung der WolframGPT-Antworten durch Nachhilfe Nehmende	61
1	Matrix: Bewertung der ChatGPT-3.5-Antworten zur Vollständigkeit durch Studierende	113
2	Matrix: Bewertung der ChatGPT-3.5-Antworten zur Vollständigkeit durch Nachhilfe Nehmende	114
3	Matrix: Bewertung der ChatGPT-4.0-Antworten zur Vollständigkeit durch Studierende	114
4	Matrix: Bewertung der ChatGPT-4.0-Antworten zur Vollständigkeit durch Nachhilfe Nahmende	115
5	Matrix: Bewertung der WolframGPT-Antworten zur Vollständigkeit durch Studierende	115
6	Matrix: Bewertung der WolframGPT-Antworten zur Vollständigkeit durch Nachhilfe Nehmende	116
7	Matrix: Bewertung der ChatGPT-3.5-Antworten zur Sprache und Terminologie durch Studierende	117
8	Matrix: Bewertung der ChatGPT-3.5-Antworten zur Sprache und Terminologie durch Nachhilfe Nehmende	117

9	Matrix: Bewertung der ChatGPT-4.0-Antworten zur Sprache und Terminologie durch Studierende	118
10	Matrix: Bewertung der ChatGPT-4.0-Antworten zur Sprache und Terminologie durch Nachhilfe Nehmenden	118
11	Matrix: Bewertung der WolframGPT-Antworten zur Sprache und Terminologie durch Studierende	119
12	Matrix: Bewertung der WolframGPT-Antworten zur Sprache und Terminologie durch Nachhilfe Nehmenden	119

Abbildungsverzeichnis

2.1	Bestandteile einer Nervenzelle, Neuron. Quelle: [23, S. 30]	14
2.2	Einfaches künstliches neuronales Netzwerk mit zwei verborgenen Schichten	15
2.3	Mathematisches Modell für ein künstliches Neuron. Quelle: [31]	16
4.1	Kreisdiagramme Korrektheit	42
4.2	Originales und manipuliertes Logikrätsel im Vergleich	44
4.3	Kreisdiagramme Vollständigkeit	49
4.4	Wolfram-Language-Antwort	51
4.5	Kreisdiagramme Sprache und Terminologie	52
4.6	Kreisdiagramme Strukturierung	53
4.7	Aufgabentypenverteilung der Umfrage	57
4.8	Umfrage: Bewertung der Korrektheit	58
4.9	Umfrage: Bewertung der Vollständigkeit	62
4.10	Umfrage: Bewertung der Sprache und Terminologie	63
4.11	Umfrage: Verständnis des Rechenwegs	65
4.12	Umfrage: Verständnis durch das Ergebnis	65
4.13	Umfrage: Verständnis durch die Erklärung	66
4.14	Umfrage: Wissenstransfer auf neue Probleme	67
5.1	ChatGPT 4.0 Aufgabenausschnitt mit Analysefehler	76
1	Umfrageaufgaben: Korrektheit	112
2	Umfrageaufgaben: Vollständigkeit	113
3	Umfrageaufgaben: Sprache und Terminologie	116
4	Umfrageaufgaben: Struktur	120
5	QR-Code für GitHub	121

Abkürzungsverzeichnis

APIs Application Programming Interfaces. 21

DL Deep Learning. 17

DSGVO Datenschutz-Grundverordnung. 27

EU Europäische Union. 27

GPT Generative Pretrained Transformer. 20

KI künstliche Intelligenz. 1

KNN künstlich neuronale Netze. 14

ML maschinelles Lernen. 13

NLP Natural Language Processing. 20

XAI Explainable Artificial Intelligence. 17

1 Einleitung

“Education is the most powerful weapon which you can use to change the world.”
(Nelson Mandela)

Zuständig für die grundlegende Vermittlung von Wissen und die Förderung der Schüler und Schülerinnen ist das Schulwesen [1]. Für die Studierenden übernimmt das Hochschulwesen diese Rolle, indem es spezialisierte und weiterführende Bildungsangebote bereitstellt. Bei einem zusätzlichen Unterstützungsbedarf ist eine gesonderte Förderung, in Form von externen Tutorinnen und Tutoren, ein zentraler Schlüssel, um den Lernerfolg zu sichern und eine dauerhafte Verbesserung zu erzielen [2]. Doch was, wenn der Schlüssel zur Bildungsförderung in die Hände eines auf künstliche Intelligenz (KI) basierenden Tutors gelegt wird? Angesichts der Tatsache, dass im internationalen Vergleich 30 % der deutschen Schülerinnen und Schüler die Mindestanforderungen im Bereich Mathematik nicht erreicht haben [3] und den dadurch resultierenden Auswirkungen auf die, auf den durch das Schulwesen vermittelte Grundlagen aufbauende, Hochschullehre, stellt sich die Frage, ob Technologien wie ChatGPT für diese Herausforderungen eine Lösung darstellen könnten.

Die Rolle und Bedeutung von KIs nimmt in der heutigen Zeit in allen Bereichen stetig zu. ChatGPT ist eine künstliche Intelligenz, welche als KI-gestützter Gesprächspartner fungieren kann. Da es natürliche Sprache verarbeiten und Antworten auf gestellte Fragen generieren kann, handelt es sich um einen Chatbot [4]. Das Wissen, welches für eine Antwort notwendig ist, bezieht dieser Chatbot aus einer an das menschliche Gehirn angelehnten digitalen Struktur, bezeichnet als künstliches neuronales Netzwerk [5, S. 170–173]. Durch das Training mit einer großen Menge von Textdaten aus dem Internet ist es imstande, Aussagen über verschiedene Fachbereiche zu treffen [4]. Die Fähigkeit, zu komplexen Sachverhalten eine individuell angepasste und verständliche Antwort zu generieren, macht ChatGPT aktuell zu einer fortschrittlichen KI [6].

Chatbots wie ChatGPT scheinen seit kurzer Zeit schnelle Fortschritte zu machen. Allerdings sind sie keine neue Errungenschaft. Bereits 1966 entwickelte Joseph Weizenbaum ein Computerprogramm, welches Gespräche simulierte [7]. Der Zweck dieses rudimentären Chatbots, bekannt als ELIZA, war es, die Rolle einer Psychotherapeutin oder eines Psychotherapeuten zu übernehmen [7]. Es analysierte die Eingaben nach Schlüsselbegriffen und antwortete gemäß dem erkannten Muster mit gescripteten

Gegenfragen [7]. Obwohl ELIZAs Fähigkeiten eng begrenzt waren, wurden ihr von einer großen Anzahl von Nutzenden menschliche Eigenschaften zugeschrieben. Dieser Effekt wird bis heute als ELIZA-Effekt bezeichnet und gewinnt in der gegenwärtigen Zeit, in der KI im Alltag weit verbreitet ist, zunehmend an Relevanz.

Vergleicht man damit den Fortschritt, welcher seitdem gemacht wurde, wird deutlich, wie weit die Technologie seit den frühen Tagen der Forschung gekommen ist. Während ELIZA nur einfache Muster erkannte, bieten moderne Chatbots weitreichende Möglichkeiten.

Im Kundenservice und -support kommen durch KI unterstützte Chatbots bereits jeden Tag zum Einsatz, um den Kundinnen und Kunden Produktempfehlungen zu geben oder Fragen zu beantworten [8]. Im Bereich der Finanzdienstleistungen und im Bankwesen helfen sie bei der Kontenverwaltung [8], [9]. Außerdem analysieren sie im Bereich der Gesundheit Symptome und geben Anregungen zur Behandlung von Patientinnen und Patienten [10]. Chatbots werden in Plattformen wie Facebook eingebunden, um die Nutzenden beim Versenden von Nachrichten und der Einstellung von Erinnerungen zu unterstützen. Sprachassistenten wie Alexa und Siri sind Chatbots, die bei der Kontrolle von Smart-Home-Geräten, der Wiedergabe von Musik und einer Vielzahl anderer alltäglicher Aufgaben helfen [11].

So ist es nicht verwunderlich, dass sich mit der zunehmenden Digitalisierung der Schulen und der verstärkten Verwendung digitaler Hilfsmittel [12] in der Lehre der Diskurs über den Einsatz von KI-basierten Hilfsmitteln entfaltet. Insbesondere Chatbots ermöglichen eine direkte Kommunikation, um Fragen schnell und unkompliziert zu beantworten [13]. KIs sind zudem jederzeit verfügbar, was einen erheblichen Vorteil gegenüber Lehrpersonal darstellt, daher eröffnen sich besonders im Bildungsbereich dadurch neue Möglichkeiten.

Die Popularität der künstlichen Intelligenz ChatGPT von OpenAI nimmt seit der Veröffentlichung im November 2022 zu [13], [4]. Bereits in den ersten zwei Monaten stieg die Abonnentenzahl auf 100 Millionen [14]. Das kostenlose Angebot der Versionen 3.5 und 4o Turbo [4], [13] kann einer der Gründe für ihre hohe Bekanntheit sein. Laut „Exploding Topics“ liegt ChatGPT auf Platz 14 der meistbesuchten Websites für Juni 2024 im globalen Ranking [15]. ChatGPT wurde aus Gründen der aktuellen Relevanz dieser künstlichen Intelligenz als Gegenstand dieser Studie ausgewählt.

1.1 Ziel und Fragestellung

ChatGPT, als modernes Sprachmodell [4], bietet als Mathematik-Tutor die Chance, Schülerinnen und Schülern bei Problemstellungen zu helfen und sie aktiv bei

der Lösungsfindung zu unterstützen [2]. Ungeachtet des erkennbaren Potenzials dieser Modelle zur Ergänzung traditioneller Lernmethoden führen die fortschreitenden Verbesserungen dieser Technologien zu einem anhaltenden Forschungsbedarf.

Dabei stellt sich die Frage, in welchem Maße die einzelnen Versionen dieser Technologie in der Lage sind, eine kompetente Erklärung auf mathematische Fragen zu geben und in der Rolle als Mathematik-Tutor den Lösungsweg angemessen zu erläutern. Der Begriff des Tutors wird im Rahmen dieser Forschungsarbeit in einer vereinfachten Form betrachtet. Der Fokus liegt auf der Untersuchung der Kompetenz zur korrekten Vermittlung mathematischer Konzepte. Die darüber hinausgehenden Fähigkeiten eines menschlichen Tutors, insbesondere die proaktive Identifikation von Schwächen der Lernenden, werden hierbei nicht betrachtet. Aus diesem Sachverhalt ergibt sich die zentrale Forschungsfrage dieser Arbeit:

Inwiefern unterscheiden sich die ChatGPT-Versionen 3.5, 4 und 4 mit der Integration von Wolfram Alpha in ihrer Effektivität als digitale Mathematik-Tutoren für Studierende, Schülerinnen und Schüler?

Um diese Frage beantworten zu können, ist es von Bedeutung, eine Bewertung der verschiedenen Versionen anhand vorab bestimmter Kriterien vorzunehmen und die daraus gewonnenen Erkenntnisse zu vergleichen. Auf diese Kriterien wird im Methodikteil in Kapitel 3 dieser Arbeit konkret eingegangen. Zusätzlich wird eine Einschätzung der verschiedenen Versionen unter Berücksichtigung dieser Kriterien von Studierenden, Schülerinnen und Schülern durchgeführt. Dies gewährleistet, dass die Ansichten der Zielgruppe nicht ignoriert werden. Die detailliertere Betrachtung dieser spezifischen Aspekte der Hauptforschungsfrage führt zu der Unterteilung in folgenden Teilfragen:

Inwiefern unterscheiden sich die Versionen ChatGPT 3.5, 4.0 und 4.0 und Wolfram in ihren grundlegenden Eigenschaften?

Wie unterscheiden sich die untersuchten ChatGPT-Versionen in Bezug auf die definierten Bewertungskriterien?

Wie wird die Verständlichkeit der durch die untersuchten Versionen von ChatGPT generierten Antworten durch die Studierenden, Schüler und Schülerinnen beurteilt?

Daher zielt diese Arbeit darauf ab, die Kompetenzen der ChatGPT-Versionen 3.5, 4.0 und 4.0 mit der Integration von Wolfram Alpha in Bezug auf ihre Eignung als digitale Lernhelfer für das Fach Mathematik zu untersuchen. Hierbei liegt der Fokus vor allem darauf, dass die Erklärung korrekt, nachvollziehbar und leicht verständlich ist. Die Zielgruppe sind hier zum einen leistungsschwächere Schülerinnen und Schüler ab der achten Klasse, welche bereits ein Förderangebot im Bereich Mathematik in Anspruch nehmen. Um die Rückmeldungen von Schülerinnen und Schülern in diesem

Bereich zu bekommen, wird eine Kooperation mit dem Nachhilfeinstitut „Studienkreis“ eingegangen. Zudem werden auch die Einschätzungen von Studierenden der Fakultät 10 der Technischen Hochschule Köln erfasst.

Weitere Anwendungsbereiche von ChatGPT und die technischen Einzelheiten der verwendeten Algorithmen werden nicht berücksichtigt. Die Studie betrifft die derzeit verfügbaren Versionen aus dem Jahr 2024 und erhebt keinen Anspruch auf die Übertragbarkeit der Ergebnisse auf zukünftige Updates oder andere geografische Regionen.

1.2 Methodik und Vorgehen

Um herauszufinden, in welcher Form sich die drei untersuchten Versionen von ChatGPT für die Erklärung mathematischer Aufgaben eignen, werden zunächst Aufgaben aus fünf unterschiedlichen Themenbereichen der Mathematik bestimmt. Als Themengebiete werden *Gleichungen*, *Integralrechnung*, *Stochastik*, *analytische Geometrie* und *Logikrätsel* gewählt. Zu jedem dieser Gebiete werden fünf Aufgaben mit unterschiedlichem Schwierigkeitsgrad erarbeitet. Insgesamt werden 25 Fragen gestellt und von ChatGPT 3.5, 4.0 und 4.0 mit Integration von Wolfram Alpha beantwortet. Um die Reproduzierbarkeit zu kontrollieren, wird jede Frage dreimal jeder dieser Versionen gestellt. Anschließend werden die gesammelten 225 Antworten hinsichtlich ihrer *Korrektheit*, *Reproduzierbarkeit*, *Sprache* und *Terminologie*, *Verständlichkeit* und *Strukturierung* bewertet.

Aus diesem Antwortgenpool wird für jedes Themengebiet zufällig eine Antwort pro KI für eine der fünf Aufgaben ausgewählt. Mit diesen Aufgaben wird eine Umfrage für Schüler und Schülerinnen ab der achten Klasse erstellt. Zudem werden Studierende und Mitarbeitende der *Fakultät für Informatik und Ingenieurwissenschaften* der *Technischen Hochschule Köln* um ihre Einschätzung gebeten.

Für die Datenerhebung wird die Durchführung einer Umfrage gewählt, um eine möglichst große Anzahl von Teilnehmenden zu erreichen. Zudem ist es bei den Schülerinnen und Schülern des Nachhilfeinstituts „Studienkreis“, die im Fach Mathematik gefördert werden, wichtig, den zeitlichen Aufwand der Befragung gering zu halten. Aus diesen Gründen erscheint die Durchführung einer Umfrage als sinnvollste Methode.

1.3 Motivation

Die Motivation, mich mit der Qualität der Erklärungen mathematischer Aufgaben durch verschiedene Versionen von ChatGPT zu beschäftigen, basiert auf zwei wesentlichen Faktoren. Zum einen habe ich im Verlauf meines Studiums der allgemeinen Informatik ein starkes Interesse an Mathematik und künstlicher Intelligenz entwickelt. Zum anderen wird mein Interesse insbesondere durch meine Tätigkeit als Mathematik-Nachhilfelehrerin beim „Studienkreis“ verstärkt. Diese Beschäftigung hat mein starkes Interesse an der Schnittstelle zwischen Bildung und künstlicher Intelligenz geweckt. Durch diese Tätigkeit habe ich wertvolle Einblicke in die Schwierigkeiten und Herausforderungen gewonnen, mit denen Schülerinnen und Schüler konfrontiert sind. Insbesondere habe ich festgestellt, dass viele Lernende von zusätzlichen, leicht zugänglichen Erklärungen und Hilfestellungen profitieren könnten.

Mathematik ist das Nachhilfefach mit der größten Nachfrage. Bundesweit werden 59 % der Nachhilfe beim „Studienkreis“ in Mathematik gegeben [16]. Schülerinnen und Schüler mit Förderbedarf benötigen auch außerhalb des Schul- und Nachhilfeunterrichts Unterstützung. Deshalb ist es für mich wichtig zu erfahren, ob eine KI wie ChatGPT in der Lage ist, für Mathematikaufgaben eine leicht verständliche und dennoch korrekte Lösung zu generieren. Der Vergleich der drei Versionen soll Aufschluss darüber geben, wo die Stärken und Schwächen der jeweiligen Versionen liegen und wie ChatGPT dazu beitragen kann, die Nachhilfe zu verbessern und Lernbarrieren zu überwinden.

1.4 Aufbau der Arbeit

In Kapitel 2 werden die grundlegenden Konzepte und Begriffe erläutert und eingeführt. Diese theoretische Grundlage dient der Bildung eines für das spätere Verständnis essenziellen Fundamentes und ermöglicht dadurch einen Überblick über die grundlegenden Eigenschaften der untersuchten Versionen von ChatGPT. Darüber hinausgehend werden die Ergebnisse anderer Forschungsarbeiten vorgestellt, auf deren Grundlage diese Arbeit aufbaut.

In Kapitel 3 wird das Vorgehen der beiden Datenerhebungsmethoden dargestellt, welche für die Beantwortung der Forschungsfrage angewendet werden. Dabei wird zuerst auf die Erhebung der in dieser Arbeit überprüften Versionen von ChatGPT eingegangen und die Auswertung dieser Daten anhand der definierten Kriterien beschrieben. Anschließend wird auf die durchgeführte Umfrage eingegangen, welche für eine Einbeziehung der Einschätzungen der befragten Nachhilfe Nehmenden und Studierenden in diese Arbeit sorgt.

In Kapitel 4 werden die Ergebnisse, welche durch die beiden durchgeführten Datenerhebungsmethoden ermittelt werden, ausführlich vorgestellt. Um die Verständlichkeit zu erleichtern, erfolgt die Darstellung dieser Daten in einer grafisch aufgearbeiteten Form. Diese Präsentation der Daten geschieht sortiert nach den überprüften Kriterien, dies soll eine übersichtliche Form der Darstellung und damit das Nachvollziehen der wesentlichen Erkenntnisse fördern.

In Kapitel 5 werden die im Forschungsergebnisteil der Arbeit vorgestellten Ergebnisse interpretiert und in den Kontext dieser Arbeit, aber auch in den der bestehenden Literatur und der theoretischen Grundlagen eingeordnet. Dabei wird die Leistung der untersuchten Versionen sowohl in Bezug auf die definierten Bewertungskriterien als auch im Hinblick auf die Einschätzungen der Befragten bewertet. Unvorhergesehene Befunde, welche während der Datenerhebung auftraten, werden beschrieben und hinsichtlich ihrer Bedeutung für den Kontext dieser Arbeit interpretiert. Zusätzlich werden die Limitationen der durchgeführten Forschung aufgezeigt und es wird ein Ausblick auf die zukünftige Forschung gegeben.

In Kapitel 6 erfolgt die Zusammenfassung der wichtigsten Erkenntnisse dieser Arbeit. Die in der Einleitung vorgestellte Forschungsfrage und die daraus untergliederten Teilfragen werden beantwortet und ein Ausblick auf zukünftige Entwicklungen gegeben.

Der Anhang dieser Arbeit enthält den Link zu dem GitHub-Repository, auf welchem die Datengrundlage, auf der die Forschungsergebnisse basieren, transparent einsehbar sind. Diese Materialien sollen die Nachvollziehbarkeit der Ergebnisse erhöhen und anderen Forscherinnen und Forschern die Möglichkeit geben, die Analysen zu reproduzieren oder weiterzuentwickeln.

2 Theoretischer Rahmen

Um die Grundlage dieser Arbeit zu schaffen, fokussiert sich dieser Abschnitt der Forschungsarbeit auf die Vermittlung der grundlegenden Konzepte. Begrifflichkeiten, welche in dieser Arbeit Verwendung finden, werden definiert und relevante Konzepte erläutert. Darüber hinaus wird die erste Teilfrage durch die Darlegung und den Vergleich der Eigenschaften der verschiedenen ChatGPT-Versionen beantwortet. Aufbauend auf diesen Grundlagen wird der aktuelle Forschungsstand dargelegt und in den Kontext der vorliegenden Arbeit eingeordnet.

2.1 Künstliche Intelligenz

Das Gebiet der Künstlichen Intelligenz spielt in dieser Arbeit eine entscheidende Rolle, da es die technologische Grundlage für die untersuchten ChatGPT-Versionen bildet. Ein fundamentales Verständnis der zugrunde liegenden Konzepte ist daher nicht nur für das Verständnis der Ergebnisse von Bedeutung, sondern auch für die Beurteilung der Fähigkeiten und Grenzen der in dieser Arbeit untersuchten ChatGPT-Versionen. Vor diesem Hintergrund wird zunächst die grundlegende Bedeutung und Entwicklung der Künstlichen Intelligenz erläutert. Anschließend werden die relevanten Konzepte behandelt, wobei der Schwerpunkt auf den für das Verständnis von ChatGPT wesentlichen Aspekten liegt.

2.1.1 Begriffserklärung und Geschichte

Das Konzept, eine Maschine denken und intelligent bzw. menschlich handeln zu lassen, ist schon seit der Antike Bestandteil von Überlegungen der Menschheit [17]. Bereits der Philosoph Aristoteles beschäftigte sich mit der Idee von selbstständigen Maschinen, die eigenständig ihre Aufgaben erledigen können. „Wenn jedes Werkzeug [...] vorausahnend [...] das ihm zukommende Werk verrichten könnte [...], so bedürfte es weder für den Werkmeister der Gehilfen noch für die Herren der Sklaven“ [17]. Mythische Figuren wie Talos von Kreta, eine erschaffene Maschine, welche der Erzählung nach die Insel Kreta beschützte, existieren schon seit langer Zeit in vielen Kulturen.

Obwohl die Menschheit sich also schon seit tausenden von Jahren mit dem Konzept von künstlicher Intelligenz beschäftigt, gibt es bis heute keine allgemeingültige Definition des Begriffes [13]. Manche bezeichnen KI als die Simulation menschlicher Intelligenz in Maschinen. Andere haben einen weiteren Blickwinkel und betrachten KI als jede Form von intelligentem Verhalten, das von Maschinen demonstriert wird. Je nach Kontext und Bezug unterscheidet sich die Interpretation davon, was künstliche Intelligenz ist. Der Begriff wurde allerdings fundamental durch zwei Ereignisse geprägt, die im Folgenden vorgestellt werden.

Der Turing-Test

Bereits 1950 beschäftigte sich der britische Mathematiker und Informatiker Alan Turing mit der Fragestellung, ob Maschinen denken können [18], [19]. Allerdings stand hierbei nicht das Interesse im Vordergrund, ob Maschinen im philosophischen Sinn denken können, sondern vielmehr, ob eine Maschine in der Lage ist, sich so intelligent zu verhalten, dass sie nicht mehr von einem Menschen unterschieden werden kann [18], [20]. Um diese Frage zu beantworten, entwickelte er einen einfachen Test, mit dem eine Prüfung dieser Fragestellung durchgeführt werden kann [18].

Bei dem Turing-Test, von Alan Turing als *imitation game* bezeichnet, kommuniziert der sogenannte Fragesteller bzw. die Fragestellerin mit zwei anderen Teilnehmenden [18]. Bei den beiden Befragten handelt es sich um einen Menschen und eine Maschine [18], [20]. Der Fragesteller bzw. die Fragestellerin hat die Aufgabe zu ermitteln, welche der beiden Befragten die Maschine ist [18]. Er/sie kommuniziert mit den Teilnehmenden ohne nonverbale Elemente über Text oder eine zusätzliche Person, damit Faktoren wie Stimme und Mimik keinen Einfluss auf das Ergebnis nehmen können [18], [19]. Nur anhand der übermittelten Antworten soll entschieden werden, welcher Teilnehmende kein Mensch ist [18]. Der Turing-Test gilt als bestanden, wenn der Fragesteller bzw. die Fragestellerin nicht in der Lage ist zu unterscheiden, welche der beiden Gesprächspartner die Maschine ist [18], [20].

Eine Maschine, welche den Test bestehen konnte, kann laut Turing als intelligent bezeichnet werden [18]. Wenn auch der Begriff *künstliche Intelligenz* zu diesem Zeitpunkt noch nicht existierte, wurde im Kontext des Turing-Tests erstmalig die Bezeichnung *Intelligenz* im Zusammenhang mit einer Maschine verwendet.

Der Turing-Test war ein wichtiger Schritt bei der Entwicklung der künstlichen Intelligenz, allerdings ist er auch Ziel von Kritik [20], [19]. Da der Test eine künstliche Intelligenz nur danach bewertet, wie gut sie vorgeben kann, ein Mensch zu sein, wird die Frage gestellt, ob der Turing-Test nicht eher eine Maschine nach der Fähigkeit bewertet, ihr Gegenüber zu täuschen [19]. In diesem Zusammenhang ist zu hinterfragen,

ob die Fähigkeit zur Täuschung Rückschlüsse auf die Intelligenz einer Maschine zulässt [19]. So schnitt beispielsweise in einem öffentlichen Online-Turing-Test ChatGPT 3.5 schlechter ab als ELIZA. Dies kann darauf zurückgeführt werden, dass ELIZA dafür entwickelt wurde, eine Therapeutin vorzuspielen, während ChatGPT mit Absicht kein menschliches Verhalten vortäuschen soll [21]. Die Befragten gaben zudem an, dass sie ihre Entscheidung aufgrund des Sprachstils und sozialer wie emotionaler Merkmale getroffen hätten [21].

Ein anderer Kritikpunkt ist die grundsätzliche Frage, ob sich menschliche Intelligenz mit der von Maschinen vergleichen lässt [20]. Im direkten Vergleich von Mensch und Maschine wird jegliche vom menschlichen Standard abweichende Form von Intelligenz außer Acht gelassen, die eine Maschine aufweisen könnte [20]. Somit könnte man sagen, dass der Turing-Test eher Menschlichkeit anstatt Intelligenz misst [20]. Trotz der Kritik hat Alan Turing die Diskussion über künstliche Intelligenz angefacht. Er hat mit seiner Arbeit die theoretische Grundlage für die Entwicklung von KI geliefert und mit dem Turing-Test ein überprüfbares Ziel für die Nachahmung menschlicher Intelligenz durch Maschinen definiert.

Die Dartmouth Conference

Die Dartmouth Conference, welche im Sommer 1956 stattfand, war das erste organisierte Treffen, das sich ausschließlich mit dem Forschungsgebiet der künstlichen Intelligenz befasste [13]. Sie prägte die Entstehungsgeschichte maßgeblich und wird offiziell als Beginn des KI-Forschungsfeldes angesehen [13]. Die Konferenz wurde von John McCarthy, Marvin Minsky, Nathaniel Rochester und Claude Shannon geplant und abgehalten, um einen wissenschaftlichen Austausch zu ermöglichen [22]. Die Konferenz zielte darauf ab, die Möglichkeiten der Entwicklung denkender Maschinen zu diskutieren und das heutige Forschungsgebiet der künstlichen Intelligenz zu definieren [22]. Drei grundlegende Annahmen wurden dabei diskutiert, die das Fundament für die Forschung in diesem Bereich legten [22]:

1. Der Denkvorgang eines Menschen kann auch außerhalb des menschlichen Gehirns stattfinden [22].
2. Das menschliche Denken lässt sich formal beschreiben [22].
3. Ein Digitalrechner ist ein geeignetes Instrument, um diesen Denkvorgang zu modellieren [22].

Diese Annahmen spiegelten die impliziten Überzeugungen der beteiligten Wissenschaftler/-innen wider und prägten die anfängliche Ausrichtung der Forschung in künstlicher Intelligenz [23].

Die an der Dartmouth Conference teilnehmenden Wissenschaftler/-innen glaubten, dass sie innerhalb einer kurzen Zeitspanne von zehn bis zwanzig Jahren signifikante Fortschritte im Bereich der KI erreichen könnten [23, S. 42]. Diese ambitionierten Erwartungen waren wahrscheinlich auf eine Unterschätzung der Komplexität menschlicher Kognition und die beschränkten technologischen Mittel der damaligen Zeit zurückzuführen [23]. Trotz dieser anfänglichen Rückschläge wird die Konferenz allgemein als die Geburtsstunde der KI-Forschung angesehen, die den Weg für zukünftige Entwicklungen in diesem Bereich ebnete [23].

KI-Definitionen

Trotz des gelegten Fundaments konnte bis heute keine Einigung auf eine allgemeingültige Definition für künstliche Intelligenz erfolgen. Dennoch wurden in den vergangenen Jahren viele Definitionen für künstliche Intelligenz aufgestellt [23].

So definiert John McCarthy KI als die Wissenschaft und Ingenieurskunst zur Herstellung intelligenter Computerprogramme, die zwar mit der Aufgabe verwandt ist, durch Computerprogramme die menschliche Intelligenz zu verstehen, sich aber nicht auf biologisch beobachtbare Methoden beschränken muss [24, S. 1]. Diese Definition hebt das Ziel der KI hervor, Maschinen zu schaffen, die Intelligenz erfordernde Aufgaben ausführen können [24]. Bei der Erreichung dieses Ziels soll jedoch keine Beschränkung auf biologische Prozesse erfolgen, sondern mit flexiblen Ansätzen vorgegangen werden können [24].

Problematisch bei McCarthys Definition ist, dass es verschiedene Interpretationen und Ansichten dazu gibt, was Intelligenz konkret bedeutet [24, S. 1 f.]. Die Beschreibung von KI als Mittel zum Verständnis der menschlichen Intelligenz birgt die Herausforderung, dass die Komplexität und Vielschichtigkeit, welche die menschliche Intelligenz auszeichnen, unterschätzt werden [24]. So ist zum Beispiel das menschliche Gehirn bis heute noch nicht vollständig erforscht worden [24, S. 244 f.]. Dies führt zwangsläufig zu Unterschieden zwischen biologischen und technologischen Prozessen [24].

Russell und Norvig beschreiben KI als den Entwurf und die Untersuchung von Agenten, die ihre Umgebung wahrnehmen und Maßnahmen ergreifen, um ihre Erfolgswahrscheinlichkeit zu maximieren [23]. In diesem Zusammenhang wird ein autonom agierendes System als Agent bezeichnet [23]. Dieser Agent nimmt seine Umgebung wahr, wertet diese aus und ergreift passende Maßnahmen, um sein Ziel zu erreichen [23]. Im Vordergrund dieser Definition steht nicht nur, einen intelligenten Agenten zu entwickeln, sondern auch, ihn eigenständig Entscheidungen treffen zu lassen [23].

Russel und Norvig ordnen Definitionen für KI mittels einer zweidimensionalen Matrix in vier Kategorien ein [23]. Dabei unterscheiden sie die Spalten in Menschliches und

Rationales und die Zeilen in Denken und Agieren. In diesem Kontext ist menschliches Handeln nicht mit irrationalen Handeln gleichzusetzen, sondern mit einer unvollkommenen Vorgehensweise [23]. Im Gegensatz dazu gilt Rationalität als ideales Konzept der Intelligenz [23].

Systeme, die wie Menschen denken	Systeme, die rational denken
<p>„Die aufregende und neuartige Anstrengung, Computern das Denken beizubringen, ... KI will die Sache selbst: Maschinen mit Verstand, im vollen und wörtlichen Sinne.“ (Haugeland, 1985)</p> <p>„Die Automatisierung von Aktivitäten, die wir dem menschlichen Denken zuordnen, Aktivitäten wie beispielsweise Entscheidungsfindung, Problemlösung, Lernen ...“ (Bellman, 1978)</p>	<p>„Die Studie mentaler Fähigkeiten durch die Nutzung programmierter Modelle.“ (Charniak und McDermott, 1985)</p> <p>„Die Studie der Programmtechniken, die es ermöglichen, wahrzunehmen, logisch zu schließen und zu agieren.“ (Winston, 1992)</p>
Systeme, die wie Menschen agieren	Systeme, die rational agieren
<p>„Die Kunst, Maschinen zu schaffen, die Funktionen erfüllen, die, wenn sie von Menschen ausgeführt werden, der Intelligenz bedürfen.“ (Kurzweil, 1990)</p> <p>„Die Studie, wie man Computer dazu bringt, Dinge zu tun, bei denen ihnen normalerweise der Mensch noch überlegen ist.“ (Rich und Knight, 1991)</p>	<p>„Computermethoden ist die Studie des Entwurfs intelligenter Agenten.“ (Poole et al., 1998)</p> <p>„KI ... beschäftigt sich mit intelligentem Verhalten in künstlichen Maschinen.“ (Nilsson, 1998)</p>

Tabelle 2.1: Definition künstlicher Intelligenz in vier Kategorien.

Durch die Zuordnung von KI in diese vier Kategorien wird die Vielfalt der Ansätze und Definitionen von KI übersichtlicher [23]. Es wird offensichtlich, dass KI aus unterschiedlichen Perspektiven betrachtet werden kann: entweder durch die Konzentration auf menschliches Verhalten, menschliches Denken, rationales Denken oder rationales Verhalten [23]. Diese Kategorisierung hilft, die unterschiedlichen Ziele und Methoden in der KI-Forschung zu verstehen und einzuordnen.

Starke und schwache KI

Eine weitere und weit verbreitete Möglichkeit ist die Unterteilung von KI in starke und schwache künstliche Intelligenz [13]. Die beiden Varianten entscheiden sich grundlegend in ihren Fähigkeiten und den ihnen zugrunde liegenden Zielen [13].

Eine schwache KI ist auf die Ausführung bestimmter Aufgaben spezialisiert und führt vorprogrammierte Funktionen aus [13]. Sie wird zur Lösung von gleichen und sich wiederholenden Problemen verwendet [13]. Sie findet bei der Spracherkennung oder beim Durchsuchen großer Datenmengen Anwendung und ist meist auf die Erkennung von Mustern trainiert [13]. Siri und Alexa sind Beispiele für eine solche schwache KI

[13]. Diese Art von KI hat kein tieferes Verständnis der Aufgaben, die sie ausführt, und funktioniert nur in dem Bereich, für den sie trainiert wurde [13]. Eine selbstständige Übertragung in andere Bereiche ist daher nicht möglich [13].

Eine starke KI hingegen hat das Ziel, die menschliche Intelligenz in allen Aspekten zu erreichen oder zu übertreffen [13]. Das würde es ihr ermöglichen, selbstständig Aufgaben zu erlernen und erlernte Fähigkeiten in verschiedenen Aufgabenkontexten anzuwenden, da sie ein Verständnis erlernter Fähigkeiten hat [13]. Sie kann flexibel auf Probleme reagieren und durch deren Untersuchung selbstständig Lösungen finden, was auch einen kreativen Lösungsansatz beinhalten kann [13]. Die Realisierung einer solchen starken KI stellt bisher eine erhebliche Herausforderung dar, deren Umsetzung bis heute nicht möglich war [13].

2.1.2 Vergleich künstlicher und menschlicher Intelligenz

Die Definition von künstlicher Intelligenz ist umstritten und gestaltet sich schwierig, da der eng verbundene Begriff der Intelligenz selbst schwer fassbar ist [23], [13]. Bei Menschen umfasst Intelligenz viele kognitive Fähigkeiten wie Lernen, Verstehen, kreatives Denken und die Anpassung an neue Situationen [25], [23]. Diese Vielfältigkeit sorgt dafür, dass nur schwer ein einheitlicher Rahmen für die Einordnung von Intelligenz gefunden werden kann [23], [13].

Menschliche Intelligenz ist dadurch beschränkt, dass nur eine begrenzte Menge an Informationen in einer bestimmten Zeitspanne in das Kurzzeitgedächtnis aufgenommen und verarbeitet werden kann [25], [6]. Das Gedächtnis eines Menschen ist eine neuronale Struktur, welche durch Alter, Bildung und Erfahrung beeinflusst wird [25]. Kognitive Verzerrungen und Emotionen können das Urteilsvermögen beeinträchtigen und dafür sorgen, dass Entscheidungen, die nicht auf reiner Logik basieren, getroffen werden [26].

Das führt zu dem zentralen Unterschied zwischen künstlicher und menschlicher Intelligenz. Während menschliche Intelligenz, die durch biologische Prozesse arbeitet, von persönlichen Erfahrungen, Glauben und Gefühlen geprägt ist und somit nicht immer auf Logik und Rationalität beruht, stützt sich künstliche Intelligenz auf Algorithmen und Logik [23], [25]. Künstlicher Intelligenz fehlen hingegen die subjektiven Erfahrungswerte, die sich das menschliche Gehirn angeeignet hat [26].

In vielen Bereichen erweist sich eine Zusammenarbeit von menschlicher und künstlicher Intelligenz als besonders zielführend [26], da KI in der Lage ist, in für Menschen nicht erreichbarer Geschwindigkeit große Datenmengen auszuwerten und Berechnungen durchzuführen [26]. Menschen verfügen hingegen über ein intuitives Verständnis und

subjektive Erfahrungswerte, was in unsicheren Situationen nützlich ist [25]. Hinsichtlich der Schwächen beider Arten von Intelligenz können diese sich ergänzen [26].

2.1.3 Grundlagen des maschinellen Lernens

Ein großes Teilgebiet der künstlichen Intelligenz ist maschinelles Lernen (ML) [23]. Das maschinelle Lernen beschäftigt sich mit der Entwicklung von Systemen, die nicht für eine spezifische Aufgabe programmiert werden müssen [27], [13]. Stattdessen fokussiert sich ML darauf, ein System so zu trainieren, dass es in der Lage ist, Muster wie Gemeinsamkeiten etwa in Bildern zu erkennen und diese korrekt zu interpretieren [27]. Dazu wird die KI mit großen problemspezifischen Datensätzen trainiert und vermag es später, auf der Basis der bereits analysierten Daten, der sogenannten Trainingsdaten, zuverlässige Vorhersagen und Entscheidungen für ähnliche Daten zu treffen [27], [13].

Maschinelles Lernen hat sich besonders bei Problemen als zielführend erwiesen, die mit Entitäten zusammenhängen, die mehrere Merkmale oder Attribute haben [27]. So kann man ML besonders gut bei Klassifizierung, Regression und Clustering anwenden [27], [13].

Grundsätzlich lassen sich die in dem folgenden Absatz beschriebenen drei Unterkategorien für die Art des Lernens finden [27]:

1. Supervised Learning: Dieser Lernansatz verwendet einen Trainingsdatensatz, der aus Paaren von Input und Output besteht [28]. Der Input enthält die als Eingabe dienenden Informationen, dabei kann es sich beispielsweise um Bilder, Texte oder Zahlen oder auch mehrere verschiedene Datenformate handeln [28], [29]. Der Output repräsentiert das korrekte Ergebnis zu den jeweiligen Eingabedaten, welches das Modell nach dem Trainingsprozess vorhersagen soll [29]. Während des Trainingsprozesses lernt das Modell, die Eingabedaten korrekt den Ausgabedaten zuzuordnen [29]. Dies geschieht durch Optimierungsmethoden, die darauf abzielen, die Abweichung von dem korrekten Output zu minimieren [29]. Dabei verwendet das Modell die vollständigen Datenpaare, bestehend aus Eingabe und zugehöriger Ausgabe [29].

2. Unsupervised Learning: Im Gegensatz zum Supervised Learning wird kein Outputdatensatz zum Training bereitgestellt [29]. Stattdessen wird es dem Modell überlassen, eigenständig Zusammenhänge innerhalb der Daten zu ermitteln [29], [27]. Die Erkennung von Strukturen und Mustern in den Daten ist das hauptsächliche Ziel [29], [27]. Unsupervised Learning findet aus diesem Grund oft dann Anwendung, wenn grundlegende Einblicke in die Datenstruktur benötigt werden oder gemeinsame Merkmale der Daten ermittelt werden sollen [29].

3. Reinforcement-Learning: Bei dem Reinforcement-Learning wird kein Trainingsdatensatz verwendet [30, S. 239 f.], [23]. Stattdessen wird einem Agenten ein konkretes Ziel definiert, welches er durch aktive Interaktion mit der Umgebung erreichen soll [23], [30, S. 239]. Das Erreichen dieses Ziels wird durch Trial and Error versucht, wobei der Agent eine möglichst optimale Strategie zum Erreichen des Zieles finden muss [30]. Die Umgebung beeinflusst den Agenten und bei der Interaktion mit ihr erhält der Agent Rückmeldungen, die sowohl positiv (Belohnungen) als auch negativ ausfallen können [23]. Die Maximierung der Gesamtsumme dieser Belohnungen ist die Hauptaufgabe des Agenten [30].

2.1.4 Künstliche neuronale Netzwerke

Künstliche neuronale Netzwerke sind Modelle des maschinellen Lernens, die auf der biologischen Grundlage des Gehirns basieren [23], [13]. Sie können bei der Lösung komplexerer und nicht linearer Problemstellungen Verwendung finden, da sie in der Lage sind, Muster in Daten zu erkennen [23]. Sie bestehen aus verschiedenen Schichten von Knoten, welche über gewichtete Verbindungen kommunizieren [13]. Der Lernprozess erfolgt über eine Anpassung dieses Gewichtes in Abhängigkeit von der Korrektheit der gegebenen Vorhersage.

Biologische Grundlage

Namensgebend für künstlich neuronale Netze (KNN) ist eine Gehirnzelle namens Neuron [23]. Es wird davon ausgegangen, dass Neuronen für die zentrale Informationsverarbeitung zuständig sind [23]. Sie sind für die Wahrnehmung und Weiterleitung von Informationen verantwortlich [23]. Das menschliche Gehirn ist ein Netzwerk aus miteinander verbundenen Neuronen [23], es umfasst ungefähr 100 Milliarden bis 1 Billion einzelne Nervenzellen.

