

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Potenziale und Grenzen von Künstlicher Intelligenz im
Unterrichtsfach Mathematik: Eine Analyse und ein Praxistest von
ChatGPT

verfasst von | submitted by
Daniel Leeb BEd

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Education (MEd)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 199 520 523 02

Studienrichtung lt. Studienblatt | Degree
programme as it appears on the student
record sheet:

Masterstudium Lehramt Sek (AB) Unterrichtsfach
Mathematik Unterrichtsfach Physik

Betreut von | Supervisor:

Assoz. Prof. Mag. Mag. Dr. Christoph Ableitinger
Privatdoz.

Abstract (Deutsch)

Die fortschreitende Digitalisierung beeinflusst immer mehr Aspekte des Lebens und Arbeitens, insbesondere auch das Bildungswesen und die Schule. In den letzten drei Jahren hat die Künstliche Intelligenz (KI) zunehmend an Bedeutung gewonnen und rückt verstärkt in den Mittelpunkt des öffentlichen Interesses. KI-gestützte Sprachmodelle bieten dabei neue Möglichkeiten für Lehr- und Lernprozesse. Im Mathematikunterricht könnte das Sprachmodell ChatGPT als digitaler Lernbegleiter dienen, der Schüler*innen individuelle Hilfestellungen bietet und Lösungswege erklärt. Dabei stellt sich jedoch die Frage, inwieweit KI mathematische Aufgaben korrekt lösen kann und welche Auswirkungen ihr Einsatz haben könnte. Auch für Lehrpersonen eröffnen sich neue Möglichkeiten in Bezug auf die Planung von Unterrichtseinheiten und die Aufgabenerstellung. Die vorliegende Arbeit analysiert, inwiefern ChatGPT in der Lage ist, mathematische Aufgaben auf Schulniveau zu lösen. Im Fokus der Untersuchung steht die Leistungsfähigkeit von ChatGPT bei der Bearbeitung der österreichischen AHS-Matura, die sich aus Aufgaben mit unterschiedlichen Antwortformaten und Inhalten verschiedener Themengebiete zusammensetzt. Zudem wird getestet, wie die Formulierung von Eingabeaufforderungen die Qualität der Ausgaben beeinflusst. Die Ergebnisse zeigen, dass ChatGPT in der Lage ist, mathematische Aufgaben korrekt zu lösen und die Prüfungen positiv zu absolvieren. Bei komplexeren Aufgabenstellungen, bei denen visuelle Analysefähigkeiten gefragt sind, stößt das Modell jedoch (vorerst) an seine Grenzen. Der Einsatz von KI im Mathematikunterricht hat Potenzial für die individuelle Unterstützung von Lernenden, Lösungswege können detailliert aufgeschlüsselt werden und Sachverhalte in einfachen Worten erklärt werden. Gleichzeitig stellt dies aber auch eine Gefahr für das eigenständige Denken von Schüler*innen dar. Lehrpersonen können bei versierter Anwendung der KI mit Zeitersparnissen in der Vor- und Nachbereitung des Unterrichts und der Prüfungserstellung und -bewertung rechnen. Die Geschwindigkeit des technologischen Fortschritts erfordert eine ständige Reevaluierung des KI-Einsatzes, um das Lehren und Lernen von Mathematik in Zukunft noch kreativer und effizienter zu gestalten.

Abstract (English)

Digitalization is increasingly influencing all areas of life and work. School and education are no exception of this trend. In recent years, the use of Artificial Intelligence (AI) has come into focus. AI-powered language models offer new possibilities for teaching and learning processes. In mathematics education, the language model ChatGPT could serve as a digital learning companion, providing students with individual assistance and explaining solutions. However, the question arises as to what extent AI can solve mathematical tasks correctly and what impact its use could have. For teachers, there are also new opportunities regarding lesson planning and task creation. This study analyzes the capabilities of ChatGPT in solving mathematical tasks at school level. For this investigation ChatGPT was confronted with the Austrian AHS-Matura, which consists of tasks with different formats and covers a variety of topics. Additionally, the study tests how the phrasing of prompts affects the quality of the outputs. The results show that ChatGPT is capable of solving mathematical tasks correctly and passing all tested exams. However, when faced with more complex tasks that require visual analysis skills, the model reaches its limits (for now). The use of AI in mathematics education has the potential to provide individualized support for learners, allowing solutions to be broken down in detail and concepts explained in simple terms. But at the same time, it also poses a risk to students' independent thinking. Teachers, when using AI proficiently, can expect saving time in the preparation and follow up work of lessons, as well as in the creation and evaluation of exams. The speed of technological progress requires continuous reevaluation of AI to make the teaching and learning of mathematics even more creative and efficient in the future.

Danksagung

Mit dem Abschluss dieser Masterarbeit geht ein bedeutender Lebensabschnitt für mich zu Ende – eine Zeit, die mich nicht nur akademisch, sondern auch menschlich sehr geformt hat. Diese Jahre waren geprägt von persönlicher Weiterentwicklung, Herausforderungen und wertvollen Erfahrungen. An dieser Stelle möchte ich all jenen danken, die mich auf diesem Weg begleitet und unterstützt haben.

Ein herzliches Dankeschön gilt meinem Betreuer Dr. Christoph Ableitinger, der mich von der Themenfindung bis zur finalen Umsetzung dieser Arbeit mit wertvollen Anregungen, hilfreichen Ratschlägen und konstruktivem Feedback unterstützt hat. Seine fachliche Expertise und seine geduldige Begleitung waren äußerst wertvoll.

Ein besonderer Dank geht an all meine Wegbegleiter*innen, insbesondere Iwona C., Marvin Z., Simon G., Sylvia B., Denise F. und Anna H. Ob an der Uni, in der Bibliothek, in der Freizeit oder unterwegs auf Reisen – eure Freundschaft, eure Unterstützung und unsere Gespräche sind von unschätzbarem Wert für mich gewesen. Ich bin sehr froh, euch in dieser prägenden Zeit an meiner Seite gehabt zu haben.

Mein tiefster Dank gilt meiner Partnerin Selina G., du warst mir stets eine große Stütze, hast mir Mut gemacht und mit deinen kreativen Ideen einen wertvollen Beitrag zu dieser Arbeit geleistet. Du bist ein wunderbarer Mensch und ich bin unendlich dankbar, dich an meiner Seite zu wissen.

Der größte Dank jedoch gebührt meiner Familie, meinen Eltern Verena und Herbert, meinem Bruder Dominik und meiner Oma Gerda. Danke, dass ihr mir diese Ausbildung ermöglicht habt, stets ein offenes Ohr hattet und immer für mich da wart. Ohne euch wäre das alles nicht möglich gewesen.

Wien, Jänner 2025

Inhaltsverzeichnis

1. Einleitung	7
1.1. Relevanz und Forschungsfrage	7
1.2. Aufbau und Überblick	8
2. Einführung in die Künstliche Intelligenz	9
2.1. Definition des Begriffs	9
2.2. Geschichte der Künstlichen Intelligenz	10
2.3. Algorithmen	13
2.4. Maschinelles Lernen (ML)	14
2.5. Neuronale Netze	16
2.6. Funktionsweise des Lernprozesses	19
2.7. Deep Learning (DL)	21
2.8. Hardware für KI	22
2.9. Natural Language Processing (NLP)	23
2.10. Prompt Engineering	27
2.10.1. Definition und Relevanz von Prompt Engineering	27
2.10.2. Eigenschaften guter Prompts	28
2.10.3. Arten von Prompts	30
2.11. ChatGPT	33
3. KI und Mathematikunterricht	34
3.1. Relevanz von KI im schulischen Unterricht	34
3.2. KI in der österreichischen Bildungslandschaft	36
3.3. Anwendungsmöglichkeiten und Kompetenzen von KI im Unterrichtsfach Mathematik	39
3.3.1. Lösungskompetenz	39
3.3.2. Planungskompetenz	44
3.3.3. Kontroll- und Bewertungskompetenz	48
3.3.4. Weitere Kompetenzen	50

4. Testdurchführung mit ChatGPT	52
4.1. Lösen von Mathematikaufgaben mit ChatGPT	52
4.1.1. Methode und Durchführung	52
4.1.2. Ergebnisse	56
4.2. Fallbeispiele gelöster Aufgaben	59
4.2.1. Fallbeispiel 1: Zahlenmengen	59
4.2.2. Fallbeispiel 2: Kosten, Erlös und Gewinn	61
4.2.3. Fallbeispiel 3: Vektoren	63
4.2.4. Fallbeispiel 4: Weg, Geschwindigkeit und Beschleunigung	64
4.2.5. Fallbeispiel 5: Einheitskreis	66
4.2.6. Fallbeispiel 6: Ereignisse eines Zufallsversuchs	68
4.2.7. Fallbeispiel 7: Median	69
4.2.8. Fallbeispiel 8: Integral und Flächeninhalt	70
4.2.9. Fallbeispiel 9: Rechtwinkeliges Dreieck	73
4.2.10. Fallbeispiel 10: Histogramm	74
5. Diskussion	76
5.1. Zusammenfassung der Erkenntnisse	76
5.2. Interpretation	77
5.3. Folgen und Nutzen für Schüler*innen und Lehrpersonen	81
5.4. Ausblick und Desiderata	83
6. Literaturverzeichnis	85
7. Abbildungsverzeichnis	94
8. Tabellenverzeichnis	95

1. Einleitung

1.1. Relevanz und Forschungsfrage

Die Digitalisierung prägt zunehmend alle Lebens- und Arbeitsbereiche, insbesondere auch Bildung und Schule. Vor allem in den vergangenen drei Jahren rückt der Einsatz von Künstlicher Intelligenz immer stärker in den Fokus. KI-gestützte Sprachmodelle wie ChatGPT eröffnen neue Möglichkeiten für Lehr- und Lernprozesse, sowohl für Schüler*innen als auch für Lehrpersonen. ChatGPT basiert auf fortschrittlichen Algorithmen des maschinellen Lernens und ist in der Lage, menschenähnliche Texte zu generieren, komplexe Fragestellungen zu beantworten und Problemlösungsstrategien zu entwickeln. Diese Eigenschaften machen das Modell auch für den Einsatz im schulischen Kontext, insbesondere im Mathematikunterricht, interessant. Der Mathematikunterricht stellt bekanntlich hohe Anforderungen an das logische Denken, sowie die Fähigkeit, komplexe Sachverhalte zu analysieren, zu interpretieren und Berechnungen durchzuführen. Gleichzeitig zählt Mathematik zu den Fächern, in denen viele Schüler*innen Schwierigkeiten haben. Hier könnte der Einsatz von KI Lernende gezielt fördern, indem individuelle Hilfestellungen geboten, Lösungswege erklärt oder alternative Herangehensweisen aufgezeigt werden. ChatGPT könnte dabei als digitaler Lernbegleiter fungieren, der rund um die Uhr verfügbar ist und flexibel auf unterschiedliche Lernniveaus eingeht.