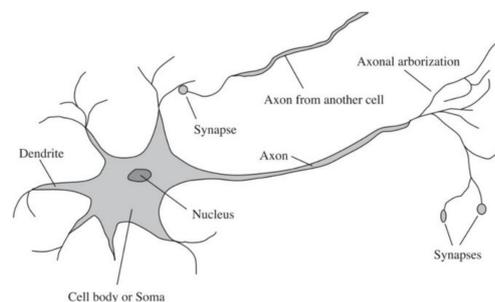


Abbildung 2.1: Bestandteile einer Nervenzelle, Neuron. Quelle: [23, S. 30]

Das Neuron besteht aus einem Zellkörper, welcher als Soma bezeichnet wird, und enthält wichtige Zellorganellen, wie den Zellkern, die für die grundlegenden Lebensfunktionen der Zelle zuständig sind [23]. Von diesem Soma gehen verzweigte Strukturen aus, die Dendriten genannt werden [23]. Ihre Aufgabe ist die Aufnahme elektrischer Impulse anderer Neuronen und deren Weiterleitung an das Soma [23]. Der Übergang zu dem Axon wird als Axonhügel bezeichnet und ist für die Sammlung, Filterung und Weiterleitung der Signale der Dendriten zuständig [23]. Das Axon transportiert die Signale zu anderen Nervenzellen. Das Ende eines Axons wird von den Synapsen gebildet, welche als Schnittstelle zu anderen Zellen dienen, die eine Übertragung von Signalen an benachbarte Zellen ermöglichen [23].

Aufbau

Bei dem Modell des KNN werden ein biologisches Neuron und dessen Signalweiterleitung als Grundlage genommen, um künstlich ein abstrahiertes neuronales Netzwerk zu erzeugen [23]. Dieses Netzwerk besteht aus Schichten von miteinander verbundenen künstlichen Neuronen, auch Knoten genannt [23]. Diese Schichten unterteilen sich in eine Eingabeschicht, eine Ausgabeschicht und beliebig viele verborgene Schichten [13].

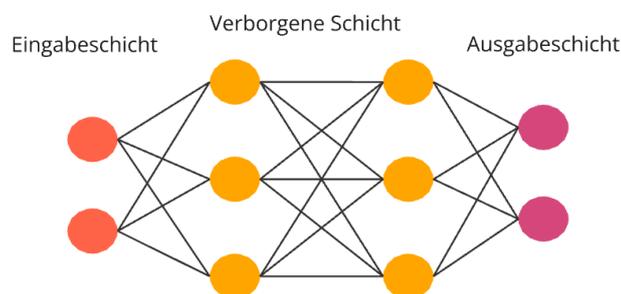


Abbildung 2.2: Einfaches künstliches neuronales Netzwerk mit zwei verborgenen Schichten

Wie das biologische Vorbild kommunizieren künstliche Neuronen miteinander, indem sie Signale weiterleiten [23]. Ein Neuron wird aktiviert und leitet das Signal weiter, wenn der Eingabewert einen bestimmten Schwellenwert überschreitet, welcher durch die gewählte Aktivierungsfunktion definiert wird [23]. Dieses Signal wird dann über gewichtete Verbindungen zu anderen Neuronen der nächsten Schicht des Netzes weitergeleitet [23]. Die Gewichtung jeder Verbindung beeinflusst die Stärke des Signals,

welches das empfangende Neuron erhält, da die Ausgabe des sendenden Neurons mit der Gewichtung der Verbindung zu dem empfangenden Neuron multipliziert wird [23]. Im empfangenden Neuron wird das gewichtete Signal mit den Signalen anderer Neuronen summiert, bevor es durch dessen eigene Aktivierungsfunktion weiterverarbeitet wird [23].

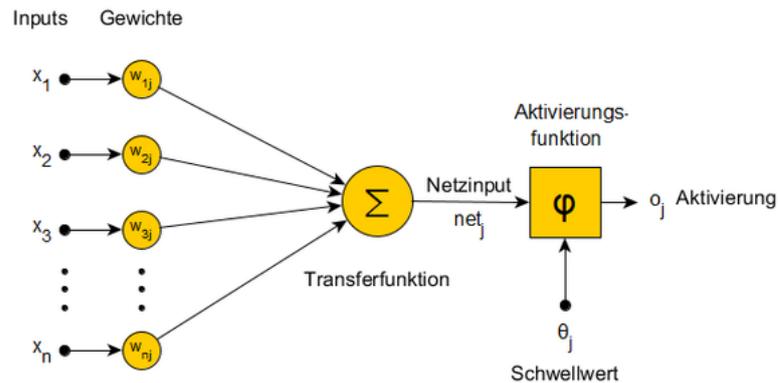


Abbildung 2.3: Mathematisches Modell für ein künstliches Neuron. Quelle: [31]

$x_{1..n}$ sind die Eingabewerte für das empfangende Neuron, welche von dem sendenden Neuron N der vorherigen Schicht gesendet werden [23]. w_{nj} steht für die numerische Gewichtung der Verknüpfung des empfangenden Neurons j mit dem sendenden Neuron n [23]. Diese Gewichte werden bei der Erstellung des neuronalen Netzwerkes zufällig festgelegt und im Verlauf des Trainingsprozesses weiter angepasst [23]. $net_j = \sum_{n=1}^z w_{nj}x_n$ ist die gewichtete Summe der Eingaben des empfangenden Neurons j , deren Berechnung nach Erhalt der Eingabe erfolgt [23]. Dabei gibt z die Anzahl dieser Eingaben an [23]. Dazu wird der Eingabewert mit dem Gewicht der Verknüpfung multipliziert und anschließend die Summe der Eingaben gebildet [23]. Danach wird die differenzierbare Aktivierungsfunktion $\Phi(net_j) = \Phi(\sum_{n=1}^z w_{nj}x_n)$ auf diese Summe angewendet [23]. Die Wahl der Aktivierungsfunktion hängt von der Anwendungsart ab [23].

Durch die Implementierung einer passenden Aktivierungsfunktion wird sichergestellt, dass das Neuron in Abhängigkeit von der gewichteten Eingabesumme einen entsprechenden Ausgabewert erzeugt [23]. Diese Ausgabe bestimmt den Aktivitätszustand des Neurons [23]. Wird ein für die Aktivierungsfunktion spezifischer Schwellenwert erreicht oder überschritten, aktiviert sich das Neuron, andernfalls bleibt es inaktiv [23]. Außerdem gewährleistet die Aktivierungsfunktion durch ihre nicht Linearität, dass das NN nicht zu einer einfachen linearen Funktion wird [23].

Deep Learning (DL)

Deep Learning baut auf der Grundidee der KNN auf und erweitert diese durch den Einsatz von mehreren verborgenen Schichten [13]. Traditionelle künstliche neuronale Netzwerke bestehen in der Regel aus wenigen verborgenen Schichten. Diese geringe Verwendung von verborgenen Schichten führt dazu, dass ihre Fähigkeit zur Erkennung komplexerer Muster eingeschränkt ist [13]. Die begrenzte Anzahl von verborgenen Schichten erleichtert aber auch die Nachvollziehbarkeit des Vorgehens innerhalb dieser Schichten. Das führt dazu, dass die Entstehung des Ergebnisses, welches von dem KNN ausgegeben wird, leichter nachzuvollziehen ist [13].

Deep-Learning-Modelle nutzen hingegen viele verborgene Schichten, welche es ihnen ermöglichen, Zusammenhänge und Muster zu erkennen, die traditionellen KNNs verborgen bleiben [13]. Das Erlernen von komplexeren Merkmalen oder Konzepten wird durch den Einsatz von dieser tieferen Struktur ermöglicht, da eine Schicht auf den Ausgaben der vorherigen Schicht aufbaut [13]. Dadurch können die genauen Vorgänge, die zu einem bestimmten Ergebnis geführt haben, aber auch schlecht bis überhaupt nicht mehr nachvollzogen werden [13]. Deshalb spricht man bei einem solchen tiefen neuronalen Netzwerk auch von einer Blackbox [13].

Die Blackbox entsteht aufgrund der vielen Schichten und der Komplexität der zwischen ihnen bestehenden Verbindungen. Diese erschweren es, die einzelnen Schritte der Informationsverarbeitung zu verfolgen und zu verstehen [13]. Die begrenzte Nachvollziehbarkeit von Deep-Learning-Modellen verringert insbesondere in Bereichen wie Recht oder Medizin, in denen Transparenz von besonderer Bedeutung ist, das Vertrauen in ihre Entscheidungen. Auch eine Fehlerdiagnose wird durch die Blackbox-Natur eines DL-Modells erschwert, so können Fehler in den Trainingsdaten unbemerkt bleiben und zu fehlerhaften Ergebnissen führen [13]. Forschungen in dem Bereich der Explainable Artificial Intelligence (XAI) setzen sich zum Ziel, die Entscheidungen von DL-Modellen transparenter zu machen und so deren Akzeptanz und Anwendbarkeit zu verbessern [32].

Transformer-Architektur

Das Transformer-Modell ist eine Architektur des Bereiches des maschinellen Lernens, welche auf dem Konzept der künstlichen neuronalen Netze basiert [5, S. 170–173], [33]. Dieses Modell ist auf die Verarbeitung natürlicher Sprache spezialisiert und somit für die effektive Verarbeitung von Sequenzen entwickelt [5, S. 170–173]. Dadurch hat es seit seiner Einführung im Jahr 2017 für große Fortschritte im Bereich der Sprachverarbeitung, aber auch bei der Audio- und Bildverarbeitung geführt [34], [35].

Die Verarbeitung von Sprache ist komplexer als diejenige numerischer Daten, da der Aufbau eines Satzes eine entscheidende Rolle für dessen Bedeutung spielt [5, S. 171 f.]. So kann die Reihenfolge, in der die Wörter in einem Satz vorkommen, die Bedeutung eines Satzes vollkommen verändern und identische Wörter haben in einem anderen Kontext verschiedene Bedeutungen [5, S. 170–173]. Auch die Zeichensetzung ist relevant und kann die Aussage eines Satzes grundlegend verwandeln [5, S. 170–173]. All diese sprachlichen Besonderheiten müssen von dem Modell berücksichtigt werden. Das Trainieren von KNNs, die Sprachsequenzen verarbeiten können, ist deshalb mit großem Aufwand verbunden [5, S. 170–173]. Das Transformer-Modell besteht primär aus zwei Komponenten, dem Encoder und dem Decoder [33], [35].

Unabhängig von dem verwendeten Modell muss beim maschinellen Lernen ein Text in eine numerische Form überführt werden, damit dieser verarbeitet werden kann [35]. Dies erfolgt bei dem Transformer-Modell in mehreren Schritten innerhalb des Encoders [33]. Zu Beginn werden die textbasierenden Daten pro Wort in einen mathematischen Vektor umgewandelt, dieser Schritt wird als Token-Embedding bezeichnet [34]. Diese Vektoren dienen dazu, die Wortbedeutung zu erkennen, wobei sich die Vektoren ähnlicher Wörter ähneln [34], [5]. Bei der Überführung der Satzsequenz in eine vektorielle Darstellung würde die Position des einzelnen Wortes im Satz verloren gehen [35]. Zur Bewahrung dieser Information wird ein „Positional Encoding“ hinzugefügt, welches die Position eines Wortvektors in der Sequenz angibt [35]. Hierzu werden oft trigonometrische Funktionen verwendet und zum Vektor addiert [35]. Anschließend wird der Zusammenhang der einzelnen Wörter innerhalb des Satzes analysiert, um den Kontext und die Bedeutung des Satzes zu erfassen [35], [5, S. 171 f.]. Dabei priorisiert das Transformer-Modell die wichtigsten Satzteile, um die Bedeutung der Wörter im Kontext zu erkennen [34]. Dieser Self-Attention-Mechanismus sorgt auch dafür, dass Wörter, die stark miteinander in Beziehung stehen, priorisiert werden [35], [34]. Durch diese Schritte kann das Transformer-Modell die wichtigsten Satzteile priorisieren und die Bedeutung der Wörter im Kontext erkennen [35], [5].

Die analysierten und für die Weiterverarbeitung vorbereiteten Satzbausteine dienen dem Decoder als Grundlage für die Generierung der Ausgabe [35]. Der Decoder generiert die Ausgabe in mehreren Schritten [33], [35]. Zuerst wird sichergestellt, dass bei der Satzgenerierung nur auf den bisher generierten Inhalt des Satzes zugegriffen werden kann [34, S. 1]. Das heißt, es soll nur der Kontext der bisher erstellten Satzsequenz benutzt werden und nicht auf die durch das Training als erwartungsgemäß eingestufte Fortsetzung des Satzes zugegriffen werden [34]. Das soll dafür sorgen, dass unerwartete Inhalte dennoch bedacht werden und keine für den Kontext falschen Antworten ausgegeben werden [35]. Dieser Schritt wird als Masked Self-Attention bezeichnet und spielt sowohl bei dem Training als auch nach dem Trainingsprozess eine fundamentale Rolle [35], [36]. Anschließend sorgt der Encoder-Decoder-Attention-Schritt dafür, dass die von dem Encoder erzeugten Informationen genutzt werden, um die generierte

Antwort nicht nur in dem Kontext der generierten Satzsequenz zu betrachten, sondern auch mit der Eingabe in Zusammenhang zu bringen [35], [33].

Bei der Verarbeitung von sequenziellen Daten wie Sprache sind Transformer-Modelle besonders leistungsfähig [5, S. 170-173]. Die Architektur erlaubt zudem, dass Berechnungen parallel durchgeführt werden können, und sorgt dafür, dass das Training in einem Bruchteil der für das Training von künstlichen neuronalen Netzen benötigten Zeit durchgeführt werden kann [35], [36]. Zudem ist die Möglichkeit zum Transfer-Learning, wo bereits vortrainierte Modelle für bestimmte kleinere Probleme feinjustiert werden können, zeiteffizient und kostensparend [36]. Allerdings benötigen Transformer-Modelle auch einen großen Trainingsdatensatz und neigen bei kleineren Trainingsdatensätzen zum Overfitting [36]. In einem solchen Fall kann das Transfer-Learning eine mögliche Lösung für dieses Problem bereitstellen [36].

2.2 ChatGPT

ChatGPT ist ein bekanntes und fortschrittliches KI-Tool, welches von OpenAI entwickelt wurde [4]. Der Name des Systems setzt sich aus mehreren Komponenten zusammen. Der Begriff *Chat* verweist auf die Benutzeroberfläche, die an bekannte Messenger wie WhatsApp und ähnliche Anwendungen erinnert und deren Bedienung allgemein geläufig ist [4]. Der Bestandteil *GPT* steht hingegen für *Generative Pretrained Transformer* [4]. *Generative* bezieht sich auf die Fähigkeit, neue Inhalte zu erstellen, indem auf bereits gelernte Muster zurückgegriffen wird [4]. *Pretrained* weist auf die großen Mengen an Daten hin, mit denen die KI trainiert wurde [4]. *Transformer* ist die zugrunde liegende Architektur, die bei der Entwicklung verwendet wurde [36], [4]. Dabei wird jedoch nicht das ursprünglich im Paper „Attention is All You Need“ vorgestellte Transformermodell mit Encoder und Decoder verwendet, sondern eine Modifikation, die ausschließlich auf der Verwendung des Decoder-Teils basiert [37, S. 32].

ChatGPT-Versionen ab 3.5 sind in der Lage, große Textmengen zu verarbeiten und Antworten in natürlicher Sprache zu generieren [4]. Die früheren Modelle konnten nur unter der Voraussetzung bedient werden, dass Benutzende ihre Eingaben in einer formalen Sprache, die an Programmiersprachen angelehnt war, schrieben [13]. Mit ChatGPT 3.5 änderte sich dies grundlegend und Nutzende waren erstmals in der Lage, einen auf Chat basierenden Dialog zu führen [13]. Die Effizienz und die Benutzerfreundlichkeit und daraus resultierende Vorteile im Vergleich zu traditionellen Suchmaschinen führten zu einer schnellen globalen Verbreitung der KI [37].

2.2.1 Entstehung

Die Grundlage, welche die Entwicklung von ChatGPT ermöglicht hat, ist die jahrzehntelange Forschung in dem Bereich der künstlichen Intelligenz. Insbesondere das Gebiet des maschinellen Lernens und der natürlichen Sprachverarbeitung (engl. Natural Language Processing (NLP)) hat großen Einfluss auf die Entwicklung von ChatGPT genommen [37]. Durch die Vorstellung der Transformer-Architektur im Jahr 2017 wurde die grundlegende Architektur für die Entwicklung der ersten GPT-Version bereitgestellt [35], [36].

Die Entwicklung von Generative Pretrained Transformer (GPT), der Urversion der heutigen Versionen von ChatGPT, war ein Durchbruch, da sie zeigte, dass KI in der Lage ist, Sprache zu verstehen und auch natürliche Sprache mit hoher Qualität zu erzeugen [37], [38]. Mit der Veröffentlichung des ersten GPT-Modells im Jahr 2018 konnten bereits qualitative Texte generiert werden [37]. Allerdings folgte bereits zwei Jahre später mit GPT-2 eine Version mit signifikanten Verbesserungen, welche durch ein Training mit einer noch größeren Menge von Textdaten und die Anpassung an spezifische Aufgaben ermöglicht wurden [13], [37]. Im Jahr 2020 wurde die Version 3 eingeführt und setzte neue Maßstäbe in der natürlichen Sprachverarbeitung, durch die verbesserten Fähigkeiten bei der Übersetzung und Zusammenfassung von Texten und der Generierung von Antworten [38], [37]. Die Fähigkeit, Texte zu verstehen und Antworten zu generieren, sorgte für eine vielseitige Einsatzfähigkeit [37].

Die Einführung von ChatGPT 3.5, welche auf der Version GPT-3 basiert, ermöglichte es erstmalig einem großen Nutzerkreis, mit dieser Version zu arbeiten [13]. Sie ist in der Lage, Dialoge zu führen und Inhalte zu erklären, wobei die Kommunikation natürlich und menschenähnlich ist [38]. ChatGPT 3.5 ist seit der Veröffentlichung für jeden frei zugänglich und hatte bereits in den ersten Wochen nach der Bereitstellung einen großen Benutzerkreis [13].

Seit ChatGPT 3.5 hat OpenAI den Chatbot stetig aktualisiert und verbessert [4]. Damit wurde die Leistung optimiert, aber auch die Sicherheit und Zuverlässigkeit wurden verbessert [13], [4]. ChatGPT 4.0 bietet eine Optimierung bei der Bearbeitung von komplexeren Anfragen und ist besser im Umgang mit fehlerhaften Eingaben [39]. Auch die Einführung der Möglichkeit, ChatGPT 4.0 mit zusätzlichen Plug-ins zu erweitern, bietet neue Funktionalitäten [40]. Dadurch ergeben sich neue Anwendungsmöglichkeiten und eine bessere Qualität der Informationen, die bereitgestellt werden [39].

2.2.2 Unterschiede ChatGPT 3.5 und 4.0

Die Unterschiede zwischen den beiden ChatGPT-Versionen mag bei oberflächlicher Betrachtung zunächst nicht offensichtlich sein, doch bei genauerer Betrachtung offenbaren sich schnell die unterschiedlichen Qualitäten der Antworten [41], [39]. Diese Unterschiede entstanden durch die technischen Verbesserungen der nachfolgenden Version [39].

So besteht ChatGPT 3.5 aus rund 175 Milliarden Parametern [13]. Obwohl es keine offiziellen Informationen über die genaue Anzahl der Parameter von ChatGPT 4.0 gibt, wird allgemein angenommen, dass die Parameteranzahl deutlich höher ist [13]. Mit Parametern sind die Gewichte der Verknüpfungen zwischen zwei Neuronen des künstlichen neuronalen Netzes gemeint, die während des Trainingsprozesses zur Optimierung des Modells angepasst werden. Je mehr Parameter ein Netzwerk hat, desto komplexer wird es und desto besser kann es bei gutem Training gute Ergebnisse liefern [13], [13]. Allerdings müssen eine höhere Komplexität und Größe eines KNNs nicht zwangsläufig mit einer besseren Qualität gleichgesetzt werden.

Ein weiterer fundamentaler Unterschied zwischen den Versionen ist, dass ChatGPT 3.5 mit Daten bis Januar 2022 trainiert wurde und daher nur auf Informationen bis zu diesem Zeitpunkt zugreifen kann [39], [13]. Das bedeutet, dass Informationen zu Ereignissen nach diesem Zeitpunkt nicht vorliegen können [13]. ChatGPT 4.0 wurde hingegen mit Daten bis 2023 trainiert und kann daher auf aktuellere Informationen zugreifen [39]. Darüber hinaus hat ChatGPT 4.0 durch die Integration von Application Programming Interfaces (APIs) die Möglichkeit, über Bing auf aktuelle Informationen im Internet zu rekurren, wodurch eine Bereitstellung aktueller Informationen möglich ist [40], [39].

Zusätzlich zu den Verbesserungen in der Textverarbeitung stellt ChatGPT 4.0 durch das Integrieren von Plug-ins erweiterte Fähigkeiten bereit, wie Bildverarbeitung und -generierung sowie Spracherkennung [42], [40]. Durch die Erweiterung der KI-Fähigkeiten mit passenden Plug-ins eröffnen sich neue Perspektiven und Anwendungsmöglichkeiten. Grundsätzlich ist ChatGPT in der Lage, mathematische Berechnungen durchzuführen [39], [41]. Bei Aufgaben, bei denen eine hohe Genauigkeit gefragt ist oder die eine komplexere Problemstellung beinhalten, ist diese Fähigkeit jedoch eingeschränkt [43]. Durch die Integration eines passenden Plug-ins wie Wolfram Alpha können die Fähigkeiten zur zuverlässigen Lösung mathematischer Aufgaben verbessert werden [43].

2.2.3 Wolfram Alpha

Wolfram Alpha ist eine seit 2005 aktive semantische Suchmaschine, welche natürliche Sprache als Eingabe verarbeitet und dadurch auf individuelle Fragestellungen reagieren und antworten kann [44], [45]. Die Verarbeitung der Eingabe erfolgt durch die Erfassung der Semantik der Frage [44]. Die Benutzenden bekommen anders als bei einer herkömmlichen Suchmaschine wie Google keine Hyperlinks zu passend eingegebenen Schlagwörtern präsentiert, aus denen sie sich die benötigten Daten selbstständig herausuchen müssen [44]. Stattdessen analysiert Wolfram Alpha die Bedeutung des eingegebenen Textes und stellt eine passende Auswahl an Informationen aus der Datenbank zur Verfügung [44]. Dabei wird auf eine vielfältige Datengrundlage zurückgegriffen, welche aus verschiedenen Bereichen stammt, wie Medizin, Mathematik, Physik etc. [46]. Die Aktualisierung der Daten, auf welche Wolfram Alpha Zugriff hat, erfolgt regelmäßig, um keine veralteten Informationen zu erhalten [44], [46]. Die verwendeten Informationen stammen laut der Wolfram-Alpha-Website aus einem breiten Spektrum von Expertenwissen aus 10 Billionen Primärquellen und 50 000 verschiedenen Arten von Algorithmen und Gleichungen [46], [45]. Die generierten Antworten basieren nicht auf Meinungen, sondern auf fundierten und objektiven Daten oder Berechnungen [44].

Wolfram Alphas Fähigkeiten beschränken sich nicht nur auf das Verständnis natürlicher Sprache als Eingabe und die Generierung dazu passender Antworten, sondern zeichnen sich auch durch die Erkennung von Zusammenhängen zwischen Daten aus [44]. Die Suchmaschine verbindet Fakten mit Berechnungen, setzt Informationen in ihren Kontext und kombiniert Erkenntnisse der Forschungen verschiedener Bereiche [44], [46]. Auch die Fähigkeit, Berechnungen durchzuführen und dabei auf mathematische Konzepte und Algorithmen zurückzugreifen, ist eine der fundamentalen Stärken [44]. In ihrem Artikel kamen Dahal, Lamichhane, Luitel und Pant zu dem Ergebnis, dass Wolfram Alpha effizient komplexe Berechnungen durchführen und mathematische Theoreme nutzen kann und dabei ein detailliertes Ergebnis zurückgibt [47]. Bei der Beantwortung der Fragen legt Wolfram Alpha den Schwerpunkt auf eine hohe Korrektheit der Aussagen, dabei erfolgt nur die Ausgabe von Antworten, die das System mit einer sehr hohen Wahrscheinlichkeit korrekt einstuft [44, S. 276 f.], [45]. Das hat zur Konsequenz, dass die Antwort nicht immer zuverlässig alle relevanten Informationen anzeigt [44]. Vor allem ist dies der Fall, wenn bei diesen Informationen nicht sichergestellt werden kann, ob sie Falschinformationen enthalten [44].

Um ein optimales Verständnis der Lösung zu garantieren und die Nachvollziehbarkeit zu gewährleisten, ist die Angabe des Lösungswegs ein wesentlicher Schritt. Die Fähigkeit, nicht nur die ermittelten Ergebnisse bereitzustellen, sondern auch die einzelnen Schritte des Lösungsprozesses transparent und nachvollziehbar darzustellen,

erweitert die Einsatzmöglichkeiten von Wolfram Alpha im Bildungskontext [48]. Die Möglichkeit, den Lösungsweg einer gestellten Frage angezeigt zu bekommen, wurde durch die `SShow Steps`Funktion bereitgestellt und ermöglicht den Nutzenden, die Lösungsfindung zu verfolgen und zu verstehen [48]. Insbesondere im Bildungsbereich ist dies von Bedeutung, da es die Bildung von Konzepten erleichtert und dabei hilft, abstrakte Resultate greifbar zu machen [48]. Die Einsatzmöglichkeiten dieser Funktion sind allerdings auf mathematische Probleme und einige andere spezifische Bereiche beschränkt [48]. Es besteht die Möglichkeit, dass bei komplexen mathematischen Problemen, die fortgeschrittene mathematische Methoden erfordern, keine detaillierten Schritte zur Verfügung gestellt werden. Die `SShow Steps`Funktion ist nicht nur über die Benutzeroberfläche von Wolfram Alpha verfügbar, sondern kann auch über APIs in andere Projekte integriert werden [49].

Die Implementierung von Wolfram Alpha erfolgte mit der Programmiersprache `Wolfram Language`, die auf symbolische Berechnungen und die Verarbeitung großer Datenmengen spezialisiert ist und eine umfassende Sammlung von Algorithmen integriert hat [45]. Die Basis von Wolfram Alpha bildet die Software `Mathematica`, welche ebenfalls auf `Wolfram Language` basiert und die Durchführung komplexer Berechnungen und Analyse von großen Datenmengen ermöglicht.

ChatGPT 4.0 und Wolfram

Wolfram Alpha ist eine auf symbolischen Verarbeitungsmethoden basierende KI, welche anders als ChatGPT, welches auf einem KNN basiert und sich die Fähigkeit zur Sprachverarbeitung durch Training angeeignet hat, durch festgelegte Regeln natürliche Sprache verarbeitet [44]. Dadurch ist es der KI nur möglich, in Sprachen angewendet zu werden, bei denen die Wortbedeutung und die grammatikalischen Regeln von den Entwicklerinnen und Entwicklern bereits definiert wurden [44]. Wolfram Alpha verwendet primär die Sprache Englisch, eine vollständige Verwendung der Suchmaschine in anderen Sprachen ist bisher nicht möglich [44].

Seit 2023 gibt es die Möglichkeit, ChatGPT 4.0 in Kombination mit Wolfram Alpha zu nutzen [43]. Die Anbindung der Wolfram-API ermöglicht es ChatGPT, mithilfe von Wolfram Alpha Antworten auf Fragen zu generieren [43]. Dazu werden die von den Nutzenden an ChatGPT 4.0 gestellten Fragen von der KI verarbeitet [43], [42]. Bei dieser Verarbeitung wird von ChatGPT 4.0 eine für Wolfram Alpha geeignete Abfrage erstellt, deren Inhalt nach dem Ermessen des Chatbots formuliert wird [43]. Diese Anfrage wird über die Schnittstelle an Wolfram Alpha weitergegeben, wo die Verarbeitung dieser Anfrage stattfindet [40], [44]. Die Ergebnisse der Berechnungen Wolfram Alphas werden an ChatGPT zurückgegeben und in eine für die Anfrage

des Benutzenden sinnvolle Antwort umgewandelt [43]. Bei diesem Schritt entscheidet ChatGPT, in welchem Maße die von Wolfram Alpha erhaltenen Informationen verwendet werden [43].

Das Interface zu Wolfram Alpha stellt für ChatGPT zwei Möglichkeiten zur Ansprache bereit [43]. Die erste Option für ChatGPT, eine Anfrage an die Wolfram-API zu übermitteln, ist die Übergabe von einer textuellen Anfrage an Wolfram Alpha [43]. Da ChatGPT ein auf natürliche Sprache spezialisiertes Modell ist, kann es eine für Wolfram Alpha passend formulierte Anfrage erstellen und übermitteln [43]. Die zweite Möglichkeit der Kommunikation mit Wolfram Alpha ist die Erstellung von Wolfram-Language-Code [43]. Diese Variante, deren Leistungsfähigkeit laut Stephen Wolfram, dem Entwickler von Wolfram, im Vergleich zu der auf natürlicher Sprache basierenden Kommunikationsmöglichkeit höher ist, ist auf Präzision ausgelegt [43]. Allerdings existieren laut Aussagen von Stephen Wolfram noch Schwierigkeiten bei der Generierung von korrektem Wolfram-Language-Code durch ChatGPT [43].

Das Zusammenführen dieser beiden KIs hat den Vorteil, dass die Stärken beider Seiten optimal ergänzt werden können [43]. ChatGPTs Stärke liegt in der natürlichen Sprachverarbeitung und dem Erfassen des Kontextes, während Wolfram Alpha über eine aktuelle und umfassende Wissensdatenbank und die Fähigkeit verfügt, komplexe Berechnungen und Analysen durchzuführen [43]. Durch diese Verbindung kann dem Effekt des Halluzinierens entgegengewirkt werden [43].

Allerdings entstehen durch die Schnittstelle auch Probleme, welche von Stephen Wolfram Alpha in einem Artikel erwähnt werden, den er anlässlich der Verbindung von ChatGPT mit Wolfram Alpha schrieb [43]. Die Erstellung der Anfrage erfolgt durch ChatGPT, somit erhält Wolfram Alpha lediglich die von dem Chatbot formulierten Informationen zu der Nutzeranfrage [43]. Ein von ChatGPT fehlerhaft formulierter Arbeitsauftrag hat dadurch auch unter Verwendung der API eine inkorrekte Antwort zum Resultat [43]. Eine Untersuchung von mathematischen und naturwissenschaftlichen Problemen ergab, dass ein Einsatz der Wolfram-Alpha-API im Vergleich zu ChatGPT 4.0 keine nennenswerten Verbesserungen erzielen konnte [50]. Als Grund für diesen Zustand wurde das Kommunikationsproblem über die Schnittstelle angegeben, welches zur Folge hat, dass Wolfram Alpha zu der Problemstellung unpassende Fragestellungen von ChatGPT 4.0 bekam [50], [43].

Eine weitere potenzielle Fehlerquelle besteht in der Generierung von Antworten für die Nutzenden auf der Basis eigener Entscheidungsmechanismen durch ChatGPT [43]. Dabei kann eine Fehlinterpretation der von Wolfram Alpha erhaltenen Ergebnisse entstehen [43]. Dieses Problem kann dann entstehen, wenn ChatGPT keine Lösungsschritte zurückgegeben werden und dadurch keinen konkreten Einblick in die Lösungsfindung von Wolfram Alpha bekommt. Aber auch die Übersetzung der von Wolfram Alpha zurückgegebenen Antwort in die für die Nutzenden verständliche

Sprache kann, sowohl bei dem Stellen als auch bei dem Erhalten der Anfrage, zu einer Missinterpretation der Ergebnisse führen. In ihrer Studie untersuchten Manakhimova et al. die Übersetzungsleistung von GPT-4 in verschiedenen Sprachpaaren, einschließlich Englisch-Deutsch und Deutsch-Englisch [51]. Die Ergebnisse zeigten, dass GPT-4 in diesen Sprachpaaren eine hohe Übersetzungsgenauigkeit erreichte und mit den besten spezialisierten NMT-Systemen vergleichbar war [51]. Allerdings wurden auch Schwächen in der Übersetzung von idiomatischen Ausdrücken und der Behandlung komplexer grammatikalischer Strukturen ermittelt [51]. Ausgehend von den Ergebnissen dieser Studie scheint das Risiko einer Fehlinterpretation der Anfrage an oder der Antwort von Wolfram Alpha aufgrund eines Übersetzungsfehlers gering zu sein [51].

Anders als bei der alleinigen Nutzung von Wolfram Alpha haben die entstehenden Lösungen das Potenzial, stark voneinander abzuweichen [43]. Der Tatsache, dass unterschiedliche Versuche verschiedene Ergebnisse erzeugen können, nähert sich somit eher ChatGPT 4.0 als Wolfram Alpha an [43], [50]. In welchem Umfang die Integration von Wolfram Alpha die Fähigkeiten zum korrekten Lösen einer mathematischen Aufgabe beeinflusst und wie sich die Verwendung auf die Erklärung der Rechnung auswirkt, soll im Rahmen dieser Arbeit betrachtet werden.

2.2.4 Verwendung und Prompten

Die grundlegende Handhabung von ChatGPT gestaltet sich intuitiv und einfach [13], [38]. Nach der Eingabe eines Textes, welcher auch als Prompt bezeichnet wird, durch die nutzende Person wird eine passende Antwort generiert [13]. Dabei wird die Qualität der Antworten stark durch die Formulierung der Eingabe beeinflusst [38]. Auch die Verwendung von unterschiedlichen Wörtern und Sprachen nimmt Einfluss auf die Ausgabe der KI [38]. Der Grund dafür ist, dass dialogbasierte KI-Sprachmodelle wie ChatGPT darauf spezialisiert sind, Eingaben der Benutzerin oder des Benutzers im Kontext zu interpretieren und darauf basierend individuelle Antworten zu generieren [38], [13]. Dabei hängt die Qualität des Ergebnisses stark von der Präzision des formulierten Prompts ab [38]. Im Gegensatz zu Suchmaschinen, wo man oft nur Schlüsselwörter eingibt, erfordern KI-Modelle eine detaillierte Aufgabenstellung und die Angabe des Kontextes und relevanter Informationen [38]. Die Verfeinerung dieser Prompts kann zur Erzielung besserer Ergebnisse beitragen [38].

Bei der Formulierung eines optimalen Prompts sind die Konkretheit und Eindeutigkeit der Fragestellung essenziell für die Erfassung des Kontexts und der Intention der/des Benutzenden [38]. Dies ermöglicht die Vermeidung von Fehlinterpretationen bei der Verarbeitung und Interpretation durch ChatGPT sowie die Generierung präziser Antworten zu dem gestellten Problem [38]. Fragen, die allgemein gehalten sind, resultieren häufig in allgemeinen und unspezifischen Antworten [38]. Die Angabe des Kontextes,

in dem die Fragestellung erfolgt, trägt zu einer spezifischeren und passenderen Antwort bei [38]. Dies kann auch dann hilfreich sein, wenn die Formulierung einer eindeutigen Frage Schwierigkeiten bereitet, da durch die Angabe des Kontextes dennoch eine präzisere Antwort ermöglicht werden kann [38].

Missverständnisse bei den KI-generierten Antworten können durch aktive Kommunikation vermieden werden. Die Fähigkeit von ChatGPT, sich auf den bisherigen Chatverlauf zu beziehen, eröffnet die Möglichkeit, durch das aktive Stellen von Nachfragen Missverständnisse auszuräumen [38]. Somit kann die Zusammenfassung des Textverständnisses den Kontext verdeutlichen und zum besseren Verständnis beitragen. Die Auseinandersetzung mit gegensätzlichen Meinungen im Dialog kann zudem zu neuen Erkenntnissen führen und die Verbreitung einseitiger Informationen vermeiden [38].

Bei der Generierung von Antworten durch die künstliche Intelligenz können Fehler auftreten, die zu Fehlinformationen führen [38]. Daher ist eine Überprüfung der Korrektheit der Antworten essenziell bei der Nutzung der KI [38]. Dies erfordert eine kritische Betrachtung und Kontrolle der Fakten durch die Hinzunahme weiterer Quellen [38]. In diesem Kontext unterscheidet sich ChatGPT nicht wesentlich von der Recherche mit digitalen Inhalten im Internet, bei der ebenfalls eine Überprüfung der Informationen notwendig ist [38].

2.2.5 Probleme und Kritik

Die Nutzung von ChatGPT bietet zahlreiche Vorteile. Die künstliche Intelligenz ermöglicht es Benutzenden, sich rasch und unkompliziert über diverse Themen zu informieren. Durch die Möglichkeit zur Interaktion mit dem System können Anwendende gezielt Fragen stellen und Themen eingehend diskutieren. Die Technologie wird trotz ihrer positiven Aspekte aber auch kritisch betrachtet. Da ChatGPT eine relativ neue Technologie ist, muss erst noch herausgefunden werden, wie man am besten mit ihr umgeht und wie sie sinnvoll in das tägliche Leben oder in die Lehre integriert werden kann [38], [13].

Dabei stellt die Verlässlichkeit der von ChatGPT bereitgestellten Informationen eine wesentliche Herausforderung dar [13]. Die künstliche Intelligenz wird mit Daten aus dem Internet trainiert, die fehlerhaft oder veraltet sein können [52]. Nach Abschluss des Trainings werden keine neuen Informationen hinzugefügt, was bedeutet, dass die Datengrundlage ab diesem Zeitpunkt nicht mehr aktualisiert wird [13], [52, S. 10]. Insbesondere in Themengebieten mit unzureichender Datengrundlage kann ChatGPT nicht immer korrekte und aktuelle Antworten liefern [13].

Ein weiteres Problem stellt das Phänomen des *Halluzinierens* dar, bei dem ChatGPT nicht zutreffende oder erfundene Informationen in einer Weise präsentiert, die sie als realistische Fakten erscheinen lässt [13], [52, S. 10]. Aufgrund der komplexen Struktur neuronaler Netzwerke in der Größe von ChatGPT ist es technisch nicht möglich, die spezifischen Quellen zu identifizieren, die für eine gegebene Antwort verantwortlich sind [23]. Daher kann ChatGPT selbst auch keine Auskunft über die Herkunft seiner Informationen geben [13]. Diese Einschränkung verhindert, dass Nutzerinnen und Nutzer die Verlässlichkeit einer Quelle beurteilen können, da sie nicht nachvollziehen können, auf welchen Daten oder Informationen eine Antwort basiert [13]. Dieses Fehlen von Transparenz erschwert es, die Glaubwürdigkeit und Zuverlässigkeit der durch die KI bereitgestellten Informationen zu überprüfen, und erhöht das Risiko von Missverständnissen oder Fehlinformationen [13], [52, S. 10]. Für Nutzende, welche sich zum ersten Mal über ein Themengebiet informieren möchten, ist eine Unterscheidung zwischen korrekten und inkorrekten Informationen ein besonderes Problem [52, S.10], [13]. Die Angabe der Informationsquelle könnte die Überprüfung der Glaubwürdigkeit der bereitgestellten Informationen erleichtern. Allerdings gewährleistet dies nicht die Korrektheit der von der KI generierten Antworten, sondern stellt lediglich eine Möglichkeit zur Überprüfung der Quellen bereit. Die Verantwortung für die Überprüfung der Richtigkeit der erhaltenen Antworten verbleibt weiterhin bei den Nutzenden. Problematisch ist in diesem Zusammenhang, dass aufgrund des Effekts des *Halluzinierens* fehlerhafte Informationen glaubhaft und schlüssig dargestellt werden können, was die Identifikation der Falschinformationen trotz der Angabe der Quellen erschwert. Es ist daher relevant, dass bei einer Recherche nicht ausschließlich eine einzige Quelle herangezogen wird, sondern zusätzlich auch weitere Quellen zum Abgleich genutzt werden.

Auch beim Datenschutz bestehen einige Bedenken, die in einigen Ländern, wie Italien, zu einem zeitweisen Verbot der KI geführt haben [13], [53]. ChatGPT nutzt die aus den Chatverläufen mit den Nutzenden gesammelten Daten zur kontinuierlichen Verbesserung. So können von den Nutzenden unbewusst vertrauliche Daten preisgegeben werden, die anschließend von der KI analysiert und verarbeitet werden. Dies kann zur Speicherung sensibler Informationen führen und die automatische Weiterverarbeitung personenbezogener Daten zur Folge haben. Die Europäische Union (EU) regelt dies durch die Datenschutz-Grundverordnung (DSGVO), die strenge Vorgaben für die Verarbeitung und Verbreitung personenbezogener Daten macht. Die kostenfreie Version 3.5 stellt keine Möglichkeit bereit, die Verarbeitung der eingegebenen Daten zu verhindern. Bei der nachfolgenden Version 4.0 wird den Nutzenden allerdings die Möglichkeit geboten, das Training mit den Chatdaten zu unterbinden [4].

Die von der KI generierten Texte sind nicht vorurteilsfrei und können einseitige Informationen enthalten [52, S. 10]. Dadurch gibt es im Bildungsbereich Bedenken, dass die Verwendung von ChatGPT einseitige und nicht vorurteilsfreie Informationen

liefert [52]. Dies könnte der Entwicklung des kritischen Denkens schaden und die Kreativität massiv beeinträchtigen [52, S. 10], [13]. Auch das selbstständige Lernen und Arbeiten kann durch die Nutzung des KI-Tools weniger gefördert werden [13]. Die Möglichkeit einer Abhängigkeit von einem KI-Tool könnte massive Folgen für die Qualität der Lehre haben [52, S. 10]. Die tatsächlichen Auswirkungen, die ChatGPT auf den Bildungsbereich haben wird, sind jedoch nur schwer vorherzusagen [13].

2.3 ChatGPT in der Lehre

Über den sinnvollen Einsatz von auf KI basierenden Tools in der Lehre wird diskutiert [54]. An manchen Bildungseinrichtungen wurde ein Verbot von ChatGPT als Hilfsmittel ausgesprochen [54]. Laut dem Ministerium für Schule und Bildung des Bundeslands Nordrhein-Westfalen kann allerdings bei der schnellen und dynamischen technologischen Entwicklung der letzten Jahre ein Verbot der Verwendung von KI-Tools im Bildungsbereich keine tragfähige Reaktion sein [54]. Stattdessen wird betont, dass Bildungseinrichtungen in der Pflicht stehen, den verantwortungsbewussten Umgang mit KI zu lehren [54]. Auf diese Weise können Bildungseinrichtungen ihren Bildungs- und Erziehungsauftrag erfüllen und die Schülerinnen und Schüler optimal auf eine zunehmend digitalisierte Welt vorbereiten [54].

Nicht nur hinsichtlich der Vermittlung zukünftiger digitaler Kompetenzen ist eine Einbindung in die Lehre sinnvoll, sondern auch bezüglich der Förderung des Lernprozesses hat künstliche Intelligenz positive Auswirkungen [55]. Der größte Vorteil ist die permanente Erreichbarkeit. Während Lehrpersonal nur zu bestimmten Zeiten für Fragen zur Verfügung steht, ist eine KI jederzeit ansprechbar [52]. Zudem bietet KI Vorteile in der individuellen Förderung [56]. Sie kann auf spezifische Probleme eingehen, sich an die Lerngeschwindigkeit der Lernenden anpassen, den Lernfortschritt überwachen und zusätzliche Übungsaufgaben erstellen [56].

Künstliche Intelligenz wie ChatGPT ist für Schülerinnen und Schüler von großem Interesse [54]. Forschungsstudien wie die von Qawaqneh et al. oder Ana Stojanov belegen, dass der Einsatz von KI in der Lehre die Motivation beim Verinnerlichen von Lerninhalten erheblich steigern kann [55],[57]. So kann durch den verantwortungsvollen Einsatz von KI in der Lehre nicht nur eine individuell anpassbare und zeitlich flexible Lernunterstützung angeboten, sondern auch die Motivation der Lernenden erhöht werden [56], [55, S. 118]. Im Bereich des Bildungswesens ermöglicht dieser Einsatz die Entstehung von neuen Möglichkeiten [56], [58, S. 3]. So betont Martínez-Sevilla, dass KI das Verständnis mathematischer Konzepte erleichtert und besonders bei der Vertiefung des Wissens und der Visualisierung unterstützen kann [56]. Es wird jedoch betont, dass KI das Lehrpersonal nicht ersetzen, sondern vielmehr ergänzen soll [56]. Die

Bereitstellung solcher unterstützenden Technologien kann das Lernen erleichtern, sollte aber als Erweiterung und nicht als Ersatz für qualifiziertes Lehrpersonal verstanden werden [56].

Mit diesem Hintergrund wird deutlich, dass die Einbindung von KI in die Lehre viele Möglichkeiten und Vorteile bringt. Der Einsatz von Tools wie ChatGPT kann zur Erweiterung und Unterstützung der traditionellen Lehre beitragen [58]. Eine Herausforderung stellt der sinnvolle und verantwortungsbewusste Umgang mit diesen Werkzeugen dar.

2.3.1 Konstruktivismus

Betrachtet man die Lerntheorie des Konstruktivismus, wird deutlich, wie ChatGPT beim Verinnerlichen von Konzepten eine sinnvolle Unterstützung bieten kann [58, S. 1 f.]. Die Theorie des Konstruktivismus basiert auf der Annahme, dass Wissen aktiv angeeignet werden muss [58, S. 1 f.]. So ist eine direkte Übertragung von Wissen nicht möglich, sondern Lernende müssen sich während des Lernprozesses ihre eigene Konzepte erstellen [58, S. 2]. Dadurch entwickelt jeder Mensch ein individuelles Bild der Realität und nutzt eigene Konzepte, um Gelerntes zu verinnerlichen und abzurufen [58]. Dies kann dazu führen, dass von anderen vermittelte Konzepte von der oder dem Lernenden missverstanden werden [58]. Erfolgreiches Lernen erfordert somit eine intensive und ausführliche Auseinandersetzung mit dem Lehrstoff [58].

Ein dynamischer Austausch ist zur Unterstützung dieses Lernprozesses von zentraler Bedeutung [58]. Bei der Entwicklung eines Konzeptes gelten Fehler als eine Möglichkeit zur Verbesserung und Weiterentwicklung [58]. Durch eine KI wird eine Umgebung geboten, in der es möglich ist, das eigene Konzept zu überprüfen und Hypothesen zu testen, ohne negatives Feedback durch menschliche Teilnehmende zu befürchten [58]. Die Möglichkeit, sein eigenes Konzept individuell zu überprüfen und aus verschiedenen Perspektiven zu betrachten, ist laut der Theorie des Konstruktivismus von besonderer Relevanz, da Lernende Wissen am besten aufnehmen, wenn sie aktiv an der Konstruktion dieses Wissens beteiligt sind und es mit ihren eigenen Erfahrungen und Interessengebieten verknüpfen können [58].

Eine chatbasierte KI kann helfen, indem sie durch direkte Kommunikation auf das Gegenüber eingeht und unterschiedliche Erklärungsansätze anbietet [58, S. 2 f.]. Durch Nachfragen können Missverständnisse ausgeräumt werden und das eigene Verständnismodell kann konstruiert und ausgebaut werden [58]. Im Unterricht ist eine solch ausführliche Kommunikation mit den einzelnen Lernenden meist nicht möglich [58].

Somit ist ChatGPT im Kontext des Konstruktivismus ein effektives Instrument für den Lernprozess [58]. Es wird den Lernenden durch die Bereitstellung verschiedener

Lösungsansätze und die Möglichkeit zum stetigen Hinterfragen und Anpassen des eigenen Verständnisses ermöglicht, ihr Wissen selbstbestimmt zu konstruieren [58]. Dieses Prinzip entspricht dem Grundgedanken des Konstruktivismus [58].

2.3.2 Forschungsstand und Ausblick

Eine Vorhersage, welche langfristigen Auswirkungen der Einsatz von KI im Bildungsbereich haben wird, ist nur schwer vornehmbar. Die Verwendung von Technologien wie ChatGPT bietet viele vorteilhafte Möglichkeiten, hat jedoch auch das Potenzial negativer Folgen. Durch die Tatsache, dass ChatGPT eine relativ junge Technologie ist, existieren noch nicht viele Daten zur Prognose der Auswirkungen und Folgen des Einsatzes in der Lehre. Daher sind weitere Forschungen in diesem Bereich notwendig [59]. Supriyadi und Kuncoro stellen im Rahmen ihrer Studie über das Sprachmodell ChatGPT fest, dass die zukünftige Gestaltung des Mathematikunterrichtes stark von der Entwicklung von Chatbots beeinflusst werden dürfte [60]. Laut Supriyadi und Kuncoro wird kritisches Denken als eine Kompetenz beschrieben, die durch den Einsatz von KI im Unterricht gefördert wird [60]. In diesem Kontext diskutieren Supriyadi und Kuncoro das Potenzial von künstlicher Intelligenz zur Revolutionierung des bisherigen Mathematikunterrichts [60, S. 315].

Vor dem Verwenden von KIs in der Lehre ist es wichtig, dass die zum Einsatz kommenden KIs hinsichtlich ihrer Leistungsfähigkeit und Qualität im Erklären mathematischer Aufgaben und Konzepte untersucht werden. Dadurch werden die verschiedenen Modelle vergleichbar und es wird ein Einblick in mögliche Auswirkungen geschaffen. Aufgrund der Aktualität des Themas wurden bereits in verschiedenen Forschungsarbeiten diverse Aspekte unterschiedlicher auf natürlicher Sprachverarbeitung basierender Modelle vergleichend beleuchtet.

So überprüfte Gugagayanan in seiner Masterarbeit die grundlegende Eignung von NLP-Chatbots zur korrekten Beantwortung mathematischer Aufgaben auf einem höheren akademischen Niveau [61]. Die Leistung des Chatbots wurde mit einer vorprogrammierten Mathematik-Bibliothek verglichen [61]. Durch den direkten Vergleich wurde eine höhere Genauigkeit in der Beantwortung der überprüften Themenbereiche durch den NLP-Chatbot festgestellt [61]. Dieses Forschungsergebnis deutet laut Gugagayanan auf eine vielversprechende Einsatzmöglichkeit von NLP-gestützten Systemen für die Lösung mathematischer Probleme hin [61].

In ihrem Konferenzpapier haben Dahal, Lamichhane, Luitel und Pant die Fähigkeiten und Grenzen von KI-Chatbots bei der Lösung mathematischer Probleme untersucht und gelangten zu der Feststellung, dass Chatbots wie ChatGPT eine hervorragende Eignung haben, leichtere bis anspruchsvollere mathematische Probleme zu lösen und

zu erklären [47]. Die Erklärungen von ChatGPT wurden als explizit und daher als geeignete Lernunterstützung eingeschätzt, wobei die Anpassung der Lerngeschwindigkeit und der Antwort an die Lernenden als Vorteil betont wurde [47]. Allerdings zeigte ChatGPT laut dem Tagungsbeitrag nur begrenzte Fähigkeiten bei der Lösung komplexer Aufgaben, sodass menschliche Interaktion weiterhin notwendig bleibt, insbesondere bei der Vermittlung tieferer mathematischer Konzepte und Fragestellungen [47]. Bei der Verwendung auf natürlicher Sprachverarbeitung basierender KI-Modelle, wie ChatGPT, wurden in der Konferenz die Möglichkeit einer Fehlinterpretation der Nutzereingaben und unzureichende Trainingsdaten als Kernprobleme benannt [47]. In dem Bericht der Konferenz wurde nur allgemein über ChatGPT berichtet [47], allerdings gibt es bei der konkreteren Betrachtung der Versionen Unterschiede in den Leistungen der Varianten dieser KI.

Im Rahmen ihrer Forschungsarbeit über die Lösungsfähigkeit von ChatGPT in Bezug auf mathematische Probleme gelangten Plevris, Papazafeiropoulos und Rios zu der Feststellung, dass die ChatGPT-Version 3.5 einfache mathematische Probleme meist korrekt löste [41]. Die dabei erzielte Antwortzeit der KI war im Vergleich zu dem Nachfolger deutlich kürzer [41]. Im Rahmen seiner Bachelorarbeit untersuchte Johnatan Paul die Qualität der generierten Antworten verschiedener Chatbots, darunter auch ChatGPT 3.5 und ChatGPT 4.0, auf mathematische Fragen [59]. Die in diesem Kontext durchgeführte Datenauswertung führte zu dem Ergebnis, dass ChatGPT 3.5 im Vergleich zu den anderen überprüften kostenfreien Chatbots die beste Konstanz, Formatierung und Korrektheit bei der Aufgabenbeantwortung aufwies, allerdings im Vergleich zu ChatGPT 4.0 schlechter abschnitt [59, S. 49]. Ana Stojanov stellte in ihrer autoethnografischen Studie zur Nutzung von ChatGPT 3.5 hingegen fest, dass die von ChatGPT 3.5 gegebenen Antworten inkonsistent, oberflächlich und manchmal falsch waren [57]. Sie räumte allerdings auch ein, eine Möglichkeit zur Steigerung der Motivation der Lernenden und zur Vermittlung eines grundlegenden Verständnisses zu verschiedenen Themengebieten zu erkennen [57]. Daher betonte Stojanov die Notwendigkeit, die von der KI gegebenen Antworten kritisch zu hinterfragen [57]. Sie empfahl eine Kombination aus ChatGPT 3.5 mit menschlichem Feedback und anderen Lernressourcen, um eine korrekte Wissensbasis zu gewährleisten [57].

Laut Plevris, Papazafeiropoulos und Rios übertraf ChatGPT 4.0 seine Vorgängerversion hinsichtlich der Genauigkeit deutlich [41]. Die Untersuchungen der Forschenden ergaben eine eindeutige Verbesserung von ChatGPT 4.0 im Vergleich zu der Version 3.5 in der Fähigkeit zur Lösung komplexerer mathematischer Aufgaben [41]. Auch in J. Pauls wissenschaftlicher Arbeit wurde ermittelt, dass ChatGPT 4.0 von den geprüften KIs die hochwertigsten Antworten für mathematische Probleme lieferte [59]. Er merkt jedoch in seiner Arbeit auch an, dass die von ChatGPT 4.0 erstellten Erklärungen nicht immer alle notwendigen Rechenschritte enthalten [59, S. 49].

Ernest Davis und Scott Aaronson kamen in ihrer Untersuchung über die Verbesserung von ChatGPT 4.0 mit den beiden Plug-ins Wolfram Alpha und Code Interpreter zu dem Schluss, dass die Verwendung von Plug-ins das Potenzial zur erheblichen Steigerung der Lösung mathematischer Probleme hat [50]. Jedoch wurde in dem Zusammenhang der Verwendung von Plug-ins auch ein häufig auftretendes Kommunikationsproblem über die Schnittstellen identifiziert, welches die Generierung von sinnvollen Antworten erschwerte [50].

Bei ihrer Konferenz gelangten Dahal, Lamichhane, Luitel und Pant zu der Feststellung, dass Wolfram Alpha als wissenschaftliche KI die besten Ergebnisse in Bezug auf die Korrektheit der generierten Aufgabenlösung erzielt [47]. Dieses Ergebnis entsteht durch die Ausgabe genauer und detaillierter Lösungen, wobei die Fähigkeit zur Generierung dieser darauf zurückgeführt wird, dass Wolfram Alpha speziell auf mathematische Berechnungen und Kalkulationen spezialisiert wurde [47]. Allerdings gestaltet sich die direkte Kommunikation mit Wolfram Alpha für die Benutzenden schwieriger als die Kommunikation über ChatGPT. Das liegt laut dem Konferenzpapier daran, dass Wolfram Alpha bei der Bearbeitung von Anfragen nicht so flexibel ist wie auf natürliche Sprachverarbeitung spezialisierte Chatbots [47].

Die Ergebnisse der Forschungsarbeiten eröffnen einen wertvollen Einblick in verschiedene auf natürlicher Sprachverarbeitung basierende KI-Systeme. Durch die Ermittlung der Stärken und Schwächen der einzelnen KIs wird ein umfassender Überblick über die Möglichkeiten moderner Chatbots gegeben. Aufbauend auf den Erkenntnissen dieser Forschungsarbeiten verfolgt diese Arbeit einen auf ChatGPT fokussierten Untersuchungsansatz. Die Wahl dieser Fokussierung auf die von OpenAI entwickelte KI ist durch die anbieterübergreifenden Vergleichsanalysen anderer Forschungsbeiträge entstanden, in denen ChatGPT als Untersuchungsgegenstand sich im Schnitt durch seine Leistungen von den anderen Systemen abhob. Das Hauptziel des Vergleichs zwischen den ChatGPT-Versionen 3.5 und 4.0, erweitert durch die Untersuchung potenzieller Verbesserungen, die durch Einsatz des Wolfram-Plug-ins erzielt werden können, ist es, deren Eignung als mathematische Lernunterstützung für Lernende zu bestimmen.

3 Methodik

Ziel der Forschungsfrage ist es zu ermitteln, inwieweit ChatGPT als eine kompetente Mathematik-Nachhilfe fungieren kann. Für die Beantwortung dieser Frage wird untersucht, wie qualitativ die Antworten von der KI auf mathematische Fragestellungen sind. Es soll dadurch bewertet werden, in welchem Ausmaß ChatGPT als digitaler Nachhilfellehrer eingesetzt werden kann, um Schülerinnen, Schüler und Studierende in dem Bereich der Mathematik Hilfestellungen zu geben.

Digitalisierung und der Einsatz von neuen Technologien spielen auch im Kontext des Bildungswesens eine immer größere Rolle. In diesem Zusammenhang nimmt auch die Bedeutung der Integration von künstlicher Intelligenz in der Lehre zu. Die Nutzung von KI hat das Potenzial, eine individuelle und zeitlich flexible Lernunterstützung bereitzustellen. Angesichts der zunehmenden Verbreitung von KI-Technologie ist es entscheidend, die Grenzen und Möglichkeiten dieser Werkzeuge zu untersuchen. Die Frage, ob eine gegebene Antwort korrekt ist und wie zuverlässig eine solche Antwort gegeben wird, soll dabei besonders berücksichtigt werden.

Aufgrund der häufigen Nutzung und der fortgeschrittenen Entwicklung in der natürlichsprachigen Kommunikation fokussiert sich die Untersuchung auf drei Versionen von ChatGPT. Diese Versionen sind ChatGPT 3.5, ChatGPT 4.0 und ChatGPT 4.0 mit Integration von Wolfram Alpha. Die Versionen weisen Verbesserungen zueinander auf. Während ChatGPT 3.5 die öffentlich frei verfügbare Version ist, enthält die nachfolgende Version ChatGPT 4.0 einige technische Optimierungen. Diese sollten es ermöglichen, die Aufgaben besser zu lösen und kompetenter zu beantworten. Durch die Integration von Wolfram Alpha wird die Möglichkeit zur Lösung von mathematischen Aufgaben noch einmal erweitert.

Die Bewertung der Antworten erfolgt durch eine Kombination aus quantitativer und qualitativer Analyse, um sowohl eine genaue Untersuchung der Antworten als auch die Meinung der potenziellen Zielgruppe zu erfassen. Diese Methodik soll aufzeigen, ob ChatGPT effektiv in pädagogische Prozesse integriert werden könnte, und liefert wertvolle Einblicke in die Möglichkeiten und Grenzen der einzelnen KIs bei der Lösung von Mathematikaufgaben.

3.1 Versuchsaufbau und Durchführung

Für die Datenerhebung wurden zwei unterschiedliche Methoden angewandt. Diese zielen darauf ab, die Ergebnisse gegenseitig zu stützen und den analytischen Blickwinkel nicht auf eine Sichtweise zu konzentrieren, sondern zu erweitern. Die Basis für die Anwendung dieser Methoden bildet die Auswahl mathematischer Aufgaben, die an die drei künstlichen Intelligenzen gestellt werden. Diese Aufgaben entstammen fünf verschiedenen Kategorien, welche auf den Anforderungen des Erwartungshorizontes für das Fach Mathematik zur allgemeinen Hochschulreife des Jahres 2024 sowie den Themenbereichen aus Bewerbungstests für Hochschulen basieren. Aus diesen Kriterien werden die Themenbereiche *Gleichungen*, *Stochastik*, *Integrale*, *analytische Geometrie und Logik* ausgewählt. Es sollte berücksichtigt werden, dass die ausgewählten Kategorien nicht alle erforderlichen Themenbereiche des Erwartungshorizontes abdecken, sondern sich thematisch und vom Anspruchsniveau her an den Inhalten orientieren, die Schülerinnen und Schüler sowie Studierende von der neunten Klasse bis zu den ersten Semestern im Studium behandeln.

Für jede der fünf Kategorien werden jeweils fünf Aufgaben unterschiedlicher Typen ausgewählt. Dieses Vorgehen gewährleistet, dass verschiedene Disziplinen des Themengebietes untersucht werden. Die Anforderungen der Schwierigkeit der Aufgaben variieren, um die Fähigkeiten der KIs in einem breiteren Spektrum zu testen. Bei der Auswahl der Aufgaben wird darauf geachtet, dass alle drei Versionen der KI-Modelle ChatGPT theoretisch in der Lage sind, die gestellten Fragen zu beantworten. Aus diesem Grund werden Aufgaben, welche spezifische Funktionen erfordern, wie das Plotten eines Graphen, bewusst ausgeschlossen, da ChatGPT 3.5 solche Anforderungen aufgrund seiner technischen Beschränkungen nicht erfüllen kann. Diese Vorgehensweise stellt sicher, dass keine KI im Vorhinein ausscheidet und somit die Testbedingungen für alle KI-Versionen gleich und die Ergebnisse direkt vergleichbar sind. Der Aufgabenpool wird aus verschiedenen Quellen zusammengestellt. So werden Abituraufgaben von verschiedenen Bundesländern Deutschlands, Bewerbungstests verschiedener internationaler Hochschulen und Universitäten sowie Schulbücher und die Lehrmaterial-Datenbank des Nachhilfeinstituts „Schülerhilfe“ als Quellen der Aufgaben gewählt. Einige Aufgaben, die im Internet öffentlich zugänglich sind, werden leicht oder stark modifiziert, um die Fähigkeit der KIs zu bewerten, auf veränderte Aufgaben zu reagieren. Zu jeder der Aufgaben existiert eine mitgelieferte Musterlösung, deren Korrektheit zusätzlich verifiziert wird. Der Umfang aller Aufgaben umfasst somit insgesamt 25 Fragen, die den KIs gestellt werden.

Zuerst wird eine inhaltliche Analyse zur Ermittlung der Qualität der Antworten durchgeführt. In diesem Kontext werden die zuvor gesammelten Aufgaben den drei zu testenden Versionen von ChatGPT gestellt. Bei der Arbeit mit ChatGPT spielt

die Formulierung der Frage eine relevante Rolle. Somit beeinflusst der Prompt die Qualität der Antwort maßgeblich. Das Ziel dieses Versuches sollte es jedoch primär sein, zu kontrollieren, wie ChatGPT sich als Tutor für Schülerinnen und Schüler sowie Studierende eignet. Angesichts der Annahme, dass die Nutzergruppe aufgrund begrenzter Erfahrung eventuell nicht mit den Anforderungen an einen detaillierten Prompt vertraut ist, wird ein vereinfachter Ansatz zur Prompt-Erstellung gewählt. Bei der Formulierung der Aufgaben wird darauf geachtet, dass der verwendete Prompt inhaltlich konsistent und einheitlich gestaltet ist. Die Aufgaben werden alle einheitlich in Deutsch verfasst, damit die Auswertung und Umfrage leichter durchgeführt werden können.

Erfahrungsgemäß zeigen Lernende mit Förderungsbedarf Schwierigkeiten bei der Formulierung von Verständnisproblemen und dem Stellen von Nachfragen. Aus diesem Grund wird bei der Stellung der Aufgaben auf die Formulierung der Nachfragen verzichtet. Hierbei stellt ein Aufgabentyp, bei dem zur Überprüfung einer möglichen Verbesserung simple Nachfragen formuliert werden, eine Ausnahme dar. Die Nachfragen werden bei dem Aufgabentyp der Integralrechnung gestellt, wenn die erhaltene Antwort offensichtliche Mängel wie einen fehlenden Rechenweg aufweist. Es wird bei der Stellung der Nachfragen darauf geachtet, dass die Nachfragen einfache und kurze Formulierungen und keine fachlichen Informationen, sondern lediglich die Anforderung eines Rechenwegs enthalten. Ausgenommen von diesen Nachfragen hat der Verzicht auf Nachfragen den Grund, dass die erhaltenen Antworten vergleichbar bleiben sollen.

Den drei Versionen der KI ChatGPT werden alle 25 Aufgaben gestellt. Zu jeder Aufgabe wird der vorformulierte Prompt hinzugefügt: „*Berechne und erkläre wie ein Lehrer.*“. Um die Reproduzierbarkeit der Antwort zu überprüfen, wird jede Frage insgesamt dreimal wiederholt. Somit ergibt sich eine Größe des Datensatzes von 225 Antworten von ChatGPT. Diese 225 Antworten werden im Hinblick auf fünf Kriterien analysiert. Die Kriterien werden vor der Untersuchung der Antworten bestimmt und die Definition wird im Vorfeld für die Bewertung festgehalten. Bei der Bewertung werden als Hauptkriterien *Korrektheit*, *Vollständigkeit*, *Sprache und Terminologie* sowie die *strukturelle Nachvollziehbarkeit* festgelegt. Abhängig von diesen Kriterien wird beobachtet, ob die Korrektheit der Antworten reproduzierbar ist und ob die KI ihre Grenzen erkennt und kommuniziert, wenn sie nicht in der Lage ist, die korrekte Antwort zu ermitteln.

Zur Einbindung relevanter Personengruppen in die Studie wird eine Umfrage durchgeführt. Diese ermöglicht es den Teilnehmenden, unter den fünf verschiedenen Aufgabentypen einen frei auszuwählen. Somit kann die Umfrage in insgesamt fünf Abwandlungen durchgeführt werden. Für den gewählten Aufgabentyp werden den Teilnehmenden drei Antworten, eine für jede zu überprüfende Version von ChatGPT, bereitgestellt, die es zu vergleichen gilt. Somit werden insgesamt 15 Antworten aus dem Fragepool in

der Umfrage behandelt. Die Wahl der Aufgaben, die für die Umfrage gewählt werden, erfolgt durch Zufall, damit durch gezielte Aufgabenwahl das Ergebnis der Umfrage nicht manipuliert werden kann. Die Beschränkung auf drei Antworten pro Aufgabentyp wird getroffen, damit die Aussagekraft der Umfrage erhalten bleibt. Durch die Bündelung der Antworten wird verhindert, dass die Ergebnisse zu stark streuen, und die Aussagen der Teilnehmenden bleiben somit vergleichbar.

Der Fragenkatalog der Umfrage ist ebenfalls zum Zweck der Vergleichbarkeit an den Kriterien der Auswertung angelehnt. Somit enthält er im ersten Teil eine Einschätzung dieser Kriterien. Im Rahmen des zweiten Teils des Fragebogens wird auf den Nutzen des Rechenwegs im Hinblick auf einen Lernerfolg eingegangen. Im letzten Teil der Erhebung wird die grundlegende Haltung der Befragten zur Anwendung von KI bestimmt. Die Fragebögen sind so konzipiert, dass die Befragten aus verschiedenen vorgegebenen Antwortmöglichkeiten eine Antwort auswählen können, wodurch die Beantwortung zeitsparend und unkompliziert wird. Zur Ermöglichung von präziseren Antworten und Anmerkungen werden zwei Freitextfelder hinzugefügt. Bei den Fragen wird auf eine neutrale Formulierung geachtet, um die Antwort nicht unterbewusst zu beeinflussen. Darüber hinaus werden die Fragen an ChatGPT ausschließlich auf Deutsch gestellt, sodass auch der gesamte Fragebogen nur auf Deutsch verfügbar ist.

Die Durchführung der Umfrage erfolgt für die Personengruppen der Studierenden und der Schülerinnen und Schüler getrennt. Dabei beantworten beide Personengruppen die gleichen Fragen.

Für die Studierenden wurde, um eine möglichst große Teilnahme zu erzielen, eine Online-Umfrage erstellt. Diese Umfrage wurde über den hochschulweiten Verteiler der Technischen Hochschule Köln an alle Studierenden und Mitarbeitenden der Fakultät für Informatik und Ingenieurwesen versendet. Um die Ehrlichkeit der Antworten zu fördern und die Bereitschaft zur Teilnahme an der Umfrage zu maximieren, war die Teilnahme an der Umfrage anonym. Bei der Umfrage wurden zeitgleich alle drei KIs bewertet, wobei die Reihenfolge der KI-Antworten bei allen Umfragen immer gleich blieb.

Parallel zu der Umfrage unter den Studierenden wurden Nachhilfe Nehmende der „Studienkreis“-Niederlassungen Seelscheid und Rösrath befragt. Die Teilnahme erfolgte durch Schülerinnen und Schüler ab der neunten Klasse, die im Fach Mathematik Förderunterricht erhielten. Die Teilnehmenden wiesen ein unterschiedliches Kursniveau auf und besuchten verschiedene Schulformen. Aufgrund des persönlichen Kontaktes zu dieser Personengruppe erfolgte die Befragung manuell auf Papierbögen, während des Unterrichts. Der Nachhilfeunterricht durfte durch die Umfrage nicht beeinträchtigt werden, weshalb die Umfrage innerhalb eines Zeitfensters von fünfzehn Minuten durchgeführt wurde. Aus diesem Grund konnten nicht in allen Fällen alle drei Antworten bearbeitet werden. In solchen Fällen wurde die Umfrage nicht für alle KIs durchgeführt

und in der folgenden Nachhilfestunde fortgesetzt. Die Fragebögen enthielten keine Angaben zu den Namen der Teilnehmenden, jedoch war die Umfrage nicht vollständig anonym, da die Personen bekannt waren. Um die Anonymität weitgehend zu wahren, wurden die Fragebögen in einen Kasten eingeworfen. Aufgrund der geringen Gruppengröße wäre eine Nachvollziehbarkeit jedoch möglich gewesen. Diese Tatsache konnte Auswirkungen auf die wahrheitsgemäße Beantwortung der Umfrage haben.

3.2 Bewertungskriterien

Um bei der Bewertung der Erklärungen ein einheitliches Vorgehen zu gewährleisten, werden die Kriterien und ihre möglichen Bewertungen im Vorhinein festgehalten. Diese Vorgehensweise soll eine konsistente und nachvollziehbare Einordnung der generierten Erklärungen gewährleisten. Zusätzlich sollen mögliche Verzerrungen bei der Bewertung der Kriterien minimiert werden und dadurch soll die Vergleichbarkeit erhalten bleiben.

Durch die Wahl der Bewertungskriterien sollen die wesentlichen Eigenschaften, die für eine verständliche Erklärung notwendig sind, abgedeckt werden. Dem Kriterium der Korrektheit kommt hierbei eine zentrale Bedeutung zu, da es die Grundlage für das Verständnis der mathematischen Konzepte und die darauf aufbauenden weiteren Überlegungen bildet. Um die Zuverlässigkeit bei der Generierung korrekter Antworten zu messen, wurde die Reproduzierbarkeit dieser Antworten untersucht. Dieses Kriterium ist wichtig, da sich die Lernenden auf die zuverlässige und wiederholbare Generierung von korrekten Aufgabenlösungen verlassen müssen. Ein vollständiger Lösungsweg ermöglicht es den Lernenden, die gesamte Problemlösung nachzuvollziehen. Das Verständnis, wie man von den Ausgangsbedingungen zur Lösung gelangt, ist essenziell, um die zugrunde liegenden Konzepte wirklich zu begreifen und nicht nur das Endergebnis zu kennen. Für eine für die Zielgruppe möglichst zugängliche und gut verständliche Erklärung ist nicht nur die Wahl einer angemessenen Sprache, sondern auch der korrekte Einsatz der Fachterminologie von Relevanz. Aus diesem Grund wurde das Kriterium der Sprache und Terminologie in die Auswertung mit einbezogen. Abschließend wurde die Strukturierung der dargestellten Antwort als Kriterium gebildet. Dieses Kriterium überprüft die Darstellungsform und Formatierung der Aufgabenlösung sowie die logische Reihenfolge der Lösungsschritte, durch die das Zurechtfinden in der dargestellten Antwort erleichtert wird.

Für die Veranschaulichung der abstrakten Definitionen der Bewertungskriterien wurden zu jedem dieser Kriterien im Abschnitt A des Anhangs jeweils eine positiv und eine negativ bewertete Antwort beigefügt. Die Darstellung dieser Beispiele soll zusätzlich das Spektrum der Antwortenbewertungen verdeutlichen. Sie zeigt ein klares Bild der

Erwartungen an die generierten Antworten, welche gemäß den im folgenden Teil näher beschriebenen Kriterien als gut oder schlecht bewertet wird.

3.2.1 Korrektheit

Das Kriterium der Korrektheit stellt sicher, dass die von der KI erstellte Antwort das gleiche Endergebnis wie die Musterlösung liefert. Zur Sicherstellung, dass die Musterlösung keine fehlerhaften Lösungen beinhaltet, wird diese zuvor überprüft. Grundsätzlich kann ein Ergebnis im mathematischen Sinn nur als richtig oder falsch gelten. Allerdings werden bei der Beurteilung dieses Kriteriums drei mögliche Bewertungen vorgesehen: *korrekt*, *teilweise korrekt* und *inkorrekt*. Bei einem vollkommen richtigen Endergebnis wird die Antwort als *korrekt* eingestuft. Wenn das Endergebnis nicht richtig ermittelt wurde, ist die Aufgabe *inkorrekt* gelöst worden. Als *teilweise korrekt* werden Aufgaben eingestuft, die eine der folgenden Bedingungen erfüllen: Aufgaben, die richtig gelöst wurden, allerdings bei der Zusammenfassung oder dem Antwortsatz mit einem anderen und falschen Ergebnis ausgegeben wurden; Aufgaben, bei denen beispielsweise durch die Nichtbeachtung des Definitionsbereichs eine falsche Lösung nicht ausgeschlossen wurde; Aufgaben, die aus mehreren Teilaufgaben bestehen, wobei die Mehrheit von diesen korrekt beantwortet wurde. Diese differenzierte Bewertung soll eine präzisere Analyse der Lösungen ermöglichen.

3.2.2 Vollständigkeit

Die Prüfung der Vollständigkeit des Lösungsweges soll beurteilen, ob die generierte Lösung alle notwendigen Schritte beinhaltet, um von der Fragestellung zur Antwort zu gelangen. Dieses Kriterium ist fundamental für das Verständnis des Rechenweges. Vor allem in dem Kontext, dass Schwierigkeiten bei der Formulierung von Nachfragen entstehen können, ist die Vollständigkeit der Lösung relevant. Zu der Bewertung der Aufgabe werden sechs Stufen gewählt: *vollständig*, *fehlendes (Zwischen-)Ergebnis*, *unvollständiger Rechenweg*, *unvollständige Erklärung*, *unvollständig*, *vollkommen unvollständig*. Eine Antwort ist *vollständig*, wenn sie alle für die Lösung notwendigen Rechenschritte klar und ausführlich genug erklärt, sodass die Lösung vollständig nachvollzogen werden kann. Außerdem werden die Rechnung und das Endergebnis angegeben. Ein *fehlendes (Zwischen-)Ergebnis* liegt vor, wenn das Endergebnis der Rechnung nicht angegeben wurde oder es zu einem Folgefehler in einer Lösung kam, die ohne diesen Fehler als vollständig bezeichnet werden könnte. Bei einem *unvollständigen Rechenweg* oder einer *unvollständigen Erklärung* wurden für das Verständnis der Aufgabe wichtige Schritte in der Rechnung oder der Erklärung ausgelassen. Eine Lösung, die sowohl bei dem Rechenweg als auch bei der Erklärung fehlende Schritte aufweist,

welche die Nachvollziehbarkeit des Lösungswegs nicht unmittelbar beeinträchtigen, wird als *unvollkommen* eingestuft. Eine *vollkommen unvollständige Lösung* liegt vor, wenn die Erklärung und Rechnung an für die Nachvollziehbarkeit der Lösung wichtigen Stellen fehlen.

3.2.3 Sprache und Terminologie

Dieses Kriterium soll überprüfen, ob die Sprache angemessen für ein leichtes Verständnis der Aufgabe ist. Es ist wichtig, dass Lernende keine kompliziert geschriebene Antwort erhalten. Dennoch müssen an den richtigen Stellen Fachbegriffe enthalten sein, die auch im richtigen Kontext verwendet werden. Allgemeine Formeln, wie sie in den Formelsammlungen zu finden sind, müssen bei der Anwendung korrekt sein. Für die Einstufung *ausgezeichnet* muss die Antwort exakt und fachlich angemessen, gleichzeitig jedoch klar und einfach genug sein, um das Verständnis der Lernenden zu fördern. Die Formulierungen sind präzise und vermeiden Verwirrung, indem sie komplizierte Konzepte zugänglich darstellen. *Gut* ist eine Antwort, wenn die Sprache überwiegend präzise und angemessen ist, jedoch gelegentliche Fachbegriffe oder Formulierungen enthalten, die klarer oder einfacher ausgedrückt werden könnten. Diese beeinträchtigen das allgemeine Verständnis nur minimal. Diese Bewertung wird auch vergeben, wenn eine ansonsten ausgezeichnete Erklärung in Englisch statt in Deutsch erfolgte. *Ausreichend* ist eine Antwort, wenn leichte terminologische Fehler auftreten oder die Erklärungen in ihrer Komplexität für die Schüler/-innen nicht vollständig auf eine angemessene Art vereinfacht sind. Diese Einstufung erfolgt zum Beispiel bei der Ausgabe von Code anstelle einer Rechnung, da nicht davon ausgegangen werden kann, dass Schülerinnen und Schüler den Code verstehen können. Die Bewertung *mangelhaft* wird vergeben, wenn sich Teile der Lösung in ihrer Aussage widersprechen oder gravierende terminologische Fehler wie eine fehlerhafte Formel auftreten. Auch können die Erklärungen übermäßig komplex und unverständlich sein, was den Lernenden das Verständnis erheblich erschwert.

3.2.4 Strukturierung

Bei der Strukturierung wird der allgemeine Aufbau der Lösung betrachtet. Es ist essenziell, dass die Rechenschritte des Lösungswegs in einer logischen Reihenfolge bearbeitet werden und nicht zwischen Themen gesprungen wird. Zudem sollte die Antwort prägnant und auf den Punkt gebracht sein und unnötige Wiederholungen sollten vermieden werden. Für die einzelnen Abschnitte sollen passende Überschriften vergeben werden, um die Erklärung in sinnvolle Unterabschnitte zu unterteilen und übersichtlich zu gestalten. Für dieses Kriterium werden vier Bewertungsstufen definiert:

ausgezeichnet strukturiert, gut strukturiert, teilweise strukturiert, unstrukturiert. Eine *ausgezeichnete Struktur* erfordert eine sinnvolle Reihenfolge einer Lösung, dabei sind die Überschriften der Absätze mit zu dem Inhalt passenden Überschriften versehen. Diese Strukturierung zielt darauf ab, den Gedankenfluss zu unterstützen und eine klare Hierarchie der Inhalte aufzubauen. Eine *gute Strukturierung* zeichnet sich dadurch aus, dass die Lösung sowohl logisch als auch zusammenhängend präsentiert wird, dahingehend jedoch kleinere Mängel aufweist. Eine *ausreichende Strukturierung* enthält für die Aufgabe unnötige und ausschweifende Erklärungen, wobei auch Wiederholungen und Sprünge in der Argumentation auftreten können. Auf der Stufe *unstrukturiert* ist die Antwort unnötig lang oder weist häufige Sprünge zwischen den verschiedenen Teilen der Aufgabe auf, was zu einer unordentlichen Struktur führt.

3.2.5 Reproduzierbarkeit

Um die Reproduzierbarkeit der korrekten Antworten auf bestimmte Fragen zu überprüfen, wird jede Frage in der Studie jeweils dreimal unter identischen Bedingungen wiederholt. Diese Wiederholung ermöglicht die Bewertung der Konsistenz der Ergebnisse und gewährleistet deren Vergleichbarkeit. Durch diese Vorgehensweise kann erfasst werden, in welchem Maße eine korrekte Antwort reproduziert werden kann. Dadurch soll auch vermieden werden, dass eine Aufgabe nur durch Zufall korrekt ist oder falsche Ergebnisse liefert. Insgesamt ist diese Überprüfung wichtig, um die Verlässlichkeit einer Antwort zu erhöhen.

4 Forschungsergebnisse

Nach der ausführlichen Erläuterung der angewandten Datenerhebungsmethoden widmet sich dieser Abschnitt der detaillierten Darstellung der daraus resultierenden Ergebnisse. Die Daten wurden aufbereitet und in einer anschaulichen Form von Diagrammen dargestellt, um ein deutliches Bild der KI-gestützten Beantwortung mathematischer Fragen zu vermitteln. Da zwei unterschiedliche Forschungsmethoden angewandt wurden, um die Qualität der Antworten zu überprüfen, werden die Ergebnisse der beiden Methoden separiert betrachtet. Anschließend erfolgt die Einordnung der Ergebnisse im Kontext der Forschungsfrage. Die mit den beiden Methoden erhobenen Daten können vollständig über das im Anhang C bereitgestellte GitHub-Repository eingesehen werden.

4.1 Ergebnisse der Datenauswertung

Die Datenauswertung der gesammelten Aufgabenerklärungen basiert auf den vorab definierten Kriterien, um eine möglichst konsistente und objektive Bewertungsgrundlage zu schaffen. Jede Antwort wurde sorgfältig gelesen, nachvollzogen und analysiert. Der beschriebene Rechenweg wurde überprüft, um potenzielle Fehlerquellen zu identifizieren und die Antwort im Gesamtkontext der Untersuchung einordnen zu können. Die Auswertung wurde ausschließlich von einer Person über einen Zeitraum von mehreren Tagen durchgeführt. Diese Entscheidung ermöglicht eine konsistente Anwendung der Bewertungskriterien, da sie ausschließlich auf einer einzelnen Perspektive basiert. Gleichzeitig muss jedoch berücksichtigt werden, dass dieser Ansatz die Möglichkeit einer unbewussten Voreingenommenheit beinhaltet, was als potenzielle Einschränkung der Studie anzusehen ist.

Zur besseren Übersichtlichkeit wurden die Daten nach der Bewertung in Kreisdiagrammen dargestellt. Dies ermöglicht eine klare und direkte Vergleichbarkeit der Anteile und erleichtert die Erfassung der Datenverteilung. Für jedes im Methodikteil beschriebene Kriterium wurden vier solcher Kreisdiagramme erstellt, um die relative Gewichtung des Bewertungskriteriums im Gesamtkontext aller drei KIs und im Kontext jeder KI einzeln abzubilden. Die Grundlage für die Diagramme bilden alle 75 von der jeweiligen KI beantworteten Fragen, während für den Gesamtkontext alle 225 Fragen betrachtet wurden.

4.1.1 Korrektheit

Die Auswertung der insgesamt 225 Fragen hat ergeben, dass bei der Betrachtung aller Antworten der drei untersuchten KI-Versionen 53% der Antworten vollständig *korrekt* beantwortet wurden. Allerdings wurden auch 32% mit einer fehlerhaften Antwort beantwortet. Diese Durchschnittswerte bilden die Grundlage für den Vergleich der Leistungen der verschiedenen KIs und zeigen deutliche Leistungsunterschiede. Dabei wird deutlich, dass ChatGPT 3.5 mit deutlichem Abstand die schlechtesten Ergebnisse in Bezug auf die Korrektheit der Antworten geliefert hat.

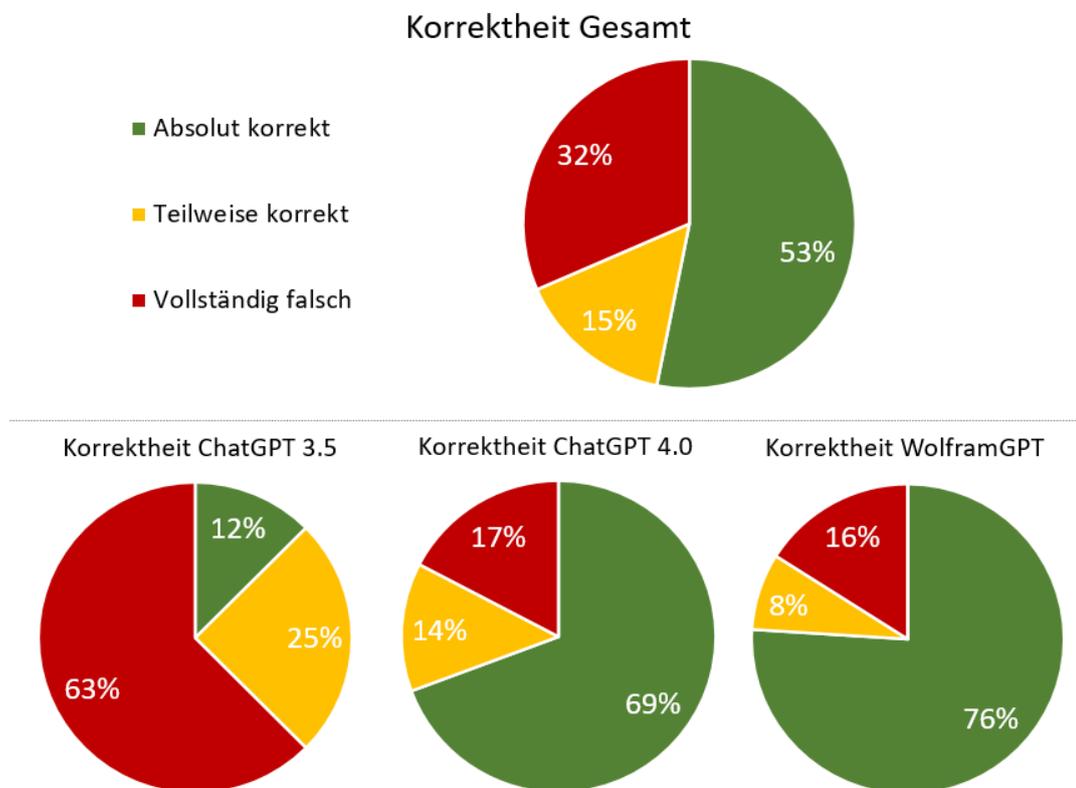


Abbildung 4.1: Kreisdiagramme Korrektheit

So lag bei ChatGPT 3.5 die Wahrscheinlichkeit einer vollkommen *korrekten* Antwort auf die gestellte Mathematikfrage lediglich bei 12%. Zudem wurde bei 60% der von dieser KI beantworteten Fragen eine *inkorrekte* Antwort generiert. Fast ein Viertel der Antworten wurde zumindest *teilweise korrekt* beantwortet. Bei drei Fragen wurde

angegeben, dass kein Ergebnis erzeugt werden kann, wodurch es zu keiner Lösung der Aufgabe kam.