Jedoch wirft der Einsatz solcher Technologien auch Fragen auf. Inwiefern kann eine KI wie ChatGPT tatsächlich mathematische Aufgaben korrekt und nachvollziehbar lösen? Welche Grenzen zeigen sich bei der Bearbeitung verschiedener Aufgabenformate? Inwieweit beeinflusst der Einsatz das selbstständige Denken und Lernen von Schüler*innen? Zudem stellt sich die Frage, wie Lehrkräfte KI sinnvoll in den Unterricht einbinden und auch für die Vorbereitung nutzen können. Die vorliegende Arbeit setzt sich das Ziel, die Leistungsfähigkeit von ChatGPT bei der Bearbeitung mathematischer Aufgaben auf Schulniveau systematisch zu untersuchen. Hierbei liegt der Fokus auf der Analyse verschiedener Aufgabenformate, darunter offene, halboffene, Multiple-Choice-, Zuordnungs-, Lückentext- und Konstruktionsaufgaben. Zusätzlich werden unterschiedliche mathematische Inhaltsbereiche wie Algebra, Geometrie, Analysis, Wahrscheinlichkeit und Statistik betrachtet, um ein umfassendes Bild der Stärken und Schwächen von ChatGPT zu zeichnen. Dazu werden die Aufgaben der österreichischen

AHS-Matura herangezogen. Ein besonderes Augenmerk liegt darauf, wie sich die Formulierung von Eingabeaufforderungen, sogenannte Prompts, auf die Qualität und Korrektheit der von ChatGPT generierten Antworten auswirkt. Die Analyse soll aufzeigen, ob bestimmte Formulierungen zu besseren Ergebnissen führen. Diese Erkenntnisse sind nicht nur für die Bewertung der Leistungsfähigkeit von ChatGPT relevant, sondern auch für die davon abhängige Anwendung in der Unterrichtspraxis. Die Ergebnisse dieser Untersuchung sollen einen Beitrag zur aktuellen Diskussion über den sinnvollen Einsatz von KI im Bildungswesen leisten. Es soll aufgezeigt werden, in welchen Bereichen ChatGPT als unterstützendes Tool im Mathematikunterricht nutzbringend eingesetzt werden kann und wo potenzielle Risiken bestehen. Darüber hinaus sollen Handlungsempfehlungen für den sinnvollen Umgang mit KI im schulischen Mathematikunterricht abgeleitet werden. Die Arbeit möchte damit sowohl Lehrenden als auch Lernenden Orientierung bieten, wie KI gezielt zur Förderung mathematischer Kompetenzen genutzt werden kann.

1.2. Aufbau und Überblick

Die vorliegende Arbeit ist in vier Hauptkapitel unterteilt, die aufeinander aufbauen und eine Analyse der Künstlichen Intelligenz und ihres Einsatzes im Mathematikunterricht ermöglichen. Kapitel 2 bildet mit einer Einführung in die Künstliche Intelligenz die theoretische Grundlage. Zunächst wird der Begriff der Künstlichen Intelligenz definiert, im Anschluss wird die Geschichte der KI nachgezeichnet, um den historischen Kontext und die Entwicklung der Technologie nachvollziehbar zu machen. Ein weiterer wichtiger Abschnitt widmet sich den technischen Grundlagen der KI, darunter Algorithmen und maschinelles Lernen (ML), die die Grundlage vieler KI-Systeme bilden. Die Funktionsweise neuronaler Netze und die Fortschritte im Bereich des Deep Learnings (DL) werden ebenso erörtert, wie die Rolle des Natural Language Processings (NLP). Dies stellt eine spezifische Anwendung dar, die für den Umgang von KI mit menschlicher Sprache von zentraler Bedeutung ist. Ein weiterer Schwerpunkt liegt auf dem Thema Prompt Engineering, das sich mit der Art und Weise befasst, wie Eingaben an die KI formuliert werden, um die Qualität der Antworten zu verbessern. Kapitel 3 ist das Bindeglied zwischen KI und Unterricht, wobei die Möglichkeiten und Herausforderungen der Integration von KI im schulischen Mathematikunterricht analysiert werden. Zunächst wird die allgemeine Relevanz von KI im Bildungswesen erörtert und die Rolle von KI in der österreichischen Bildungslandschaft untersucht. Der Schwerpunkt dieses Kapitels liegt jedoch

auf den konkreten Anwendungsmöglichkeiten von KI im Fach Mathematik. Hier wird aufgezeigt, wie KI Lehrkräfte in verschiedenen Bereichen unterstützen kann, von der Lösung von Aufgaben über die Planung von Unterrichtseinheiten bis hin zur Kontrolle von Schüler*innenantworten. Die eigens durchgeführte Testung mit ChatGPT in Kapitel 4 bildet das Herzstück der Arbeit. Die Leistungsfähigkeit von ChatGPT bei der Lösung von Mathematikaufgaben wird hiermit evaluiert. Zunächst wird die Methodik der Testdurchführung erläutert, wobei die Auswahl der Aufgaben, die Durchführung und die Auswertung beschrieben werden. Anschließend werden die Ergebnisse der Testdurchführung präsentiert und anhand von Fallbeispielen illustriert. Die Analyse der Ergebnisse zeigt auf, in welchen Bereichen ChatGPT bereits gute Leistungen erbringt und wo noch Schwächen bestehen. Auf Grundlage dieser Erkenntnisse werden in Kapitel 5 die Anwendungsmöglichkeiten für Lehrpersonen diskutiert. Zudem wird auch der Nutzen für Schüler*innen thematisiert sowie untersucht, in welchen Bereichen sie die KI vorteilhaft einsetzen können. Im Ausblick werden offene Forschungsfragen und mögliche Weiterentwicklungen der KI-Technologie skizziert, die für den Mathematikunterricht von Bedeutung sein könnten.

2. Einführung in die Künstliche Intelligenz

2.1. Definition des Begriffs

Künstliche Intelligenz, kurz KI, oder Artificial Intelligence (AI) aus dem Englischen, bezeichnet ein Teilgebiet der Informatik, das in den vergangenen Jahren in der breiten Öffentlichkeit stark thematisiert wurde. Bereits jetzt wird sie als die Kerntechnologie des 21. Jahrhunderts bezeichnet, die bedeutende Teilbereiche unseres Lebens verändern wird (Kitzmann, 2022). Da bereits der Begriff der Intelligenz keine eindeutige und allumfassende Beschreibung zulässt, gilt selbiges auch für die Künstliche Intelligenz. Mögliche Definitionen gibt es aber dennoch:

„Künstliche Intelligenz ist der Versuch, rationale bzw. kognitive menschliche Intelligenz auf (technischen) Maschinen zu simulieren, um sie für den Menschen gewinn- und nutzbringend einzusetzen“ (Otte, 2021, S. 2).

Im Vordergrund steht die Nachahmung der menschlichen Intelligenz. Fähigkeiten wie logisches und kreatives Denken, Planen und Lernen sollen von technischen Geräten durchgeführt werden, um Probleme zu lösen und bestimmte Ziele zu erreichen. Hervorgehoben wird zudem die Anpassungsfähigkeit des Systems auf Basis vorheriger Eingaben (Europäisches Parlament, 2020). Forschungseinrichtungen und IT-Firmen wie Google fassen den Begriff technischer auf:

„Künstliche Intelligenz (KI) umfasst eine Reihe von Technologien, mit denen Computer eine Vielzahl erweiterter Funktionen ausführen können, darunter die Möglichkeit, gesprochene und geschriebene Sprachen zu erkennen, zu verstehen und zu übersetzen, Daten zu analysieren, Empfehlungen auszugeben und mehr“ (Google Cloud, o.D.).

Die Eingaben, meist in Form von Text, sollen erkannt und sortiert werden und auf Basis von Programmierung oder maschinellern Lernen hilfreiche Ausgaben erzeugen. In den letzten Jahren hat sich das maschinelle Lernen erheblich weiterentwickelt, was hauptsächlich auf die gestiegene Verfügbarkeit großer Datenmengen und die erhöhte Rechenleistung zurückzuführen ist (Fraunhofer-Institut für Kognitive Systeme, 2024).

2.2. Geschichte der Künstlichen Intelligenz

In diesem Kapitel angeschnittene Begriffe und wichtige Konzepte werden in den folgenden Kapiteln noch präziser erläutert, um die Basis der KI-Technologie besser zu verstehen. Der Begriff der Künstlichen Intelligenz ist allerdings alles andere als neu. Blickt man zurück, stößt man auf Alan Turing, der bereits 1947 Vorträge zum Thema hielt und drei Jahre später den Artikel „Computing Machinery and Intelligence“ publizierte. In diesem stellte er erstmals den Turing-Test vor und behandelte Konzepte wie maschinelles und verstärkendes Lernen, sowie genetische Algorithmen. Er war der Meinung, dass es einfacher sei, Algorithmen zu entwickeln, die das Lernen ermöglichen, anstatt die Intelligenz einer Maschine manuell zu programmieren (Russel et al., 2016). Als Geburtsstunde des modernen KI-Begriffs gilt der von John McCarthy, Marvin Minsky, Nathaniel Rochester und Claude Shannon initiierte zweimonatige Forschungsworkshop „Summer Research Project on Artificial Intelligence“ am Dartmouth College im Jahr 1956. Dessen Ziel war es, herauszufinden, inwiefern Maschinen

Sprache verwenden, Abstraktionen bilden, komplexe Probleme lösen und sich selbst verbessern können. Der Workshop selbst blieb wenig erfolgreich. Im folgenden Jahrzehnt wurden allerdings ansehnliche Fortschritte erzielt, das veröffentlichte Computerprogramm „Logic Theorist“ konnte eine Vielzahl der Theoreme der Principia Mathematica, einem Grundlagenwerk der axiomatischen Mathematik, beweisen. Weiters wurden Computerprogramme und -sprachen wie LISP, ELIZA und SHRDLU veröffentlicht, die Texteingaben in menschlicher Sprache verstehen und interpretieren konnten, was als „Natural Language Processing“ (NLP) bezeichnet wird. LISP wurde 1958 am MIT AI Lab von McCarthy entwickelt und stieg zur dominierenden KI-Programmiersprache für die kommenden 30 Jahre auf (Heinrichs et al., 2022). Bis in die 1980er Jahre herrschte im Forschungsbereich eine gewisse Euphorie, was vor allem mit der rasanten Entwicklung der Computertechnologie zu tun hatte. Als jedoch der wirtschaftliche Erfolg bestimmter KI-Systeme ausblieb, wurden die Fördermittel erheblich gekürzt, was zum Rückgang der Forschung in diesem Bereich führte (De Florio-Hansen, 2020).

Im Jahr 1980 entwarf der amerikanische Philosoph John Searle das Gedankenexperiment des Chinesischen Zimmers. Er stellte damit den Turing-Test infrage, also das Kriterium, ab wann man einer Maschine Bewusstsein oder Intelligenz zuschreiben könne. Der Ablauf des Experiments geht wie folgt: Man stellt sich vor, in einem geschlossenen Raum zu sitzen, ohne Chinesisch zu verstehen. Außerhalb des Raums befinden sich Personen, die chinesische Fragen auf Papier in den Raum schieben. Im Raum gibt es ein Buch mit detaillierten Anweisungen (z.B. auf Deutsch), das erklärt, wie man auf jede chinesische Frage eine passende chinesische Antwort schreibt. Man bearbeitet die Fragen, indem man die Symbole entsprechend der Anleitung manipuliert, bis man eine Antwort erzeugt, die man wieder durch den Schlitz nach draußen schiebt. Von außen betrachtet scheint es, als würde die Person im Raum chinesisch verstehen, da die Antworten für die Außenstehenden korrekt erscheinen (Otte, 2021). Die reine Symbolmanipulation ist allerdings als syntaktisch zu betrachten, von einem Sprachverständnis der Person im Zimmer kann nicht die Rede sein. Genauso verhält es sich mit derzeit verfügbaren Programmen mit Künstlicher Intelligenz, sie können als Analogon zum Chinesischen Zimmer aufgefasst werden (Heinrichs et al., 2022).