Bei Verwendung der ChatGPT-Version 4.0 wurde im Vergleich zu der Vorgängerversion eine deutliche Verbesserung der Ergebnisse in Bezug auf die Korrektheit erzielt. So wurden 69 % der Fragen *korrekt* beantwortet und weitere 14 % lieferten zumindest *teilweise korrekte* Ergebnisse. Allerdings erfolgte bei 17 % der Aufgaben die Ausgabe eines falschen Ergebnisses.

Die Verwendung der Integration von Wolfram Alpha in ChatGPT 4.0 erzielte weitere leichte Verbesserungen. Obwohl sich die Anzahl der *inkorrekten* Antworten mit 16 % nur minimal um einen Prozentpunkt verringert hat, wurde bei den *teilweise korrekten* Antworten eine Verringerung um sechs Prozentpunkte erzielt. Somit sanken die *teilweise korrekt* beantworteten Aufgaben auf 8 % und die vollkommen *korrekten* Antworten stiegen auf 76 %. Dadurch erzielte ChatGPT mit der Anbindung von Wolfram Alpha die besten Ergebnisse hinsichtlich der korrekten Beantwortung der Aufgaben.

Grundsätzlich lässt sich bei dem Vergleich von ChatGPT 3.5 mit der Version 4.0 in Bezug auf die Korrektheit der Aufgabenlösung eine deutliche Steigerung erkennen. Mit der Integration von Wolfram Alpha konnte bei allen Aufgabentypen eine weitere leichte Steigerung der Korrektheit festgestellt werden. Die Ergebnisse wurden präziser, was sich in einem Rückgang der *teilweise korrekten* Lösungen bei einer Steigerung der *korrekten* Antworten zeigt. Die Anzahl der falschen Ergebnisse bleibt jedoch nahezu unverändert. Dabei konnte beobachtet werden, dass Aufgaben, die von ChatGPT 4.0 falsch beantwortet wurden, oft auch mit der Integration von Wolfram Alpha fehlerhaft waren.

Im Bereich der Logikaufgaben wurden von allen drei KIs die schlechtesten Leistungen erbracht. Aber auch bei den leicht modifizierten Aufgaben der einzelnen Aufgabentypen traten vermehrt Fehler auf. Bei dieser Modifizierung wurden leichte Änderungen an öffentlich zugänglichen Aufgaben vorgenommen, die allerdings zu einem anderen Ergebnis führten. Alle KIs zeigten Schwierigkeiten bei der Beantwortung dieser manipulierten Aufgaben und es wurde häufig das von der ursprünglichen Aufgabe geforderte Ergebnis ausgegeben.

Beispiel: Logikaufgabe

So wurden beispielsweise bei einer Logikaufgabe, welche ursprünglich aus einer Aufnahmeprüfung der Universität Oxford stammt, die Farben zweier Figuren geändert. Die Farbe eines Dreiecks wurde mit der Farbe eines Kreises vertauscht. Diese geringe Änderung führte dazu, dass die Form der Ergebnisfigur auch vertauscht wurde. Dennoch gab jede Version von ChatGPT das ursprünglich korrekte Ergebnis mindestens

einmal aus. Das korrekte Ergebnis der veränderten Aufgabe wurde nur einmal genannt, wobei die KI, welche das korrekte Ergebnis ausgab, ChatGPT 4.0 mit Wolfram Alpha war.

Katrin und Klaus haben die Chance, ein Geschenk von Kim zu bekommen. Sie müssen dafür nur herausfinden, unter welcher der farbigen Figuren Kim das Geschenk versteckt hat. Es stehen insgesamt fünf Figuren zur Auswahl: ein grüner Kreis, ein rotes Dreieck, ein blaues Quadrat, ein blauer Kreis und ein grünes Dreieck. Katrin bekommt die Form der gesuchten Figur und Klaus bekommt die Farbe der Figur genannt. Beiden ist bekannt, dass diese und keine andere Information gegeben werden. Kim: „Weiß jemand von euch, wo das Geschenk ist?“ Beide verneinen. Kim: „Wisst ihr nun, wo sich das Geschenk verbirgt?“ Beide schütteln den Kopf. Kim: „Und jetzt?“ Beide nicken. Wo ist das Geschenk versteckt?

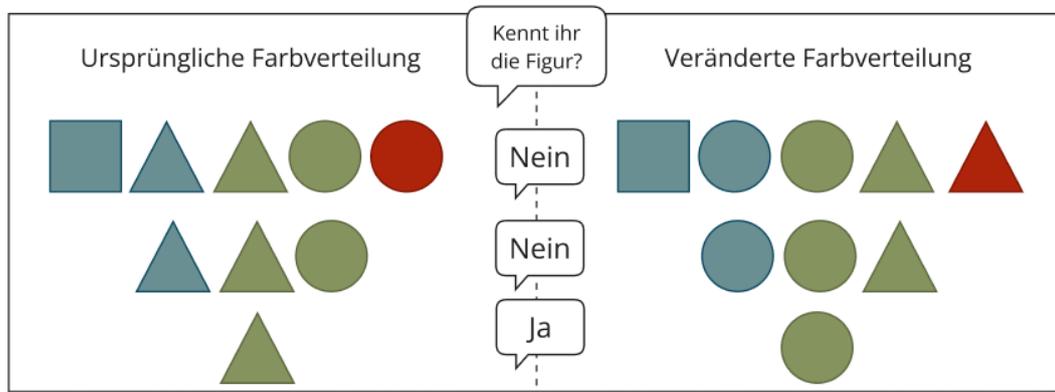


Abbildung 4.2: Originales und manipuliertes Logikrätsel im Vergleich

Beispiel: Manipulierte Integrationsaufgabe

Unabhängig von der eigentlichen Studie wurde ein weiteres Experiment durchgeführt. Dieses diente zur Überprüfung der Reaktion der Systeme für vorgegebene, jedoch durch Manipulation der Ausgangsbedingungen inkorrekte Ergebnisse. Durch die Veränderung der Formel dieser spezifischen Aufgabe wurde das zu überprüfende Musterergebnis *inkorrekt*. Die Abänderung der Formel $v_A(t) = 0,25t + 10 \cdot (1 - e^{-t})$ zu $v_A(t) = 0,25t + 10 \cdot (1 * e^{-t})$ führte zu folgender Aufgabenstellung:

Bei einem 100-Meter-Lauf treten Läufer A und B gegeneinander an. Die Geschwindigkeitsfunktionen der Läufer lauten wie folgt:

- Läufer A sprintet mit der Geschwindigkeitsfunktion $v_A(t) = 0,25t + 10 \cdot (1 * e^{-t})$.
[...]

Hierbei ist t die Zeit in Sekunden ab dem Start des Laufes und $v(t)$ die Geschwindigkeit der Läufer in Meter pro Sekunde. Zeige, dass Läufer A ungefähr 9,8 Sekunden benötigt, um die 100 Meter zu absolvieren.

Die Ergebnisse demonstrieren ein unterschiedliches Verhalten der einzelnen KIs bei dieser Aufgabe. ChatGPT 3.5 gab in allen Versuchen als Antwort, dass Läufer A etwa 9,8 Sekunden für die 100 Meter benötigen würde. Auch bei einer deutlichen Abweichung des berechneten Ergebnisses in einem Durchlauf zu dieser Annahme wurde ausgegeben: „[...] Läufer A benötigt also ungefähr 40,74 Sekunden, was nahe an 9,8 Sekunden liegt. [...]“. Diese Bemerkung zeigt eine auffällige Diskrepanz zwischen der tatsächlichen Berechnung und der interpretierten Antwort durch ChatGPT 3.5.

Auch ChatGPT 4.0 gab in einer der drei Wiederholungen die in der Aufgabenstellung getroffene Annahme als korrektes Ergebnis aus. Bei einem Versuch wurde jedoch auch angegeben, dass diese Aufgabe nur mittels eines numerischen Verfahrens gelöst werden könnte. In diesem Fall wurde keine Antwort gegeben, sondern nur ein Rechenansatz geliefert. Im verbleibenden Durchlauf wurde die korrekte Lösung ermittelt und festgestellt, dass die berechnete Lösung im Widerspruch zu der erwarteten Lösung steht.

Bei der Durchführung dieses Versuches mit der Integration von Wolfram Alpha in ChatGPT 4.0 wurde ausnahmslos bei allen drei Durchläufen erkannt, dass die angenommene Lösung nicht korrekt sein kann. Die tatsächliche Lösung wurde berechnet und mit der Annahme abgeglichen. Bei der Abgleichung kam es in den Versuchen aufgrund der Abweichung der Ergebnisse zu einer wiederholten Kontrolle des Ergebnisses. Dabei wurde in einem Versuch die eigene Rechnung aufgrund dieser Abweichung von der KI zehnmal überprüft.

Die durchgeführten Versuche konnten die Unterschiede in der Korrektheit und in dem Verhalten bei abgeänderten oder fehlerhaften Aufgabenstellungen zeigen. Dadurch wird nicht nur die Zuverlässigkeit zur korrekten Beantwortung einer mathematischen Aufgabe gezeigt, sondern auch die Toleranz bei fehlerhaften Eingaben. Für die Einsatzfähigkeit im akademischen Kontext ist eine mathematische Genauigkeit von zentraler Bedeutung.

Reproduzierbarkeit

Um die Reproduzierbarkeit zu testen, wurde jede Aufgabe dreimal wiederholt. Dadurch war eine Ermittlung der Anzahl *korrekt* ausgegebener Antworten möglich. Durch die

dreifache Wiederholung der insgesamt 25 Fragen ergibt sich eine erste grundlegende Einschätzung der Konsistenz der Antworten unter identischen Testbedingungen. Die Versuchsgröße kann jedoch möglicherweise zum Ziehen allgemeingültiger Schlüsse nicht ausreichend gewählt worden sein. Die Interpretation der Ergebnisse bezieht sich somit auf den Kontext dieses Versuches und ist als Verallgemeinerung mit Einschränkungen und Vorsicht zu betrachten. Alle Fragen wurden hinsichtlich des Ergebnisses untersucht und einer der drei Korrektheitsstufen zugeordnet. In der dokumentierten Tabelle steht Rot/0 für *inkorrekt*, Gelb/0.5 für *teilweise korrekt* und Grün/1 für *korrekt* beantwortet.

ChatGPT 3.5					
	Aufgabe 1	Aufgabe 2	Aufgabe 3	Aufgabe 4	Aufgabe 5
Integralrechnung	0	0	0	0	0
	1	0	0	0	1
	0.5	0	0	0	0
Gleichungen	0.5	0	0.5	1	0
	0	0	0.5	0.5	0
	0	0	0	0.5	0
Stochastik	0	0.5	0.5	0.5	0
	0.5	0	0.5	0.5	0
	0.5	0	0	0	1
analy. Geometrie	0	0.5	1	0	0.5
	0	0.5	1	0	0.5
	0	0.5	0	1	1
Logik	0	0	0	1	0
	0	0	0	0	0
	0	0	0	0	0

Tabelle 4.1: Korrekte Antworten aller Fragen von ChatGPT 3.5

Bei ChatGPT 3.5 wurde bei keiner der drei Wiederholungen jedes Mal eine *korrekte* Antwort geliefert. Bei einer Aufgabe wurden zumindest zwei der drei Wiederholungen korrekt beantwortet. Bei sieben Aufgaben wurde zumindest eine *korrekte* Antwort gegeben und bei 17 der Aufgaben war keine der Wiederholungen vollkommen *korrekt*. Die dargestellten Ergebnisse zeigen in Bezug auf die Reproduzierbarkeit *korrekter* Antworten bei mathematischen Aufgabenstellungen durchgehend schlechte Ergebnisse. Eine *korrekte* Antwort kann nur selten erneut generiert werden, was zur Inkonsistenz der gelieferten Antworten führt und zeigt, dass eine zuverlässige Verwendung in diesem Kontext in einer Form nicht möglich ist.

ChatGPT 4.0 liefert im direkten Vergleich zu ChatGPT 3.5 deutlich bessere Ergeb-

nisse bei der Beantwortung mathematischer Aufgabenstellungen. Insgesamt 13 der 25 gestellten Aufgaben wurden bei allen drei Wiederholungen *korrekt* beantwortet. Bei vier Aufgaben wurde zweimal und bei fünf Aufgaben einmal die korrekte Lösung genannt. Allerdings konnte auch bei drei Aufgaben in keinem der Versuche überhaupt eine korrekte Lösung gefunden werden. Diese stark verbesserte Reproduzierbarkeit der Lösungen weist auf eine höhere Genauigkeit und gute Zuverlässigkeit dieser KI im Kontext der Aufgaben hin. Probleme bei der zuverlässigen Beantwortung der Fragen traten vor allem bei dem Aufgabentyp Logik bei den beiden veränderten Aufgaben auf. Auch im Bereich der Gleichungen kam es bei zwei Aufgaben zu Problemen. Dennoch wird durch die Ergebnisse des Versuches deutlich, dass *korrekte* Antworten von ChatGPT 4.0 grundsätzlich gut reproduzierbar sind. Fehlerhafte Ausgaben traten punktuell bei einzelnen Aufgaben auf und sind auf einen fehlerhaften Rechenansatz zurückzuführen, der meist auch bei der Reproduktion beibehalten wurde. Die Ergebnisse der Untersuchung werden in der folgenden Tabelle dargestellt.

ChatGPT 4.0					
	Aufgabe 1	Aufgabe 2	Aufgabe 3	Aufgabe 4	Aufgabe 5
Integralrechnung	1	1	1	1	1
	1	0.5	0	1	1
	1	1	1	1	1
Gleichungen	1	0	0	1	1
	1	0	0	1	1
	0.5	1	0	1	1
Stochastik	1	1	0.5	0.5	1
	1	1	1	0.5	1
	1	1	0.5	0.5	1
analy. Geometrie	0	1	1	1	0.5
	1	0.5	1	1	1
	0	1	1	1	0.5
Logik	0	1	1	1	1
	0	1	1	1	0
	0	1	1	1	0

Tabelle 4.2: Korrekte Antworten aller Fragen von ChatGPT 4.0

Die in der folgenden Tabelle 4.3 dargestellten Ergebnisse zeigen die Korrektheit der Antworten von ChatGPT 4.0 mit der Integration von Wolfram Alpha. Die Verwendung von Wolfram Alpha hat nur kleine Auswirkungen auf die Reproduzierbarkeit der Lösung im Vergleich zu ChatGPT 4.0. Es konnte 12-mal die *korrekte* Antwort in allen drei Versuchen reproduziert werden, was eine kleine Verschlechterung zu den mit ChatGPT 4.0 erzielten Ergebnissen darstellt. Allerdings wurde mit der Ausgabe von zwei *korrekten* Ergebnissen in acht Fällen eine Verdopplung der von ChatGPT 4.0 erzielten Antworten

erreicht. Fünf der Aufgaben hatten bei der dreimaligen Wiederholung zumindest eine richtige Antwort. Der Fall, dass keine korrekte Antwort gefunden werden konnte, trat in dem Versuch nicht auf. Dies zeigt, dass es ChatGPT 4.0 mit Wolfram Alpha als einziger KI möglich war, die von den anderen beiden Versionen nicht lösbaren Aufgaben zumindest einmal *korrekt* zu beantworten.

ChatGPT 4.0 mit Wolfram Alpha					
	Aufgabe 1	Aufgabe 2	Aufgabe 3	Aufgabe 4	Aufgabe 5
Integralrechnung	1	1	1	1	1
	1	1	1	1	1
	1	0.5	0	1	1
Gleichungen	1	1	1	1	1
	1	0	0	1	1
	0.5	0	0	1	1
Stochastik	1	1	1	1	1
	1	0.5	1	0.5	1
	1	1	1	1	1
analy. Geometrie	0.5	0.5	1	1	1
	1	1	1	1	1
	0	1	1	0	1
Logik	1	1	1	1	0
	0	1	1	1	1
	0	1	0	1	0

Tabelle 4.3: Korrekte Antworten aller Fragen von ChatGPT 4.0 mit Wolfram Alpha

4.1.2 Vollständigkeit

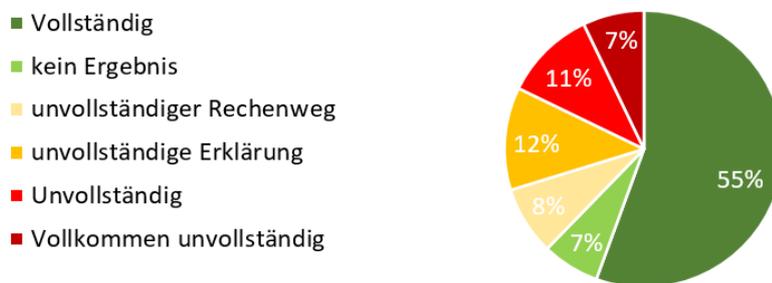
Die Auswertung der Fragen ergab in Bezug auf die Vollständigkeit der Antworten, dass alle drei ChatGPT-Versionen zusammen durchschnittlich 55% der Antworten *vollständig* erklärt haben. Kleine Mängel, wie *fehlende (Zwischen-)Ergebnisse*, traten bei durchschnittlich 7% der Aufgaben auf. Ein *unvollständiger Rechenweg* konnte in 8% und eine *unvollständige Erklärung* in 12% der Fälle festgestellt werden. 11% der Antworten wurden als *unvollständig* eingestuft, da sie größere Mängel aufwiesen, und 7% der Antworten waren für die Fragestellung *vollkommen unvollständig*.

ChatGPT 3.5 erzielte bezüglich der Vollständigkeit der Antworten die schlechtesten Ergebnisse. Mit 44% *vollständigen* Erklärungen und 6% Antworten mit *fehlenden Ergebnissen* ist die Hälfte der Antworten solide beantwortet worden. Allerdings lag diese KI mit 17% *vollkommen unvollständigen* Antworten unter dem Durchschnitt aller KIs. 5% der Antworten wurden mit *unvollständigem Rechenweg* und 17% mit

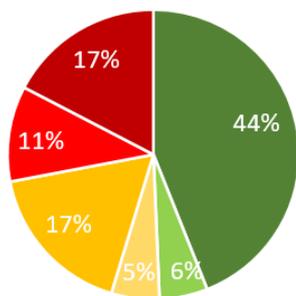
unvollständigen Erklärungen gegeben. Die Ergebnisse der prozentualen Verteilung der eingeordneten Stufen zeigen eindeutig, dass diese Version in Bezug auf die Vollständigkeit der Antworten die schwächste Leistung erbrachte. Dies wird besonders deutlich bei der Betrachtung des Rückgangs bei den guten und der Steigerung der schlechten Ergebnisse.

Die nachfolgende Version 4.0 erreicht in diesem Vergleich bessere Ergebnisse, da bei dieser Version schon 62 % der Antworten mit einer *vollständigen* Erklärung generiert wurden. *Fehlende (Zwischen-)Ergebnisse* und *unvollständige Rechenwege* traten jeweils in 8 % der Antworten auf, während 7 % der Fragen mit einer *unvollständigen Erklärung*, 12 % *unvollständig* und 3 % *vollkommen unvollständig* waren.

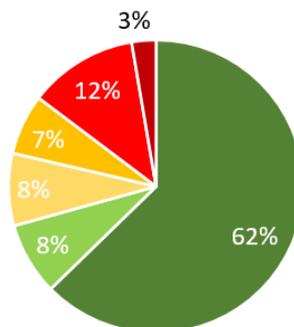
Vollständigkeit Gesamt



Vollständigkeit ChatGPT 3.5



Vollständigkeit ChatGPT 4.0



Vollständigkeit WolframGPT

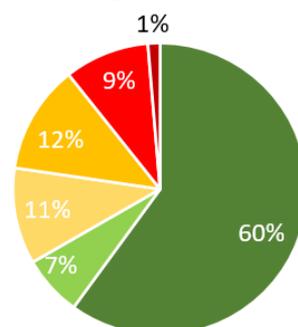


Abbildung 4.3: Kreisdiagramme Vollständigkeit

Mit der Integration von Wolfram Alpha haben sich die beiden schlechtesten Kategorien zusammen um fünf Prozentpunkte reduziert. So wurden nur 1 % der Antworten als vollkommen und 9 % als *unvollständig* eingestuft. Allerdings traten mit 12 % eine

unvollständige Erklärung und mit 11 % ein *unvollständiger Rechenweg* häufiger auf. 7 % der Erklärungen hatten *fehlende (Zwischen-)Ergebnisse*, während 60 % der Antworten alle relevanten Rechenschritte und eine vollständige Erklärung enthielten.

Die Resultate des Versuchs ergaben somit, dass die Wahrscheinlichkeit für eine *vollständige* Antwort bei ChatGPT 4.0 am größten war. Allerdings lieferte diese Version auch mehr *unvollständige* Antworten im Vergleich zur Kombination mit Wolfram Alpha. Durch die Möglichkeit, Nachfragen zu stellen und konkrete Erklärungen zwischen Schritten anzufordern, müssen diese Ergebnisse allerdings relativiert betrachtet werden.

4.1.3 Sprache und Terminologie

Bei der Überprüfung der Sprache und Terminologie wurden von allen KIs grundsätzlich gute Ergebnisse erzielt. Die verwendete Sprache der KI-Systeme war überwiegend angemessen, und terminologische Fehler traten selten auf. Dennoch sind während des Versuches bestimmte Verhaltensmuster aufgefallen, die aus terminologischer Sicht oder im Hinblick auf didaktische Erklärungen für Schülerinnen, Schüler und Studierende problematisch sein könnten.

Im Durchschnitt haben alle drei KIs bei 57 % der Antworten keine sprachlichen oder terminologischen Mängel gezeigt. Weitere 21 % wiesen lediglich geringfügige Mängel auf und wurden daher als *gut* eingestuft. Insgesamt erhielten 10 % die Stufe *ausreichend* und 12 % wurden der Stufe *mangelhaft* zugeordnet.

Bei der Beantwortung mathematischer Aufgaben erreichte ChatGPT 3.5 in 36 % der Fälle ein für Schülerinnen, Schüler und Studierende nicht bemängelbares Niveau. Weitere 24 % der Antworten enthielten nur kleine Fehler und wurden deshalb als *gut* eingestuft. Die Bewertung *ausreichend* wurde für 12 % der Antworten vergeben, während 28 % als *mangelhaft* klassifiziert wurden, da sie erhebliche terminologische oder sprachliche Fehler aufwiesen. Insgesamt lieferte ChatGPT 3.5 die unzureichendsten Ergebnisse, was möglicherweise auf die Tendenz der KI zurückgeführt werden kann, falsche Ergebnisse und falsches Vorgehen auf fehlerhafte Erklärungen und Berechnungsmethoden oder mathematische Annahmen zu stützen.

Ein Beispiel hierfür ist ein Lösungsansatz, der das systematische Ausprobieren für die Gleichung $c \cdot (c + p) = h + 120$ vorsieht, wobei gilt: $c, p, h \in \mathbb{P}$. Der Lösungsansatz des Ausprobierens wird mit folgendem Satz begründet. [...] *Da es nicht viele Primzahlen unter 120 gibt, können wir einfach durchprobieren.* [...] Weil es nur 30 Primzahlen unter 120 gibt, kann diese Aussage je nach Sichtweise als praktikabel erscheinen. Allerdings existiert schon bei dieser kleinen Zahl eine nicht unerhebliche Anzahl von $\binom{30}{3} = \frac{30!}{3!(30-3)!} = 4060$ Kombinationsmöglichkeiten. Für eine KI stellt diese

Anzahl an möglichen Kombinationen kein Hindernis dar, für einen Menschen ist diese Anzahl an Überprüfungen aber nicht gering und mit einem für die Aufgabe unangemessenen Aufwand verbunden. Gravierendere Fehler traten auf, wenn Formeln oder Schlussfolgerungen so angepasst wurden, dass sie zu dem erzielten Ergebnis passten. Dies resultierte in unlogischen und widersprüchlichen Argumentationen, wie beispielsweise in einer Antwort auf das im Kapitel 4.1.1 vorgestellte Logikrätsel. Dort wurde die Erklärung so gestaltet, dass sie zur ursprünglichen Antwort passte, was jedoch eine inkonsistente und widersprüchliche Argumentation zur Folge hatte: [...] *Da Katrin weiß, dass das Geschenk nicht unter dem grünen Kreis ist, muss es unter dem grünen Dreieck sein, da dies die einzige Figur ist, die eine grüne Form hat. Klaus weiß auch, dass das Geschenk nicht grün ist, daher kann er schließen, dass es unter dem blauen Kreis sein muss. Also ist das Geschenk unter dem grünen Dreieck versteckt.*

ChatGPT 4.0 erzielte in diesem Bereich die besten Ergebnisse, da 73 % der Antworten eine *ausgezeichnete* Sprache und Terminologie für Lernende aufwiesen. Trotz kleiner Ungenauigkeiten erzielten 17 % *gute* Ergebnisse. 7 % der Antworten wurden *ausreichend* bewertet und nur 3 % wurden als *mangelhaft* eingeordnet. Das gelegentliche Generieren von Antworten auf Englisch war bei allen drei Versionen ein bemerkenswertes Phänomen. Da ChatGPT 4.0 dieses Verhalten allerdings nur unwesentlich häufiger zeigte als die anderen Versionen, kann dieses Phänomen nicht spezifisch einer bestimmten Version zugeschrieben werden, da es bei allen dreien auftrat.

Die Kombination von ChatGPT 4.0 mit Wolfram Alpha erzielte geringfügig schwächere Ergebnisse als ChatGPT 4.0. So wurden nur 62 % der Antworten als *ausgezeichnet* und 21 % als *gut* eingestuft. Weitere 12 % waren *ausreichend* und 5 % zeigten erhebliche Defizite. Diese Ergebnisse können auf die spezifische Eigenheit von ChatGPT 4.0 mit Wolfram Alpha zurückgeführt werden, die Ausgabe der Ergebnisse vorzugsweise in Wolfram Alpha Language Code darzustellen.

Berechnungen mit Wolfram Language

```
wolfram Code kopieren
(* a) Wahrscheinlichkeit, dass alle ankommenden Gäste untergebracht werden können *)
probUntergebracht = CDF[BinomialDistribution[200, 0.9], 180]

(* b) Wahrscheinlichkeit, dass zwei oder mehr Gäste nicht untergebracht werden können *)
probNichtUntergebracht = 1 - probUntergebracht

(* c) Erwartungswert und Standardabweichung der Anzahl der zu entschädigenden Gäste *)
erwartungswertX = 200 * 0.9
standardabweichungX = Sqrt[200 * 0.9 * 0.1]

(* d) Erwarteter Gewinn bei 200 Buchungen *)
gewinnProZimmer = 50
verlustProNichtVerfuegbaresZimmer = 50
expectedGewinn = 50 * expectedBelegteZimmer - 50 * expectedEntschaedigteGaeste
```

Abbildung 4.4: Wolfram-Language-Antwort

Dieser Code enthält meist den Rechenweg, kann jedoch von Benutzenden ohne entsprechende Vorkenntnisse schwer verstanden werden. Der Einsatz von Code als Erklärungsmittel scheint angesichts der Erwartung, dass Erklärungen auf eine Weise erfolgen sollten, die der einer Lehrerin oder eines Lehrers ähnelt, denkbar ungeeignet.

Wie auch bei den vorherigen Kriterien sind die Ergebnisse dieses Versuches im Kontext betrachtet worden, dass Unklarheiten durch gezielte Nachfragen ausgeräumt werden können. Allerdings stellen irreführende Formulierungen eine bedeutende Fehlerquelle dar und können gerade dann, wenn fachliches Wissen fehlt, zu Missverständnissen führen. Die Gefahr liegt darin, dass die Aussagen oft auf den ersten Blick schlüssig erscheinen und Fehler nicht unmittelbar auffallen.

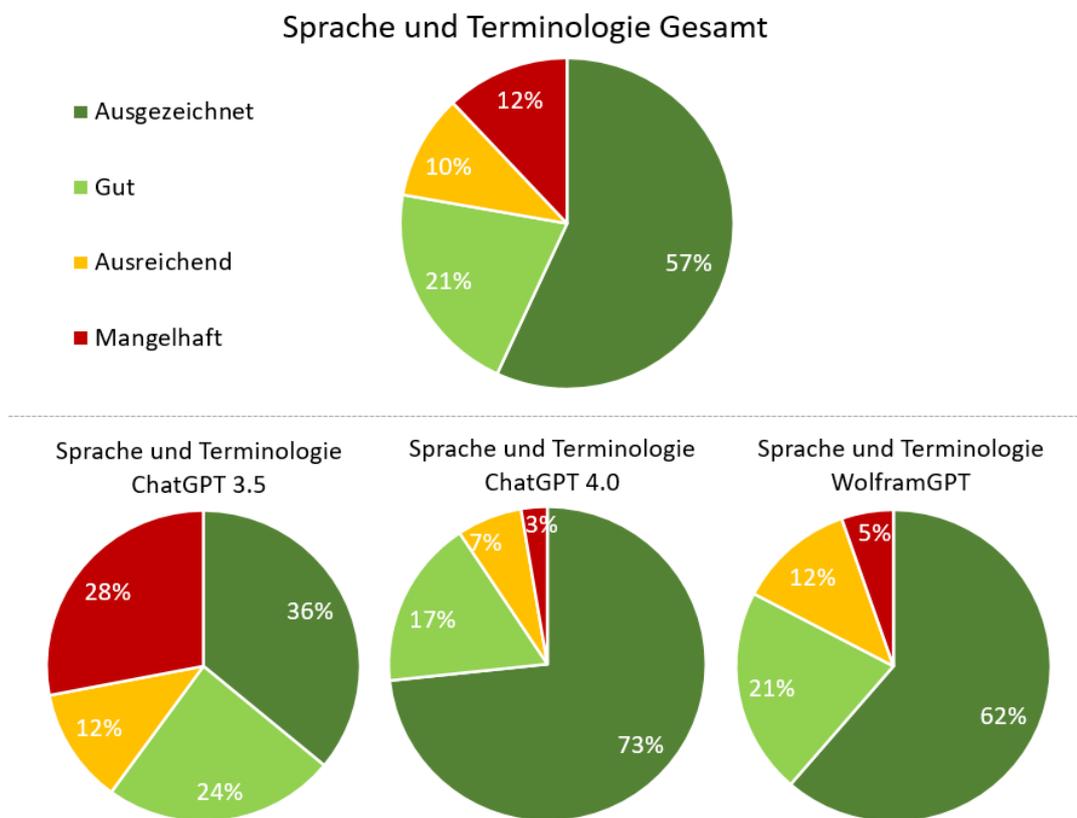


Abbildung 4.5: Kreisdiagramme Sprache und Terminologie

4.1.4 Strukturierung

Bei diesem Kriterium traten bei den getesteten Versionen von ChatGPT keine bemerkenswerten Unterschiede auf. Die Struktur einer Antwort war zu 70 % übersichtlich und sinnvoll, weshalb sie als *ausgezeichnet* klassifiziert wurde. Bei dem kleinsten Anteil fielen nur leichte Mängel bei der Struktur auf, wodurch 6 % aller Antworten der Stufe *gut* zugeordnet wurden. 16 % der Antworten hatten eine teilweise sinnvolle Struktur, wodurch diese Unterteilung am zweithäufigsten auftrat. Die verbleibenden 8 % wurden als unstrukturiert eingeteilt.

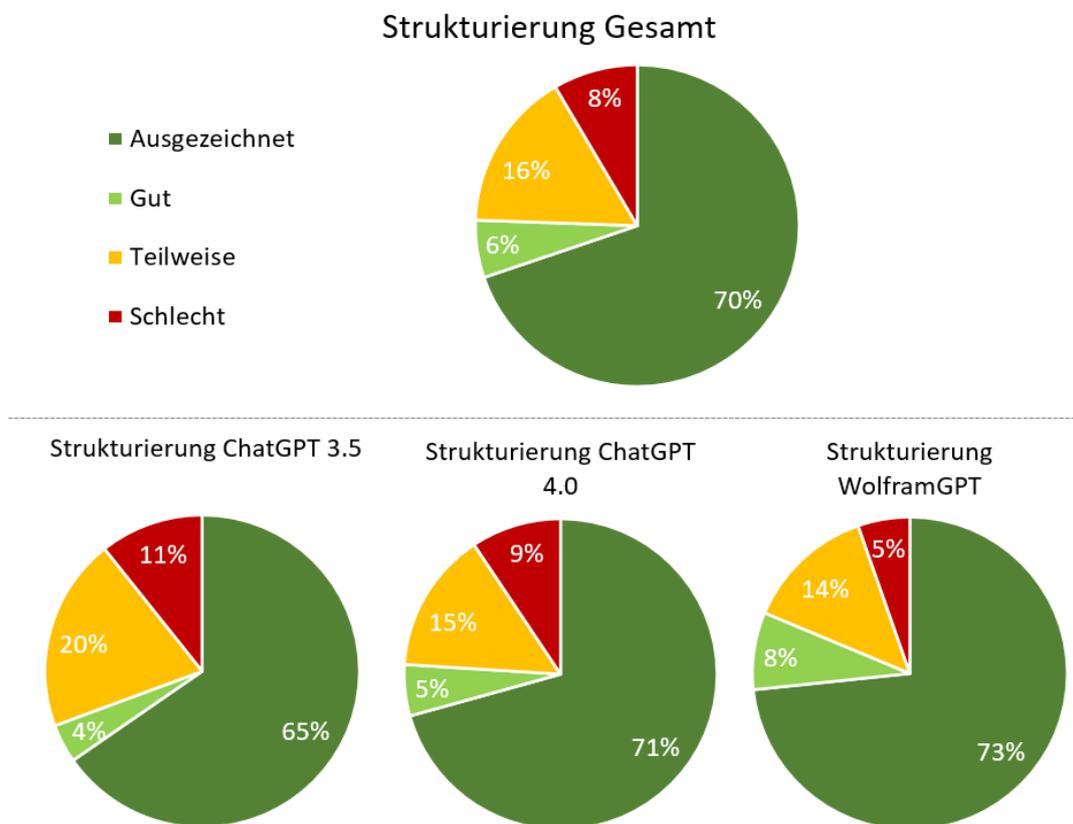


Abbildung 4.6: Kreisdiagramme Strukturierung

So generierte ChatGPT 3.5 eine *ausgezeichnete* Struktur bei 65 % der Antworten, was im Vergleich zum Durchschnitt eine etwas niedrigere Rate darstellt. Antworten der Klassifizierung *gut* wurden mit einer Wahrscheinlichkeit von 6 % generiert, was die niedrigste Wahrscheinlichkeit unter den betrachteten KIs auf dieser Stufe abbildet.

Teilweise sinnvoll strukturierte Antworten sind mit 20 % vertreten, während Antworten, die als schlecht bewertet wurden, 11 % ausmachen.

Mit ChatGPT 4.0 wurden 71 % der Antworten als *ausgezeichnet* kategorisiert, was eine leichte Verbesserung gegenüber der vorherigen Version darstellt und ungefähr dem Gesamtdurchschnitt entspricht. Es gibt mit 5 % *guten* Antworten nur einen geringen Anteil dieser Kategorie, dafür macht die *teilweise effiziente Strukturierung* 15 % der Antworten aus. 9 % der Antworten wurden als schlecht bewertet. Somit liegen die Ergebnisse von ChatGPT 4.0 dem Durchschnitt nahe.

Unter den Systemen zeigt Wolfram Alpha mit nur geringem Unterschied die besten Ergebnisse. Mit 73 % *ausgezeichneten* und 8 % *guten* Antworten erreicht diese Version in beiden Stufen leicht bessere Ergebnisse als der Durchschnitt. Eine *teilweise gute Struktur* wird in 14 % der Fälle erreicht und eine *schlechte Struktur* tritt mit 5 % Wahrscheinlichkeit bei einer Antwort auf.

Auffällig bei allen Versionen war die Tendenz zum Springen zwischen verschiedenen Teilen einer Aufgabe, dabei wurden diese Teile zwar mit passenden Überschriften versehen, die Erklärung wurde allerdings durch die nachfolgenden Schritte unterbrochen. Dies sorgte dafür, dass der Lesefluss und der Prozess des Verstehens immer wieder davon gehemmt wurden, dass ein weiterer Schritt hinzukam.

Dieses Problem wird exemplarisch in der Struktur einer Erklärung von ChatGPT 4.0 deutlich. Bei der aus einem Schulbuch stammenden Aufgabe soll zunächst bewiesen werden, dass drei Punkte ein rechtwinkliges Dreieck bilden. Anschließend soll der Flächeninhalt dieses Dreiecks bestimmt und durch das Hinzufügen eines dritten Punktes zu einem Quadrat erweitert werden. Die Antwort war wie folgt gegliedert:

Beispiel: Strukturierung

Schritt 1: Seitenlängen des Dreiecks berechnen

Schritt 2: Rechtwinkligkeit prüfen

Schritt 3: Flächeninhalt des Dreiecks

Schritt 4: Ergänzung zum Quadrat

Rechnerische Umsetzung

- *Ergebnisse der Berechnungen mit den Unterpunkten:*
 - *Seitenlängen; Überprüfung auf Rechtwinkligkeit und Gleichschenkligkeit; Flächeninhalt des Dreiecks; Ergänzung zum Quadrat*

- *Erklärung* mit den Unterpunkten:
 - *Seitenlängen; Rechtwinkligkeit; Flächeninhalt; Quadratergänzung*

Auch wenn die Antwort eine gewisse Struktur aufweist, bewirkt die Unterteilung in Arbeitsauftrag, Rechnung und Erklärungsblock eine wiederholt auftretende Unterbrechung der momentanen Erklärung von den jeweils folgenden Schritten. Ein weiteres Problem ergibt sich aus der Formatierung der Überschriften: Während zu Beginn die einzelnen Schritte gut zu identifizieren sind, gestaltet sich der zentrale Kern der Antwort, nämlich die Rechnung und Erklärung, unübersichtlich. Diese Gestaltung führt dazu, dass sich die Rechnung und die Erklärung als Unterpunkt eines Unterpunktes, dessen Überschrift lediglich als fettgedruckter Text ohne Absatz erkenntlich ist, schwer identifizieren lassen, was die Lesbarkeit und das schnelle Auffinden relevanter Informationen behindert.

4.1.5 Zusammenfassung der Ergebnisse

Bei einer übergreifenden Betrachtung der Leistungen der untersuchten ChatGPT-Versionen in den verschiedenen Kategorien zeigt sich, dass verschiedene KI-Versionen in unterschiedlichen Kategorien ihre Stärken aufweisen. Dabei zeigten diese in den diversen Aufgabentypen zum Teil unterschiedliche Ergebnisse. In einer spezifischen Betrachtung der verschiedenen Aufgabentypen kann eine bestimmte Version somit bessere Ergebnisse erreicht haben als die Version, welche in der Kategorie insgesamt die stärksten Leistungen gezeigt hat.

Unter Betrachtung dieses Aspekts kann zusammenfassend festgehalten werden, dass ChatGPT 4.0 unter Verwendung der Wolfram-Alpha-Integration die besten Ergebnisse lieferte, wenn es um die korrekte Lösung eines Ergebnisses ging. Durch die Abhängigkeit eines *korrekten* Ergebnisses von der Reproduzierbarkeit zeigte sich dadurch auch eine Verbesserung in diesem Bereich. Bei dem Kriterium der Vollständigkeit sorgte die Verwendung von Wolfram Alpha für keine nennenswerten Verbesserungen im Vergleich zu ChatGPT 4.0. Betrachtet man hingegen die Sprache und Terminologie, weist ChatGPT 4.0 einen klaren Vorteil gegenüber den anderen beiden Versionen auf. Bei der Strukturierung bringt die Verwendung von ChatGPT 4.0 mit Wolfram Alpha eine leichte Verbesserung, allerdings ist auch hier der Unterschied zu ChatGPT 4.0 nicht groß. Somit lässt sich sagen, dass die ChatGPT-Version 3.5 die schlechtesten Ergebnisse erzielt hat. Die Unterschiede, die mit der Verwendung von Wolfram Alpha mit der Version 4.0 erreicht werden können, erstrecken sich hingegen hauptsächlich auf die Korrektheit des Ergebnisses.

4.2 Ergebnisse der Umfrage

In diesem Unterkapitel werden die Ergebnisse der Umfrage in den fünf verschiedenen Kategorien vorgestellt und mit der eigenen Datenanalyse verglichen. Die Auswertung der Umfragen der Schülerinnen, Schülern und Studierenden gestaltet sich komplexer als die Auswertung der gestellten Fragen, da bei der Betrachtung der Ergebnisse die Kontextualisierung der vorgelegten Aufgaben essenziell ist. Eine konkrete Interpretation der Ergebnisse ist daher in dem Kontext der gestellten Aufgaben zu sehen. Für diesen Bezug ist die Kenntnis über die Ergebnisse der Datenauswertung, der in der Umfrage untersuchten Aufgaben relevant. Die spezifische Auswertung der in der Umfrage enthaltenen Aufgaben wird daher als Diagramm in Anhang B präsentiert.

Das Kriterium der Korrektheit ist für die Verwendbarkeit der generierten Antwort im Bildungskontext von besonderer Bedeutung, da sie maßgeblich zum Verständnis und zur Vermittlung des Wissens beiträgt. Falsche Antworten können zu Missverständnissen führen, was die Effektivität des Lernprozesses beeinträchtigt. Obwohl auch andere Kriterien von Bedeutung sind, nimmt die Korrektheit deshalb im Bildungskontext eine fundamentale Rolle ein, weshalb sie in dieser Arbeit besonders intensiv betrachtet wird. Für das Kriterium der Korrektheit erfolgt in Kapitel 4.2.1 der Vergleich zwischen Datenauswertung und Umfrageergebnissen mittels Konfusionsmatrizen. Diese Darstellungsmethode bot sich an, da die Bewertungskriterien der Umfrage mit denen der Datenanalyse übereinstimmte. Bei den weiteren Kriterien wurden Abstufungen mit teilweise speziellen Eigenschaften definiert, was die Einbindung dieser Abstufungen in der Umfrage weniger intuitiv macht. Dies erschwert einen genauen Vergleich der Einordnungen in Form einer solchen Matrix. Um dennoch einen Vergleich zu ermöglichen, erfolgt im Anhang B eine Gegenüberstellung der durchgeführten Datenanalyse und Umfrage. Die detaillierte Gestaltung des verwendeten Umfragebogens ist über das GitHub-Repository einsehbar, welches in Anhang C verlinkt ist.

Die Verteilung der Aufgabentypen innerhalb der Umfrage erfolgte zudem nicht gleichmäßig, sondern konnte von den Befragten gewählt werden. Die dadurch entstandene variable Verteilung der Aufgaben bildet die Präferenzen der Befragten für bestimmte Aufgabenbereiche wider. Bei der Befragung der 38 Studierenden zeigte sich die Bevorzugung des Themenbereichs der Gleichungen, während von den 13 Nachhilfe Nehmenden die meisten die Frage des Bereichs Logik beantworteten. Wie im vorherigen Kapitel 3 erläutert, wurden von den Nachhilfe Nehmenden nicht zwangsläufig alle der drei generierten Antworten evaluiert, was zu einer unterschiedlichen Verteilung der KIs innerhalb dieser Gruppe führte. So kam es bei der Durchführung der Umfrage bei dieser Gruppe zu einer unterschiedlichen Anzahl von Umfrageergebnissen für die einzelnen untersuchten ChatGPT-Versionen. Für die Versionen 3.5 und 4.0 mit der Wolfram Alpha Integration wurden jeweils 10 Datensätze erhoben. Für die

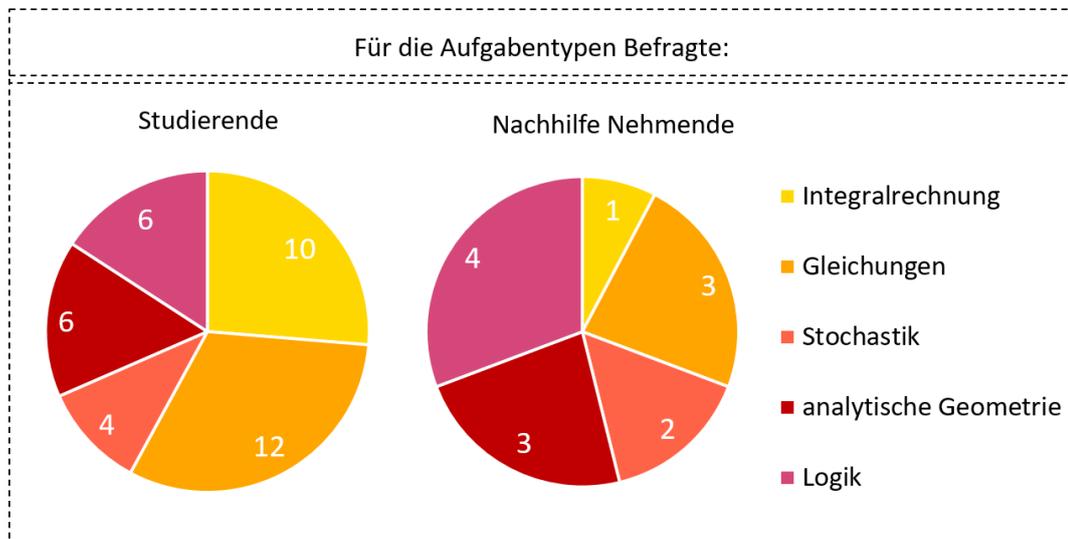


Abbildung 4.7: Aufgabentypenverteilung der Umfrage

Version 4.0 ohne zusätzliche Integration lagen hingegen 12 Umfrageergebnisse vor. Die folgende Abbildung zeigt die Anzahl der durchgeführten Umfragen in den einzelnen Themengebieten.

Einstellung der Befragten zur Verwendung von KI in der Lehre

Die persönliche Einstellung der Befragten bezüglich des Einsatzes künstlicher Intelligenz in der Bildung kann die Antworten in der Umfrage erheblich beeinflussen. Aus diesem Grund erfolgte eine Stellungnahme jeder befragten Person zu diesem Thema. Die Meinungstendenzen zwischen den beiden Gruppen zu diesem Thema unterschieden sich nur geringfügig. So gaben 69,2% der Nachhilfe Nehmenden an, neutral zu dem Einsatz von KI zur Lösung mathematischer Probleme im Unterricht zu stehen. Dahingegen tendierte der Personenkreis der Studierenden eher zu einer positiveren Einstellung.

Diese Überzeugungen zeigten sich auch in den Antworten auf die Frage, ob eine KI-basierte Erklärung ein tieferes Verständnis der mathematischen Konzepte fördern kann. Diese Frage wurde von 69,2% der Nachhilfe Nehmenden mit *unsicher* beantwortet, während 68,4% der Studierenden diese Frage bejahten.

4.2.1 Korrektheit

Bei der zuvor durchgeführten Auswertung der von ChatGPT generierten Antworten wurde festgestellt, dass die Erklärungen und die von den KIs vorgestellten Ergebnisse nicht korrekt sein müssen. Vor diesem Hintergrund erfolgte im Rahmen der Umfrage eine Einschätzung zur Korrektheit der gegebenen Antwort durch die Teilnehmenden.

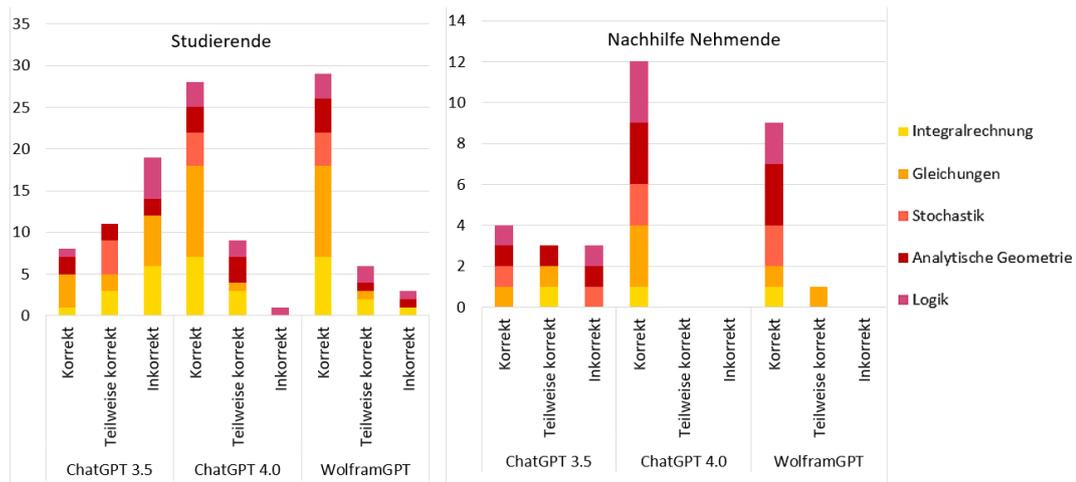


Abbildung 4.8: Umfrage: Bewertung der Korrektheit

Das obige Diagramm liefert einen Überblick über die Bewertung durch die befragten Studierenden und die Nachhilfe Nehmenden. Für jede der drei ChatGPT-Versionen wird die Anzahl der Stimmen für die drei Antwortoptionen dargestellt. Die fünf Aufgabentypen können durch die Farbunterschiede innerhalb der Säule ermittelt werden.

Die zwei Diagramme verschaffen somit einen Einblick in das Vertrauen in die Korrektheit der durch die KIs gestellten Lösungen und ermöglichen einen allgemeinen Vergleich der beiden Gruppen. Aus dem Diagramm geht hervor, dass die Schülerinnen und Schüler tendenziell eher die von den KIs gegebenen Antworten als *korrekt* eingestuft haben als die Studierenden.

Bei dem Lernprozess ist es essenziell, dass das verinnerlichte Wissen auch korrekt ist. Daher ist es wichtig, dass Anwendende in der Lage sind, falsche Aufgabenlösungen für die gestellte mathematische Frage zu erkennen. Um eine grundsätzliche Einschätzung dieser Fähigkeit zu erleichtern, werden die Antworten der Umfrage mit dem geprüften Korrektheitsgrad in den folgenden Konfusionsmatrizen verdeutlicht. Dabei wird eine

korrekte Erkennung in Grün und eine vollständig falsche Zuordnung in Rot dargestellt.

ChatGPT 3.5	Geprüfte Korrektheit			
Angabe der Befragten		Korrekt	Teilweise korrekt	Inkorrekt
	Korrekt	0	0	8
	Teilweise korrekt	0	4	7
	Inkorrekt	0	0	19

Tabelle 4.4: Konfusionsmatrix: Bewertung der ChatGPT-3.5-Antworten durch Studierende

ChatGPT 3.5	Geprüfte Korrektheit			
Angabe der Befragten		Korrekt	Teilweise korrekt	Inkorrekt
	Korrekt	0	1	3
	Teilweise korrekt	0	0	3
	Inkorrekt	0	1	2

Tabelle 4.5: Konfusionsmatrix: Bewertung der ChatGPT-3.5-Antworten durch Nachhilfe Nehmende

Keine der von ChatGPT 3.5 ausgegebenen Erklärungen, die in der Umfrage verwendet wurden, beinhaltet vollständig *korrekte* Ergebnisse. Die Umfrageergebnisse der Studierenden zeigen eine Tendenz zur Identifizierung der *inkorrekten* Ergebnisse bei den Antworten. Bei der Gruppe der Nachhilfe Nehmenden zeigt sich allerdings eine gleichmäßigere Verteilung in allen Antwortoptionen. Die genauere Analyse der Antworten ergibt, dass Studierende 55,9% der *inkorrekten* Antworten als inkorrekt erkannt haben, aber auch 23,5% der *inkorrekten* Antworten für korrekt gehalten wurden. Die Schülerinnen und Schüler erkannten hingegen nur ein Viertel der falschen Antworten als vollkommen falsch und hielten 37,5% trotz ihres nicht korrekten Ergebnisses für vollkommen richtig.

Bei der Betrachtung der Antworten von der Version 4.0, bei denen nur die Lösung des Aufgabentyps Logik inkorrekt war, wird die grundsätzliche Tendenz der Befragten deutlich, Antworten dieses ChatGPT-Modells als korrekt anzusehen. Die Gruppe der Studierenden nimmt bei der Einschätzung der Antworten einen geringfügig kritischeren Blickwinkel ein als die Gruppe der Schülerinnen und Schüler, welche alle Antworten der KI als *korrekt* einstuft. Aus der Analyse der Daten geht hervor, dass beide Gruppen meist zuverlässig korrekte Antworten erkennen. Bei der Identifikation *inkorrekt*er Antworten zeigen sich jedoch Defizite. Die Studierenden waren imstande, 16,67%

der *inkorrekten* Antworten zu erkennen, allerdings wurden auch 50 % der *inkorrekten* Antworten für korrekt gehalten. Die Nachhilfe Nehmenden haben hingegen alle falschen Antworten als korrekt angesehen und somit keine der falschen Aufgabenlösungen erkannt.

ChatGPT 4.0	Geprüfte Korrektheit			
Angabe der Befragten		Korrekt	Teilweise korrekt	Inkorrekt
	Korrekt	25	0	3
	Teilweise korrekt	7	0	2
	Inkorrekt	0	0	1

Tabelle 4.6: Konfusionsmatrix: Bewertung der ChatGPT-4.0-Antworten durch Studierende

ChatGPT 4.0	Geprüfte Korrektheit			
Angabe der Befragten		Korrekt	Teilweise korrekt	Inkorrekt
	Korrekt	9	0	3
	Teilweise korrekt	0	0	0
	Inkorrekt	0	0	0

Tabelle 4.7: Konfusionsmatrix: Bewertung der ChatGPT-4.0-Antworten durch Nachhilfe Nehmende

Die Aufgaben der Umfrage, die von ChatGPT 4.0 unter Verwendung des Wolfram-Alpha-Plug-ins bearbeitet wurden, waren in den Aufgabenbereichen Logik und analytische Geometrie inkorrekt beantwortet worden. Somit gab es im Vergleich zu den Antworten von ChatGPT 4.0 weniger korrekt gelöste Aufgabenbereiche. Dennoch verringerte sich die Anzahl der als *korrekt* eingestuft Antworten nicht, sondern wuchs weiter an. Das wachsende Vertrauen in die Erklärung von WolframGPT zeigt sich dadurch, dass 58,33 % der Studierenden fälschlicherweise einer *inkorrekten* Antwort Glauben schenken, während von den Schülerinnen und Schülern wie auch bei der ChatGPT-Version 4.0 alle falschen Antworten als korrekt eingestuft wurden.

Das Vertrauen in die Korrektheit der Erklärungen der Antworten von allen ChatGPT-Versionen ist bei beiden Gruppen ausgeprägt. Dabei ist die Glaubhaftigkeit der angebotenen Lösung bei der Version 4.0 größer und steigert sich unter Verwendung von Wolfram Alpha noch einmal. Grundsätzlich scheint die Gruppe der Studierenden die Antworten etwas mehr zu hinterfragen als die der Schüler/-innen.

Wolfram Alpha	Geprüfte Korrektheit			
Angabe der Befragten		Korrekt	Teilweise korrekt	Inkorrekt
	Korrekt	22	0	7
	Teilweise korrekt	3	0	3
	Inkorrekt	1	0	2

Tabelle 4.8: Konfusionsmatrix: Bewertung der Wolfram-Antworten durch Studierende

WolframGPT	Geprüfte Korrektheit			
Angabe der Befragten		Korrekt	Teilweise korrekt	Inkorrekt
	Korrekt	4	0	6
	Teilweise korrekt	0	0	0
	Inkorrekt	0	0	0

Tabelle 4.9: Konfusionsmatrix: Bewertung der WolframGPT-Antworten durch Nachhilfe Nehmende

4.2.2 Vollständigkeit

Die Untersuchung der Umfrageergebnisse, welche sich auf den Bereich der von den Befragten eingestuften Vollständigkeit der Antwort bezieht, zeigt Unterschiede zwischen der Einschätzung der Befragten und der Einschätzung, wie sie im Rahmen der Auswertung aller Fragen getroffen wurden. Bei der umfrageunabhängigen allgemeinen Auswertung kam es bei den drei von ChatGPT 3.5 generierten Antworten der Aufgabentypen Gleichungen, Stochastik und analytische Geometrie bezüglich ihrer Vollständigkeit zu einer massiven Bemängelung der Qualität. Dies machte sich in einer Zuteilung dieser Aufgabentypen in die zwei untersten Stufen bemerkbar. Außerhalb dieser generierten Antworten kam es bei der allgemeinen Datenauswertung zu keinen bemerkenswerten Beanstandungen bezüglich der Vollständigkeit der Erklärungen.

Die Umfrageergebnisse der beiden Gruppen bestätigen diese Einstufung jedoch nicht vollumfänglich. Bei der Betrachtung des folgenden Säulendiagramms fällt auf, dass beide befragte Gruppen die Vollständigkeit der von der ChatGPT-Version 3.5 gegebenen Erklärungen als am schlechtesten einschätzten. Selbst mit Erzielung dieser schwächeren Leistung wird deutlich, dass die Befragten die Einschätzung der Vollständigkeit bei dieser KI deutlich positiver vornahmen als im Ergebnis der Datenauswertung. Eine besondere Auffälligkeit ergab sich aus der teilweise eher positiven Einstufung von den Antworten der Themenbereiche analytische Geometrie und Stochastik beider Gruppen. Beide Aufgaben wurden im Rahmen der Datenauswertung als *eher unvollständig* kategorisiert, jedoch zeigen die Umfrageergebnisse keine Widerspiegelung dieser Einschätzung. Auch das Themengebiet der Gleichungen erhält durch die Umfrage

teilweise positivere Zuordnungen. Unvorhergesehen erfolgte in der Umfrage durch die beiden Gruppen eine eher negative Einschätzung der in der Datenauswertung mit guten Bewertungen eingestuftem Aufgabentypen Integralrechnung und Logik.

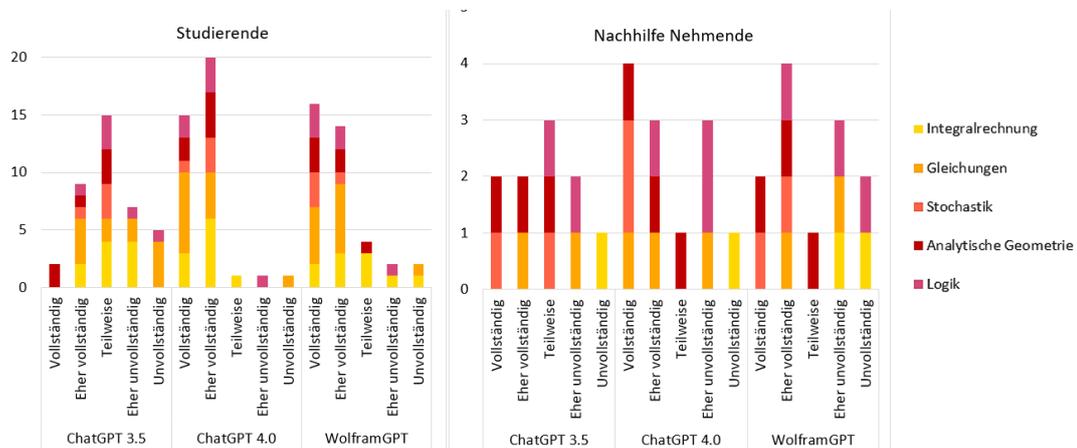


Abbildung 4.9: Umfrage: Bewertung der Vollständigkeit

Wie in der oberhalb abgebildeten Grafik zu erkennen ist, befand nahezu die gesamte Menge der befragten Studierenden die von ChatGPT 4.0 generierte Erklärung für *vollständig* oder zumindest *eher vollständig*. Diese Einschätzung der Studierenden deckt sich zu einem Großteil mit der vorab durchgeführten Datenauswertung. Die Antworten der Gruppe der Nachhilfe Nehmenden streuen sich breiter, dennoch wird eine leichte Tendenz zu einer positiven Einordnung der Erklärung erkennbar. Die breitere Verteilung der Einschätzungen der Schülerinnen und Schüler in Bezug auf die Vollständigkeit ergibt bei den im Rahmen der Datenauswertung mit einer guten Bewertung eingestuftem Aufgabentypen zum Teil eine negativere Einschätzung seitens der Befragten.

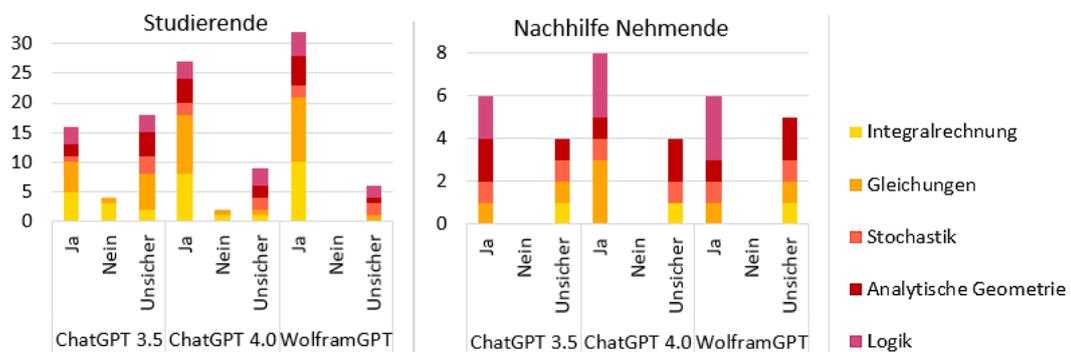
Der bei ChatGPT 4.0 bereits erkennbare Trend setzt sich auch mit der Integration von Wolfram Alpha fort. Beide Gruppen empfinden die gegebenen Erklärungen für die Fragestellung mehrheitlich eher als *vollständig*, die Eindeutigkeit dieser Einschätzung ist jedoch geringer, da bei beiden befragten Gruppen eine Abnahme zu erkennen ist. Die Einstufung als *vollständig* ist die vorherrschende Bewertung der Studierenden bei dieser Version, wobei auch die Klassifikation *eher vollständig* häufig vergeben wurde. Trotz des erstmaligen Überwiegens des höchsten Bewertungskriteriums ist auch ein leichter Anstieg bei den schlechteren Stufen erkennbar. Die Einschätzungen der Schülerinnen und Schüler zeigen im Kontrast zu den Studierenden eine klare Verschlechterung. Für Antworten wurde die Stufe *eher vollständig* favorisiert vergeben, wobei die beiden schlechteren Stufen verglichen mit der Version 4.0 nur einen leichten

Zuwachs erhalten.

Grundsätzlich erhielt die Version 4.0 von beiden Gruppen in der Umfrage und auch der Datenauswertung die besten Bewertungen bezüglich der Vollständigkeit. Bei der Gruppe der Schülerinnen und Schüler kommt es zu teilweise großen Unterschieden in der Einschätzung dieses Kriteriums, dabei ist die kleine Umfragegröße von 13 Befragten ein ausschlaggebender Faktor.

4.2.3 Sprache und Terminologie

Korrektheit der Fachbegriffe



Verständlichkeit der Sprache

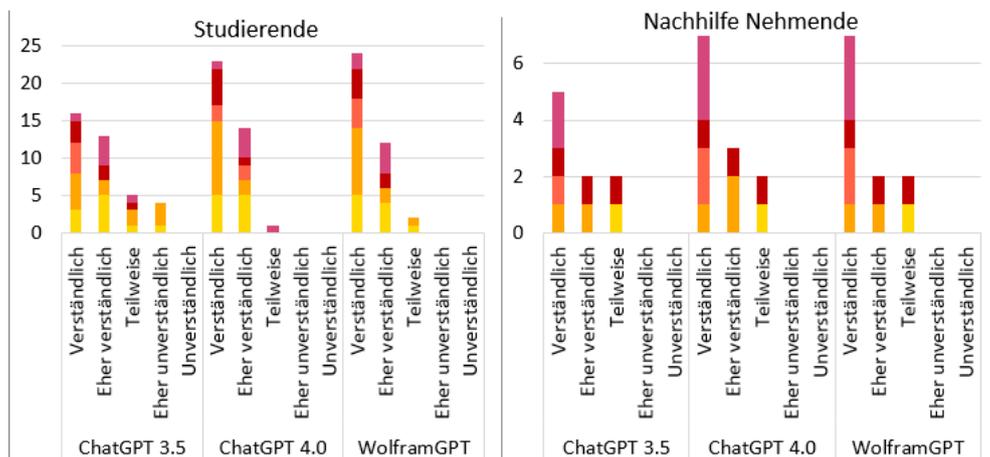


Abbildung 4.10: Umfrage: Bewertung der Sprache und Terminologie

Die in der Umfrage Verwendung findenden Aufgaben wurden in Bezug auf ihre sprachliche und terminologische Eignung für den Einsatz als didaktische Erklärung von mathematischen Aufgaben positiv eingestuft. Diese positive Einstufung spiegelt sich durch die Ergebnisse der Umfrage wider. Die verwendeten Fachbegriffe wurden zum Großteil als *korrekt* eingestuft, wobei sich auch eine Unsicherheit beider Befragungsgruppen bei dieser Festlegung zeigt. Bezüglich der korrekten Verwendung der Terminologie nimmt diese Unsicherheit bei der Festlegung in der Gruppe der Studierenden bei ChatGPT 4.0 und ChatGPT 4.0 mit der Verwendung des Wolfram-Alpha-Plug-ins ab. Der Großteil der Nachhilfe Nehmenden schätzte die von ChatGPT 4.0 verwendeten Fachbegriffe als *korrekt* ein, wobei beachtet werden muss, dass von den Schülerinnen und Schülern bei ChatGPT 4.0 auch mehr Umfrageergebnisse vorliegen als zu den anderen Modellen.

Die verwendete Sprache hingegen wurde von beiden Gruppen grundsätzlich als *verständlich* eingestuft, wobei Bewertungen unter der Stufe *eher verständlich* bei beiden Gruppen nur vereinzelt auftraten. Die Umfrageergebnisse aller Befragten deuten darauf hin, dass die verwendete Sprache und die Formulierungen von ChatGPT 4.0 eine deutlich bessere Verständlichkeit aufweisen als in der Vorgängerversion. Die Verwendung von Wolfram Alpha führt laut den Umfrageergebnissen nur noch zu einer geringen Steigerung der Verständlichkeit. Die in der Umfrage Verwendung findenden Aufgaben wurden im Rahmen der Datenauswertung durchgehend als verständlich eingeordnet, dadurch decken sich die aus der Umfrage gewonnenen Ergebnisse nur bezüglich der eher positiven Einschätzungen der Sprache. Eine Verbesserung der sprachlichen und terminologischen Formulierungen von KI-Version zu KI-Version konnte bei allen in der Umfrage vorgestellten Aufgaben durch die Datenauswertung nicht ermittelt werden.

4.2.4 Verständnis

Für die Feststellung der Eignung der untersuchten Modelle als Mathematik-Tutor ist es von besonderer Relevanz, dass die gegebenen Antworten von der Zielgruppe verstanden werden. Aus diesem Grund wurden innerhalb der Umfrage vier Fragen gestellt, welche unterschiedliche Aspekte dieses Bereiches abdecken.

Für eine gelungene Erklärung ist das Verständnis des Rechenwegs essenziell, deshalb wurden die Befragten um eine Bewertung ihres Verständnisses bezüglich des vorgestellten Rechenwegs gebeten. Die Ergebnisse der Umfrage sind in den Diagrammen unten zu erkennen. ChatGPT 4.0 erzielt beim Aspekt des verständlichen Rechenwegs bessere Ergebnisse als das Vorgängermodell 3.5. Die Verwendung von Wolfram Alpha führt zu keiner Verbesserung des Verständnisses, hat aber auch keine nennenswerte Verschlechterung zur Konsequenz. Die Rechenwege für die mathematischen Aufgaben

werden von den Studierenden grundsätzlich etwas besser verstanden als von den Schülerinnen und Schülern.

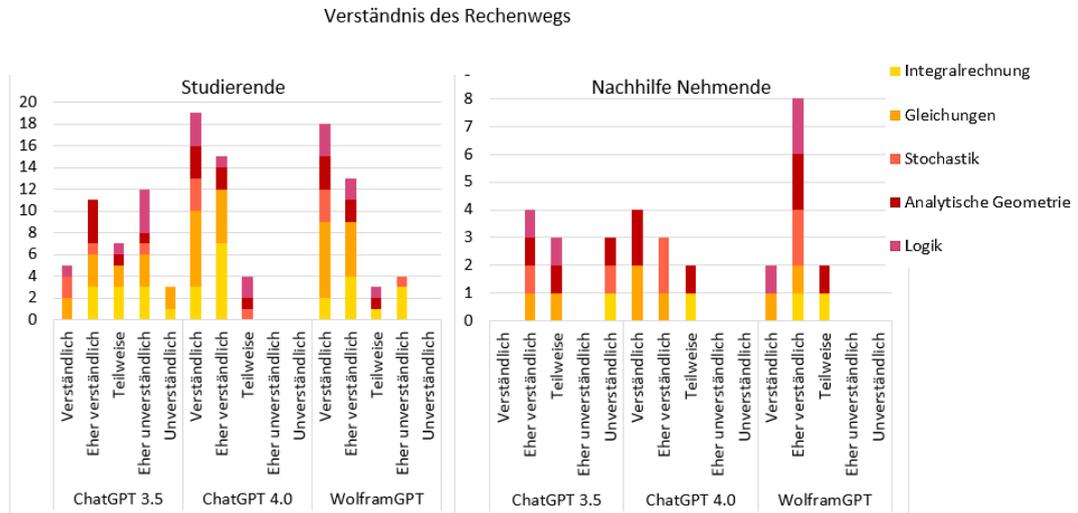


Abbildung 4.11: Umfrage: Verständnis des Rechenwegs

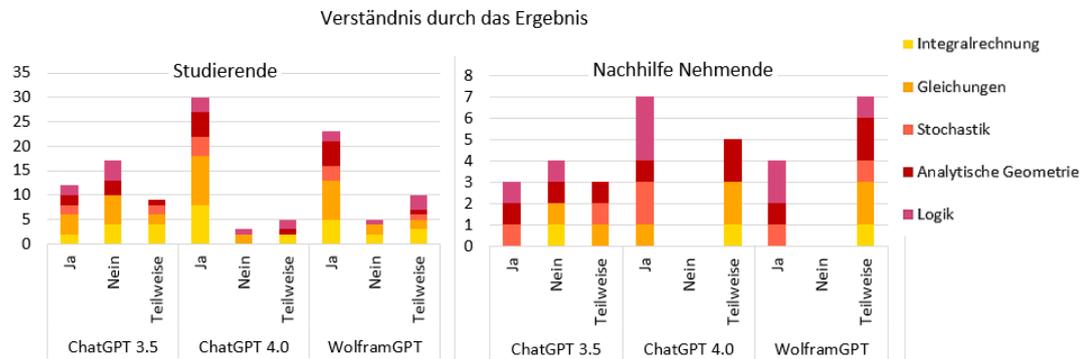


Abbildung 4.12: Umfrage: Verständnis durch das Ergebnis

Ein weiterer in der Umfrage untersuchter Aspekt betrifft die Verbesserung des Verständnisses der mathematischen Aufgabe mittels des durch die jeweilige KI generierten Ergebnisses. Der Fokus liegt dabei darauf, wie die gegebene Lösung ein Verständnis der Problemstellung fördert. Die erhobenen Daten zeigen eine Tendenz, die den Untersuchungsergebnissen bezüglich des Verständnisses des Rechenwegs ähneln. Dieses Ergebnis war zu erwarten, da beide Fragen ähnliche Themengebiete abdecken. Bemerkenswerterweise tragen laut den Umfrageergebnissen auch inkorrekte Ergebnisse zu

dem Verständnis der Aufgabenstellung bei. Dies ist insbesondere bei ChatGPT 4.0, sowohl mit als auch ohne das Wolfram-Alpha-Plug-in, der Fall. Fehlerhafte Lösungen bei Aufgabentypen wie Logik, die von allen Versionen fehlerhaft behandelt wurde, sowie bei analytischer Geometrie, die lediglich von ChatGPT 4.0 korrekt bearbeitet wurde, können laut Angabe der Befragten zu einer Verbesserung des Verständnisses beitragen. Bei den beiden auf ChatGPT 4.0 basierenden Versionen kam es zu geringer Verneinung dieser Frage, die vor allem die fehlerhaften Ergebnisse primär ausließ.

Die Umfrageergebnisse zeigen, dass die Erklärung maßgeblichen Einfluss auf das Verständnis der Aufgabe und von deren Rechenweg hat. Dabei fördern die nicht auf ChatGPT 3.5 basierenden Modelle das Verständnis der Aufgabe durch ihre Erklärung stärker. Dies zeigt sich in einer zunehmend positiveren Bewertung der Version 4.0 und Wolfram Alpha. Die befragten Studierenden zeigen einen Zusammenhang zwischen dem Verständnis der Aufgabe durch die Erklärung und der Einschätzung der Vollständigkeit dieser Erklärung. So wurde eine als *unvollständig* eingordnete Erklärung von Aufgabentypen als nicht hilfreich eingestuft. Die Ergebnisse der Schülerinnen und Schüler bestätigen diesen Zusammenhang jedoch nicht.

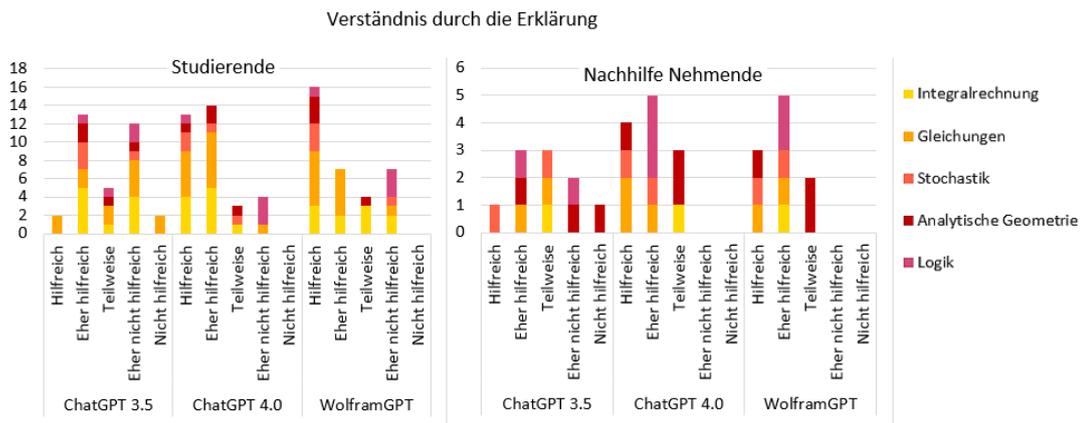


Abbildung 4.13: Umfrage: Verständnis durch die Erklärung

Für einen gelungenen und nachhaltigen Lerneffekt ist der Transfer des Gelernten auf neue Aufgaben ein wichtiger Schritt. Aus diesem Grund wurden die Befragten um eine Einschätzung der Übertragbarkeit des durch die KIs vorgestellten Lösungswegs auf andere Aufgaben gebeten. Diese Einschätzung fiel für ChatGPT 3.5 erwarteterweise schlechter aus als für die anderen beiden Versionen. Dennoch ist in Anbetracht der Unvollständigkeit der Erklärung und der durchgehend falschen Ergebnisse ein besseres Ergebnis erzielt worden, als erwartet wurde. Die Selbsteinschätzung der Befragten

war bei ChatGPT 4.0 besonders gut und nahm bei der zusätzlichen Anbindung von Wolfram Alpha nur wenig ab.

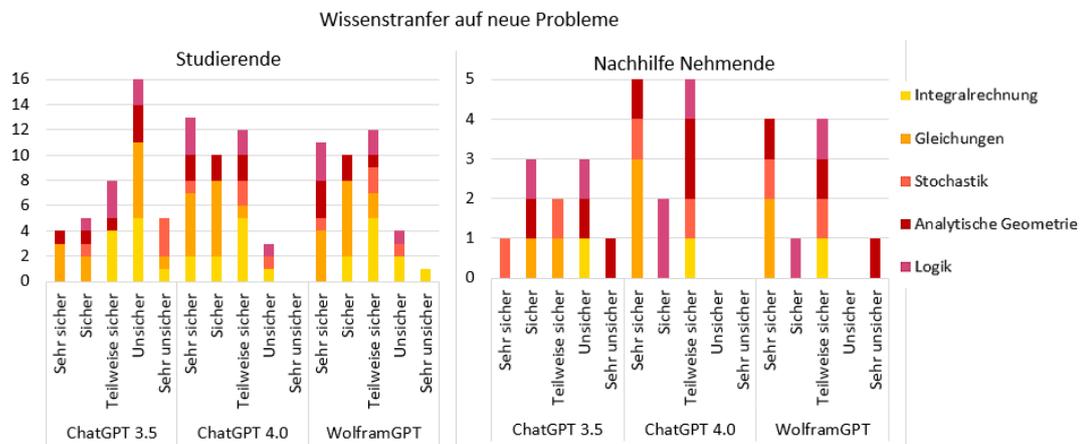


Abbildung 4.14: Umfrage: Wissenstransfer auf neue Probleme

In Bezug auf das Verständnis wurden besonders bei ChatGPT 4.0 und auch Wolfram Alpha gute Rückmeldungen erzielt. In diesem Bereich sind sich die beiden befragten Gruppen in ihrer Meinung einig und haben kaum merkliche Abweichungen. Die Anwendung von WolframGPT führt laut der Umfrage zu keiner Verbesserung bezüglich des Verständnisses.

4.2.5 Zusammenfassung der Ergebnisse

Die Ergebnisse der durchgeführten Datenauswertung und der Umfrage geben einen Hinweis auf die Unterschiede der drei getesteten Versionen von ChatGPT. Durch die Unterteilung in die verschiedenen Kriterien eröffnet sich die Möglichkeit einer genaueren Analyse der Vor- und Nachteile der KIs in den unterschiedlichen Bereichen. Die ermittelten Unterschiede geben einen Ausblick auf die Einsatzmöglichkeiten der KIs im Bereich der Bildung.

Die durchgeführte Datenerhebung zeigt eine klare Steigerung der von ChatGPT 4.0 erzielten Leistung im direkten Vergleich zu seinem Vorgängermodell 3.5. Diese Verbesserung erstreckt sich auf alle im Rahmen dieser Arbeit bearbeiteten Bereiche. Die Verbesserungen, welche unter kombinierter Verwendung von ChatGPT 4.0 mit Wolfram Alpha erzielt werden können, finden sich laut der Datenauswertung aller Fragen hauptsächlich bei der Korrektheit des erzielten Ergebnisses und dessen Reproduzierbarkeit. Im Kontext der Vollständigkeit der Erklärung und der Angemessenheit der

verwendeten Sprache und Terminologie waren die ohne das Plug-in erzielten Ergebnisse besser.

Die Umfrageergebnisse zeigen ein großes Vertrauen der Befragten in die Korrektheit der KIs. Die Ergebnisse der beiden auf ChatGPT 4.0 basierenden Modelle wurden von beiden Gruppen häufiger als *korrekt* eingestuft. Dabei wurden allerdings bei allen verwendeten KIs falsche Ergebnisse für korrekt gehalten. Die relative Quote der fälschlicherweise für korrekt gehaltenen Lösungen nahm von ChatGPT 3.5 bis ChatGPT 4.0 mit der Integration von Wolfram Alpha zu. Dabei muss bedacht werden, dass bei der Verwendung der Version 4.0 absolut weniger *inkorrekte* Lösungen auftraten.

Die in der Erklärung der drei untersuchten Modelle Verwendung findende Sprache und die entsprechenden Fachbegriffe wurden im Rahmen der Umfrage als *korrekter* und *verständlicher* eingestuft, wenn es sich um eines der neueren Modelle handelte. Somit wurden die Fähigkeiten von ChatGPT 3.5 schlechter bewertet. Auch in Bezug auf die grundlegende Verständlichkeit erzielte ChatGPT 3.5 schlechtere Ergebnisse als der Nachfolger. In dem Zusammenhang des Transfers der gelernten Inhalte auf ähnliche Aufgaben kam es zu einer schlechten Einschätzung der von ChatGPT 3.5 vermittelten Inhalte, während ChatGPT 4.0 von den getesteten KI-Modellen die besten Ergebnisse erzielte.

Die beiden in der Umfrage getesteten Personenkreise zeigen unterschiedliche Einstellungen zur Verwendung von KI in der Lehre. Während die Studierenden eher eine positive Haltung gegenüber der Verwendung von künstlicher Intelligenz im Bildungskontext aufweisen, offenbaren die Schüler/-innen meist eine neutrale Haltung. Dies wirkte sich auch auf die Einschätzung der Eignung von KI zur Förderung des Verständnisses von mathematischen Konzepten aus, die bei der deutlichen Mehrheit der Studierenden positiv war, während die Nachhilfe Nehmenden überwiegend die Auswahl *unsicher* trafen.

5 Diskussion

Zur Beantwortung der zentralen Forschungsfrage wurden die Unterschiede der ChatGPT-Versionen 3.5, 4.0 und 4.0 mit der Integration von Wolfram Alpha in Bezug auf ihre Eignung in den Kriterien Korrektheit, Vollständigkeit, Strukturierung sowie Sprache und Terminologie untersucht. Die Analyse der von diesen Versionen generierten Antworten auf die mathematischen Fragen verschiedener Themengebiete wurde durch die Durchführung einer Umfrage stichprobenartig mit der Meinung der für den Einsatz in der Lehre relevanten Personengruppen abgeglichen. Die in dem vorherigen Kapitel vorgestellten Ergebnisse der beiden durchgeführten Datenerhebungsverfahren boten bereits erste Einblicke in die Eignung der verschiedenen Versionen zum Einsatz in dem Bildungskontext. Ausgehend von dieser Grundlage ist die Interpretation der gewonnenen Ergebnisse vorgesehen, wobei diese im Kontext der theoretischen Grundlagen in den bereits vorhandenen Forschungsstand eingeordnet werden sollen. Dabei sollen nicht nur die einzelnen Leistungen der KIs analysiert, sondern es soll darüber hinaus auch die praktische Relevanz für den Einsatz im Bildungskontext betrachtet werden.

Für das fundamentale Verständnis sind eine korrekte Erklärung und Lösung essenziell, da sie die Basis für ein dauerhaftes Verständnis der vermittelten Konzepte bilden. Inkorrekte Lösungen oder Ungenauigkeiten können zu Misstrauen in Bezug auf die Antworten führen und dadurch die Lernmotivation negativ beeinflussen. Zudem können fehlerhafte Erklärungen zum Erlernen von Fehlkonzepten führen, welche im Nachhinein nur schwer korrigierbar sind. Aus diesen Gründen wird in dieser Forschungsarbeit dem Kriterium der Korrektheit der KI-generierten Lösungen im Vergleich zu den anderen Kriterien die größte Bedeutung beigemessen.

5.1 ChatGPT 3.5

Bei dem Kriterium der Korrektheit wurde von ChatGPT 3.5 die höchste Anzahl von inkorrekten Ergebnissen erzielt. So wurden bei der Analyse der Antworten von ChatGPT 3.5 nicht nur Probleme bei der korrekten Ausführung von Rechnungen, sondern auch bei der Ermittlung korrekter Rechenansätze offenbart. Diese beiden Defizite, welche auch in Kombination aufgetreten sind, sorgten primär für die hohe Quote von Falschaussagen. Der Lernfortschritt kann durch inkorrekte Antworten

negativ beeinflusst werden und auch das Vertrauen in zukünftige Lösungen dieser KI kann durch die Fehleranfälligkeit massiv vorbelastet sein.