Mitte der 1980er Jahre wurde der Backpropagation-Algorithmus, ein Verfahren zum Training künstlicher neuronaler Netze, wiederentdeckt und fand breite Anwendung in der Informatik.

Aufgrund der Schwächen früherer Systeme setzte man zunehmend auf Wahrscheinlichkeit, maschinelles Lernen und experimentelle Methoden, der Fokus lag nun mehr auf praxisrelevanten Anwendungen. Die Integration von Daten, statistischer Modellierung, Optimierung und maschinellem Lernen führte zur allmählichen Wiedervereinigung von Teilbereichen wie Computer Vision, Robotik, Spracherkennung und Verarbeitung natürlicher Sprache, die sich über die Jahre vom Kernforschungsbereich entfernt hatten. Die Zeiten rund um die Jahrtausendwende waren von massiven Fortschritten der Computerrechenleistung und dem Ausbau des Internets geprägt. Damit einher ging die Sammlung von Unmengen an Daten, egal ob in Form von Text, Bild, Video oder Sprache. Man spricht in diesem Zusammenhang von „Big Data“. Die Datensätze werden wiederum für das maschinelle Lernen verwendet (Russel et al., 2016).

Nicht unerwähnt lassen sollte man die frühen Erfolge der Künstlichen Intelligenz, die die Aufmerksamkeit der Öffentlichkeit erregten und ihre Fähigkeiten aufzeigten. 1997 schlägt der von IBM entwickelte Schachcomputer Deep Blue den zu dieser Zeit amtierenden Schachweltmeister Gary Kasparov mit 3,5 zu 2,5. Vierzehn Jahre später brilliert die KI auch in der US-Fernsehquizshow Jeopardy, bei der es auf das Sprachverständnis von Wortwitzen und Sarkasmus ankommt. Die IBM-Software besiegt bisherige Rekordgewinner der Show. 2016 kommt es zum Duell zwischen dem von Google entwickelten Programm Alpha-Go und dem weltbesten Go-Spieler. Die KI fertigt den Südkoreaner Lee Sedol mit 4 zu 1 ab. Go ist ein strategisches Brettspiel aus Asien, bei dem zwei Spieler abwechselnd schwarze und weiße Steine auf einem Gitter platzieren, um durch das Umzingeln von Gebieten auf dem Brett möglichst viel Raum zu kontrollieren. Allgemeinhin gilt Go als komplexer und schwieriger zu meistern als Schach (De Florio-Hansen, 2020).

Im vergangenen Jahrzehnt gab es große Fortschritte im Bereich der Sprach- und Bilderkennung, was vor allem auf „Deep Learning“ zurückzuführen ist. Beim ImageNet-Wettbewerb 2012, bei dem die teilnehmenden Systeme mehrere Millionen gegebene Bilder in über tausend Kategorien einordnen mussten, konnte sich das System aus der Deep Learning-Gruppe von Geoffrey Hinton an der Universität Toronto, mit einer ca. 10% niedrigeren Fehlerrate von den anderen bestehenden Systemen deutlich abheben. Seitdem haben Deep Learning-Systeme in einigen Bereichen der Bildverarbeitung die menschliche Leistung übertroffen, in anderen wiederum hinken sie hinterher. Ähnliche Fortschritte wurden auch in der Spracherkennung, der

maschinellen Übersetzung und der medizinischen Diagnose erzielt (Russel et al., 2016). Im November 2022 veröffentlichte das US-Unternehmen OpenAI das Chat- und Dialogsystem „ChatGPT“, welches sich durch seine vielseitigen Anwendungsmöglichkeiten von bisherigen Systemen abhob. Offen formulierte Eingaben von Nutzer*innen konnte das System in den meisten Fällen passend bearbeiten und beantworten (Helfrich-Schkarbanenko, 2023). Innerhalb kürzester Zeit verzeichnete das Programm einen starken Anstieg der Zugriffszahlen, Ende Jänner 2023 waren es laut Reuters 100 Millionen monatlich aktive Nutzer*innen. In diesem Monat gab es durchschnittlich rund 13 Millionen einzelne Besucher*innen täglich. Damit ist es die bisher am schnellsten wachsende (Internet-)Anwendung für Endverbraucher*innen (Hu, 2023).

2.3. Algorithmen

Algorithmen sind entscheidend für eine Künstliche Intelligenz, da sie die Schritte bestimmen, mit denen eine Maschine Daten analysiert und daraus lernt. Sie können als deren „mathematische Basis“ aufgefasst werden. Ein Algorithmus ist eine strukturierte, schrittweise Abfolge, die ein Problem systematisch löst. Diese Abfolge von definierten Schritten basiert auf mathematischen und logischen Grundsätzen und ermöglicht die Umwandlung von Eingabedaten in gewünschte Ausgabedaten (Otte, 2021). Weiters existieren ähnliche Begriffsdefinitionen:

„Ein Algorithmus ist eine endliche Folge von eindeutig bestimmten Elementaranweisungen, die den Lösungsweg eines Problems exakt und vollständig beschreiben“ (Ziegenbalg et al., 2016, S. 26).

Einfach ausgedrückt können auch wir Menschen eine Folge von Handlungsanweisungen befolgen, Kochrezepte und Verkehrsregeln basieren beispielsweise auf Algorithmen. Der Algorithmus zum Starten eines Autos lautet beispielsweise wie folgt: einsteigen, anschnallen, Schlüssel ins Zündschloss, Kupplung steigen, ersten Gang einlegen, Schlüssel im Zündschloss drehen (De Florio-Hansen, 2020).

Die Wurzeln des Algorithmus-Begriffs liegen in den Arbeiten des persischen Mathematikers Al-Charismi im 9. Jahrhundert, dessen Werk die Rechenmethoden der arabischen Zahlen in

Europa einführte (Russell et al., 2016). Algorithmen sind heute allgegenwärtig und bilden die Basis für viele Anwendungen in der digitalen Welt, von einfachen mathematischen Berechnungen bis hin zu hochkomplexen Entscheidungsprozessen der Künstlichen Intelligenz. Sie sorgen etwa in Suchmaschinen für die Sortierung von Ergebnissen oder die Personalisierung von Werbung und Empfehlungen (de Witt et al., 2023).

2.4. Maschinelles Lernen (ML)

Maschinelles Lernen ist das Herzstück einer KI und entscheidend für ihr Verständnis. Hierbei kommen Algorithmen zum Einsatz, die die Leistung eines Systems kontinuierlich optimieren. Diese Modelle benötigen zunächst umfangreiche Trainingsdatensätze, anhand derer sie charakteristische Merkmale und Zusammenhänge identifizieren. Beim Training anhand von Algorithmen entsteht ein mathematisches Modell, das diese Muster abbildet und zur Unterstützung von Vorhersagen oder Entscheidungen verwendet wird. Bekannte KI-Systeme wie ChatGPT funktionieren nach diesem Prinzip und lassen sich als fortschrittliche „Vorhersagemaschinen“ beschreiben (Shah, 2023). Beispielsweise können Roboter eigenständig erlernen, wie sie Objekte richtig greifen und bewegen sollen. Dies gelingt ihnen, indem sie durch wiederholte Versuche herausfinden, welche Griffe funktionieren und welche nicht. Der Prozess ist schematisch in Abbildung 1 dargestellt. Das dabei gewonnene Wissen basiert auf Rückmeldungen über erfolgreiche und weniger erfolgreiche Versuche, was zu einer stetigen Optimierung führt (Fraunhofer-Institut für Kognitive Systeme, 2024).

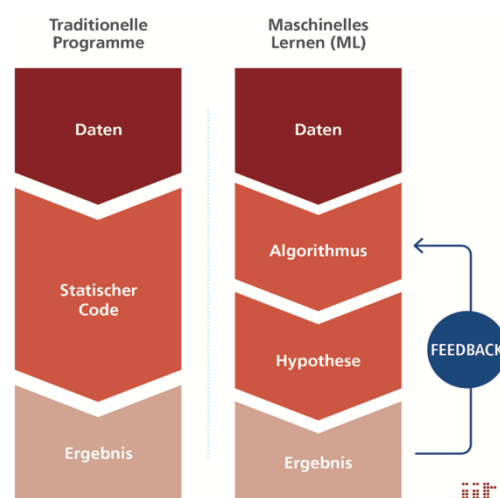


Abbildung 1: Traditionelle Programme und maschinelles Lernen (Wittpahl, 2019, S. 25)

Die Modellbildung basiert auf dem Prinzip der Induktion, aus gegebenen Messdaten soll auf allgemeine Aussagen bzw. Hypothesen geschlossen werden. Dieser Ansatz ist jedoch nicht immer fehlerfrei. Angenommen man sieht in Europa stets weiße Schwäne, könnte man daraus die Hypothese ableiten: „Alle Schwäne sind weiß“. Diese Hypothese erscheint zunächst sinnvoll und wird so lange als gültig betrachtet, bis sie durch Beobachtungen widerlegt wird, wie etwa durch die Entdeckung schwarzer Schwäne in anderen Regionen. Induktiv gewonnenes Wissen, also Wissen, das auf Mustern oder Häufigkeiten basiert, bleibt stets eine Annahme, die unter gewissen Umständen falsch sein kann. Solche Hypothesen haben Wahrscheinlichkeitscharakter. Erst durch deduktive Beweise, die eine Hypothese in ein strenges logisches Gefüge einbetten, kann eine Aussage als Wahrheit akzeptiert werden. Ohne deduktiven Nachweis bleibt eine Hypothese immer eine Wahrscheinlichkeitsaussage und nicht endgültig richtig (Otte, 2021). Die Algorithmen und Modelle sind nicht statisch. Mit jeder weiteren Anwendung und neuen Daten verbessern sie sich und können immer präzisere oder individuellere Ergebnisse liefern. So wird KI für vielfältige Anwendungsbereiche optimiert, von Sprachverarbeitung über Bildanalyse bis hin zu komplexen Entscheidungsprozessen (Shah, 2023).

Bei überwachten Lernverfahren wird für jede Eingabe die gewünschte Ausgabe klar definiert, dazu werden Methoden wie jene der Klassifikation oder Regression verwendet. Zum Beispiel soll ein System lernen, Schriftzeichen anhand ihrer visuellen Merkmale zu identifizieren. Besonders herausfordernd ist die Erkennung handgeschriebener Zeichen, da diese je nach Person in Form, Neigung und Größe stark variieren. Es gibt keine universelle Definition eines Buchstabens wie „A“, die alle Varianten einbezieht. Um diese Vielfalt zu bewältigen, greift das maschinelle Lernen auf zahlreiche handschriftliche Beispiele zurück und erkennt daraus die charakteristischen Merkmale eines Zeichens. So wird das „A“ nicht durch eine festgelegte Definition, sondern durch die häufig vorkommenden Merkmale und Muster von Strichen im Schriftbild identifiziert. Legt man dem System nun ein Bild des Buchstaben „A“ vor, so soll die Antwort „A“ eine hohe Aktivität zeigen bzw. die Antwort mit der höchsten Wahrscheinlichkeit sein. Solche Aufgaben werden als Klassifikationen bezeichnet (Alpaydin, 2019). Bei Vorhersageaufgaben mit kontinuierlichen Werten, wie beispielsweise der Prognose von Börsenkursen, werden dagegen Regressionsverfahren eingesetzt. Regressionsmethoden sind statistische Verfahren, um Beziehungen zwischen einer abhängigen Variable und einer oder mehreren unabhängigen Variablen zu erfassen und zu

modellieren. Ziel ist es, diese Verbindungen zu verstehen und damit Werte der abhängigen Variable vorherzusagen. Je nach Art der Daten und Problemstellung gibt es verschiedene Regressionsarten. Bei der linearen Regression mit einer unabhängigen und einer abhängigen Variablen wird beispielsweise versucht, eine Gerade durch eine Menge von Punkten zu legen. Die Abstände der einzelnen Punkte zur Gerade sollen dabei möglichst klein sein (Fischer et al., 2024).