Dieses Ergebnis entspricht den aus anderen Forschungsergebnissen ableitbaren Erwartungen. So kamen Plevris, Papazafeiropoulos und Rios im Rahmen ihrer Forschungsarbeit zu dem Ergebnis, dass ChatGPT 3.5 schlechter abschnitt als seine Nachfolgeversion 4.0 [41]. Bei der Datenerhebung der beschriebenen Forschungsarbeit erfolgte eine Unterteilung in zwei Datensets, frei zugängliche Aufgaben aus dem Internet und eigens entwickelte Aufgaben [41]. Die Auswertung dieser beiden Gruppen ergab mit ChatGPT 3.5 bei dem ersten Aufgabenset zu einem Anteil von 53,33% und bei dem zweiten Aufgabenset zu 42,22% korrekten Lösungen [41]. Um die Vergleichbarkeit der genannten Forschung mit der vorliegenden Arbeit zu erhöhen, ist es sinnvoll, diese beiden Datensets zusammen zu betrachten, da die Datenerhebung dieser Arbeit auf einer Kombination aus im Internet verfügbaren und nicht im Internet zugänglichen Aufgaben besteht [41]. Bei einem Vergleich der Ergebnisse der beiden Forschungsarbeiten ist zu beachten, dass die Datenerhebung dieser Arbeit anders als bei der Forschungsarbeit von Plevris, Papazafeiropoulos und Rios unterschiedliche Anteile an Aufgaben aus öffentlichen und nicht öffentlichen Quellen umfasst. Der durch die Ergebnisse ihrer Forschungsarbeit berechenbare Durchschnittswert von 47,78% steht deutlich im Kontrast zu der Quote von 12%, die den Anteil der durchschnittlich korrekten Antworten in dieser Forschungsarbeit darstellt.

Auch das von J. Paul durchgeführte Experiment liefert mit einer Quote von 44,44% korrekter Antworten ähnliche Ergebnisse [59, S. 15] wie die Forschungsarbeiten von Plevris, Papazafeiropoulos und Rios. Die gestellten Aufgaben entstammen auch verschiedenen Themenbereichen und wurden hauptsächlich aus dem Lernmaterial des Studiengangs Wirtschaftsinformatik der Technischen Hochschule Köln entnommen [59, S. 10 f.].

Die Abweichung der beiden Forschungsergebnisse von den in dieser Bachelorarbeit gewonnenen Erkenntnissen könnte auf einen Unterschied im Schwierigkeitsgrad der in der vorliegenden Studie eingesetzten mathematischen Aufgaben zurückzuführen sein. Die in den anderen Forschungen verwendeten Aufgaben weisen einen stärker gemischten Schwierigkeitsgrad auf und liegen somit teilweise deutlich unter dem Niveau von Abituraufgaben. So enthält die Studie von Plevris, Papazafeiropoulos und Rios beispielsweise einfache Prozentrechenaufgaben: *“A decade ago, the population of a city was 55,182 people. Now, it is 170% larger. What is the city’s current population?”* [41]. Zudem bestehen die Aufgabenstellungen in der vorliegenden Arbeit aus primär umfangreicheren Textaufgaben, bei denen der Schwerpunkt weniger auf der korrekten Berechnung, sondern vielmehr auf der Entwicklung eines geeigneten Rechenansatzes liegt. Demgegenüber enthalten die in den beiden beschriebenen Forschungen verwendeten Aufgaben eine klarere und direktere Aufgabenstellung, welche oft schon einen

eindeutigen Arbeitsauftrag enthält. Durch diesen Arbeitsauftrag wird meist schon direkt auf den anzuwendenden Lösungsansatz hingewiesen.

Ein weiterer Grund für den deutlichen Unterschied in den Quoten der Korrektheit kann in der in dieser Arbeit verwendeten Definition selbst begründet liegen. In der im Rahmen dieser Arbeit durchgeführten Studie wurden Aufgabenlösungen, die kleinere Fehler in der Ausgabe enthielten, als teilweise korrekt eingestuft. Dazu zählten mitunter auch Aufgaben, die beispielsweise das korrekte Ergebnis generierten, dieses allerdings in einer zusätzlichen Zusammenfassung am Ende nicht ausgaben, sondern dort ein anderes Ergebnis lieferten. Auch fallen hierin Aufgaben, bei denen eine Teilaufgabe nicht korrekt gelöst wurde. Würden diese Fehler nicht betrachtet, näherte sich das Ergebnis dieser Datenerhebung den Resultaten der anderen Studien deutlich an.

Unabhängig von den Unterschieden zu den Ergebnissen von früheren Studien zeigt die Forschung die mitunter großen Defizite dieser KI in Bezug auf die Ausgabe korrekter Lösungen. Die ausgegebenen Lösungen waren nicht nur mit einer hohen Wahrscheinlichkeit inkorrekt, sondern enthielten unlogische Schlussfolgerungen und Widersprüche in den Erklärungen und Aussagen. So kam es bei der Beantwortung teils zu Sätzen wie: „... *da b gerade ist, kann b nicht 2 sein* ...“ oder „... *b muss eine andere gerade Primzahl außer 2 sein* ...“. Die widersprüchlichen oder falschen Informationen sind nicht immer offensichtlich, sondern fügen sich mitunter glaubhaft in den Text ein. Das sorgt für eine gelegentlich schwierige Erkennbarkeit dieser Informationen. Vor allem für Personen, welche ihren Kenntnisstand über das Themengebiet noch festigen müssen, kann die Identifikation von unlogischen Schlussfolgerungen Schwierigkeiten bereiten. Die Aneignung der inkorrekten Konzepte kann sich zudem negativ auf den Lernerfolg auswirken.

Zudem ist beobachtbar, dass Lösungen und Begründungen an das von der jeweiligen KI als richtig identifizierte vorgegebene Ergebnis angepasst wurden. Dieses Phänomen wird besonders bei Aufgaben deutlich, die einen Beweis für ein bestimmtes Ergebnis fordern. Das Auftreten dieses Vorgehens liegt primär in der Konstruktion des Modells der KI begründet, welches auf die Erkennung von Mustern in den Daten ausgelegt ist und die Sequenz der Wörter in der Antwort basierend auf der Wahrscheinlichkeit generiert. Dies kann dazu führen, dass ein Ergebnis, welches mit hoher Wahrscheinlichkeit als korrekt eingestuft wurde, beispielsweise durch das Überspringen von logischen Schritten in der Erklärung oder unlogische Argumente unterstützt wird. Durch die Interpretation des ermittelten Ergebnisses im Einklang mit der in der Aufgabe als korrekt vorgegebenen Musterlösung wird der Fehler nicht basierend auf objektiven Fakten festgestellt und berichtet, sondern die fehlerhafte Angabe rationalisiert. So wurde eine von der KI bestimmte Zeitspanne von 40,75 Sekunden als „... *nahe an 9,8 Sekunden* ...“ beschrieben. ChatGPT ist nicht auf die Durchführung mathematischer und logischer Operationen spezialisiert und gibt die Lösungswege gemäß der durch

die Trainingsdaten geschaffenen Memorisierung aus. Das führt zur Bevorzugung des Rechenwegs, welcher in den Trainingsdaten stärker vertreten war. Problematisch an diesem Vorgehen ist primär die schlechte Erkennbarkeit des fehlerhaften Rechenwegs aufgrund des korrekten Ergebnisses der Aufgabe durch die Lernenden. Die dadurch resultierende Erhöhung des Risikos der Einprägung des vermeintlich richtigen Lösungswegs begünstigt das Erlernen von nicht korrekten Schritten und Methoden zur Erlangung dieser Antwort.

Von der Aufgabe bereitgestellte und zu beweisende Ergebnisse oder Aussagen wurden zusätzlich von ChatGPT 3.5 nicht zur Erbringung des Beweises verwendet. Dies führte zu einer Erhöhung der Komplexität der Berechnungen dieser Aufgaben. Dies wiederum resultierte entweder in der Nichtlösbarkeit der Aufgabe für diese KI oder zu umständlichen Erklärungen und Berechnungen, welche für Verständnisprobleme bei den Erklärungen für die Lernenden sorgen. In Anbetracht der Einschränkungen der numerischen Lösungsfähigkeit dieser ChatGPT-Version stellt die Wahl des Lösungsweges ein signifikantes Hindernis dar. Die KI wählt den Lösungsweg nicht entsprechend der höchsten Sicherheit für die Erlangung einer korrekten Lösung. Dies liegt in der Architektur der KI begründet, welche nicht die Korrektheit der Aussagen priorisiert, sondern basierend auf in den Trainingsdaten erkannten Mustern antwortet.

Auch die Plausibilität der durch die Nutzenden gegebenen Angaben oder der Aufgabenstellung wurde von der KI im Rahmen dieses Versuchs nicht hinterfragt. So wurde bei fehlerhafter oder nicht eindeutiger Aufgabenstellung weder eine Korrektur vorgenommen noch eine Nachfrage gestellt. Stattdessen wurde die generierte Antwort an die unlogische Aussage des Nutzenden angepasst, selbst wenn diese fehlerhaft war. Dies zeigt, dass die Angaben der Nutzenden von der KI scheinbar als korrekt und fehlerfrei akzeptiert werden. Eine Überprüfung der inhaltlichen Sinnhaftigkeit findet entweder nicht statt oder ermittelte Abweichungen werden übergangen, wodurch den fehlerhaften Vorgaben gefolgt wird und eine auf der Basis der Nutzerangaben passende Antwort generiert wird.

ChatGPT 3.5 versuchte gemäß der im Rahmen der Arbeit durchgeführten Datenauswertung bei mehreren Aufgaben, diese durch willkürliches Einsetzen von Werten zu lösen. Dieses Vorgehen ist ineffektiv und sorgt für lange und unübersichtliche Antworten, da bei allen Antworten, welche innerhalb des Versuchs auf dieser Lösungsmethode aufbauten, mehrere Versuche notwendig waren. Zusätzlich war die Lösungsmethode bei den gestellten Aufgaben nicht zielführend und lieferte keine korrekten Ergebnisse. Dies ist auf den Schwierigkeitsgrad der mathematischen Probleme zurückzuführen, bei denen Raten als Lösungsansatz unbrauchbar ist. Eine auf willkürlichem Raten basierende Problemlösung unterstützt nicht die Nachvollziehbarkeit der Antwort und bringt für den Einsatz als Mathematik-Tutor in diesem Zusammenhang keinen Mehrwert.

In dem durchgeführten Versuch war ChatGPT 3.5 die einzige KI, welche es angegeben hat, wenn sie eine Aufgabe nicht lösen konnte. Allerdings erfolgte eine solche Angabe durch die KI nur vereinzelt und lediglich unter der Begründung, dass die gestellte Aufgabe zu komplex sei, um sie nichtnumerisch zu lösen. Eine solche Angabe ist nützlich, da sie die Antwort auf die Frage, welche trotz dieser Angabe generiert wurde, besser einschätzbar macht. Diese Aussage trat allerdings auch bei einer Aufgabe auf, welche aus einem hilfsmittelfreien Teil entstammte und durchaus mit einfachen mathematischen Berechnungen und Umformungen lösbar war. So wurde bei einer Aufgabe, welche die Lösungen der Gleichung $2^{(x+2)} + 4^x = 4 + 8^x$ unter der Verwendung zweier verschiedener Lösungsansätze verlangte, folgende Antwort gegeben: „. . . *Leider ist es schwierig, diese Gleichung algebraisch weiter zu lösen. Man müsste wahrscheinlich numerische Methoden verwenden, um eine Näherungslösung zu finden.*“ Die Aufgabe sollte das grundlegende Verständnis der Potenzrechnung überprüfen und hatte aufgrund der Hilfsmittelfreiheit die sich in \mathbb{N}_0 befindenden Ergebnisse $x_1 = 0 \vee x_2 = 1$. Das Umformen der Potenzen ermöglichte die Lösung der Aufgabe ohne komplizierte mathematische Berechnungen im Kopf.

Aus der Umfrage geht grundsätzlich hervor, dass die von ChatGPT 3.5 bestimmten inkorrekten Ergebnisse eher als inkorrekt erkannt wurden. Dies zeigt, dass die von dieser KI generierten Antworten von den Befragten grundsätzlich kritisch hinterfragt wurden. Die Gesamtheit der Befragten hielt dennoch 26,19% der falschen Antworten für korrekt. Dabei erkannte die Gruppe der Studierenden mit 23,53% weniger falsche Aufgaben nicht als die Nachhilfe Nehmenden mit 37,5%. Dieser Unterschied kann darauf hindeuten, dass die Studierenden mehr Übung bei der Erkennung von Ungereimtheiten oder Fehlern in Antworten haben. Allerdings kann dies auch am Aufbau des Versuches liegen. Da den Nachhilfe Nehmenden die Aufgaben einzeln und teilweise an verschiedenen Tagen vorgelegt wurden, hatten sie anders als die Studierenden meist keine Vergleichsmöglichkeit mit den Lösungen der anderen KIs. Dennoch wurde bei beiden Gruppen ein erheblicher Anteil der inkorrekten Lösungen nicht als falsch erkannt. Bei der Verwendung einer KI in der Rolle eines Mathematik-Tutors ist es essenziell, dass die Anwendenden in der Lage sind, falsche Informationen zu erkennen. Wenn in die Lösung der KI mehr Vertrauen gelegt wird als in die eigene Lösungsfähigkeit, kann dies zu einer Beeinträchtigung des eigenen mathematischen Verständnisses zu dem Sachverhalt führen.

ChatGPT hat großes Potenzial als unterstützendes Werkzeug, den Lernprozess von Studierenden zu begleiten, da es die Möglichkeit bietet, eigens auf die Problemstellung zugeschnittene Antworten zu generieren und sich dabei direkt und individuell an die Bedürfnisse der Fragenden anzupassen. Die Möglichkeit, Fragen zu stellen, ohne eine Beurteilung durch andere befürchten zu müssen, kann essenziell zur Förderung einer offenen Kommunikation beitragen. Bei der Kommunikation mit einer KI anstelle eines menschlichen Partners oder einer Partnerin ist somit die Hemmschwelle, vermeidlich

„dumme“ Fragen zu stellen, deutlich niedriger, da eine negative Beurteilung durch die KI nicht zu befürchten ist. Stojanov stellte in ihrer Studie fest, dass der durch den Einsatz von ChatGPT 3.5 erzielte Lerneffekt für introvertierte Lernende größer ist als bei der Zusammenarbeit mit einem menschlichen Tutor oder einer Tutorin [57]. Dies kann darauf zurückzuführen sein, dass sich die Fragenden freier fühlen, ihre Unsicherheiten auszudrücken, und mehr Fragen stellen, die sie sonst möglicherweise zurückhalten würden. Jedoch lässt die hohe Wahrscheinlichkeit des Auftretens von Fehlern eine Empfehlung für den Einsatz von ChatGPT 3.5 in der Lehre nur stark beschränkt zu, da ein Risiko der Aneignung falscher Konzepte zur Lösung der Aufgaben besteht. Die Umfrage zeigt zusätzlich eine Unsicherheit der Befragten über die Fähigkeit zur Anwendung des durch die KI vorgestellten Lösungswegs auf ähnliche Aufgaben. Dies macht zusätzlich deutlich, dass das Ziel eines nachhaltigen Lerneffekts durch die von dieser KI generierten Antworten nicht erreicht wurde.

5.2 ChatGPT 4.0

Die Weiterentwicklung der ChatGPT-Version 4.0 zu der Vorgängerversion sorgte für eine höhere Erwartungshaltung bezüglich der Verbesserung der Qualität der Antworten. Die Datenauswertung bestätigt diese durch andere Forschungsergebnisse bekräftigte Vermutung. Allerdings ist auch beobachtbar, dass bei der Analyse von ChatGPT 3.5 identifizierte Eigenschaften teilweise auch bei ChatGPT 4.0 auftreten. Das Erscheinen dieser bereits für ChatGPT 3.5 beschriebenen Fehlerquellen zeigt sich bei ChatGPT 4.0 jedoch nur in stark abgeschwächter Form.

Die von ChatGPT 4.0 erzielte Rate der Korrektheit ist im Vergleich zu der Version 3.5 deutlich höher. Verglichen mit dem von Plevris, Papazafeiropoulos und Rios ermittelten durchschnittlichen Anteil von 20 % inkorrekten Lösungen [41] nähert sich die in dieser Arbeit festgestellte Fehlerquote von 17 % deutlich den bereits vorliegenden Forschungsergebnissen an. Dies zeigt eine signifikant größere Zuverlässigkeit bei der Rückgabe richtig gelöster Aufgaben und bestätigt dadurch die erhöhte Leistungsfähigkeit des nachfolgenden Modells im Vergleich zu seinem Vorgänger. ChatGPT 4.0 ist bei der Lösung der Aufgaben dazu fähig, die eigenen Rechnungen genauer auszuführen und sinnvolle Lösungswege zu ermitteln.

Trotz der signifikant gestiegenen Quote der Korrektheit der Antworten wurde bei der Untersuchung der Antworten ein punktuell auftretendes inkorrekt resultierendes Ergebnis beobachtet. Die Analyse der Aufgabenlösungen hat die Abhängigkeit des Auftretens falscher Antworten von der Aufgabe verdeutlicht. Besonders bemerkenswert ist die Konzentration der Fehler in diesen Durchgängen. So traten bei den dreifachen Aufgabenwiederholungen meist keine oder bei zwei bis drei der Wiederholungen Fehler

auf. Dieses Fehlercluster zeigt sich konsistent über die verschiedenen Aufgaben hinweg. Eine genauere Betrachtung der problembehafteten Aufgaben macht die offenkundige Schwierigkeit der KI mit im Internet frei verfügbaren, aber leicht abgeänderten Aufgaben deutlich.

Allerdings zeigt sich auch, dass im Gegensatz zu dem Vorgängermodell keine großen Probleme mit der korrekten Berechnung der Ergebnisse auftraten. Fehler ergaben sich hauptsächlich begründet durch die falsche Wahl des Rechenansatzes oder Lösungswegs. Die Wahl von unpassenden Lösungsansätzen ist dabei der primäre Grund für die fehlerhafte Lösung der Aufgaben. Der rechnerische Ansatz bildet das Fundament für eine gelingende mathematische Lösung und Erklärung. Eine korrekte Rechnung ist zwar nicht unwichtig, aber baut auf einem gut durchdachten Rechenansatz auf. Der Rechenansatz ist essenziell für die Lernenden, um ein Verständnis der Lösung der Aufgabe aufzubauen, da sie dadurch die Kompetenz erwerben, auch ähnliche Aufgaben zu lösen. Die KI wählte in der Datenerhebung bei Fehlern oft den falschen Rechenansatz. Dieses Problem könnte allerdings durch eine kommunikativere Bedienung abgeschwächt oder vermieden werden. So könnte die Aufgabenerklärung durch einen präziseren Prompt oder gezielte Nachfragen korrigiert werden. Dieses Vorgehen entspricht eher der für die KI als Standard zugrunde liegenden Kommunikationsform des Chats. Das legt den Schluss nahe, dass ChatGPT auch die besten Leistungen erzielt, wenn diese Form der Kommunikation verwendet wird. Dieser Schluss kann allerdings durch die durchgeführte Studie nicht bestätigt werden, da, wie im Kapitel 3 beschrieben, die Aufgabenstellung mit einem einheitlichen Prompt direkt übermittelt wurde. Durch die Art der Kommunikation, die in der Datenerhebung gewählt wurde, entstand kein längerer Dialog mit der jeweiligen KI.

Die bei der Vorgängerversion auftretenden widersprüchlichen und unlogischen Schlussfolgerungen und Erklärungen zeigten sich bei der Version 4.0 weitaus seltener. Allerdings waren die in dem Versuch auftretenden Argumentationsfehler schwerer zu identifizieren, da sie sich fast natürlich in die Erklärung einfügten und von richtigen Sachverhalten unterstützt wurden. Beispielsweise kam es bei einer Aufgabe, in der ermittelt werden sollte, an welchem der Eckpunkte das Dreieck seinen rechten Winkel hat, zu folgender Erklärung: „... *Das Skalarprodukt $\vec{CA} * \vec{AB}$ ist gleich null, was bedeutet, dass die Vektoren \vec{CA} und \vec{AB} orthogonal zueinander sind. Da diese Vektoren bei den Punkten A und C beginnen, in B aufeinandertreffen, ist der Punkt B der Scheitelpunkt des rechten Winkels im Dreieck. ...*“ Die professionelle Wirkung, welche die generierten Erklärungen durch den Einsatz von Fachbegriffen und die vorher korrekt durchgeführten Berechnungen erzielen, führt zu einer schwereren Entdeckbarkeit dieser fehlerhaften Lösung. Dies kann insbesondere bei dem Einsatz in der Lehre zu erheblichen Problemen führen, da es die Einschätzbarkeit der Aufgabe durch die Lernenden erschwert.

Inkorrekte oder mathematisch unsinnige Aufgabestellungen bereiteten dieser Version der KI zudem weiterhin Probleme. Wenn in der Aufgabe ein Beweis für ein konkretes Ergebnis gefordert wurde, welches aufgrund einer Manipulation an der Formel nicht bestätigt werden konnte, kam es dennoch zu einer Bestätigung dieses Ergebnisses durch die KI. Die häufige Wiederholung des Lösungsversuches durch andere Ansätze zeigt, dass ChatGPT 4.0 bei der Lösung dieses Sachverhaltes an seine Grenzen stieß. Auffällig bei der Wiederholung der Analyse ist, dass durchschnittlich zwei gescheiterte Analyseversuche unternommen wurden, bis die KI zu dem Schluss kam, dass das aus der Aufgabestellung hervorgehende Ergebnis korrekt wäre. Anders als ChatGPT 3.5 gab diese Version keine inkorrekte Begründung an, in der die Argumente an die Lösung angepasst wurden, und gab auch nicht den Rechenweg aus, der zu der Bestätigung der Ergebnisse geführt hatte. Dennoch wurde in zwei der drei Durchläufe das Kontrollergebnis eindeutig als korrekt kommuniziert, während bei dem verbleibenden Versuch nicht weiter auf diesen Teil der Aufgabe eingegangen wurde. Diese Beobachtung verdeutlicht, dass seitens der KI eine Problematik bei der Durchführung der Berechnung erkannt und auch in einem gewissen Maß kommuniziert wird, dennoch anschließend die zu erzielende Antwort bestätigt wird. Die KI legt augenscheinlich nicht primär den Wert auf eine zu den erzielten Ergebnissen passende Schlussfolgerung, sondern priorisiert die erwartete Ausgabe. Die nachstehende Abbildung zeigt einen Ausschnitt einer Aufgabe, auf die in diesem Abschnitt Bezug genommen wurde.

Lassen Sie uns diese Integration durchführen und die Gleichung nach t auflösen, um zu sehen, ob $t \approx 9,8$ Sekunden.

! Fehler bei der Analyse ▾

Die direkte Lösung der Gleichung für t , bei der $s(t) = 100$ Meter, ist mit den Standardmethoden nicht direkt umsetzbar. Dies weist auf die Komplexität der Gleichung hin, die aufgrund der Kombination von exponentiellen und polynomialen Termen schwer lösbar ist. Um dennoch eine Näherungslösung für t zu finden, bei der Läufer A 100 Meter zurücklegt, können wir numerische Methoden anwenden.

Ich werde jetzt einen numerischen Ansatz verwenden, um zu bestimmen, wie lange Läufer A für die 100 Meter benötigt, und überprüfen, ob dies ungefähr 9,8 Sekunden entspricht. [-]

! Fehler bei der Analyse ▾

Es scheint, dass ich einen Fehler bei der Anwendung der numerischen Methode gemacht habe, indem ich die Funktion inkorrekt für die Integration verwendet habe. Die korrekte Vorgehensweise wäre, die bereits integrierte Form von $s(t)$ direkt für die numerische Lösung zu verwenden, anstatt sie erneut zu integrieren. Ich werde nun einen korrekten Ansatz verwenden, um die Zeit zu berechnen, die Läufer A benötigt, indem ich die Formel für $s(t)$ direkt nutze.

Lassen Sie uns den Ansatz korrigieren und die erforderliche Zeit für Läufer A erneut berechnen.

[-]

! Fehler bei der Analyse ▾

Abbildung 5.1: ChatGPT 4.0 Aufgabenausschnitt mit Analysefehler

Grundsätzlich ist die Version 4.0 im Gegensatz zu ihrem Vorgängermodell in der Lage, bei Aufgaben eigene Fehler bei der Berechnung zu erkennen und diese zu kommunizieren. Die Kommunikation des festgestellten Problems bei der Bestimmung des Ergebnisses erhöht die Transparenz der generierten Erklärung. Diese Transparenz bei der Beantwortung ist für die Nutzenden von besonderer Relevanz, da sie die generierte Erklärung besser nachvollziehbar macht und das Vertrauen in die Antwort stärken kann. Nach der Feststellung eines Fehlers wurde das eigene Vorgehen durch die KI eigenständig überdacht und eine alternative Herangehensweise gewählt. Diese Selbstüberwachung und Anpassung des Vorgehens sind ein wichtiger Schritt in der Weiterentwicklung der KI, da durch sie Fehlerquellen erheblich reduziert werden können. Die erhöhte Quote an korrekten Antworten im Vergleich zu dem Vorgängermodell kann eng mit dieser Fähigkeit in Zusammenhang gebracht werden.

Bei der Strukturierung von Aufgaben mit mehrschrittigen Berechnungen oder Teilaufgaben neigte dieser Chatbot dazu, zwischen diesen Elementen der Aufgabe zu springen. Das sorgte oftmals für eine unordentliche Struktur und ein schlechtes Zurechtfinden in der Erklärung. Dies wird vor allem von dem in Kapitel 4.1.4 beschriebenen Hintergrund bestärkt, dass die Überschriften bei manchen Erklärungen nicht einheitlich formatiert waren, wodurch bestimmte Teile der generierten Antwort nur schlecht zu finden waren. Allerdings kann die schrittweise Erklärung auch als Vorteil für den Lerneffekt gesehen werden, wenn Lernenden erst ein kurzer Überblick über alle Teile der Aufgabe gegeben und anschließend schrittweise an die Aufgabenlösung herangeführt wird. Dadurch wird nicht die gesamte Lösung direkt präsentiert, sondern die Möglichkeit geboten, nur Abschnitte zu lesen und durch diese schrittweise präsentierten Hinweise selbstständig zu der Lösung zu gelangen.

Neben der ermittelten Verbesserung aller im Rahmen der Datenauswertung analysierten Kriterien war auch in der Umfrage eine deutlich positivere Einschätzung durch die Befragten zu erkennen. So schnitt ChatGPT 4.0 in der Umfrage in allen Bereichen besser ab als die Vorgängerversion. Die korrekt gelösten Aufgaben wurden mit einer hohen Wahrscheinlichkeit von den Befragten als richtig identifiziert. Die erheblich höhere Anzahl an korrekten Erklärungen in der Umfrage beeinflusste den Anstieg der korrekt eingeschätzten Aufgaben direkt und könnte auch auf die positivere Einschätzung anderer Faktoren Einfluss genommen haben. Die Tatsache, dass die Ergebnisse von Wolfram Alpha noch einmal besser eingeschätzt wurden, obwohl mehr fehlerhaft gelöste Aufgaben in der Umfrage vorkamen, zeigt allerdings, dass dieser Faktor keinen großen Einfluss ausgeübt hat.

Allerdings zeigt sich neben der überwiegend präzisen Erkennung von korrekten Lösungen auch die Tendenz dazu, inkorrekte Lösungen für richtig zu halten. Der prozentuale Anteil der fälschlicherweise für korrekt gehaltenen Lösungen ist im Vergleich zu der vorherigen Version stark gestiegen. So wurden insgesamt $\frac{2}{3}$ der falschen Antworten als

vollkommen richtig eingeschätzt. Bei der Betrachtung dieser Gesamtmenge wird zudem auch ein Unterschied zwischen den beiden Gruppen der Befragten deutlich: So erkannte die Gruppe der Studierenden die falschen Ergebnisse mit 50 % zuverlässiger als die Nachhilfe Nehmenden, bei denen keine der Antworten als falsch eingeschätzt wurde. Die Fähigkeit, die generierten Antworten bezüglich ihrer Korrektheit einzuschätzen, ist elementar für eine zielgerichtete und sichere Verwendung dieser Hilfsmittel in der Lehre. Die Erkennung von Fehlern in Erklärungen kann zu einem positiven Lerneffekt führen, da die Lernenden den Lösungsweg dazu nachvollziehen und überprüfen müssen. Der Prozess des Überdenkens und Korrigierens kann zu einer Festigung der mathematischen Konzepte führen, baut jedoch auf einem Grundverständnis der von der Aufgabe und der KI erstellten Lösung auf. Die Nichterkennung birgt das Risiko, dass falsche Konzepte übernommen und verfestigt werden, die später nur schwer korrigiert werden können. Dies kann nicht nur bei den in der Aufgabe behandelten mathematischen Themengebieten zu Problemen führen, sondern auch das Verständnis zukünftiger Themengebiete beeinträchtigen, da viele mathematische Themen aufeinander aufbauen.

Abgesehen von der Einschätzung der Korrektheit wird anhand der Umfrage die deutliche Steigerung des Verständnisses der Befragten durch die generierte Erklärung und das Ergebnis ersichtlich. Die Einschätzungen der Antworten dieser Version fielen in allen Aspekten der Umfrage positiv aus. Sie führten dazu, dass beide Gruppen ihr Verständnis der Aufgabe so einschätzten, dass sie das durch diese Antwort vermittelte Wissen auf weitere Aufgaben ähnlicher Art anwenden können. Diese Ergebnisse verdeutlichen, dass die Erklärungen von ChatGPT 4.0, basierend auf der Selbsteinschätzung der Befragten, das Hauptziel der nachhaltigen Vermittlung mathematischer Konzepte und der Hilfestellung bei mathematischen Aufgaben erreicht haben.

5.3 ChatGPT 4.0 mit Wolfram Alpha

Wolfram Alpha ist eine wissensbasierte Rechenmaschine, die in der Lage ist, präzise Antworten auf in natürlicher Sprache eingegebene Fragen zu geben. Da Wolfram Alpha die Priorität auf die Korrektheit der Ausgabe legt und Ergebnisse, bei denen Unsicherheiten hinsichtlich der Korrektheit bestehen könnten, nicht ausgibt, war die Erwartung an die Kombination von ChatGPT 4.0 und Wolfram Alpha eine Steigerung der Korrektheit der Antworten. Durch die Integration von Wolfram Alpha in ChatGPT 4.0 ist die Nutzung der Rechenleistung von Wolfram Alpha auch mit deutscher Sprachausgabe möglich, da ChatGPT die Übersetzungsarbeit übernimmt. Die Bereitstellung von Antworten in anderen Sprachen ist für Lernende besonders vorteilhaft, da selbst bei guten Sprachkenntnissen in Englisch mathematische Erklärungen oft leichter in der Muttersprache verständlich sind. Eine an die Nutzenden angepasste Sprachausgabe

kann somit zu einer erheblichen Unterstützung für das Verständnis und den Erfolg bei dem Lernen führen.

Durch die Notwendigkeit der Übersetzung der in deutscher Sprache formulierten Anfragen ins Englische sowie der Rückübersetzung der englischsprachigen Antworten ins Deutsche bei der Verwendung der Wolfram Alpha-Integration ergibt sich eine potenzielle zusätzliche Fehlerquelle. Dieser Übersetzungsprozess kann zu Ungenauigkeiten in der Interpretation und Wiedergabe der Ergebnisse führen, wodurch die Präzision der Ausgaben beeinträchtigt werden könnte. Die sprachliche Vermittlung zwischen den beiden Systemen birgt somit ein erhöhtes Risiko für Missverständnisse und semantische Abweichungen. Wie in Kapitel 2.2.3 beschrieben, wies ChatGPT 4.0 in der von Manakhimova et al. durchgeführten Forschung eine hohe Präzision bei der Übersetzung zwischen den Sprachen Englisch und Deutsch auf [51]. Obwohl ChatGPT 4.0 in der von Manakhimova et al. durchgeführten Forschung eine hohe Präzision bei der Übersetzung zwischen Englisch und Deutsch zeigte [51], gibt es bislang nur wenige Forschungsarbeiten, die diesen spezifischen Aspekt der Sprachübersetzung in Verbindung mit der Wolfram Alpha-Integration untersuchen. Aufgrund der begrenzten Forschungslage ist bisher keine verlässliche Bewertung der Zuverlässigkeit und Präzision von Übersetzungen in diesem Kontext möglich.

Die Datenanalyse zeigt entgegen den Erwartungen nur einen leichten Anstieg der Korrektheit der Antworten, welcher sich größtenteils durch eine Reduktion des Anteils solcher teilweise korrekten Lösungen ergibt. Bemerkenswert ist, dass die Fehler bei denselben Aufgaben wie bei ChatGPT 4.0 auftraten. Bei genauer Betrachtung der Antworten wird deutlich, dass die primäre Fehlerquelle der unvorteilhaft gewählte Lösungsweg für die Aufgabe war. Das legt den Schluss nahe, dass Wolfram Alpha bereits den inkorrekten Lösungsweg von ChatGPT 4.0 übermittelt bekommen hatte und somit die Aufgabe falsch gelöst wurde. Davis und Aaronson gelangen in ihrer Forschungsarbeit zu ähnlichen Ergebnissen und führen die geringe Verbesserung der Resultate auf ein von ihnen als „Schnittstellenproblem“ bezeichnetes Phänomen zurück [50]. Somit wäre eine Verbesserung durch das Plug-in nur beschränkt möglich, da es von ChatGPT bereits inkorrekte Lösungsansätze übermittelt bekommt und nicht die ursprüngliche Aufgabe. Das Potenzial der Zusammenführung von ChatGPT und Wolfram Alpha wird somit nicht voll ausgenutzt und bleibt ausbaufähig.

Auch Stephen Wolfram führt die geringe Verbesserung der korrekten Lösungsfähigkeit durch die Integration von Wolfram Alpha an ChatGPT 4.0 auf eine nicht optimal gestaltete Kommunikation über die Schnittstelle zurück [43]. Er äußerte zu den Ungenauigkeiten bei der Bedienung von Wolfram Alpha durch ChatGPT die Erwartung, dass ChatGPT die Fähigkeit zur Formulierung von Anfragen mit der Zeit durch Training verbessern werde [43]. Zusätzlich sei es notwendig, weitere Entwicklungsarbeit in die Fähigkeit von ChatGPT zum Erstellen eigener Tests mit dem

Ziel zur Selbstkontrolle zu investieren [43]. Mittlerweile dürfte der Großteil dieser Arbeit bereits implementiert worden sein, dennoch verdeutlicht dies, dass die Probleme erkannt wurden und eine Ausbesserung priorisiert wurde.

Dieser Schluss entsteht allerdings nicht nur bei den durch ChatGPT übermittelten Daten, sondern punktuell auch bei den durch Wolfram Alpha zurückgegebenen Ergebnissen. ChatGPT konnte bei manchen Aufgaben auch nach expliziter Aufforderung keinen expliziten Rechenweg angeben, um zu zeigen, wie die Lösung zustande gekommen ist. Dies könnte daran liegen, dass die von Wolfram Alpha zurückgegebenen Ergebnisse eventuell nicht für diese explizite Erklärung ausreichen. Anstelle einer Rechnung wurden häufig verschiedene Nutzungsmöglichkeiten von Wolfram Alpha als Lösungsweg vorgestellt oder es wurde der programmierte Wolfram Language Code zurückgegeben (siehe Abb. 4.4). Beides ist für die lehrerähnliche Beantwortung einer mathematischen Aufgabe nicht geeignet, um die Lösungsfähigkeit der Lernenden zu verbessern. Das Auftreten dieser Ausgaben führte primär zu der Verschlechterung der Bewertung der sprachlichen Erklärungselemente, da nicht davon ausgegangen werden kann, dass die Nutzergruppe den ausgegebenen Wolfram Language Code nachvollziehen kann. Zudem fehlte durch die Angabe dieser Elemente häufig ein funktionierender mathematischer Rechenweg oder eine für Lernende verständliche Erklärung. Dies sorgte primär für die schlechtere Bewertung der Vollständigkeit dieser Version. Beide Probleme könnten voraussichtlich durch gezielte Nachfragen seitens der Anwendenden gelöst werden. Eine Verbesserung in der beidseitigen Kommunikation über die Schnittstelle würde somit zu einer Optimierung in mehreren Bereichen führen.

Die Anbindung von Wolfram Alpha verbessert die Fähigkeit von ChatGPT 4.0, Berechnungen durchzuführen, und führt durch die Erkennung von Ungereimtheiten zu einer Selbstkorrektur. Auch Fehler in der Nutzereingabe wurden zielsicher erkannt und sogar angesprochen. Dadurch war diese Version die erste, welche bei manipulierten Aufgabenstellungen erkannte, dass das mitgelieferte Kontrollergebnis falsch war, und dies mit einem Fehler in der Aufgabenstellung in Zusammenhang brachte. Dies ist ein erheblicher Vorteil gegenüber den anderen Versionen, da sowohl eine inkorrekte Bestätigung des Kontrollergebnisses vermieden als auch die Fehlerquelle zielsicher identifiziert wurde. So kam es zu Aussagen wie: „... *Nach erneuter Berechnung mit exakten Koeffizienten zeigt sich, dass Läufer A ungefähr 26,83 Sekunden benötigt, um die 100-Meter-Distanz zu überwinden. Dies widerspricht der ursprünglichen Annahme, dass Läufer A etwa 9,8 Sekunden benötigt. Es scheint, als ob ein Missverständnis oder ein Fehler in der Annahme vorliegt. Auf Basis der gegebenen Geschwindigkeitsfunktion $v_a(t) = 0,25t + 10 \cdot (1 \cdot e^{-t})$ ergibt die Berechnung eine deutlich längere Zeit ...*“ Durch diese zielsichere Erkennung wird das Vertrauen in die Antworten der KI gestärkt und unnötige Fehler werden vermieden. Zudem scheint die Verwendung von Wolfram Alpha eine unkritische Bestätigung des erwarteten Ergebnisses erstmals zu verhindern.

Die Ergebnisse der Umfrage zeigen eine tendenziell positivere Einschätzung der fachlichen Kompetenz von ChatGPT 4.0 mit der Integration von Wolfram Alpha im Vergleich zu einer Nutzung ohne dieses Plug-in. Diese Bewertung spiegelt sich besonders in dem Anstieg der positiven Einschätzung zur Korrektheit des Ergebnisses wider. Diese Einschätzung ist allerdings unter Berücksichtigung der höheren Anzahl von in der Umfrage bei dieser Version vorkommenden fehlerhaften Antworten problematisch. Eine korrekte Einschätzung der Aufgabenlösungen durch die Befragten hätte zu einem Rückgang der als korrekt eingestuften Ergebnisse führen müssen, allerdings wurden insgesamt 72,22 % der inkorrekten Ergebnisse nicht für falsch gehalten. Bei einer Unterscheidung in die beiden Untergruppen ergibt sich, dass 58,33 % der Studierenden eine inkorrekte Antwort für korrekt hielten. Die Nachhilfe Nehmenden haben alle fehlerhaften Antworten als korrekt eingeordnet. Damit wurden weniger falsche Antworten identifiziert als bei den von ChatGPT 4.0 generierten Antworten. Die weitere Zunahme der bei ChatGPT 4.0 schon hohen Falschidentifikation der Antworten kann auf verschiedene Gründe zurückzuführen sein.

So erschwerte die zunehmend plausiblere Darstellung der Antworten durch die KIs möglicherweise die Identifizierbarkeit der falschen Antworten. Im Kontext der Verwendung des Wolfram Alpha-Plug-ins ist dieser Ansatz allerdings fragwürdig, da das Plug-in nicht für die Generierung der Antwort an die Nutzenden zuständig ist. Obwohl das Plug-in auf die fachlich fundierte Generierung von Antworten spezialisiert ist, ist es nicht primär für die Art der Präsentation verantwortlich und modifiziert somit auch die Darstellung der Antwort nicht.

Des Weiteren liegt bei der Betrachtung des weiteren Anstiegs bei der Anbindung von Wolfram Alpha der Schluss nahe, dass den fortgeschrittenen Versionen von dem Nutzerkreis eine höhere Kompetenz zugeschrieben wird. So können die der Version 4.0 insbesondere bei der Kombination mit einer fundierten Wissensmaschine wie Wolfram Alpha zugeschriebenen Kompetenzen aufgrund des verwendeten Systems unterschiedlich stark wahrgenommen werden. Diese wahrgenommenen Kompetenzen führen zu einem höheren Vertrauen in die Einschätzung der Korrektheit der Antworten, selbst wenn diese nicht richtig sind. Durch die überzeugende Informationsdarstellung durch ChatGPT 4.0 und die Verwendung einer auf wissenschaftlichen und fundierten Daten basierenden Plattform wird eine Vertrauensbasis geschaffen, welche bei den Nutzenden eine Neigung zu weniger kritischen Bewertungen der präsentierten Antworten hervorrufen kann.

5.4 Einschränkungen der Datenerhebung

Die Ergebnisse dieser Forschungsarbeit sind im Kontext der mit den durchgeführten Forschungs- und Datenerhebungsmethoden verbundenen Einschränkungen zu betrachten. Um die einheitliche Bewertung bei der Datenanalyse zu gewährleisten, wurden vor der Durchführung die verschiedenen Kriterien und deren Bewertungsstufen definiert. Während der Durchführung der Datenauswertung wurden jedoch bis dahin unbekannte Eigenheiten der verschiedenen Versionen von ChatGPT ermittelt, deren passende Zuordnung zu den Kriterienstufen eine nachträgliche Anpassung dieser Stufen notwendig machte. Durch diese Anpassung wurde eine nachträgliche Überarbeitung der bis dahin bereits zugeordneten Aufgaben notwendig. Die durchgeführte Überarbeitung sollte eine einheitliche Bewertung der Aufgaben und damit die Vergleichbarkeit dieser Ergebnisse gewährleisten.

Die Möglichkeit einer unbewussten Beeinflussung der durchgeführten Datenanalyse durch die persönliche Meinung stellt ein Risiko dar. Für die Wahrung der Objektivität der Ergebnisse und der Selbstkontrolle wurde die Arbeit mit anderen Forschungsergebnissen verglichen. Zudem wurde eine von der Datenauswertung unabhängige Umfrage durchgeführt, deren Ergebnisse neben der primären Funktion der Gewinnung neuer Einsichten auch der Ermittlung großer Abweichungen dienten. Diese Schritte trugen zu der Sicherstellung einer nicht alleine auf den subjektiven Perspektiven der Forschenden basierten Interpretation bei.

Der stark beschränkte Kreis der Teilnehmenden, in dem die Umfrage durchgeführt wurde, stellt eine große Einschränkung für die Gültigkeit der Ergebnisse dar. Die Gruppe der Studierenden bestand hauptsächlich aus Mitgliedern der Technischen Hochschule Köln, welche an der Fakultät für Informatik und Ingenieurwissenschaften studierten. Allerdings konnten durch die fakultätsweite Versendung per E-Mail auch Mitarbeitende an dieser Umfrage teilnehmen. Somit wird durch diese Gruppe eine eingeschränkte und nicht allgemeingültige Sichtweise dargestellt, was eine Bezugnahme auf andere Bereiche, außerhalb dieser Hochschule, stark einschränkt. Die Gruppe der Schülerinnen und Schüler stammte aus unterschiedlichen Schulformen und Klassen, hatte allerdings gemein, dass sie Förderunterricht in dem Fach Mathematik in Anspruch nahmen. Der kleine Kreis und die regionale Nähe der Schulen sorgen allerdings auch hier für keine große Vielfalt. Die Studie eignet sich in dieser Hinsicht somit nicht für eine allgemeingültige Schlussfolgerung. Die Fragen wurden auf Deutsch gestellt; die Auswirkungen, welche die verwendete Sprache auf die generierte Antwort hat, sind nicht klar, was eine Eingrenzung der Forschungsergebnisse auf den deutschsprachigen Raum notwendig macht.

Die in der Datenerhebung gestellten Fragen wurden mit einem einfachen Prompt verfasst und es wurde auf Nachfragen verzichtet. Dies geschah aus den in Kapitel 3

beschriebenen Gründen. Allerdings stellt sich die Frage, wie sich eine direkte Kommunikation mit der KI auf die Antwortqualität auswirkt. Zukünftige Forschungsarbeiten sollten diesen Punkt stärker betrachten, da diese Art der Nutzung von ChatGPT das Potenzial hat, effizienter zu sein und qualitative und individuelle Antworten zu generieren. Der schnelle Fortschritt in der Entwicklung dieser Technologie macht eine Beobachtung der Veränderung dieses Bereiches notwendig. Es gibt noch nicht genügend Forschungsarbeiten in diesem Feld, was die Einschätzung des Einflusses erschwert, welchen der Einsatz von Chatbots wie ChatGPT auf die Bildung und den zukünftigen Prozess des Lernens hat.

6 Fazit

Diese Arbeit widmete sich der zentralen Frage, wie die Versionen von ChatGPT als Mathematik-Tutor fungieren und dadurch Schüler/-innen und Studierende beim Lernprozess unterstützen können. Diese Fragestellung ging aus dem Titel dieser Arbeit hervor und bildet den Ausgangspunkt für den weitergehenden Vergleich der Qualität und Effektivität der generierten Antworten. Unter Berücksichtigung der in Kapitel 2.3.1 beschriebenen Lerntheorie des Konstruktivismus wird deutlich, warum ChatGPT als gelungene Ergänzung zum Lernprozess betrachtet werden kann. Der Konstruktivismus geht davon aus, dass Wissen nicht passiv erlernt werden kann, sondern aktiv und individuell von den Lernenden konstruiert werden muss. So müssen alle Lernenden auf der Grundlage ihrer eigenen Erfahrungen und ihres Vorwissens ihr neues Wissen aufbauen. Ausgehend von diesem Ansatz stellt ChatGPT eine gelungene Unterstützung bei dem Lernprozess dar, da die KI in der Lage ist, auf die Lernenden zugeschnittene Antworten zu geben. Dabei ist der zeitliche Faktor beim Stellen der Frage unerheblich, da ChatGPT jederzeit verfügbar ist. Diese beiden ausschlaggebenden Faktoren ermöglichen es den Lernenden, den Lernprozess an ihre eigene Geschwindigkeit und ihre eigenen Bedürfnisse anzupassen, und machen ChatGPT somit zu einem vielversprechenden Werkzeug. Für den Einsatz bei der eigenständigen Vertiefung im Selbststudium bietet ChatGPT eine einfache Möglichkeit, auf unterstützende Ressourcen zuzugreifen, welche aktiv bei der Klärung von Verständnisproblemen Unterstützung bieten. Die dauerhafte und zeitunabhängige Verfügbarkeit dieser Ressource stellt zusätzlich einen großen Vorteil dar.

Aufgrund der größer werdenden Rolle von künstlicher Intelligenz im Bildungsbereich ist es von zentraler Bedeutung, die Qualität der generierten Antworten und die Meinungen und Einschätzungen der Zielgruppe zu untersuchen. Die dafür durchgeführte Antwortenuntersuchung und Umfrage ermöglichten die Analyse der durch die verschiedenen Versionen von ChatGPT gegebenen Antworten hinsichtlich der in der Einleitung vorgestellten Forschungsfrage. Das Ziel dieser Arbeit war somit die Ermittlung der Unterschiede in der Effektivität der verschiedenen ChatGPT-Versionen, insbesondere in Kombination mit der Integration von Wolfram Alpha, als digitale Mathematik-Tutoren für die Zielgruppe der Schüler, Schülerinnen und Studierenden.

Die Analyse der Antworten zu den an ChatGPT gestellten mathematischen Aufgaben zeigte eine klare Leistungssteigerung in der Qualität der erklärenden Antworten

von der Version 3.5 zu der Version 4.0. ChatGPT 4.0 erzielte in allen geprüften Bereichen bessere Ergebnisse als die Vorgängerversion, wobei besonders stark die Korrektheit der Aufgabenlösung sowie die angemessene und korrekte Verwendung der Sprache und Terminologie hervortraten. Durch die deutliche Verringerung der bei ChatGPT 3.5 auftretenden Problemquellen stellt ChatGPT 4.0 im Vergleich zu ChatGPT 3.5 eine klare Verbesserung dar und eignet sich besser für den Einsatz als mathematische Lernunterstützung. Allerdings führt das seltenere Auftreten der Fehler und vor allem das passendere Integrieren von falschen Interpretationen und Schlussfolgerungen zu einem Risiko bei dem Einsatz von dieser KI in der Lehre. Die Version 3.5 gestaltete das Erkennen von unlogischen oder falschen Aussagen einfacher, da sich diese meist weniger gut in die Antwort integrierten als bei der Nachfolgeversion 4.0. Die durchgeführte Umfrage verdeutlichte dieses Problem, da die Befragten die falschen Ergebnisse der Version 4.0 deutlich schlechter als falsch identifizierten als bei deren Vorgänger. Eine Steigerung dieses Phänomens ist bei der Integration von Wolfram Alpha erkennbar geworden, wodurch die Schlussfolgerung entstand, dass das Vertrauen der Befragten in die KI maßgeblich durch die Wahrnehmung der vermeintlich gesteigerten Leistungsfähigkeit beeinflusst wird. Diese Steigerung des Vertrauens kann in der Annahme begründet liegen, dass die Anbindung von Wolfram Alpha, welches als kompetentes und wissenschaftliches System bekannt ist, automatisch zu verlässlicheren Auskünften führe. Eine Verstärkung dieser Annahme erfolgt durch die Gleichsetzung von neueren Technologien, in diesem Fall ChatGPT 4.0 gegenüber der älteren Version 3.5, mit einer automatisch höheren Korrektheit der Lösungen. Die Steigerung des Vertrauens in die Version 4.0, insbesondere mit der Integration von Wolfram Alpha, führt zu einer Neigung der Benutzenden, die Resultate eher zu akzeptieren, und zu einer geringeren kritischen Überprüfung der gegebenen Antworten.

Die Umfrage zeigt deutlich, dass die Gruppe der Nachhilfe Nehmenden größere Schwierigkeiten hatte, die fehlerhaften Antworten der KIs zu identifizieren. Eine mögliche Erklärung für dieses Ergebnis kann in der aufgrund des jüngeren Altersdurchschnitts dieser Gruppe noch weniger starken Ausprägung der Fähigkeit zum kritischen Denken begründet liegen. Zusätzlich ist bei dieser Gruppe von einem schlechteren fundamentalen Grundwissen in dem Bereich der Mathematik auszugehen, da Nachhilfe in dem Großteil der Fälle erst bei einem Defizit in dem betreffenden Fach in Anspruch genommen wird. Dahingegen hat die Gruppe der Studierenden bereits einen höheren Bildungsgrad erreicht und es kann von fortgeschritteneren Mathematikkenntnissen ausgegangen werden. Bei der Erkennung von Fehlern in der Aufgabenlösung stellen diese mathematischen Kenntnisse einen großen Vorteil dar. Die grundlegende Aneignung neuer mathematischer Lehrinhalte sollte vor diesem Hintergrund nicht alleine durch ChatGPT erfolgen. Die Ergebnisse dieser Forschung zeigen, dass die Begleitung des Lernprozesses durch Lehrpersonal unerlässlich ist, um ein korrektes und grundlegendes Verständnis zu bilden. Um zu verhindern, dass sich unbemerkt falsche Konzepte setzen,

ist eine menschliche Begleitung, wie beispielsweise in Form von Unterricht, sinnvoll, da dadurch eine frühzeitige Erkennung, Besprechung und Korrektur dieser Fehler ermöglicht werden.

Die besten Ergebnisse in Bezug auf die Korrektheit der gegebenen Antworten generierte ChatGPT 4.0 mit der Wolfram Alpha-Integration. Dabei wies der Anteil der korrekten Lösungen nur geringfügige Veränderungen gegenüber der Version 4.0 ohne dieses Plug-in auf. Die Ermittlung der nur geringen Verbesserung durch die Anbindung des Plug-ins entsprach nicht den Erwartungen an das Ergebnis der Datenerhebung, da von einer deutlichen Steigerung der Rechenleistung und mathematischen Lösungsfähigkeit hinsichtlich der mathematischen Aufgaben ausgegangen wurde. Das Nichteintreten dieser deutlichen Verbesserung könnte möglicherweise auf eine nicht optimal genutzte Kommunikation über die Schnittstelle dieser beiden Anwendungen zurückzuführen sein. Dadurch könnte ChatGPT 4.0 einen ungünstigen Lösungsansatz oder eine Fehlinterpretation der Aufgabenstellung an Wolfram Alpha übertragen, wodurch die zurückgegebene Antwort für das über die Aufgabenstellung übermittelte Ausgangsproblem unpassend war. Obwohl diese ChatGPT-Version nur geringfügige Verbesserungen bei der Korrektheit aufwies, konnte durch diese KI jede der gestellten Fragen bei der dreifachen Wiederholung mindestens einmal korrekt gelöst werden. Somit gab es in dem in dieser Studie verwendeten Aufgabenpool keine Fragestellung, die für diese KI unlösbar gewesen wäre, was im Vergleich zu den anderen überprüften KIs zu einer verbesserten Reproduzierbarkeit führte.

Allerdings verdeutlicht die in der Datenerhebung ermittelte Fehlerquote die Einschränkungen der Verlässlichkeit bei den durch die KI ermittelten Aufgabenlösungen. Der Anteil der fehlerhaften Ergebnisse war bei keiner der getesteten Versionen gering genug, um diese als ausreichend zuverlässig für den vollkommen sicheren Einsatz im Bildungsbereich einzustufen. Das Ignorieren oder ein Verbot der Anwendung dieser Technologie im Kontext der mathematischen Bildung sollte jedoch nicht die aus dieser Feststellung resultierende Konsequenz darstellen. ChatGPT ist mittlerweile weit verbreitet und über das Internet leicht zugänglich, was eine selbstständige Nutzung durch Lernende fördert. Stattdessen sollte die verantwortliche Verwendung dieser Technologie im Vordergrund stehen und die grundlegenden Kompetenzen, welche für den Gebrauch von Medien wie ChatGPT fundamental notwendig sind, sollten gefördert werden. Gerade die Möglichkeit, dass die von der KI zurückgegebenen Antworten Fehler enthalten, macht die Schulung des kritischen Hinterfragens bei den Lernenden umso wichtiger. Für den sicheren Einsatz in der Bildung und die Erzielung eines nachhaltigen Lerneffekts ist diese Fähigkeit von fundamentaler Bedeutung. Die Identifizierung der durch die verschiedenen Versionen unterschiedlich „gut“ in die Antworten integrierten falschen Informationen ist entscheidend, da das vertrauenswürdige Erscheinungsbild dieser Informationen das Risiko birgt, Missverständnisse bei der Bildung mathematischer Konzepte hervorzurufen. Das aus einer fehlerhaften Antwort entstehende Verständnis

kann langfristig die Entwicklung mathematischer Kenntnisse der Lernenden gefährden. Darum ist es von Relevanz sicherzustellen, dass Lernende den Lösungsweg nachvollziehen und überprüfen, da der Prozess der Analyse der generierten Antwort nicht nur das tiefergehende Verständnis mathematischer Prinzipien fördert, sondern auch die Fähigkeit stärkt, eigenständig komplexere Probleme zu lösen.

Die Vermittlung einer grundlegenden Medienkompetenz und damit auch der Umgang mit ChatGPT sollten stärker in die Lehre eingebunden werden, um den sicheren und effizienten Umgang, aber auch die Risiken des Gebrauchs von ChatGPT zu vermitteln. Die Begleitung durch ChatGPT kann positive Auswirkungen haben, der Umgang muss allerdings separat geschult werden, um eine sichere Verwendung zu gewährleisten. Eine solche Schulung des korrekten und zielführenden Umgangs mit ChatGPT ist von Relevanz, da die freie Verfügbarkeit von digitalen Hilfsmitteln tendenziell zu deren Nutzung führt. Ein angeleiteter und bewusster Umgang mit diesen Hilfsmitteln bietet Chancen für die zukünftige Gestaltung eines individualisierten Lernprozesses. Um dies zu erreichen, sind eine zusätzliche Förderung des kritischen Hinterfragens im Umgang mit diesen Medien und eine kontrollierende Unterstützung von Lehrpersonen notwendig, um die Vorteile der digitalen Hilfsmittel optimal auszuschöpfen, ohne die schutzlose Aussetzung gegenüber potenziellen Risiken in Kauf zu nehmen.

Die von den KIs generierten Antworten enthielten nicht immer eine vollständige Erklärung oder Berechnung aller als notwendig angesehenen Schritte der Aufgabe. Dies wurde allerdings von den Befragten nicht als Hindernis für das Verständnis angesehen. Die durchgeführte Umfrage offenbarte eine vornehmlich für ChatGPT 4.0, sowohl unter der Verwendung von Wolfram Alpha als auch ohne, positive Einschätzung des durch die Antwort gewonnenen Lerneffekts und Verständnisses. Die gegebenen Erklärungen zu den gestellten Aufgaben wurden bei diesen Versionen überwiegend als vollständig angesehen, was die Fähigkeit der KI hervorhebt, wesentliche Schritte der Erklärung zu erkennen und zu thematisieren. Zugleich bietet die Chatfunktion eine kontinuierliche Möglichkeit zum Stellen von Nachfragen, was eine individuell zugeschnittene Vertiefung des Verständnisses ermöglicht und sowohl Verständnisprobleme beheben als auch Beschreibungen weiter ausführen kann. Die hohe Sicherheit der Befragten zum Transfer des vermittelten Wissens auf ähnliche Aufgaben zeigt, dass die fundamentale Umsetzung der Aufgabenerklärung durch die KI gelungen ist. Somit erreichen vor allem die beiden Varianten von ChatGPT 4.0 das grundlegende Ziel der verständlichen Erklärung mathematischer Aufgaben. Die Auswahl der Sprache und der terminologischen Begrifflichkeiten erfolgte laut der Beurteilung durch die Befragten klar und verständlich. Dies macht die grundsätzliche Eignung dieses Systems zur verständlichen Vermittlung mathematischer Konzepte deutlich. Eine über die Selbsteinschätzung der Befragten hinausgehende Überprüfung der Validität dieser getroffenen Aussagen wurde nicht vorgenommen, was zu einer Einschränkung der Verwertbarkeit dieser Aussagen führt.

Die für die Verwendung von ChatGPT vorgesehene Art der Kommunikation ist der Gebrauch per Chat. Folglich sollte die KI unter Verwendung dieses Kommunikationsmediums die besten Ergebnisse erzielen. Die Möglichkeiten, die durch eine längere Kommunikation entstehen, wurden in der in dieser Forschungsarbeit durchgeführten Datenerhebung nicht berücksichtigt, da nicht von einer natürlichen Simulation eines Kommunikationsablaufes zwischen der Zielgruppe und den KIs ausgegangen werden konnte. Die Kommunikation mit der KI hat das Potenzial, eine Steigerung der Korrektheit der Antworten zu ermöglichen. Zudem gibt sie die Möglichkeit, Verständnisprobleme zu erörtern und das eigene Wissen in die Konstruktion eines Lösungsweges einfließen zu lassen. So kann beispielsweise die Angabe des im Kontext der Aufgabe behandelten Schulthemas schon einen Hinweis auf mögliche Lösungswege bieten. Die Auswirkungen einer solchen Kommunikation auf die Qualität der Antworten sind allerdings nur spekulativ und bedürfen weiterer Nachforschungen.

Auch die Frage, in welchem Umfang die Notwendigkeit der sprachlichen Übersetzung zwischen der Nutzersprache und der englisch formulierten Anfrage an Wolfram Alpha die Qualität der zurückgegebenen Ergebnisse beeinflusst, bedarf weiterer Forschung. Der aktuelle Forschungsstand zur Übersetzungsfähigkeit von ChatGPT 4.0 gibt zwar erste Einblicke, reicht jedoch nicht aus, um fundierte Schlussfolgerungen zu ziehen. Zudem ist weitgehend unerforscht, inwieweit sich die Qualität der generierten mathematischen Lösungen verändert, wenn identische Aufgaben in verschiedenen Sprachen gestellt werden. In zukünftigen Studien sollte die Beeinflussung dieser sprachlichen und kontextuellen Faktoren auf die Antwortenqualität und damit auf die Einsatzfähigkeit von ChatGPT mit der Wolfram Alpha Integration für die Beantwortung mathematischer Fragestellungen untersucht werden. Dies würde es ermöglichen, Unterschiede in der Wirksamkeit dieser Technologien für Lernende verschiedener Sprachen zu vergleichen, um festzustellen, inwiefern alle Lernenden unabhängig von ihrer Sprache von diesen Technologien profitieren können.

Die kontinuierliche Weiterentwicklung und schnellen Fortschritte von ChatGPT seit der Einführung der Version 3.5 Ende November 2022 demonstriert die großen Fortschritte, welche diese Technologie in dieser kurzen Zeitspanne erzielt hat und lassen eine zukünftige weitere Steigerung der Antwortenqualität erwarten. Die Forschungsergebnisse zeigen die grundsätzliche Eignung von ChatGPT für die Erklärung von mathematischen Aufgaben auf Abitur- und Hochschulniveau, führen jedoch zur Identifizierung von Verbesserungsbedarf bei den generierten Antworten. Durch diese laufenden Veränderungen ändern sich auch die Grenzen der Einsatzmöglichkeiten und die Qualität der Antworten der verschiedenen Versionen von ChatGPT stetig. Dies ist vor allem mit dem Hintergrund des großen Potenzials, welches diese Technologie für den Einsatz im Bildungsbereich hat, ein wichtiger Faktor. Um dieses Potenzial vollumfänglich auszuschöpfen, sind die Beobachtung der weiteren Entwicklung und gezielte Forschungsarbeiten notwendig. In Anbetracht der kontinuierlichen Fortschritte und

der damit verbundenen ständigen Veränderungen dieses Bereiches besteht ein großes Potenzial für die zukünftige Einsetzbarkeit dieser Technologie.

Literatur

- [1] H. E. Klein und S. Schöpfer-Grabe, *Was ist Grundbildung? Bildungstheoretische und empirische Begründung von Mindestanforderungen an die Ausbildungsreife*. Institut der deutschen Wirtschaft (IW), Köln, 2012, S. 5, 8.
- [2] D. C. Rheinheimer, B. Grace-Odeleye, G. E. Francois und C. Kusorgbor, „Tutoring: A Support Strategy for At-Risk Students.“ *The Learning Assistance Review*, Jg. 15, S. 23–34, 2010.
- [3] D. Schulportal. „PISA-Studie: Die wichtigsten Ergebnisse und Reaktionen.“ Zugriff am 18. Juni 2024. (2023), Adresse: <https://deutsches-schulportal.de/bildungswesen/die-zehn-wichtigsten-ergebnisse-der-pisa-studie/#die-ergebnisse-der-pisa-studie-2022>.
- [4] OpenAI. „OpenAI: Introduction.“ Zugriff am 19. Juni 2024. (2024), Adresse: <https://platform.openai.com/docs/introduction>.
- [5] T. Holoyad, T. Schmid und W. Hildesheim, *Künstliche Intelligenz managen und verstehen: Der Praxis-Wegweiser für Entscheidungsträger, Entwickler und Regulierer* (Beuth Innovation), 1st ed. Beuth Verlag, 2023, ISBN: 3-410-31457-1.
- [6] J. Korteling, G. C. van de Boer-Visschedijk, R. A. Blankendaal, R. C. Boonekamp und A. R. Eikelboom, „Human- versus Artificial Intelligence,“ *Frontiers in artificial intelligence*, 2021.
- [7] J. Weizenbaum, „ELIZA—a computer program for the study of natural language communication between man and machine,“ *Communications of the ACM*, Jg. 9, Nr. 1, S. 36–45, Jan. 1966, ISSN: 0001-0782. DOI: 10.1145/365153.365168. Adresse: <https://doi.org/10.1145/365153.365168>.
- [8] M. Adam, M. Wessel und A. Benlian, „AI-based chatbots in customer service and their effects on user compliance,“ *Electronic Markets*, 2020. DOI: 10.1007/s12525-020-00414-7.
- [9] S. Kim, J. Goh und S. Jun, „The use of voice input to induce human communication with banking chatbots,“ in *Companion of the 2018 ACM/IEEE International Conference*, 2018, S. 151–152.

-
- [10] V. Velasco, K. D. Setiawan, R. R. Sanjaya, M. S. Anggreainy und A. Kurniawan, „AI Chatbot Technology to Predict Disease: A Systematic Literature Review,“ in *2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IEEE, 2023, S. 97–101. DOI: 10.1109/AiDAS60501.2023.10284717.
- [11] S. Malodia, N. Islam, P. Kaur und A. Dhir, „Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants?“ *IEEE Transactions on Engineering Management*, Jg. 71, S. 491–505, 2024. DOI: 10.1109/TEM.2021.3117884.
- [12] K. Kiesl und N. Bödeker. „Konzept iPad-Klasse.“ Zugriff am 21. Juni 2024. (2015), Adresse: <https://ipadklassedotorg.files.wordpress.com/2012/05/konzept-ipad-klasse-2015.pdf>.
- [13] B. Engelke und U. Engelke, *ChatGPT – Mit KI in ein neues Zeitalter: Wie KI-Tools unser Leben und die Gesellschaft verändern*, 2024. Aufl. Frechen: mitp-Verlag, 2023, ISBN: 9783747507582.
- [14] C. Leiter, R. Zhang, Y. Chen u. a., „Chatgpt: A meta-analysis after 2.5 months,“ *Machine Learning with Applications*, Jg. abs/2302.13795, 2023. DOI: 10.48550/arXiv.2302.13795.
- [15] J. Howarth, *Most Visited Websites In The World*, Zugriff am 23. Juni 2024, 1. Juni 2024. Adresse: <https://explodingtopics.com/blog/most-visited-websites>.
- [16] Studienkreis. „Verteilung der Studienkreis-Schüler nach Fächern.“ Zugriff am 23. Juni 2024. (), Adresse: <https://www.studienkreis.de/unternehmen/presse/zahlen-und-fakten/studienkreis-zahlenmaterial/>.
- [17] Aristotle, *Politics*. Kitchener: Batoche Books, 1999, Translated by Benjamin Jowett.
- [18] A. M. Turing, „I.—Computing machinery and intelligence,“ *Mind*, Jg. LIX, Nr. 236, S. 433–460, 1950, ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433. eprint: <https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf>. Adresse: <https://doi.org/10.1093/mind/LIX.236.433>.
- [19] E. Neufeld und S. Finnestad, „The mismeasure of machines,“ in *Advances in Artificial Intelligence: 29th Canadian Conference on Artificial Intelligence, Canadian AI 2016, Victoria, BC, Canada, May 31-June 3, 2016. Proceedings 29*, Springer, 2016, S. 58–63. DOI: 10.1007/978-3-319-34111-8_8.
- [20] A. Drozdek, „Human intelligence and Turing Test,“ *AI & SOCIETY*, Jg. 12, S. 315–321, 1998. DOI: 10.1007/BF01179803.
- [21] C. Jones und B. Bergen, „Does GPT-4 pass the Turing test?“ *arXiv preprint arXiv:2310.20216*, 2023. DOI: 10.48550/arXiv.2310.20216. eprint: abs/2310.20216.

-
- [22] J. McCarthy, M. L. Minsky, N. Rochester und C. E. Shannon, „A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955,“ *AI magazine*, Jg. 27, S. 12–14, 2006. DOI: 10.1609/aimag.v27i4.1904.
- [23] S. Russel und P. Norvig, *Künstliche Intelligenz -Ein moderner Ansatz*, 2. Auflage. Pearson Education Deutschland, 2004, ISBN: 978-3-8273-7089-1.
- [24] W. Ertel, *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung* (Springer), 5. Aufl. Wiesbaden: Springer Fachmedien Wiesbaden, 2016. DOI: 10.1007/978-3-658-32075-1. Adresse: <https://doi.org/10.1007/978-3-658-32075-1>.
- [25] R. Colom, S. Karama, R. E. Jung und R. J. Haier, „Human intelligence and brain networks,“ *Dialogues in Clinical Neuroscience*, Jg. 12, Nr. 4, S. 489–501, 2010. DOI: 10.31887/DCNS.2010.12.4/rcolom. Adresse: <https://doi.org/10.31887/DCNS.2010.12.4/rcolom>.
- [26] M. H. Jarrahi, „Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making,“ *Business Horizons*, 2018. DOI: 10.1016/J.BUSHOR.2018.03.007.
- [27] C. Janiesch, P. Zschech und K. Heinrich, „Machine learning and deep learning,“ *Electronic Markets*, 2021.
- [28] Z.-H. Zhou, „A brief introduction to weakly supervised learning,“ *National Science Review*, Jg. 5, Nr. 1, S. 44–53, Aug. 2018, ISSN: 2095-5138. DOI: 10.1093/nsr/nwx106. eprint: <https://academic.oup.com/nsr/article-pdf/5/1/44/31567770/nwx106.pdf>.
- [29] geeksforgeeks, *Supervised and Unsupervised learning*, Zugriff am 27. Juni 2024. Adresse: <https://www.geeksforgeeks.org/supervised-unsupervised-learning/>.
- [30] L. P. Kaelbling, M. L. Littman und A. W. Moore, „Reinforcement learning: A survey,“ *Journal of artificial intelligence research*, Jg. 4, S. 237–285, 1996. DOI: <https://doi.org/10.1613/jair.301>.
- [31] Technische Universität Dresden, *Neuronale Netze*, Zugriff am 2024-09-09. Adresse: https://methpsy.elearning.psych.tu-dresden.de/mediawiki/index.php/Neuronale_Netze.
- [32] R. Sewada, A. Jangid, P. Kumar und N. Mishra, „Explainable Artificial Intelligence (XAI),“ *Journal of Nonlinear Analysis and Optimization*, 2023. DOI: 10.36893/jnao.2022.v13i02.041-047.
- [33] Z. Nan, W. Yang und S. Zhen, „Research on Information Intelligent Service Terminal Based on Big Data Intelligent ChatGPT Technology,“ in *2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, 2023, S. 1137–1141. DOI: 10.1109/ICSECE58870.2023.10263438.

-
- [34] R. E. Turner, „An Introduction to Transformers,“ *ArXiv*, Jg. abs/2304.10557, 2023. DOI: 10.48550/arXiv.2304.10557.
- [35] A. Vaswani, N. Shazeer, N. Parmar u. a., „Attention is all you need,“ *Advances in Neural Information Processing Systems*, 2017.
- [36] F. A. Acheampong, H. Nunoo-Mensah und W. Chen, „Transformer models for text-based emotion detection: a review of BERT-based approaches,“ *Artificial Intelligence Review*, Jg. 54, Nr. 8, S. 5789–5829, 2021. DOI: 10.1007/s10462-021-09958-2.
- [37] S. Seufert und S. Handschuh, *Generative Künstliche Intelligenz: ChatGPT und Co für Bildung, Wirtschaft und Gesellschaft*, 1. Aufl. Planegg: Schäffer-Poeschel, 2024, ISBN: 9783791062228. Adresse: <https://content-select.com/de/portal/media/view/6572fe3b-e9e0-4e05-b310-a7caac1b000f>.
- [38] O. Schwartz, M. Gebert und P. Marvi, *30 Minuten ChatGPT*, 1. Auflage. Offenbach am Main: GABAL, 2024, ISBN: 9783967403824.
- [39] OpenAI. „GPT-4 Research.“ Accessed: 2024-08-21. (März 2023), Adresse: <https://openai.com/index/gpt-4-research/>.
- [40] OpenAI, *Introducing GPTs*, Zugriff am: 24.07.2024, OpenAI, 2023. Adresse: <https://openai.com/index/introducing-gpts/>.
- [41] V. Plevris, G. Papazafeiropoulos und A. J. Rios, „Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard,“ *AI*, Jg. 4, Nr. 4, S. 949–969, 2023. DOI: 10.3390/ai4040048. Adresse: <https://www.mdpi.com/journal/ai>.
- [42] OpenAI, *Introduction to GPT Actions*, Zugriff am: 24.07.2024, OpenAI. Adresse: <https://platform.openai.com/docs/actions/introduction>.
- [43] S. Wolfram, *ChatGPT Gets Its Wolfram Superpowers*, Zugriff am: 24.07.2024, März 2023. Adresse: <https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/>.
- [44] J. B. Cassel, „Wolfram|Alpha: A Computational Knowledge “Search” Engine,“ in *Google It: Total Information Awareness*. New York, NY: Springer New York, 2016, S. 267–299, ISBN: 978-1-4939-6415-4. DOI: 10.1007/978-1-4939-6415-4_11. Adresse: https://doi.org/10.1007/978-1-4939-6415-4_11.
- [45] W. Alpha, *About Wolfram Alpha: Making the world’s knowledge computable*, Accessed: 2024-07-22, 2024. Adresse: <https://www.wolframalpha.com/about>.
- [46] W. Alpha, *What Is Wolfram|Alpha?* Accessed: 2024-07-22, 2024. Adresse: <https://www.wolframalpha.com/tour>.

-
- [47] N. Dahal, B. R. Lamichhane, B. C. Luitel und B. P. Pant, „AI Chatbots as Math Algorithm Problem Solvers: A Critical Evaluation of Its Capabilities and Limitations,“ in *Proceedings of the 28th Asian Technology Conference in Mathematics (ATCM 2023)*, Asian Technology Conference in Mathematics, Pattaya, Thailand, Dez. 2023. Adresse: https://www.researchgate.net/publication/375522509_AI_Chatbots_as_Math_Algorithm_Problem_Solvers_A_Critical_Evaluation_of_Its_Capabilities_and_Limitations.
- [48] Wolfram Alpha Team, *Step-by-Step Math in Wolfram/Alpha*, Accessed: 2024-08-28, 2009. Adresse: <https://blog.wolframalpha.com/2009/12/01/step-by-step-math/>.
- [49] Wolfram Alpha Team, *Wolfram/Alpha Show Steps API Reference*, Accessed: 2024-08-28. Adresse: <https://products.wolframalpha.com/show-steps-api/documentation>.
- [50] E. Davis und S. Aaronson, „Testing GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on math and science problems,“ *ArXiv*, Jg. abs/2308.05713, 2023. DOI: 10.48550/arXiv.2308.05713.
- [51] S. Manakhimova, E. Avramidis, V. Macketanz, E. Lapshinova-Koltunski, S. Bagdasarov und S. Möller, „Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT?“ In *Proceedings of the Eighth Conference on Machine Translation*, 2023, S. 224–245.
- [52] C. K. Lo, „What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature,“ *Education Sciences*, Jg. 13, Nr. 4, S. 410, 2023. DOI: 10.3390/educsci13040410.
- [53] X. Wu, R. Duan und J. Ni, „Unveiling Security, Privacy, and Ethical Concerns of ChatGPT,“ *ArXiv*, Jg. abs/2307.14192, 2023. DOI: 10.48550/arXiv.2307.14192.
- [54] Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen. Zugriff am 09. Juli 2024. (), Adresse: <https://www.schulministerium.nrw/textgenerierende-ki>.
- [55] H. Qawaqneh, F. B. Ahmad und A. R. Alawamreh, „The Impact of Artificial Intelligence-Based Virtual Laboratories on Developing Students’ Motivation Towards Learning Mathematics,“ *International Journal of Emerging Technologies in Learning*, Jg. 18, Nr. 14, S. 105–121, 2023. DOI: 10.3991/ijet.v18i14.39873. Adresse: <https://doi.org/10.3991/ijet.v18i14.39873>.
- [56] Á. Martínez-Sevilla und S. Alonso, „AI and Mathematics Interaction for a New Learning Paradigm on Monumental Heritage,“ in *Mathematics Education in the Age of Artificial Intelligence: How Artificial Intelligence Can Serve Mathematical Human Learning*, Springer International Publishing, 2022, S. 107–136.

-
- [57] A. Stojanov, „Learning with ChatGPT 3.5 as a more knowledgeable other: an autoethnographic study,“ *International Journal of Educational Technology in Higher Education*, Jg. 20, Nr. 1, S. 35, Juni 2023, ISSN: 2365-9440. DOI: 10.1186/s41239-023-00404-7. Adresse: <https://doi.org/10.1186/s41239-023-00404-7>.
- [58] S. Grubaugh, G. Levitt und D. Deever, „Harnessing AI to Power Constructivist Learning: An Evolution in Educational Methodologies,“ *EIKI Journal of Effective Teaching Methods*, 2023. DOI: 10.59652/jetm.v1i3.43.
- [59] J. Paul, „Untersuchung über die Eignung von KI-Modellen als Tutor in der Mathematikausbildung,“ Bachelorarbeit, TH Köln, Köln, Deutschland, Apr. 2024.
- [60] E. Supriyadi und K. S. Kuncoro, „Exploring the Future of Mathematics Teaching: Insight with ChatGPT,“ *Union: Jurnal Ilmiah Pendidikan Matematika*, Jg. 11, Nr. 2, S. 305–316, 2023. DOI: 10.30738/union.v11i2.14898. Adresse: <https://doi.org/10.30738/union.v11i2.14898>.
- [61] J. Gugagayanan, „The Effects of a Chatbot Solving Mathematical Equations using NLP,“ Downloaded from: <http://ray.yorks.j.ac.uk/id/eprint/8523/>, Masters thesis, York St John University, Dez. 2022. Adresse: <http://ray.yorks.j.ac.uk/id/eprint/8523/>.

Anhang

In diesem Anhang werden ergänzende Informationen und Materialien zu den im Rahmen dieser Arbeit durchgeführten Datenerhebungen bereitgestellt. Dies dient der Transparenz und ermöglicht eine tiefergehende Auseinandersetzung mit den dieser Arbeit zugrunde liegenden Datensätzen. Die Bereitstellung der Inhalte erfolgt in einer für die Struktur dieser Arbeit logischen Reihenfolge.

In Abschnitt *A. Beispiele für die Bewertungskriterien* wird ein Positiv und ein Negativ Beispiel für die Bewertung der einzelnen Kriterien gezeigt.

In Abschnitt *B. Einstufungen der Umfrageaufgaben* werden die Einstufungen der in der Umfrage verwendeten Aufgaben in Form von Säulendiagrammen bezüglich der definierten Kriterien abgebildet.

A. Beispiele für die Bewertungskriterien

Um einen besseren Überblick über die Bewertung der Kriterien zu bekommen, werden in diesem Abschnitt für jedes der Kriterien *Korrektheit*, *Sprache und Terminologie*, *Verständlichkeit und Strukturierung* ein positiv und ein Negativ Beispiel, der in dem folgenden GitHub-Repo abgebildeten Datensammlung abgebildet.

Um die Darstellung der Chatverläufe in diesem Anhang möglichst übersichtlich und leserfreundlich zu gestalten, erfolgt die Darstellung der Chatverläufe in Textform, anstatt in einer Abbildung des Screenshots. Durch diese Methode werden die relevanten Dialoge klar und strukturiert dargestellt, ohne das Layout der Seiten zu beeinträchtigen. Der benötigte Platzbedarf der Screenshots hätte zu einer unübersichtlichen Gestaltung des Anhangs geführt und die Lesbarkeit und den Fluss der Dokumentation beeinträchtigt. Den jeweiligen Chatverläufen sind die Daten beigefügt, die zum Auffinden der zugehörigen Screenshots im GitHub-Repository notwendig sind.

Korrektheit

Die folgenden Beispiele enthalten jeweils eine Aufgabe mit einer korrekten und einer Aufgabe mit einer inkorrekten Endlösung. Der Lösungsweg, der zu diesen Ergebnissen geführt hat, wird in diesem Kontext nicht betrachtet. Für die Bewertung in diesem Kriterium zählt ausschließlich das korrekte Endergebnis.

Korrekte Antwort:

Aufgabentyp 4 - Frage 1 (ChatGPT 4.0, 2. Wiederholung)

Frage: Der Parameter u soll so bestimmt werden, dass die beiden Vektoren $(0, u, -1)$ und $(-2, -2, 0)$ den Winkel 85° einschließen. Berechne und erkläre, wie ein Lehrer.

Antwort: Um den Winkel zwischen zwei Vektoren zu bestimmen, der 85° beträgt, nutzen wir das Skalarprodukt der Vektoren sowie die Länge jedes Vektors. Der mathematische Zusammenhang zwischen dem Skalarprodukt zweier Vektoren \mathbf{a} und \mathbf{b} , dem eingeschlossenen Winkel θ und den Längen (oder Normen) dieser Vektoren ist durch die folgende Formel gegeben:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$$

Gegebene Vektoren:

$$\mathbf{a} = (0, u, -1)$$

$$\mathbf{b} = (-2, -2, 0)$$

1. Schritt: Skalarprodukt berechnen Das Skalarprodukt $\mathbf{a} \cdot \mathbf{b}$ berechnet sich wie folgt:

$$\mathbf{a} \cdot \mathbf{b} = 0 \cdot (-2) + u \cdot (-2) + (-1) \cdot 0 = -2u$$

2. Schritt: Längen der Vektoren Die Länge (Norm) eines Vektors \mathbf{a} wird berechnet durch:

$$\|\mathbf{a}\| = \sqrt{0^2 + u^2 + (-1)^2} = \sqrt{u^2 + 1}$$

Die Länge von \mathbf{b} ist:

$$\|\mathbf{b}\| = \sqrt{(-2)^2 + (-2)^2 + 0^2} = \sqrt{4 + 4} = 2\sqrt{2}$$

3. Schritt: Winkelbeziehung nutzen Setzen wir nun die Daten in die Gleichung des Skalarprodukts ein, um u zu bestimmen:

$$-2u = \sqrt{u^2 + 1} \cdot 2\sqrt{2} \cdot \cos 85^\circ$$

Der Kosinus von 85° ist etwa 0.08716 (dies kann man entweder einem Taschenrechner oder einer Tabelle entnehmen).

$$-2u = \sqrt{u^2 + 1} \cdot 2\sqrt{2} \cdot 0.08716$$

Diese Gleichung kann man jetzt nach u auflösen. Dafür isolieren wir zuerst u auf einer Seite und führen dann schrittweise die Berechnung durch.

Berechnung von u :

$$\begin{aligned} -2u &= \sqrt{u^2 + 1} \cdot 2\sqrt{2} \cdot 0.08716 \\ u &= \frac{\sqrt{u^2 + 1} \cdot 0.17432\sqrt{2}}{2} \end{aligned}$$

Nun lösen wir diese Gleichung, um den Wert von u zu finden. Wir verwenden eine numerische Methode, um die Gleichung zu lösen, da eine exakte algebraische Lösung schwierig sein kann. Lassen Sie uns das mithilfe einer Berechnung tun.

Der Wert für u , der sicherstellt, dass die beiden Vektoren $(0, u, -1)$ und $(-2, -2, 0)$ einen Winkel von 85° einschließen, beträgt ungefähr -0.124 . Dies bedeutet, dass der Vektor $(0, -0.124, -1)$ zusammen mit $(-2, -2, 0)$ den gewünschten Winkel bildet.

Wenn du weitere Fragen zu dieser Berechnung hast oder eine andere Erklärung benötigst, lass es mich wissen!

Inkorrekte Antwort:

Aufgabentyp 4 - Frage 1 (ChatGPT 4.0, 3. Wiederholung)

Frage: Der Parameter u soll so bestimmt werden, dass die beiden Vektoren $(0, u, -1)$ und $(-2, -2, 0)$ den Winkel 85° einschließen. Berechne und erkläre, wie ein Lehrer.

Antwort: Um den Parameter u zu finden, sodass die beiden Vektoren $\mathbf{a} = (0, u, -1)$ und $\mathbf{b} = (-2, -2, 0)$ einen Winkel von 85° einschließen, können wir das Skalarprodukt der beiden Vektoren verwenden.

Das Skalarprodukt zweier Vektoren \mathbf{a} und \mathbf{b} ist definiert als:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos(\theta)$$

wobei θ der Winkel zwischen den Vektoren und $\|\mathbf{a}\|$ und $\|\mathbf{b}\|$ die Beträge (Längen) der Vektoren sind.

Schritt 1: Skalarprodukt der Vektoren berechnen Für $\mathbf{a} = (0, u, -1)$ und $\mathbf{b} = (-2, -2, 0)$ ist das Skalarprodukt:

$$\mathbf{a} \cdot \mathbf{b} = 0 \cdot (-2) + u \cdot (-2) + (-1) \cdot 0 = -2u$$

Schritt 2: Beträge der Vektoren berechnen Der Betrag eines Vektors $\mathbf{a} = (a_1, a_2, a_3)$ ist gegeben durch:

$$|\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}$$

Für $\mathbf{a} = (0, u, -1)$:

$$|\mathbf{a}| = \sqrt{0^2 + u^2 + (-1)^2} = \sqrt{u^2 + 1}$$

Für $\mathbf{b} = (-2, -2, 0)$:

$$|\mathbf{b}| = \sqrt{(-2)^2 + (-2)^2 + 0^2} = \sqrt{4 + 4} = \sqrt{8} = 2\sqrt{2}$$

Schritt 3: Gleichung für den Winkel einrichten Setzen wir $\theta = 85^\circ$ und benutzen die Formel für das Skalarprodukt:

$$-2u = \sqrt{u^2 + 1} \cdot 2\sqrt{2} \cdot \cos(85^\circ)$$

Der Kosinus von 85° ist etwa $\cos(85^\circ) \approx 0.0872$. Die Gleichung wird also zu:

$$-2u = \sqrt{u^2 + 1} \cdot 2\sqrt{2} \cdot 0.0872$$

Schritt 4: Die Gleichung nach u auflösen Um den Wert für u zu finden, vereinfachen wir die Gleichung und lösen sie auf:

$$\begin{aligned} -2u &= \sqrt{u^2 + 1} \cdot 0.1744 \quad (\text{da } 2\sqrt{2} \approx 2.8284) \\ \frac{-2u}{0.1744} &= \sqrt{u^2 + 1} \end{aligned}$$

Quadrieren wir beide Seiten der Gleichung:

$$\left(\frac{-2u}{0.1744} \right)^2 = u^2 + 1$$

Jetzt berechnen wir den genauen Wert von u mithilfe eines numerischen Verfahrens, um die Quadratwurzelgleichung zu lösen. Lass mich das für dich machen.