Beim unüberwachten Lernen hingegen erkennt ein System Muster oder Strukturen in den Daten, explizites Feedback in Form von korrekten Vergleichsdaten ist dabei nicht notwendig. Es wird verwendet, wenn keine definierte Erwartung an die Daten besteht oder diese zu komplex sind, um einzelne Kategorien zu benennen (Shah, 2023). Eine zentrale Methode ist das Clustering, bei dem der Algorithmus nützliche Gruppen ähnlicher Datenpunkte identifiziert. Zum Beispiel kann ein Bildverarbeitungssystem, das Millionen Bilder analysiert, eigenständig eine Gruppe identifizieren, die wir Menschen als Katzen erkennen würden. Diese Methode eignet sich besonders gut, wenn keine vorher festgelegten Kategorien vorhanden sind. In der Astronomie nutzen Forscher*innen das unüberwachte Clustering, um verschiedene Arten von Sternen wie „rote Riesen“ und „weiße Zwerge“ zu klassifizieren. Eine Unterscheidung, die erst durch die Analyse von Sterndaten und Spektren entstanden ist. Der Algorithmus gruppiert die Sterne anhand ihrer Spektralintensitäten und ermittelt so, ob verschiedene Typen existieren und welche Merkmale diese jeweils haben (Russel et al., 2016).

Beim bestärkenden Lernen, oder auch Reinforcement Learning, führt das System eigenständig Aktionen aus und erhält dabei positives oder negatives Feedback. Ziel des Systems ist es die Belohnung zu maximieren (Fischer et al., 2024). Diese Methode wird etwa in Spielen eingesetzt, bei denen die KI durch wiederholte Züge lernt, zu gewinnen. Häufig wird das Lernen auch durch menschliches Feedback unterstützt, das die Qualität der KI-Ausgabe bewertet und zur Optimierung des Modells beiträgt (Shah, 2023).

2.5. Neuronale Netze

Künstliche Neuronale Netze (KNN) sind das zentrale Element des maschinellen Lernens und orientieren sich an den Prinzipien des menschlichen Gehirns. Sie bestehen aus vielen einzelnen Neuronen, die in einem Netzwerk durch gewichtete Verbindungen, sogenannte Synapsen,

miteinander verknüpft sind (Otte, 2021). Auch wenn künstliche neuronale Netze vereinfachte Nachbildungen der biologischen Neuronen sind, beruhen viele ihrer Grundprinzipien auf natürlichen Mechanismen. Das Verständnis über biologische Neuronen hilft daher, die Funktionsweise und Struktur künstlicher Neuronen besser nachvollziehen zu können (Fischer et al., 2024).

Jedes Neuron im menschlichen Körper setzt sich grundsätzlich aus einem Zellkörper, einem Axon und einer unterschiedlichen Anzahl von Synapsen und Dendriten zusammen. Letztere wirken als Eingangskanäle eines Neurons, die Signale von anderen Nervenzellen über Synapsen aufnehmen und zum Zellkörper weiterleiten. Dort werden die Signale summiert, überschreiten sie einen Schwellenwert, feuert das Neuron ein Signal. Dieses Ausgangssignal wird über das Axon, die Ausgangsleitung, zu den Synapsen transportiert und löst dort eine chemische Reaktion aus. Dies wird präsynaptische Aktivität genannt. Von hier gelangen neue elektrische Impulse zum nächsten Neuron. Wenn diese Impulse das Zielneuron zum Feuern anregen, spricht man von postsynaptischer Aktivität. Im Gehirn äußert sich ein Lernprozess durch Anpassungen der Synapsenstärke zwischen Neuronen (Ziegenbalg et al., 2016). Die Hebb'sche Lernregel, aufgestellt von Donald Hebb im Jahr 1949, gilt als eine der grundlegendsten und elegantesten Lernprinzipien neuronaler Netze. Hebb nahm an, dass Lernen dann stattfindet, wenn zwei Neuronen gleichzeitig aktiv sind, wodurch die Übertragungsstärke der Synapsen zwischen ihnen gestärkt wird. Dies ist der Fall, wenn prä- und postsynaptische Erregung gleichzeitig vorhanden sind (Otte, 2021). Das menschliche Gehirn unterscheidet sich grundlegend von einem Computer. Statt eines einzelnen Prozessors nutzt es Milliarden von Neuronen, die parallel arbeiten. Während Computer Verarbeitung und Speicherung strikt trennen, sind diese Funktionen im Gehirn eng miteinander verknüpft. Diese dichte Vernetzung und parallele Struktur sind für die beeindruckende Leistung des Gehirns verantwortlich, obwohl die einzelnen Neuronen vergleichsweise langsamer arbeiten als ein Prozessor (Alpaydin, 2019). Im Jahr 1957 stellte Frank Rosenblatt das Perzeptron-Modell vor, ein grundlegendes Modell für künstliche Neuronen, das bis heute die Basis für moderne künstliche neuronale Netzwerke bildet. Dieses Modell ermöglicht durch die Verschaltung einzelner Neuronen den Aufbau beliebiger logischer Schaltkreise und ist damit

die theoretische Grundlage für den Aufbau eines Computers auf Basis neuronaler Strukturen (Fischer et al., 2024).

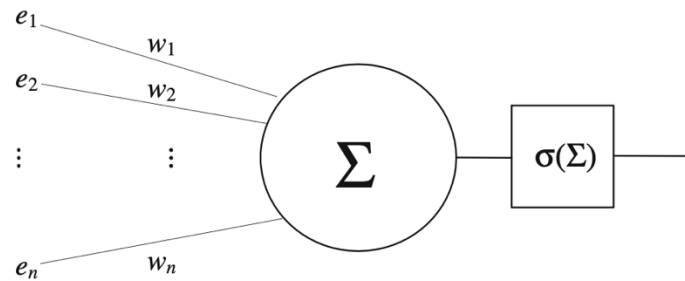


Abbildung 2: Schematisches Modell eines künstlichen Neurons
(Ziegenbalg et al., 2016, S. 324)

In einem mathematischen Modell eines neuronalen Netzwerks werden grundlegende Elemente des biologischen Vorbilds verwendet. Ein vereinfachtes Schema eines künstlichen Neurons ist in Abbildung 2 zu sehen. Die Eingänge e_1 bis e_n stellen entweder direkte sensorische Eindrücke oder Signale anderer Neuronen dar. Diese Eingangssignale werden mit Gewichten w_1 bis w_n versehen. Diese wirken, ähnlich wie reale Synapsen, bei negativen Werten hemmend oder bei positiven Werten verstärkend. Im Zellkörper Σ werden die gewichteten Eingänge summiert und ergeben die Summe $S = \sum_{i=1}^n w_i e_i$. Die Aktivierungsfunktion $\sigma(\Sigma)$ entscheidet, ob das Neuron einen Ausgangsimpuls generiert (Ziegenbalg et al., 2016). Verschiedene Aktivierungsfunktionen können dafür verwendet werden, jede hat ihre Vor- und Nachteile und wahlentscheidend sind das zu lösende Problem und die Struktur des neuronalen Netzwerks. Üblicherweise kommen Funktionen wie die Heavisidefunktion, lineare Funktionen, die Sigmoidfunktion, die Tangens-Hyperbolicus-Funktion oder die Rectified Linear Unit (ReLU) zum Einsatz. Sigmoid- und Tangens-Hyperbolicus-Funktionen eignen sich besonders für Klassifizierungsaufgaben und für Ausgaben mit begrenztem Wertebereich. Die ReLU-Funktion hingegen wird im Deep Learning bevorzugt, da sie durch ihre Einfachheit eine höhere Recheneffizienz bietet (Fischer et al., 2024).

Jedes Neuron führt einfache Berechnungen durch, doch die Leistung des Systems entsteht durch die parallele und gleichzeitige Datenverarbeitung einer Vielzahl solcher Neuronen. Durch die hohe Vernetzung können künstliche neuronale Netze Muster und Strukturen in großen Datenmengen erkennen und komplexe Aufgaben lösen, die für klassische Programmierung oft

zu komplex wären. Die Entwicklung hat inzwischen zahlreiche Arten von neuronalen Netzen hervorgebracht, die für spezifische technische Probleme optimiert sind. Diese Modelle simulieren Aspekte der Informationsverarbeitung des menschlichen Gehirns, wobei jede neue Netzstruktur neue Potenziale für die Datenverarbeitung eröffnet (Otte, 2021).

2.6. Funktionsweise des Lernprozesses

Um den (überwachten) Lernprozess eines solchen neuronalen Netzes besser nachvollziehen zu können, folgen wir einem Beispiel von Ziegenbalg et al. (2016), bei dem die Boole'sche AND-Funktion gelernt werden soll. Die Funktion mit den Eingabewerten x_1 und x_2 ist wie folgt definiert:

$$0 \wedge 0 = 0$$

$$0 \wedge 1 = 0$$

$$1 \wedge 0 = 0$$

$$1 \wedge 1 = 1$$

Die Funktion liefert also nur ein „wahres“ Ergebnis, also eine 1, wenn beide Eingabewerte x_1 und x_2 „wahr“ sind, also beide 1 sind. In den anderen Fällen erhält man ein „falsches“ Ergebnis, also eine 0. Das in Abbildung 3 gezeigte neuronale Netz soll die Modellierung vornehmen, die Anfangsgewichtung sei $w_1 = w_2 = 1$.

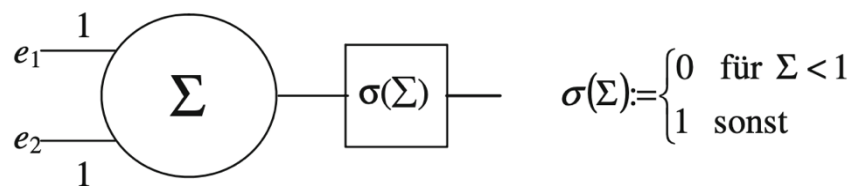


Abbildung 3: Neuronales Netz mit gegebener Anfangsgewichtung (Ziegenbalg et al., 2016, S. 329)

Wählen wir als erste Eingabewerte zweimal die 0, sprich $e_1 = e_2 = 0$, erhalten wir $\sigma(e_1 w_1 + e_2 w_2) = \sigma(0 \cdot 1 + 0 \cdot 1) = \sigma(0) = 0$. Das Ergebnis entspricht damit der Soll-Ausgabe nach $0 \wedge 0 = 0$. Für dieses Eingabepaar muss das Modell also nicht verändert werden.

Betrachten wir nun aber die Eingabewerte $e_1 = 0$ und $e_2 = 1$, erhalten wir $a_{Ist} = \sigma(0 \cdot 1 + 1 \cdot 1) = 1$. Wegen $0 \wedge 1 = 0$ ist die Soll-Ausgabe allerdings 0 ($a_{Soll} = 0$). Die Ist- und Soll-Ausgabe stimmen damit nicht überein und eine Anpassung der Gewichtung ist erforderlich. Die Lernregel, die hier zur Anwendung kommt, wird als Delta-Regel bezeichnet und sieht eine Veränderung der Gewichtung proportional zur Differenz der Soll- und Ist-Ausgabe vor:

$$w_i^{neu} = w_i^{alt} + \alpha \cdot (a_{Soll} - a_{Ist}) \cdot e_i \text{ mit } (i = 1, 2)$$

Die Proportionalitätskonstante α bestimmt die Lernrate des neuronalen Netzes. Eine höhere Lernrate bewirkt stärkere Anpassungen der Gewichtungen in Richtung des aktuellen Trainingsbeispiels, was jedoch das Risiko birgt, dass bereits korrekte Zuordnungen für andere Eingaben wieder „verlernt“ werden. Um das allgemeine Ausgabeverhalten des Netzes nicht zu stark zu beeinträchtigen, ist üblicherweise eine moderate Lernrate zu wählen. Mithilfe der Delta-Regel kann nun ein gezielter Lernschritt ausgeführt werden. Wir setzen dafür $\alpha = 0,5$:

$$\begin{aligned} w_1^{neu} &= 1 + 0,5 \cdot (0 - 1) \cdot 0 = 1 \\ w_2^{neu} &= 1 + 0,5 \cdot (0 - 1) \cdot 1 = 0,5 \end{aligned}$$

Man bekommt für $e_1 = 0$ und $e_2 = 1$ damit die gewollte Ausgabe $a_{Ist} = \sigma(0 \cdot 1 + 1 \cdot 0,5) = \sigma(0,5) = 0$. Eine weitere Trainingseingabe mit $e_1 = 1$ und $e_2 = 1$ verändert die Gewichte auf $w_1 = w_2 = 0,5$. Das Netz ist jetzt so eingestellt, dass es alle vier möglichen Eingaben lösen kann und die erwarteten Ausgaben liefern kann. Die Eingabe $e_1 = e_2 = 1$ wurde beispielsweise nicht erlernt und wird mit $a_{Ist} = \sigma(1 \cdot 0,5 + 1 \cdot 0,5) = \sigma(1) = 1$ korrekt gelöst (Ziegenbalg et al., 2016).

Die im Beispiel verwendete Delta-Lernregel zeigt den Lernprozess anschaulich, dennoch birgt sie für größere neuronale Netze gewisse Einschränkungen. Erstens dürfen mit ihr nur lineare Aktivierungsfunktionen verwendet werden, wie bereits erwähnt gibt es aber eine Vielzahl an möglichen und teils auch günstigere Arten solcher Funktionen. Weiters sind damit auch lediglich stückweise lineare Approximationen möglich. Die Nichtlinearität ermöglicht es ausreichend großen Netzwerken, beliebige Funktionen zu repräsentieren (Russel et al., 2016). Ein zweites Problem ist, dass das Neuron, auf welches die Lernregel angewendet werden soll,

ein Ausgabeneuron sein muss. Dies hat zur Folge, dass das Netz keine Zwischenschichten von Neuronen aufweisen darf (Otte, 2021).

2.7. Deep Learning (DL)

In der Praxis ist eine einzelne Schicht von Neuronen oft nicht ausreichend, um komplexe Probleme zu lösen. Dieses Manko lässt sich durch die Kombination mehrerer lernfähiger Schichten beheben. Mehrschichtige bzw. tiefe neuronale Netze bestehen aus einer Eingabeschicht, einer Ausgabeschicht und mehreren hierarchisch angeordneten, versteckten Schichten. Im Englischen werden diese als „hidden layers“ bezeichnet. In Abbildung 4 ist ein schematischer Vergleich von einschichtigen und mehrschichtigen Netzen gegeben (Ziegenbalg et al., 2016). Es gibt bisher keine festgelegte Anzahl an Schichten zwischen Eingabe- und Ausgabeschicht, die erforderlich ist, um ein maschinelles Lernmodell als Deep Learning-Modell zu klassifizieren (De Florio-Hansen, 2020).

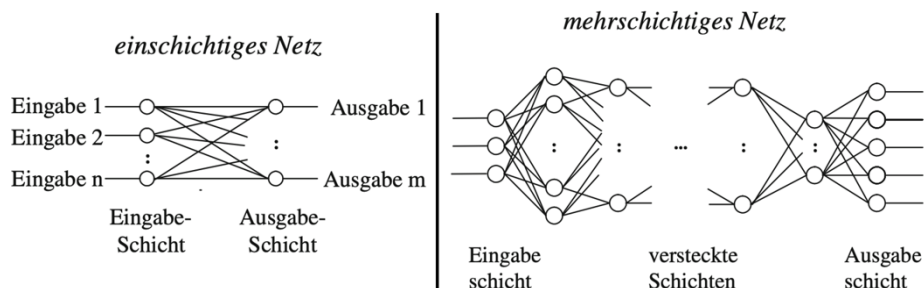


Abbildung 4: Vergleich neuronaler Netze (Ziegenbalg et al., 2016, S. 331f)

Die Backpropagation-Lernregel stellt eine Verallgemeinerung der Delta-Lernregel für nichtlineare Aktivierungsfunktionen dar. Sie ist die Grundlage für die modernen Anwendungen im Bereich des Deep Learning, da sie das Lernen deutlich schneller machen. (Fischer et al., 2024). Backpropagation basiert auf der Kettenregel der Differentiation. Dabei wird der Gradientenfehler von der Ausgabeschicht zur Eingabeschicht zurückverfolgt, um die Gewichtsadjustierungen zu berechnen und die Differenz zwischen Ist- und Soll-Ausgabe zu minimieren. Sobald die Fehler durch alle Schichten berechnet wurden, kann das Netzwerk die nötigen Änderungen an den Gewichten vornehmen, um seine Vorhersageleistung zu verbessern.

Die Erkenntnis, dass Backpropagation funktioniert, wurde in den 1970er und 1980er Jahren unabhängig von mehreren Forschungsgruppen entdeckt. Die darauf basierenden Methoden haben bereits große Fortschritte in der Sprach-, Objekt- und Bilderkennung, sowie in Bereichen wie Medikamentenforschung und Genomik erzielt (LeCun et al., 2015).

Doch wie funktioniert nun die Ausgabe bzw. die „Entscheidung“ für ein bestimmtes Ergebnis? Angenommen das neuronale Netz soll auf Basis optischer Verfahren ein Stoppschild erkennen. Die Merkmale eines Stoppschildes werden in einem Bild von den Neuronen analysiert und seine achteckige Form, die leuchtend rote Farbe, die markanten Buchstaben, die Größe des Verkehrsschildes und die Bewegung oder das Fehlen von Bewegung gelernt. Die Aufgabe des neuronalen Netzwerks besteht darin, zu entscheiden, ob es sich um ein Stoppschild handelt oder nicht. Die bereits beim Lernprozess erwähnten Gewichte spielen hier eine entscheidende Rolle. Das Netz berechnet eine Wahrscheinlichkeit, die eine fundierte Schätzung auf Grundlage der optimierten Gewichte darstellt. In diesem Beispiel könnte das System mit 83 % Wahrscheinlichkeit sicher sein, dass es sich um ein Stoppschild handelt, mit 9 %, dass es ein Geschwindigkeitsbegrenzungsschild ist und mit 8 %, dass es ein etwaiges anderes Objekt ist (Copeland, 2016).

2.8. Hardware für KI

Trotz der diskutierten Vorteile neuronaler Netze, wurden sie bis vor nicht allzu langer Zeit von der KI-Forschung weitgehend gemieden. Sie existieren seit den frühesten Tagen der KI, haben aber wenig „Intelligenz“ hervorgebracht. Das Problem war, dass selbst die einfachsten neuronalen Netzwerke äußerst rechenintensiv waren, was sie in den Augen der Forscher*innen zu unpraktikabel erscheinen ließ. Dennoch hielt eine kleine Forschungsgruppe unter der Leitung von Geoffrey Hinton an der Universität Toronto daran fest. Sie parallelisierten die Berechnungen und Algorithmen für Supercomputer und bewiesen so, dass das Konzept funktioniert (Copeland, 2016). Doch erst mit dem Einsatz rechenstarker Grafikprozessoren (GPUs), konnte das volle Potenzial ausgeschöpft werden. Ursprünglich für die Beschleunigung von 3D-Grafiken entwickelt, wurden GPUs bald für ihre Fähigkeit zur parallelen Datenverarbeitung geschätzt (Pandey et al., 2022). Die Leistung moderner GPUs ist im Vergleich zu CPUs, der eigentlichen zentralen Recheneinheit eines Computers, stark gewachsen. Beides sind Endprodukte der Halbleitertechnologie und unterliegen für gewöhnlich

dem Mooreschen Gesetz. Dieses besagt, dass sich die Rechenleistung von Mikroprozessoren etwa alle zwei Jahre verdoppelt. Der Grund dafür ist, dass immer mehr Transistoren auf gleichem Raum integriert werden können. Eine der weltweit größten Entwickler von GPUs, das amerikanische Unternehmen NVIDIA, gab bei einer Firmenpräsentation im Jahr 2018 bekannt, dass GPUs aufgrund ihrer rasanten Entwicklung einem eigenen Gesetz unterliegen. Konzernchef Huang erklärte, dass die heutigen GPUs von NVIDIA 25-mal so schnell sind wie vor fünf Jahren. Wenn ihre Entwicklung jedoch dem Mooreschen Gesetz gefolgt wäre, hätte sich die Geschwindigkeit nur auf das Zehnfache erhöht (Perry, 2018).

Ursprünglich nutzten GPUs sogenannte Shader-Kerne für spezifische Teilaufgaben in der Grafikdarstellung, wie etwa Farb- und Geometrieberechnungen. Heutige GPUs setzen auf universelle Shader, die flexibel programmierbar sind und so als kleine Universalprozessoren jede geforderte Funktion ausführen können. Dieser Wandel ermöglichte die Nutzung von GPUs nicht nur zur Grafikverarbeitung, sondern auch für allgemeine Berechnungen (Wittpahl, 2019). Auch softwareseitig unterstützte NVIDIA die Entwicklung von Deep Learning. Mit cuDNN, stellte der Konzern 2014 eine GPU-beschleunigte Bibliothek für tiefe neuronale Netzwerke vor, die speziell für das Training und das Testen von Deep-Learning-Algorithmen optimiert war (Pandey et al., 2022).

Obwohl einzelne GPU-Kerne im Vergleich zu CPU-Kernen weniger leistungstark sind und eine niedrigere Taktfrequenz haben, besitzen moderne GPUs oft Tausende solcher Kerne, sodass sie komplexe Aufgaben in großem Stil parallel abarbeiten können. Ein weiterer Vorteil ist die hohe Speicheranbindung von GPUs, die rund zehnmal so schnell wie jene von CPUs ist. Das ist besonders wertvoll, wenn große Datenmengen schnell verarbeitet werden müssen, wie bei wissenschaftlichen Simulationen, KI-Training und anderen datenintensiven Aufgaben (Wittpahl, 2019).