Die Lösung der Gleichung ergibt zwei mögliche Werte für u , nämlich $u \approx -0.0875$ und $u \approx 0.0875$. Beide Werte sind gültig und führen dazu, dass die beiden Vektoren einen Winkel von 85° einschließen.

Wenn wir also den Parameter u so wählen, dass $u \approx \pm 0.0875$ ist, erreichen wir, dass die Vektoren $(0, u, -1)$ und $(-2, -2, 0)$ den gewünschten Winkel von 85° einschließen.

Vollständigkeit

Nachfolgend wird jeweils ein positives und ein negatives Beispiel für das Kriterium der Vollständigkeit gegeben. Bei der Bewertung der Vollständigkeit wurde besonders darauf geachtet, dass keine Schritte im Lösungsweg übersprungen wurden. Die Korrektheit der Rechnung und des Ergebnisses war hierbei nicht von Bedeutung. Jeder gewählte Lösungsweg wurde akzeptiert, solange er vollständig erklärt und ohne logische Sprünge durchgeführt wurde. Dies ist im ersten Beispiel der Fall. Grammatikalische Fehler oder unlogische Formulierungen wie "Da das einzige ungerade Primzahl 2 ist, muss $b = 2$ sein..." haben keinen Einfluss auf die Bewertung dieses Kriteriums.

Das zweite Beispiel ist unvollständig, da wichtige Erklärungen zur Durchführung der einzelnen Schritte fehlen. Zunächst wird nicht klar definiert, was eine Primzahl ist, was eine notwendige Grundlage für das Verständnis des Lösungswegs darstellt. Weiterhin wird in Schritt drei zwar gezeigt, was in die umgestellte Gleichung eingesetzt werden soll, jedoch fehlt die exemplarische Durchführung dieses Einsetzens. Zudem wird erwähnt, dass überprüft wird, ob die Zahl 587 eine Primzahl ist, jedoch wird nicht erklärt, wie diese Überprüfung durchgeführt werden kann. Die fehlenden Details führen dazu, dass das Beispiel den Anforderungen an die Vollständigkeit nicht ausreicht.

Vollständige Antwort:

Aufgabentyp 2 - Frage 2 (ChatGPT 3.5, 1. Wiederholung)

Frage: $5a + b = 5 \times (132 - a^2) + 2$ Löse die Gleichung. a und b sind Primzahlen. Erkläre mir den Lösungsweg Schritt für Schritt, wie ein Lehrer.

Antwort: Natürlich, ich helfe gerne! Zuerst sollten wir die Gleichung $5a + b = 5 \times (132 - a^2) + 2$ lösen. Wir wissen, dass a und b Primzahlen sind, was bedeutet, dass sie nur durch 1 und sich selbst teilbar sind.

Schritt 1: Multipliziere die rechte Seite Die rechte Seite aus $5 \times (132 - a^2) + 2$ ist der erste Ausdruck auf der rechten Seite der Gleichung. Lassen Sie uns diesen Ausdruck berechnen.

$$5 \times (132 - a^2) + 2 = 660 - 5a^2 + 2 = 662 - 5a^2$$

Das vereinfacht sich zu: $660 - 5a^2 + 2 = 662 - 5a^2$

Also haben wir $662 - 5a^2$ auf der rechten Seite.

Unsere Gleichung lautet also jetzt:

$$5a + b = 662 - 5a^2$$

Schritt 2: Bringe alles auf eine Seite Um eine quadratische Gleichung zu erhalten. Wir möchten die Gleichung so umstellen, dass sie der Form $ax^2 + bx + c = 0$ entspricht. Also ziehen wir $5a$ von beiden Seiten ab und addieren $5a^2$ zu beiden Seiten:

$$5a^2 + 5a - b - 662 = 0$$

Schritt 3: Prüfe, ob die Lösungen Primzahlen sind Nun müssen wir diese quadratische Gleichung lösen. Aber bevor wir das tun, sollten wir einen Trick verwenden: Da a eine Primzahl ist, kann $5a$ nicht eine andere Primzahl minus b sein. Das bedeutet, dass wir eine ungerade Zahl $5a$ von einer geraden Zahl 662 abziehen und trotzdem eine ungerade Zahl erhalten müssen. Das kann nur geschehen, wenn b eine ungerade Zahl ist.

Da das einzige ungerade Primzahl 2 ist, muss $b = 2$ sein.

Also lautet unsere Gleichung jetzt:

$$5a^2 + 5a - 2 - 662 = 0$$

Schritt 4: Löse die quadratische Gleichung Um diese quadratische Gleichung zu lösen, können wir die quadratische Formel verwenden:

$$a = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Hier ist $a = 5$, $b = 5$ und $c = -664$. Setzen wir diese Werte in die Formel ein:

$$\begin{aligned} a &= \frac{-5 \pm \sqrt{5^2 - 4 \times 5 \times (-664)}}{2 \times 5} \\ a &= \frac{-5 \pm \sqrt{25 + 13280}}{10} \\ a &= \frac{-5 \pm \sqrt{13305}}{10} \end{aligned}$$

Da $\sqrt{13305}$ keine ganze Zahl ist, gibt es keine ganzzahligen Lösungen für a .

Daher hat die Gleichung keine ganzzahligen Lösungen für a und b , die beide Primzahlen sind.

Unvollständige Antwort:**Aufgabentyp 2 - Frage 2 (WolframGPT, 3. Wiederholung)**

Frage: $5a + b = 5 \times (132 - a^2) + 2$ Löse die Gleichung. a und b sind Primzahlen. Erkläre mir den Lösungsweg Schritt für Schritt, wie ein Lehrer.

Antwort: Um die Gleichung $5a + b = 5 \times (132 - a^2) + 2$ zu lösen, wobei a und b Primzahlen sind, können wir folgende Schritte unternehmen:

1. **Vereinfache die Gleichung:** Zuerst formen wir die Gleichung um, um sie einfacher zu gestalten. Wir können die rechte Seite der Gleichung umformen:

$$5 \times (132 - a^2) + 2 = 660 - 5a^2 + 2 = 662 - 5a^2$$

Also wird die Gleichung zu:

$$5a + b = 662 - 5a^2$$

2. **Drücke b aus:** Löse die Gleichung nach b auf:

$$b = 662 - 5a^2 - 5a$$

3. **Berücksichtige die Primzahlen:** a und b müssen beide Primzahlen sein. Daher müssen wir Werte für a finden, die sowohl b als auch a als Primzahlen ergeben.

Um die möglichen Primzahlen zu finden, die diese Bedingungen erfüllen, und zu überprüfen, ob b auch eine Primzahl für diese a -Werte ist, werden wir den Wertebereich von a durchlaufen und die Primzahligenschaft von b testen.

Hier bietet sich die Verwendung von Wolfram Alpha an, da es die Berechnung und Überprüfung auf Primzahlen effizient handhaben kann. Ich werde diese Berechnungen durchführen und Ihnen die möglichen Lösungen mitteilen.

Fehler beim Gespräch mit Wolfram Alpha: Es scheint, dass meine ursprüngliche Anfrage an Wolfram Alpha zu komplex war, weil ich versucht habe, die Gleichung direkt mit der Einschränkung zu lösen, dass a und b Primzahlen sind. Ich werde stattdessen einen anderen Ansatz versuchen, indem ich für verschiedene mögliche Primzahlenwerte von a den Wert von b berechne und dann prüfe, ob b ebenfalls eine Primzahl ist. Ich werde dies für ein paar kleine Primzahlen durchführen und schauen, ob wir eine gültige Lösung finden können.

Für $a = 2$ ergibt sich aus der Gleichung, dass $b = 622$.

Wir prüfen nun, ob 622 eine Primzahl ist.

Die Zahl 622 ist keine Primzahl, da sie die Primfaktoren 2 und 311 hat. Wir müssen also weitere Werte für a überprüfen, um eine Lösung zu finden, bei der sowohl a als auch b Primzahlen sind.

Lassen Sie uns weitere kleine Primzahlen für a durchprobieren: $a = 3$, $a = 5$, $a = 7$, und $a = 11$. Wir berechnen für jeden dieser Werte b und überprüfen, ob b eine Primzahl ist.

Die Ergebnisse für die weiteren Primzahlenwerte von a (3, 5, 7, 11) wurden nicht direkt geliefert. Es scheint ein Missverständnis in der Anfrage gegeben zu haben. Ich werde die Berechnung für jeden dieser Werte einzeln durchführen, um die entsprechenden b -Werte zu erhalten. Lassen Sie uns mit $a = 3$ beginnen.

Für $a = 3$ ergibt sich $b = 587$. Wir überprüfen nun, ob 587 eine Primzahl ist.

Die Zahl 587 ist eine Primzahl. Daher ist eine Lösung für die ursprüngliche Gleichung, bei der sowohl a als auch b Primzahlen sind, $a = 3$ und $b = 587$.

Sprache und Terminologie

Die unteren Beispiele vermitteln eine Vorstellung von einer hinsichtlich Sprache und Terminologie positiven und einer negativen Antwort. Das positive Beispiel zeichnet sich durch eine klare und verständliche Sprache aus, bei der die Erklärungen logisch miteinander verknüpft sind. Fachbegriffe werden korrekt verwendet und stehen im richtigen Zusammenhang. Es gibt keine Missverständnisse oder unklare Formulierungen.

Im Gegensatz dazu weist das negative Beispiel widersprüchliche Aussagen auf, und einige Sätze sind in sich unverständlich. Die Erklärungen wirken chaotisch, schwer nachvollziehbar und missverständlich. Ein typisches Beispiel hierfür wäre der Satz: „... Da Clara lügt und Luca die Lakritz hat ...“. Dieser Satz widerspricht sich, da zunächst angenommen wird, dass Clara lügt, aber anschließend diese Aussage als wahr angesehen wird und auf dieser Basis weiter argumentiert wird, was zu Verwirrung führt und den Lösungsweg unklar macht.

Ausgezeichnete Antwort:**Aufgabentyp 5 - Frage 3 (ChatGPT 4.0, 1. Wiederholung)**

Frage: Ein Bauer liebt Primzahlen und ist total begeistert als er beim Zählen seiner Tiere feststellt, dass die Anzahlen seiner Pferde, Kühe und Hühner drei verschiedene Primzahlen sind. Außerdem fällt ihm auf, dass die Anzahl der Kühe multipliziert mit der Summe aus der Anzahl der Kühe und der Anzahl der Pferde um 120 größer ist als die Anzahl der Hühner. Finde heraus wie viele Pferde, Hühner und Kühe der Bauer hat. Berechne und erkläre, wie ein Lehrer.

Antwort: Um das Problem zu lösen, beginnen wir mit der Definition der Variablen:

- p sei die Anzahl der Pferde - c sei die Anzahl der Kühe - h sei die Anzahl der Hühner

Alle diese Variablen repräsentieren unterschiedliche Primzahlen. Die gegebene Bedingung lautet:

$$c \times (c + p) = h + 120$$

Wir müssen nun eine Menge von Primzahlen finden, die diese Gleichung erfüllen. Da es viele Primzahlen gibt, beginnen wir mit kleineren Werten und überprüfen systematisch die möglichen Kombinationen.

Zunächst generieren wir eine Liste von Primzahlen. Für den Anfang können wir eine Liste der Primzahlen bis etwa 100 erstellen und dann überprüfen, welche Kombinationen die Bedingung erfüllen.

Lassen Sie uns zuerst die benötigte Liste der Primzahlen erstellen:

Wir haben eine Liste von Primzahlen bis 100 erstellt. Nun prüfen wir systematisch die Kombinationen dieser Primzahlen, um die Beziehung $c \times (c + p) = h + 120$ zu erfüllen. Wir müssen sicherstellen, dass p , c und h unterschiedliche Werte haben.

Ich werde nun die Primzahlen aus dieser Liste durchgehen und überprüfen, welche Kombinationen die Gleichung erfüllen. Wir gehen systematisch vor, indem wir für jedes c und p den Wert von $h + 120$ berechnen und dann überprüfen, ob dieser Wert minus 120 ebenfalls eine Primzahl ist und nicht c oder p entspricht.

Nach der Überprüfung der Primzahlenkombinationen finden wir eine Lösung, die die Bedingung erfüllt:

- Anzahl der Pferde (p) = 2 - Anzahl der Kühe (c) = 11 - Anzahl der Hühner (h) = 23

Diese Werte erfüllen die Gleichung $c \times (c + p) = h + 120$ folgendermaßen:

$$11 \times (11 + 2) = 23 + 120 \quad 11 \times 13 = 143 \quad 143 = 143$$

Die Berechnung bestätigt, dass die Kombination korrekt ist. Der Bauer hat also 2 Pferde, 11 Kühe und 23 Hühner. Alle sind unterschiedliche Primzahlen und erfüllen die vorgegebene mathematische Bedingung.

Mangelhafte Antwort:

Aufgabentyp 5 - Frage 5 (ChatGPT 3.5, 1. Wiederholung)

Frage: Clara, Maximilian und Luca besitzen jeder eine Süßigkeit. Finde heraus, wer die Gummibärchen, wer die Tüte Lakritz und wer die Eiscreme hat, wenn zwei von ihnen lügen und keine Süßigkeit doppelt vorkommt. Clara: "Luca hat die Lakritz." Maximilian: "Clara hat die Lakritz nicht." Luca: "Maximilian hat keine Eiscreme." Löse und erkläre, wie ein Lehrer.

Antwort: Um diese Aufgabe zu lösen, müssen wir die Aussagen der drei Personen analysieren und logisch schlussfolgern, wer welche Süßigkeit besitzt.

1. Clara sagt: "Luca hat die Lakritz." 2. Maximilian sagt: "Clara hat die Lakritz nicht." 3. Luca sagt: "Maximilian hat keine Eiscreme."

Zuerst betrachten wir die Aussage von Clara. Wenn Clara die Wahrheit sagt, dann hat Luca die Lakritz. Wenn Luca die Lakritz hat, dann kann Maximilian nicht die Eiscreme haben, da laut Luca Maximilian keine Eiscreme hat. Das würde bedeuten, dass Maximilian die Gummibärchen hat. Aber das widerspricht Maximilians Aussage, dass Clara die Lakritz nicht hat. Also lügt Clara.

Daraus können wir schließen, dass Luca die Lakritz hat.

Da Clara lügt und Luca die Lakritz hat, muss Clara die Eiscreme haben. Das bedeutet, dass Maximilian die Gummibärchen hat, da keine Süßigkeit doppelt vorkommt und Maximilian weder die Lakritz noch die Eiscreme haben kann.

Das bedeutet, dass die einzige Süßigkeit, die übrig bleibt, die Gummibärchen sind, die Maximilian nicht haben kann. Also hat Clara die Gummibärchen.

Zusammengefasst: - Clara hat die Gummibärchen. - Maximilian hat die Eiscreme. - Luca hat die Lakritz.

Die Lösung ergibt sich aus der Logik der Aussagen und der Tatsache, dass keine Süßigkeit doppelt vorkommt.

Struktur

Das Kriterium ist im Nachhinein schwer nachvollziehbar, da beim erneuten Aufruf der archivierten Chats die Struktur teilweise verschoben wurde. Überschriften erscheinen nun teilweise als Satzanfang eines Absatzes, und Zeilenumbrüche wurden oftmals entfernt. Alle Überschriften, die in den folgenden Beispielen fett gedruckt waren, waren ursprünglich korrekt über dem jeweiligen Absatz. Nach der Archivierung gingen jedoch an vielen Stellen die Zeilenumbrüche verloren. Die ursprünglichen Screenshots, die aufgrund ihrer unzureichenden Qualität eine Neuaufnahme erforderlich machten, zeigen jedoch, dass die Rechnungen und Überschriften ursprünglich korrekt strukturiert waren.

Die Beispiele für das Kriterium der Struktur verdeutlichen den unterschiedlichen Einsatz von Absätzen und Überschriften. Das positive Beispiel zeigt eine logische Struktur, bei der nur notwendige Überschriften verwendet werden und jeder Absatz angemessen gefüllt ist. Das negative Beispiel hingegen weist unnötig viele Überschriften auf, was durch häufige Sprünge zwischen Aufgabenerklärung, Berechnung und Erklärung entsteht. Zusätzlich sind die Überschriften, die wichtige Bereiche abdecken, schlecht auffindbar und enthalten teilweise nur einen einzigen Satz, was die Übersichtlichkeit weiter erschwert.

Ausgezeichnet strukturierte Antwort:

Aufgabentyp 4 - Frage 2 (ChatGPT 4.0, 3. Wiederholung)

Frage: Gegeben sind die Punkte $A(1, 2, 3)$, $B(3, 3, 1)$ und $C(5, 1, 2)$. a) Zeigen Sie, dass das Dreieck ein gleichschenkliges, rechtwinkliges Dreieck ist. Bestimmen Sie den Flächeninhalt dieses Dreiecks. b) Zeigen Sie, dass man das Dreieck ABC zu einem Quadrat ABCD ergänzen kann. Bestimmen Sie die Koordinaten von D. Berechne und erkläre, wie ein Lehrer.

Antwort: Um zu prüfen, ob das Dreieck ABC gleichschenkelig und rechtwinklig ist, müssen wir die Längen der Seiten AB, AC und BC berechnen und prüfen, ob zwei Seiten gleich lang sind und ob ein rechter Winkel vorliegt.

Schritt 1: Längen der Seiten berechnen Die Länge der Seite zwischen zwei Punkten in einem dreidimensionalen Raum wird mit der Distanzformel berechnet:

$$\text{Distanz} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Seite AB zwischen $A(1, 2, 3)$ und $B(3, 3, 1)$:

$$AB = \sqrt{(3 - 1)^2 + (3 - 2)^2 + (1 - 3)^2} = \sqrt{2^2 + 1^2 + (-2)^2} = \sqrt{4 + 1 + 4} = \sqrt{9} = 3$$

Seite AC zwischen $A(1, 2, 3)$ und $C(5, 1, 2)$:

$$AC = \sqrt{(5 - 1)^2 + (1 - 2)^2 + (2 - 3)^2} = \sqrt{4^2 + (-1)^2 + (-1)^2} = \sqrt{16 + 1 + 1} = \sqrt{18} = 3\sqrt{2}$$

Seite BC zwischen $B(3, 3, 1)$ und $C(5, 1, 2)$:

$$BC = \sqrt{(5 - 3)^2 + (1 - 3)^2 + (2 - 1)^2} = \sqrt{2^2 + (-2)^2 + 1^2} = \sqrt{4 + 4 + 1} = \sqrt{9} = 3$$

Schritt 2: Überprüfung der Dreieckseigenschaften

- **Gleichschenklig:** Die Seiten AB und BC sind beide 3 Einheiten lang. Also ist das Dreieck gleichschenkl. - **Rechtwinklig:** Um zu überprüfen, ob es rechtwinklig ist, nutzen wir das Skalarprodukt der Vektoren AB und BC. Wenn das Skalarprodukt 0 ist, steht der Winkel zwischen AB und BC im rechten Winkel.

Vektor AB = $(3-1, 3-2, 1-3) = (2, 1, -2)$ **Vektor BC** = $(5-3, 1-3, 2-1) = (2, -2, 1)$

$$\text{Skalarprodukt } AB \cdot BC = 2 \times 2 + 1 \times (-2) + (-2) \times 1 = 4 - 2 - 2 = 0$$

Da das Skalarprodukt 0 ist, ist das Dreieck rechtwinklig.

Flächeninhalt des Dreiecks Der Flächeninhalt eines rechtwinkligen Dreiecks ist:

$$\text{Fläche} = \frac{1}{2} \times \text{Seite}_1 \times \text{Seite}_2$$

Für AB und BC als Katheten:

$$\text{Fläche} = \frac{1}{2} \times 3 \times 3 = 4.5 \text{ Quadrat-Einheiten}$$

b) Ergänzung zu einem Quadrat Um das Quadrat zu ergänzen, fügen wir einen Punkt D hinzu, sodass die Seiten $AB = BC = CD = DA$ sind und die Winkel zwischen den Seiten rechtwinklig sind.

Wir wissen, dass AB und BC gleich lang sind und rechtwinklig zueinander stehen. Wir können D finden, indem wir den Vektor BC von B nach C auf A anwenden und zu A addieren:

****Vektor BC = (2, -2, 1)**** ****Punkt A = (1, 2, 3)****

$$D = A + BC = (1 + 2, 2 - 2, 3 + 1) = (3, 0, 4)$$

Zusammenfassung

- Das Dreieck ABC ist gleichschenkelig und rechtwinklig mit einem Flächeninhalt von 4.5 Quadrat-Einheiten. - Das Dreieck kann zu einem Quadrat ABCD ergänzt werden, wobei die Koordinaten von D (3, 0, 4) sind.

Mangelhaft strukturierte Antwort:

Aufgabentyp 4 - Frage 2 (ChatGPT 4.0, 2. Wiederholung)

Frage: Gegeben sind die Punkte $A(1, 2, 3)$, $B(3, 3, 1)$ und $C(5, 1, 2)$. a) Zeigen Sie, dass das Dreieck ein gleichschenkliges, rechtwinkliges Dreieck ist. Bestimmen Sie den Flächeninhalt dieses Dreiecks. b) Zeigen Sie, dass man das Dreieck ABC zu einem Quadrat ABCD ergänzen kann. Bestimmen Sie die Koordinaten von D. Berechne und erkläre, wie ein Lehrer.

Antwort: Um zu zeigen, dass das Dreieck ABC ein gleichschenkliges, rechtwinkliges Dreieck ist und den Flächeninhalt zu berechnen, gehen wir folgendermaßen vor:

Schritt 1: Seitenlängen des Dreiecks berechnen Die Seitenlängen eines Dreiecks im dreidimensionalen Raum können mit der Distanzformel berechnet werden:

$$AB = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Für die Punkte:

- $A(1, 2, 3)$ - $B(3, 3, 1)$ - $C(5, 1, 2)$

Berechnen wir die Seitenlängen:

$$- AB = \sqrt{(3-1)^2 + (3-2)^2 + (1-3)^2} - BC = \sqrt{(5-3)^2 + (1-3)^2 + (2-1)^2} - CA = \sqrt{(5-1)^2 + (1-2)^2 + (2-3)^2}$$

Schritt 2: Rechtwinkligkeit prüfen Um zu prüfen, ob das Dreieck rechtwinklig ist, nutzen wir das Skalarprodukt der Vektoren. Wenn das Skalarprodukt zweier Seitenvektoren null ist, steht der entsprechende Winkel rechtwinklig.

Die Vektoren sind:

$$- \vec{AB} = (2, 1, -2) - \vec{BC} = (2, -2, 1) - \vec{CA} = (4, -1, -1)$$

Schritt 3: Flächeninhalt des Dreiecks Der Flächeninhalt eines rechtwinkligen Dreiecks ist einfach zu berechnen:

$$\text{Fläche} = \frac{1}{2} \times \text{Kathete}_1 \times \text{Kathete}_2$$

Schritt 4: Ergänzung zum Quadrat Um das Dreieck zu einem Quadrat zu ergänzen, addieren wir die Differenz zweier Vektoren zu einem der Punkte hinzu. Der neue Punkt D ist derjenige, der durch Translation entsteht.

Rechnerische Umsetzung Jetzt setzen wir diese Berechnungen in einer Python-Umgebung um, um die genauen Werte für Seitenlängen, Winkel und die Koordinaten des Punktes D zu bestimmen.

Ergebnisse der Berechnungen

Seitenlängen

$$- AB = 3.0 - BC = 3.0 - CA = 4.2426 \text{ (ungefähr)}$$

Überprüfung auf Rechtwinkligkeit und Gleichschenkligkeit Das Skalarprodukt von \vec{AB} und \vec{BC} ist 0, was bedeutet, dass der Winkel zwischen AB und BC rechtwinklig ist. Die Längen von AB und BC sind gleich, was bestätigt, dass das Dreieck ABC rechtwinklig und gleichschenkelig ist.

Flächeninhalt des Dreiecks Der Flächeninhalt des Dreiecks ABC beträgt 4.5 Quadrat-Einheiten.

Ergänzung zum Quadrat Wir können das Dreieck zu einem Quadrat ergänzen, indem wir den Punkt D wie folgt berechnen:

$$D = B + (A - C) = (-1, 4, 2)$$

Dies ergibt die Koordinaten von D.

Erklärung

1. **Seitenlängen:** Durch Berechnen der euklidischen Distanzen zwischen den Punkten haben wir festgestellt, dass AB und BC gleich lang sind.
2. **Rechtwinkligkeit:** Das Skalarprodukt von \overrightarrow{AB} und \overrightarrow{BC} ist 0, was bestätigt, dass das Dreieck rechtwinklig ist.
3. **Flächeninhalt:** Bei einem rechtwinkligen Dreieck ist der Flächeninhalt die Hälfte des Produkts der Längen der beiden Katheten, also $\frac{1}{2} \times 3 \times 3$.
4. **Quadratergänzung:** Durch Addition des Vektors $A - C$ zu Punkt B erhalten wir einen neuen Punkt D, der das Dreieck ABC zu einem Quadrat ABCD ergänzt.

Die Koordinaten von D und der rechtwinklige, gleichschenklige Charakter des Dreiecks ABC wurden damit bestätigt.

B. Einstufung und Vergleich der Umfrageaufgaben

Dieser Teil des Anhangs dient der detaillierteren Darstellung der zugeordneten Kriterienstufen, die den in der Umfrage behandelten Aufgaben zugeordnet wurden. Die Darstellung erfolgt für die fünf untersuchten Aufgabentypen, die abhängig von der antwortenden KI-Version in die entsprechenden Bewertungsstufen des jeweiligen Kriteriums eingeordnet sind. Diese Zuordnung wird in Form eines Säulendiagramms visualisiert, bei dem die unterschiedlichen Aufgabentypen farblich voneinander abgegrenzt sind. Zusätzlich werden die in der Datennalyse gesetzten Einstufungen mit den Angaben der in der Umfrage Befragten verglichen. Die Abweichungen in den Kriterienstufen der Vollständigkeit sowie der Sprache und Terminologie von den in der Umfrage angewendeten Abstufungen erschweren einen direkten Vergleich. Dennoch erhält man durch die Darstellung einen Überblick über die Antworttendenz und hat die Möglichkeit eindeutige Diskrepanzen zwischen den geprüften und den von den Befragten Angegebenen Einstufungen zu ermitteln.

Korrektheit

Die Forschungsergebnisse zu dem Kriterium der Korrektheit werden im Rahmen dieser Bachelorarbeit in Kapitel 4.2.1 vorgestellt. Da dieses Kriterium für die Verwendbarkeit der Aufgabenerklärung grundlegend ist, fokussiert sich diese Arbeit besonders auf dieses Kriterium. Weswegen dieses Kriterium bereits außerhalb des Anhangs mit den Umfrageergebnissen verglichen wurde. Aufgrund seiner eindeutigen Überprüfbarkeit ist dieses Kriterium unabhängig von der persönlichen Einschätzung und eindeutig

bewertbar. In dem in Abbildung 1 dargestellten Diagramm wird die Bewertung der verschiedenen in der Umfrage verwendeten Antworten dargestellt, um eine Basis für den Vergleich der Ergebnisse der Datenanalyse und der Umfrage zu schaffen.

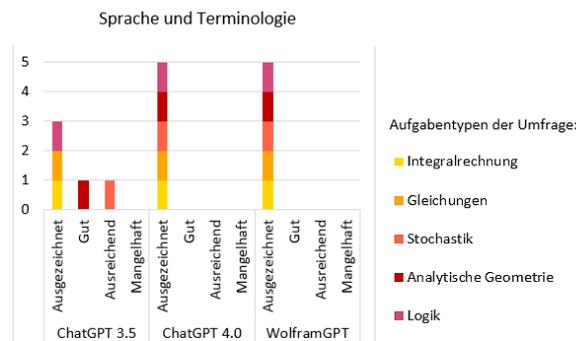


Abbildung 1: Umfrageaufgaben: Korrektheit

Vollständigkeit

Die Vollständigkeit einer Erklärung stellt einen entscheidenden Faktor für die Nachvollziehbarkeit der Antwort dar. Allerdings ermöglichen die Versionen von ChatGPT durch die Chatfunktion das Übermitteln von Nachfragen, was die Möglichkeit bietet, Unklarheiten direkt zu adressieren. Diese Interaktivität ist besonders wichtig, da die Beurteilung der Vollständigkeit einer Antwort stark von subjektiven Faktoren, wie dem Vorwissen und den individuellen Erwartungen der befragten Person, beeinflusst wird. So kann es bei der Bewertung zu unterschiedlichen Einschätzungen kommen. Zudem sind die in der Datenanalyse und in der Umfrage verwendeten Bewertungsstufen nicht identisch. Die in den Matrizen abgebildeten Daten zeigen deutliche Abweichungen in den Bewertungen der Vollständigkeit für die Version 3.5 von ChatGPT zwischen den beiden Befragtengruppen. Diese Diskrepanzen könnten auf einen Fehler bei der Einordnung der Datenanalyse zurückzuführen sein. Allerdings kann dieser Unterschied auch in dem unterschiedlichen Schwierigkeitsgrad der Aufgaben begründet liegen, da erkennbar wird, dass die schwereren Aufgabentypen eine grundsätzlich schlechtere Bewertung bei der Vollständigkeit der Erklärung bekommen haben als die leichteren.

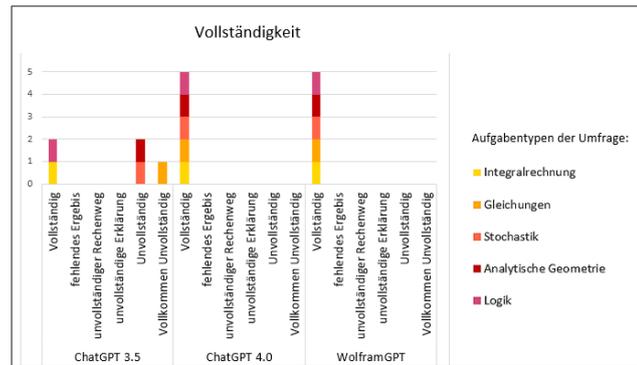


Abbildung 2: Umfrageaufgaben: Vollständigkeit

		Geprüfte Vollständigkeit					
		Vollständig	Fehlendes Ergebnis	Unvollständiger Rechenweg	Unvollständige Erklärung	Unvollständig	Vollkommen Unvollständig
Angaben der Befragten	Vollständig	0	0	0	0	2	0
	Eher Vollständig	3	0	0	0	2	4
	Teilweise	7	0	0	0	6	2
	Eher unvollständig	5	0	0	0	0	2
	Unvollständig	1	0	0	0	0	4

Tabelle 1: Matrix: Bewertung der ChatGPT-3.5-Antworten zur Vollständigkeit durch Studierende

Sprache und Terminologie

Die Bewertung der in der Umfrage verwendeten Aufgaben wird in dem nachfolgenden Diagramm aufgabenspezifisch für die fünf verschiedenen Aufgabentypen dargestellt. Diese Darstellung ermöglicht eine präzisere Gegenüberstellung der Ergebnisse der Datenanalyse mit denen der Umfrage. Hierbei wird ersichtlich, dass es zu keiner signifikanten Abweichungen kommt. Es zeigt sich jedoch, dass Abweichungen vornehmlich

ChatGPT 3.5		Geprüfte Vollständigkeit					
		Vollständig	Fehlendes Ergebnis	Unvollständiger Rechenweg	Unvollständige Erklärung	Unvollständig	Vollkommen Unvollständig
Angaben der Befragten	Vollständig	0	0	0	0	2	0
	Eher Vollständig	0	0	0	0	1	1
	Teilweise	1	0	0	0	0	0
	Eher unvollständig	0	0	0	0	0	0
	Unvollständig	2	0	0	0	0	1

Tabelle 2: Matrix: Bewertung der ChatGPT-3.5-Antworten zur Vollständigkeit durch Nachhilfe Nehmende

ChatGPT 4.0		Geprüfte Vollständigkeit					
		Vollständig	Fehlendes Ergebnis	Unvollständiger Rechenweg	Unvollständige Erklärung	Unvollständig	Vollkommen Unvollständig
Angaben der Befragten	Vollständig	15	0	0	0	0	0
	Eher Vollständig	20	0	0	0	0	0
	Teilweise	1	0	0	0	0	0
	Eher unvollständig	1	0	0	0	0	0
	Unvollständig	1	0	0	0	0	0

Tabelle 3: Matrix: Bewertung der ChatGPT-4.0-Antworten zur Vollständigkeit durch Studierende

bei Aufgaben höheren Schwierigkeitsgrades auftreten, beispielweise bei jenen, die die Themengebiete Integralrechnung und Gleichungen repräsentieren. Das Kriterium der Sprache und Terminologie wurde im Rahmen der Umfrage in zwei separate Fragen

ChatGPT 4.0		Geprüfte Vollständigkeit					
		Vollständig	Fehlendes Ergebnis	Unvollständiger Rechenweg	Unvollständige Erklärung	Unvollständig	Vollkommen Unvollständig
Angaben der Befragten	Vollständig	4	0	0	0	0	0
	Eher Vollständig	2	0	0	0	0	0
	Teilweise	2	0	0	0	0	0
	Eher unvollständig	1	0	0	0	0	0
	Unvollständig	3	0	0	0	0	0

Tabelle 4: Matrix: Bewertung der ChatGPT-4.0-Antworten zur Vollständigkeit durch Nachhilfe Nimmende

WolframGPT		Geprüfte Vollständigkeit					
		Vollständig	Fehlendes Ergebnis	Unvollständiger Rechenweg	Unvollständige Erklärung	Unvollständig	Vollkommen Unvollständig
Angaben der Befragten	Vollständig	16	0	0	0	0	0
	Eher Vollständig	14	0	0	0	0	0
	Teilweise	4	0	0	0	0	0
	Eher unvollständig	2	0	0	0	0	0
	Unvollständig	2	0	0	0	0	0

Tabelle 5: Matrix: Bewertung der WolframGPT-Antworten zur Vollständigkeit durch Studierende

aufgeteilt, was in der Darstellung zur Folge hat, dass sowohl für die Einschätzung der Terminologie als auch der Sprache separate Matrizen aneinander gefügt wurden. Dieses Auseinanderziehen erschwert den Vergleich der Kriterien.

WolframGPT		Geprüfte Vollständigkeit					
		Vollständig	Fehlendes Ergebnis	Unvollständiger Rechenweg	Unvollständige Erklärung	Unvollständig	Vollkommen Unvollständig
Angaben der Befragten	Vollständig	2	0	0	0	0	0
	Eher Vollständig	4	0	0	0	0	0
	Teilweise	2	0	0	0	0	0
	Eher unvollständig	1	0	0	0	0	0
	Unvollständig	2	0	0	0	0	0

Tabelle 6: Matrix: Bewertung der WolframGPT-Antworten zur Vollständigkeit durch Nachhilfe Nehmende

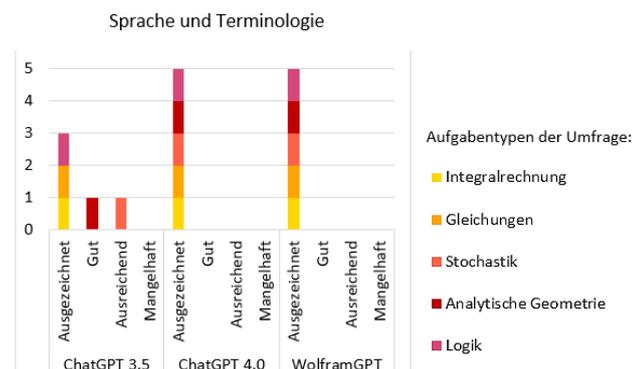


Abbildung 3: Umfrageaufgaben: Sprache und Terminologie

Struktur

Die übersichtliche Strukturierung der Antworten ist von zentraler Bedeutung für deren Übersichtlichkeit und Lesbarkeit. Dieses Kriterium fand jedoch keine explizite Berücksichtigung in der Umfrage, und stattdessen wurde die Entscheidung getroffen, den Fokus auf die ausführlichere Behandlung des Verständnisses der Antworten zu legen. Ungeachtet dieser Entscheidung erfolgt eine Darstellung der in der Datenanalyse klassifizierten Kriterien im anschließenden Diagramm.

ChatGPT 3.5		Geprüfte Sprache und Terminologie			
		Ausgezeichnet	Gut	Ausreichend	Mangelhaft
Angabe über Terminologie	Korrekt	13	2	1	0
	Unsicher	11	4	3	0
	Inkorrekt	4	0	0	0
Angabe über verwendete Sprache	Verständlich	9	3	4	0
	Eher verständlich	11	2	0	0
	Teilweise	4	1	0	0
	Eher unverständlich	4	0	0	0
	Unverständlich	0	0	0	0

Tabelle 7: Matrix: Bewertung der ChatGPT-3.5-Antworten zur Sprache und Terminologie durch Studierende

ChatGPT 3.5		Geprüfte Sprache und Terminologie			
		Ausgezeichnet	Gut	Ausreichend	Mangelhaft
Angabe über Terminologie	Korrekt	3	2	1	0
	Unsicher	2	1	1	0
	Inkorrekt	0	0	0	0
Angabe über verwendete Sprache	Verständlich	3	1	1	0
	Eher verständlich	1	0	1	0
	Teilweise	1	0	1	0
	Eher unverständlich	0	0	0	0
	Unverständlich	0	0	0	0

Tabelle 8: Matrix: Bewertung der ChatGPT-3.5-Antworten zur Sprache und Terminologie durch Nachhilfe Nehmende

ChatGPT 4.0		Geprüfte Sprache und Terminologie			
		Ausgezeichnet	Gut	Ausreichend	Mangelhaft
Angabe über Terminologie	Korrekt	27	0	0	0
	Unsicher	9	0	0	0
	Inkorrekt	2	0	0	0
Angabe über verwendete Sprache	Verständlich	23	0	0	0
	Eher verständlich	14	0	0	0
	Teilweise	1	0	0	0
	Eher unverständlich	0	0	0	0
	Unverständlich	0	0	0	0

Tabelle 9: Matrix: Bewertung der ChatGPT-4.0-Antworten zur Sprache und Terminologie durch Studierende

ChatGPT 4.0		Geprüfte Sprache und Terminologie			
		Ausgezeichnet	Gut	Ausreichend	Mangelhaft
Angabe über Terminologie	Korrekt	8	0	0	0
	Unsicher	4	0	0	0
	Inkorrekt	0	0	0	0
Angabe über verwendete Sprache	Verständlich	7	0	0	0
	Eher verständlich	3	0	0	0
	Teilweise	2	0	0	0
	Eher unverständlich	0	0	0	0
	Unverständlich	0	0	0	0

Tabelle 10: Matrix: Bewertung der ChatGPT-4.0-Antworten zur Sprache und Terminologie durch Nachhilfe Nehmenden

WolframGPT		Geprüfte Sprache und Terminologie			
		Ausgezeichnet	Gut	Ausreichend	Mangelhaft
Angabe über Terminologie	Korrekt	32	0	0	0
	Unsicher	6	0	0	0
	Inkorrekt	0	0	0	0
Angabe über verwendete Sprache	Verständlich	24	0	0	0
	Eher verständlich	12	0	0	0
	Teilweise	2	0	0	0
	Eher unverständlich	0	0	0	0
	Unverständlich	0	0	0	0

Tabelle 11: Matrix: Bewertung der WolframGPT-Antworten zur Sprache und Terminologie durch Studierende

WolframGPT		Geprüfte Sprache und Terminologie			
		Ausgezeichnet	Gut	Ausreichend	Mangelhaft
Angabe über Terminologie	Korrekt	6	0	0	0
	Unsicher	5	0	0	0
	Inkorrekt	0	0	0	0
Angabe über verwendete Sprache	Verständlich	5	0	0	0
	Eher verständlich	2	0	0	0
	Teilweise	2	0	0	0
	Eher unverständlich	0	0	0	0
	Unverständlich	0	0	0	0

Tabelle 12: Matrix: Bewertung der WolframGPT-Antworten zur Sprache und Terminologie durch Nachhilfe Nehmenden

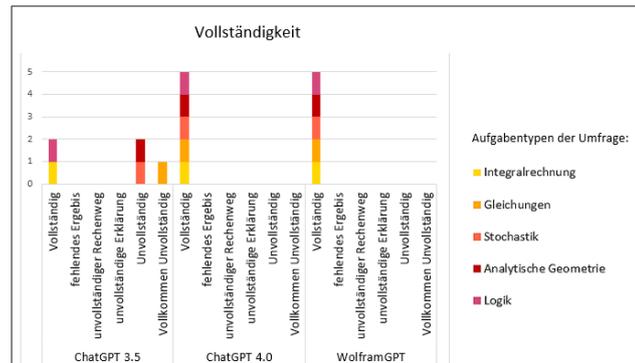


Abbildung 4: Umfrageaufgaben: Struktur

C. Datensammlung

Dieser Anhang dient dem Zugriff auf die vollständige Sammlung der Daten, die über zwei unterschiedliche Methoden für diese Forschungsarbeit erhoben wurden. Zur Gewährleistung von Transparenz und Nachvollziehbarkeit der Forschungsergebnisse sind diese Informationen in einem öffentlich zugänglichen GitHub-Repository hinterlegt. In diesem Repository sind alle durch die drei untersuchten ChatGPT-Versionen generierten Antworten enthalten. Zusätzlich wurde für jede der gestellten Fragen die erstellte Musterlösung bereitgestellt. Ebenfalls zugänglich sind der verwendete Umfragebogen sowie die Ergebnisse der Umfrage.

Das Repository kann über den folgenden Link erreicht werden:
https://github.com/slangre/ChatGPT_Math_Tutor



Abbildung 5: QR-Code für GitHub

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Gummersbach, 09.09.2024
Ort, Datum


Unterschrift