2.9. Natural Language Processing (NLP)

Die Computerlinguistik, auch als natürliche Sprachverarbeitung oder Natural Language Processing (NLP) bekannt, ist ein Teilbereich der Informatik, der sich mit Methoden der Analyse, dem Verständnis und der Erzeugung menschlicher Sprache durch Computer

beschäftigt. In den letzten 20 Jahren hat sie sich auch als praxisnahe Technologie etabliert, die zunehmend in Produkte für Endverbraucher integriert wird. Beispiele dafür sind Apples Siri oder Amazons Alexa (Hirschberg & Manning, 2015). Auch der Chatbot ChatGPT von OpenAI hat bei Benutzer*innen eine Welle der Begeisterung hervorgerufen. ChatGPT basiert auf einem großen neuronalen Sprachmodell, welches mithilfe einiger Milliarden Parametern trainiert wurde. Allgemein kann man sagen, dass die Parameterzahl die „Größe“ und in weiterer Folge auch die potenzielle Leistungsfähigkeit des Sprachmodells angibt. Es zeigt durch die Vorhersage des jeweils nächsten Wortes eindrucksvolle Leistungen in der Sprachverarbeitung. Aufgrund seiner Größe kann es komplexe Antworten erzeugen und wird daher auch als „large language model“ (LLM) bezeichnet (Fischer et al., 2024). Ein starker Anstieg der Rechenleistung, die Verfügbarkeit großer Mengen an Sprachdaten und die Weiterentwicklung maschineller Lernverfahren, haben die Entwicklung der natürlichen Sprachverarbeitung ermöglicht. In den ersten Jahrzehnten der Forschung versuchten Wissenschaftler*innen, Wortschätze und grammatikalische Regeln menschlicher Sprachen für Computer zu kodieren. Diese Aufgabe erwies sich jedoch als äußerst komplex, da natürliche Sprache durch Mehrdeutigkeit und kontextabhängige Bedeutungen geprägt ist. Ein Beispiel hierfür ist das Wort „Bank,“ welches sowohl für eine Sitzmöglichkeit als auch das Geldinstitut stehen kann (Hirschberg & Manning, 2015). Einige Methoden der Sprachverarbeitung von Computern werden im Folgenden kurz erläutert.

Die Bag-of-Words-Kodierung stellt die Häufigkeit jedes Wortes in einem Text als Vektor dar, dessen Dimension der Anzahl der Wörter im Wörterbuch entspricht. Diese Methode erfasst jedoch nicht die Wortreihenfolge, wodurch der genaue Kontext verloren geht. So wird eine Verneinung durch „nicht“ zwar erkannt, das verneinte Element ist jedoch unklar (Fischer et al., 2024). Diese Einschränkung verhindert eine präzise Textrekonstruktion, ermöglicht jedoch eine grobe thematische Klassifikation von Eingaben. Verbesserungen über diese einfachen Ansätze hinaus sind oft nur schwer zu erreichen (Hirschberg & Manning, 2015). Größeres Potenzial bieten die Algorithmen Continuous Bag-of-Words (C-BOW) und Skip-Gram, die auf sehr umfangreiche Texte trainiert werden. Das C-BOW-Modell sagt ein Zielwort basierend auf den umgebenden Kontextwörtern (z. B. zwei links und zwei rechts) voraus. Im Gegensatz dazu versucht das Skip-Gram-Modell, aus einem gegebenen Wort die benachbarten Wörter vorherzusagen. Beide Modelle erfassen so semantische Beziehungen zwischen Wörtern und verbessern die Qualität der generierten Texte (Fischer et al., 2024).

Während ChatGPT als Chatbot und Oberfläche fungiert, mit der Endverbraucher*innen interagieren, basiert dessen Funktionalität auf dem GPT-Modell. Es kann als komplexes Sprachmodell, welches Text verstehen und generieren kann, definiert werden (Helfrich-Schkarbanenko, 2023). Wir fassen ein Sprachmodell grob gesagt als eine Wahrscheinlichkeitsverteilung auf, die die Wahrscheinlichkeit für jede mögliche Zeichenkette angibt. Ein solches Modell sollte beispielsweise eine hohe Wahrscheinlichkeit für die Zeichenkette „Das Auto ist rot.“ haben, während „ist Auto Das rot?“ vom Modell als unwahrscheinlich deklariert werden sollte. Man kann also vorhersagen, welche Wörter als nächstes in einem Text erscheinen könnten, und somit Vorschläge für die Vervollständigung einer E-Mail oder Textnachricht machen. Alternativ kann ein Sprachmodell die wahrscheinlichste Übersetzung eines Satzes berechnen oder durch die Verwendung von Frage-Antwort-Paaren aus Trainingsdaten die wahrscheinlichste Antwort auf eine Frage ermitteln. Da natürliche Sprachen komplex sind, wird jedes Sprachmodell bestenfalls nur eine Annäherung darstellen (Russel et al., 2016).

Ein weiterer wichtiger Schritt im Bereich der natürlichen Sprachverarbeitung war die Entwicklung des Transformer-Modells. Es wurde 2017 in einem Artikel von acht Google-Mitarbeitern publiziert und ist für die Verarbeitung sequenzieller Daten, wie zum Beispiel Sprache, ausgelegt (Fischer et al., 2024). Seit der Veröffentlichung des Artikels mit dem Titel „Attention is All You Need“ hat sich das Transformer-Modell als Grundbaustein für Anwendungen wie ChatGPT, Bildgeneratoren wie DALL-E und Midjourney etabliert (Levy, 2024). Es basiert auf dem sogenannten „Attention-Mechanismus“, einem Verfahren, das es dem Modell ermöglicht, selektiv auf bestimmte Teile der Eingabesequenz zu fokussieren. Dadurch wird die Effizienz erheblich gesteigert (Fischer et al., 2024). Weiters werden globale Zusammenhänge innerhalb von Textmustern erfasst, wodurch inhaltlich relevante und kohärente Texte erzeugt werden können (Balasubramaniam et al., 2024).

Das Transformer-Modell wird im Folgenden grob und in Teilen erläutert, um lediglich ein Gefühl für die Funktionsweise zu bekommen. Die genaue Erläuterung würde aufgrund der Komplexität den Rahmen sprengen. Im Transformer-Modell wird die Eingabesequenz zunächst in Wortvektoren umgewandelt, ähnlich wie bei der bereits oben erwähnten Bag-of-Words-Kodierung. Bei der Vektorisierung geht allerdings die Information über die Position der einzelnen Wörter innerhalb der Sequenz verloren. Um dieses Problem zu beheben, wird im

nächsten Schritt ein „Positional Encoding“ eingeführt. Hierbei werden positionsabhängige Sinus- und Cosinus-Funktionen zu den Wortvektoren addiert. Dieses Verfahren ermöglicht es dem Modell, die Reihenfolge der Wörter in der Eingabesequenz zu berücksichtigen und so die Kontextinformationen zu bewahren (Fischer et al., 2024). Der nächste Schritt im Modell, ist bereits der Selbstaufmerksamkeits-Mechanismus. Die einfachste Methode, Selbstaufmerksamkeit anzuwenden, ist das Skalarprodukt der Wortvektoren mit sich selbst. Dies wird in der Regel ein zahlenmäßig großes Ergebnis liefern, was zu erhöhter Selbstaufmerksamkeit führt (Russel et al., 2016). Die Eingabewortvektoren, durchlaufen dreifach den Aufmerksamkeitsmechanismus. Dies ist notwendig, da der Mechanismus, ähnlich wie bei einer Internetsuche, drei verschiedene Argumente benötigt. Die „query“ (Abfrage), den „key“ (Schlüssel), mit dem die Abfrage abgeglichen wird, und den „value“ (Wert), der das Ergebnis der Suche darstellt. Abgekürzt werden diese mit Q, K und V. Diese drei Objekte sind Matrizen, da mehrere Wortvektoren zusammen eine Matrix bilden (Fischer et al., 2024). Selbstaufmerksamkeit ist nur eine der Komponenten des Transformer-Modells. Jede Transformer-Schicht besteht aus mehreren Unterschichten, wobei in jeder Schicht zunächst Selbstaufmerksamkeit angewendet wird. Es wurde bisher nur auf einen Teil des Modells, den sogenannten Encoder, eingegangen. Der Encoder ist besonders für die Textklassifikation der Eingabe geeignet, da aber auch Ausgaben erzeugt werden sollen, benötigt es einen Decoder. Der Encoder und der Decoder sind nahezu identisch, wobei der Decoder eine Variante der Selbstaufmerksamkeit verwendet, bei der jedes Wort nur auf die vorhergehenden Wörter zugreifen kann, da Text von links nach rechts generiert wird (Russel et al., 2016).

Die Performance des Transformer-Modells lässt den Eindruck entstehen, dass die Systeme Texte tatsächlich verstehen können. Da deren Funktionsweise allerdings rein auf statistischen und mathematischen Methoden basiert, kann dies jedoch verneint werden. Für die Praxis ist diese Frage ohnehin irrelevant und eher von philosophischer Natur (Ertel & Black, 2016). Wie man sehen kann, sind die möglichen Verwendungszwecke von natürlicher Sprachverarbeitung vielfältig: Sie können die Kommunikation zwischen Menschen unterstützen (z. B. durch maschinelle Übersetzung), die Interaktion zwischen Mensch und Maschine erleichtern oder Daten aus der großen Menge an sprachlichen Inhalten im Internet auswerten, um daraus Erkenntnisse zu gewinnen (Hirschberg & Manning, 2015). Im nächsten Kapitel soll diskutiert werden, mit welchen Texteingaben Sprachmodelle aufgrund der erläuterten

Funktionsweise gut hantieren können, um so die Ausgaben für die Benutzer*innen möglichst nutzbringend zu gestalten.

2.10. Prompt Engineering

2.10.1. Definition und Relevanz von Prompt Engineering

Um den Begriff des Prompt Engineering zu verstehen, muss zuerst geklärt werden, was man unter einem Prompt versteht. Ein Prompt bedeutet auf Deutsch übersetzt „Aufforderung“ oder „Eingabeaufforderung“. Bezogen auf die KI und die bereits behandelten Sprachmodelle kann der Begriff wie folgt beschrieben werden.

„Prompts sind Eingabeaufforderungen oder Anweisungen, die dem Modell gegeben werden, um eine bestimmte Art von Antwort oder Ausgabe zu generieren. In einfachen Worten, ein Prompt ist der Text, den Sie in das Modell eingeben, um eine Antwort zu erhalten. [...] Das Modell nimmt den Prompt entgegen und generiert darauf basierend eine Antwort, die auf seinem Training und den Daten, mit denen es trainiert wurde, basiert“ (Helfrich-Schkarbanenko, 2023, S. 7).

Prompts sind Eingaben oder Hinweise, die dazu dienen, eine bestimmte Antwort von einem System zu erzeugen. Der Ursprung dieser Idee liegt in frühen Systemen, die auf festen Regeln und Mustern beruhten. Mit dem Aufkommen des maschinellen Lernens übernahmen Datensätze diese Rolle. Sie fungierten als Prompts für Algorithmen, indem sie Eingabedaten und gewünschte Ausgaben zur Verfügung stellten. So lernten Modelle Muster zu erkennen und konnten schließlich Vorhersagen für neue Eingaben treffen. Die heutige Verwendung von Prompts, besonders bei Sprachmodellen wie ChatGPT, beruht auf der Fähigkeit, aus einem Textinput passende und kontextbezogene Antworten zu erzeugen (Khan, 2024). Nicht nur LLMs (large language models) wie ChatGPT arbeiten mit Textprompts, sondern auch Bildgenerierung-KIs wie das Tool „Midjourney“. Lediglich die Ausgaben sind unterschiedlich (Phoenix & Taylor, 2024). Prompts sind nicht nur Befehle, sondern kreative Anstöße, die die Richtung der Modellantwort lenken. Das Ziel ist die KI-Modelle sinnvoll zu steuern und mit ihnen zu interagieren, um nützliche Ausgaben zu erhalten (Khan, 2024). Die hohe

Leistungsfähigkeit von LLMs legt nahe, dass sie in vielen Fachbereichen als unterstützende Werkzeuge regelmäßig eingesetzt werden könnten (Frieder et al., 2024). Weiters verdeutlicht die Nutzung auch schnell, dass die Qualität der Ausgabe stark von der Eingabe abhängt. Hieraus entwickelte sich das sogenannte „Prompt Engineering“, eine Disziplin zur Verbesserung der Prompts und in weiterer Folge auch der bereitgestellten Ausgaben (Phoenix & Taylor, 2024).

„Prompt-Engineering bezieht sich auf die Wissenschaft, effektive Eingabeaufforderungen (Prompts) für Sprachmodelle zu entwerfen, um gewünschte und genaue Antworten oder Ausgaben zu erhalten. Im Kontext von Sprachmodellen, insbesondere solchen wie GPT [...], spielt die Formulierung des Prompts eine entscheidende Rolle für die Qualität und Relevanz der Antwort des Modells“ (Helfrich-Schkarbanenko, 2023, S. 10).

Eine bedeutende Entwicklung im Bereich des Prompt Engineerings ist die Verbesserung der Prompts anhand menschlichen Feedbacks. Forschende identifizieren die Probleme der Prompts und optimieren die Prompt-Formulierung schrittweise. Die Prompt-Nutzung wird damit benutzerfreundlicher und kontextbezogener, was sowohl die Zufriedenheit der Nutzer als auch die Leistung erhöht (Balasubramaniam et al., 2024).

Robin Li, der CEO des chinesischen Technologiekonzerns Baidu, prognostizierte, dass in zehn Jahren die Hälfte aller Jobs weltweit im Bereich des Prompt Engineerings angesiedelt sein werden (Duranton, 2023). Es wird jedoch erwartet, dass Prompting sich zu einer gefragten Fähigkeit entwickeln wird, anstatt eine eigenständige Berufsbezeichnung zu werden. Vergleichen kann man dies mit der allgemeinen Kompetenz einen PC oder bestimmte Programme, wie Excel oder Photoshop, bedienen zu können. Dieser Wandel stellt traditionelle Konzepte der Informatik infrage, da KI-Modelle nicht-deterministische Ergebnisse liefern, sprich mehrere mögliche Ausgaben zu ein und derselben Eingabe liefern können (Phoenix & Taylor, 2024).

2.10.2. Eigenschaften guter Prompts

Gut gestaltete Prompts ermöglichen es, konstant hilfreiche und nutzbringende Ausgaben zu erzeugen. Dennoch können zwei gut formulierte Prompts für dieselbe Aufgabe unterschiedliche Ergebnisse liefern. Aus einer informellen Perspektive erinnern diese an gutes Coaching im

Sport: Beide führen für gewöhnlich zum gewünschten Ergebnis (Campesato, 2023). In den folgenden Abschnitten sollen die Eigenschaften guter Prompts aus gegebener Literatur erläutert werden.

Kontext:

Khan (2024) spricht davon, dass der Kontext das Fundament eines jeden Prompts darstellt und von äußerster Wichtigkeit ist. Ein klar definierter Kontext kann die Wahrscheinlichkeit einer Fehlinterpretation des Prompts durch die KI erheblich verringern. Wenn ein Nutzer nach „Java“ fragt, könnte sich die Antwort auf die Programmiersprache oder auf die Insel in Indonesien beziehen. Wird jedoch der Kontext „Java-Programmierung“ angegeben, konzentriert sich die KI auf das relevante Thema. Im Wesentlichen ist der Kontext wie ein Kompass für KI-Systeme, der ihnen die Richtung weist. Auch laut Helfrich-Schkarbanenko (2023) liefert ein gut strukturierter Prompt dem Modell den notwendigen Kontext, um eine Frage zu verstehen oder eine Aufgabe korrekt auszuführen. Dies ist besonders nützlich, wenn das Modell Informationen in einem speziellen Format oder aus einer bestimmten Perspektive bereitstellen soll. Ein Vorschlag von Phoenix und Taylor (2024) ist, dass Richtung und Kontext auch in Form von Personen oder Rollen vorgegeben werden können, das Sprachmodell soll beispielsweise in die Rolle eines Wissenschaftlers oder eines Lehrers schlüpfen.

Klarheit:

Klarheit ist ein weiterer entscheidender Faktor bei der Erstellung von KI-Prompts. Sie stellt sicher, dass das KI-Modell die Absicht des Nutzers ohne Unklarheiten versteht, was zu genaueren und nützlicheren Ergebnissen führt. Anstatt lange Rückfragen an die KI zu stellen, kann ein klar formulierter Prompt oft zu einer gewünschten Antwort in einer einzigen Interaktion führen (Khan, 2024). Präzise formulierte Prompts steuern das Verhalten des Sprachmodells. So können spezifische Informationen gezielt generiert werden und Antworten in einem gewünschten Stil und Ton erstellt werden (Helfrich-Schkarbanenko, 2023). Alternativ können der KI auch Regeln vorgegeben werden, nach denen eine Analyse oder Generierung einer Antwort geschehen soll. Die Struktur der Ausgabe kann ebenfalls vorgegeben werden (Phoenix & Taylor, 2024).

Prägnanz:

Prägnanz bezeichnet die Eigenschaft einen Prompt mit möglichst wenigen Worten zu formulieren, ohne dabei an Verständlichkeit zu verlieren. KI-Modelle arbeiten am besten mit klaren und prägnanten Anweisungen. Ein kurzer Prompt führt oft zu einer schnelleren und genaueren Antwort, da das Modell keine überflüssigen Informationen verarbeiten muss (Khan, 2024). Auch Shah (2023) hebt die Notwendigkeit von prägnanten Prompts hervor, mehrdeutige Formulierungen und nicht notwendige Informationen sollten nicht verwendet werden. Besonders in geschäftlichen oder professionellen Kontexten, in denen Zeit knapp ist, hilft ein präziser Prompt dabei, die Produktivität zu steigern. Kurze Prompts benötigen zudem weniger Rechenleistung, was das System effizienter und schneller macht. Jedes zusätzliche Wort kann zu Missverständnissen führen. Durch prägnante Prompts verringert sich das Risiko von Fehlinterpretationen (Khan, 2024).

Angabe von Beispielen:

Eine weitere Methode zur möglichen Verbesserung von Prompts ist die Angabe von Beispielen. Möchte man neuen Mitarbeiter*innen oder Schüler*innen eine neue Aufgabe erklären, verwendet man oft Beispiele, die eine frühere erfolgreiche Umsetzung zeigen. Bei der Arbeit mit KI ist dies ähnlich. Die Qualität eines Prompts hängt häufig von den verwendeten Beispielen ab. Beispiele zu liefern ist oftmals einfacher, als die konkreten Eigenschaften der gewünschten Ausgabe zu beschreiben. Diese Methode ist besonders effektiv, wenn man kein Experte im jeweiligen Fachgebiet ist, in dem die Aufgabe durchgeführt werden soll (Phoenix & Taylor, 2024). Man sollte der KI weiters auch Hintergrundinformationen oder Einschränkungen geben, um präzisere und relevantere Ergebnisse zu erzielen. Dabei kann man sich vorstellen, wie viele Details man Schüler*innen geben müsste, um eine Aufgabe korrekt zu erfüllen (Shah, 2023).

2.10.3. Arten von Prompts

Ein Prompt ohne Angabe eines Beispiels wird als Zero-Shot-Prompt bezeichnet. Das Modell hat damit keinen Anhaltspunkt für eine etwaige Lösung. Zum Beispiel könnte man es bitten, einen englischen Text wie „Hello, I am Harry.“ ins Französische zu übersetzen, ohne ihm eine andere korrekte Beispielübersetzung im Prompt selbst vorzugeben. Ein Few-Shot-Prompt hingegen enthält einige Beispiele, die das Modell leiten. Wenn man zwei oder drei Übersetzungen, in Form von englischen und entsprechenden französischen Sätzen, angibt und

dann einen neuen Satz hinzufügt, nutzt das Modell diese Beispiele, um den neuen Satz zu übersetzen. Das gewünschte Ausgabeformat wird anhand der Beispiele abgeleitet (Khan, 2024).

Schon ein einziges Beispiel (One-Shot) hilft dem Modell oft weiter. Ohnehin ist es bei der Leistungsevaluierung üblich die Modelle mit mehreren Beispielen zu testen (Few-Shot). Die Studie „Language Models are Few-Shot Learners“ von Brown et al. (2020), deren Ergebnisse in Abbildung 5 gezeigt werden, hat festgestellt, dass das Hinzufügen eines einzigen Beispiels die Genauigkeit in Aufgaben von 10 % auf nahezu 50 % steigern kann (Phoenix & Taylor, 2024). Das Few-Shot-Lernen verbessert sich ebenfalls deutlich mit zunehmender Sprachmodellgröße, die mit der Anzahl von Parametern angegeben wird. Auch wenn die Ergebnisse in diesem Fall besonders eindrucksvoll sind, zeigt sich in gegebener Literatur ein allgemeiner Trend bei den meisten untersuchten Aufgaben (Brown et al., 2020). Ausreichend Beispiele anzugeben, bringt jedoch nicht nur Vorteile mit sich. Mit drei bis fünf Beispielen steigt die Verlässlichkeit der Ergebnisse, jedoch auf Kosten der Kreativität. Je mehr Beispiele gegeben werden und je ähnlicher diese sind, desto stärker wird die Antwort in die Richtung der Beispiele gelenkt. Wenn das Prompten mit einigen Beispielen nicht die gewünschten Resultate liefert, kann es hilfreich sein, die Formulierungen zu variieren oder alternative Phrasen auszuprobieren. Verschiedene Formulierungen können zu unterschiedlichen Antworten führen. Zudem kann es sinnvoll sein, das Modell ausdrücklich anzuweisen, die Beispiele nur als „Formatvorlage“ zu verwenden und nach Aufgabenverständnis zu ignorieren, um eine Überanpassung an die Beispiele zu vermeiden (Phoenix & Taylor, 2024).

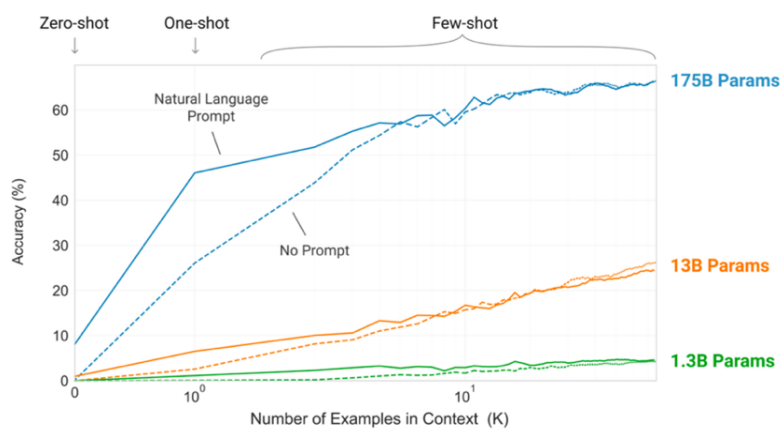


Abbildung 5: Ausgabe­genauigkeit von Zero, One- und Few-Shot-Prompts (Brown et al., 2020, S. 4)

Beispiele für einen Zero-Shot- und einen Few-Shot-Prompt, übernommen von Schorcht et al. (2023, S. 14) sind im Folgenden angeführt:

Zero-Shot-Prompt:

„Frage: Ein Jongleur kann mit 16 Bällen jonglieren. Die Hälfte der Bälle sind Golfbälle, und die Hälfte der Golfbälle sind blau. Wie viele blaue Golfbälle sind es?“

Few-Shot-Prompt:

„Frage: Roger hat 5 Tennisbälle. Er kauft 2 weitere Dosen Tennisbälle. Jede Dose enthält 3 Tennisbälle. Wie viele Tennisbälle hat er jetzt? Antwort: 11.

Frage: Ein Jongleur kann mit 16 Bällen jonglieren. Die Hälfte der Bälle sind Golfbälle, und die Hälfte der Golfbälle sind blau. Wie viele blaue Golfbälle sind es?“

Eine weitere Art ist der sogenannte Chain-of-Thought-Prompt (CoT), auch als Gedankenkette zu bezeichnen. Dieser kann ebenfalls die Problemlösungsfähigkeiten von Sprachmodellen steigern. Dabei wird das Modell angeleitet, eine Reihe von Zwischenschritten zu erzeugen, die schrittweise zur Lösung eines mehrstufigen Problems führen (Helfrich-Schkarbanenko, 2023). So lässt sich sicherstellen, dass die Ausgabe in jedem Schritt korrekt ist, bevor man zum nächsten Schritt übergeht. Diese Methode ermöglicht es zudem, längere Ausgaben zu erstellen, indem man kleine Ausgaben nacheinander generiert (Shah, 2023).

Natürlich kann die Methode des CoT-Promptings auch auf Zero-Shot-Prompts oder Few-Shot-Prompts angewendet werden. Bei Letzterem wird mindestens ein Beispiel der zu lösenden Frage und der CoT-Schrittfolge vorangestellt. Alternativ kann man mit Zero-Shot-Prompting eine ähnliche Argumentationsstruktur und Leistungssteigerung erreichen, indem man den Prompt einfach mit der Phrase „Lass uns Schritt für Schritt vorgehen.“ ergänzt (Helfrich-Schkarbanenko, 2023). Bei arithmetischen Textaufgaben in der Mathematik konnten unter der Verwendung von CoT-Prompting bessere und häufiger korrekte Lösungen festgestellt werden (Wei et al., 2023).

Eine weitere Methode stellt der Ask-me-Anything-Prompt (AmA) dar. Dabei wird das KI-Modell dazu gebracht, Fragen zu stellen. Arora et al. (2023) gaben abwechselnd Behauptungen und Fragen ein, um das System auf das Stellen von Fragen zu trainieren, sodass das Modell am Ende anstelle einer Antwort selbst eine Frage zur Behauptung formuliert. Für mathematische Problemlösungen wird in dieser Methode ein Ansatz gesehen, das Sprachmodell aktiv in den Lösungsprozess einzubinden. Der Prompt wird dabei durch folgende Aufforderung ergänzt: „Stelle notwendige Fragen, die du zur Lösung des Problems benötigst.“ Dadurch ändert sich die Interaktion, da nun das KI-Modell aktiv Fragen an die Nutzer*innen stellt, anstatt umgekehrt. Dies führt zu einem dynamischen Frage-Antwort-Prozess, der als Abfrageschleife bezeichnet werden kann. Die Schleife wird wiederholt ausgeführt, bis ein zufriedenstellendes Ergebnis erreicht wird (Schorcht et al., 2023). Die Methode kann wiederum in Kombination mit den anderen bereits vorgestellten Prompting-Techniken verwendet werden. Ein Zero-Shot-Prompt in Kombination mit CoT und AmA, übernommen von Schorcht et al (2023, S. 19), könnte also wie folgt aussehen:

„Max ist 78 cm groß. Wenn er auf eine 20 cm hohe Kiste steigt, ist er genauso groß wie Klaus. Paul ist 15 cm kleiner als Klaus. Wie groß ist Paul? Lass uns Schritt für Schritt vorgehen, um sicherzustellen, dass wir die richtige Antwort haben. Stelle notwendige Fragen, die du zur Beantwortung der Frage benötigst.“

2.11. ChatGPT

ChatGPT ist eine Chatbot-Anwendung, die auf dem GPT-Sprachmodell basiert. Während das GPT-Modell selbst die zentrale Technologie für die Textgenerierung bildet, weist ChatGPT als online abrufbare Website, eine benutzerfreundliche Oberfläche auf, über die Benutzer*innen das Modell ansteuern können (Helfrich-Schkarbanenko, 2023). GPT steht für „Generative Pre-Trained Transformer“, der Wortteil „Bot“ leitet sich von „Roboter“ ab und bezeichnet in diesem Kontext ein Computerprogramm, das bestimmte Aufgaben autonom und in kommunikativer Form ausführt (Hischer, 2023). Der Begriff „Generativ“ beschreibt die Fähigkeit des Modells, neue Texte zu erstellen, die menschlichem Schreiben ähneln. „Pre-trained“ bedeutet, dass das Modell auf einer großen Datenmenge trainiert wurde. Durch das Vortraining auf umfangreichen Texten lernt das Modell allgemeine Sprachmuster, die es ihm ermöglichen, später mit weniger

spezifischen Daten zu arbeiten. „Transformer“ bezieht sich auf das Transformer-Modell, das bereits in Kapitel 2.9. diskutiert wurde. Dieses bildet die Grundlage moderner Sprachmodelle wie GPT (Helfrich-Schkarbanenko, 2023). Das amerikanische Unternehmen OpenAI entwickelte die Sprachmodelle der GPT-Reihe, die nacheinander veröffentlicht wurden: GPT im Jahr 2018, GPT-2 im Jahr 2019 und GPT-3 im Jahr 2020, mit jeweils 117 Millionen, 1,5 Milliarden und 175 Milliarden Parametern. Im März 2023 folgte GPT-4, dessen Parameteranzahl öffentlich nicht bekannt ist, jedoch auf etwa 1,76 Billionen geschätzt wird (Campesato, 2023).

Die aktuelle und fortschrittlichste Version ist GPT-4o. Dieses neue Modell ermöglicht es ChatGPT 50 verschiedene Sprachen mit gesteigerter Geschwindigkeit und Qualität zu verarbeiten. GPT-4o besitzt Wissen bis Oktober 2023, kann aber bei Bedarf auf das Internet zugreifen, um aktuelle Informationen abzurufen. Das „o“ im Namen steht für „omni“ und meint damit die universelle Eingabe von Text, Bildern und Sprache (Field, 2024). GPT-4o ist kostenlos zugänglich, wobei die Anzahl der Nachrichten, die innerhalb eines fünfstündigen Zeitfensters gesendet werden können, je nach aktueller Auslastung, beschränkt ist. Wurde die Grenze ausgereizt, wechselt das System automatisch auf das kleinere Modell GPT-4o mini. Nutzer*innen der kostenpflichtigen Variante ChatGPT-Plus haben allerdings eine fünfmal so hohe Anfragengrenze, bevor auf dieses Modell gewechselt wird (OpenAI, o.D.).

3. KI und Mathematikunterricht

3.1. Relevanz von KI im schulischen Unterricht

Wie bereits in den vorangegangenen Kapiteln erläutert, ist der potenzielle Nutzen der Technologie für den Menschen und die Gesellschaft enorm, das gilt auch für den Bildungsbereich. Daher soll die KI zunächst im Anwendungsfeld von Schule und Unterricht beleuchtet werden. Wie bei jeder neuen Entwicklungen gibt es aber auch mögliche Gefahren, die hier nicht unerwähnt bleiben sollen. Die Integration digitaler Medien in Bildungseinrichtungen wird immer wichtiger, da KI-Tools Prozesse in der Wissensproduktion und beim Lösen von Aufgabenstellungen verändern. Schulen sollten daher ihre Lehrpläne erweitern, die über

digitale Alltagskompetenzen hinaus technische und informatische Inhalte vermitteln. Dies bedeutet, dass technische Hintergründe und Funktionsweisen von KI vermittelt werden sollen, um ein tieferes Verständnis der Auswirkungen und Möglichkeiten solcher Technologien zu fördern. Pädagogisch reicht die Vermittlung von digitalen Fähigkeiten jedoch allein nicht aus. Der Fokus muss auf dem kritischen und reflektierten Umgang mit KI liegen (Knaus et al., 2023).

Wie kann KI bzw. ChatGPT nun konkret den Unterricht und das Lernen verbessern? Die Ergebnisse von Lo (2023) zeigen, dass ChatGPT sowohl Lehrkräften als auch Schüler*innen als Assistent dienen kann. Für Lehrkräfte werden die Einsatzmöglichkeiten von ChatGPT in zwei Bereiche unterteilt: Unterrichtsvorbereitung und -bewertung. Im Bereich der Unterrichtsvorbereitung kann ChatGPT nützliche Vorschläge machen, die Lehrkräfte unterstützen. Beispielsweise kann es mögliche Gliederungen von Themen und Vorschläge für Stundenplanungen liefern (Tlili et al., 2023). Was den Bereich der Bewertung betrifft, kann ChatGPT Lehrer*innen dabei helfen, Übungen, Quiz und Fragen zu erstellen. Laut Al-Worafi (2023) sollte man jedoch kontrollieren, ob die von ChatGPT vorgeschlagenen Bewertungsaufgaben auch tatsächlich alle intendierten Lernziele abdecken. Die Aufgaben wurden in diesem Fall allerdings auf universitärem Niveau erstellt. Empfohlen wird dennoch, ChatGPT eher als Hilfestellung bei der Vorbereitung von Prüfungen zu nutzen, statt Ausgaben vollständig zu übernehmen.

Für Schüler*innen kann die KI als virtueller Tutor dienen, der ihr Lernen unterstützt. Die möglichen Funktionen von ChatGPT können hier in zwei Bereiche unterteilt werden. Erstens kann der Lernprozess selbst unterstützt werden, zum Beispiel mit der Beantwortung von Fragen oder dem Zusammenfassen von Informationen. Zweitens kann die Bewertungsfunktion angeführt werden. Die KI unterstützt hierbei beispielsweise bei der Überprüfung von Texten oder Konzepten und der Prüfungsvorbereitung (Lo, 2023). Rudolph et al. (2023) schlagen vor, dass ChatGPT verschiedene Szenarien für die Lernenden erstellen zu lassen, um Gruppenarbeiten und dessen Dynamik zu fördern. Die KI kann dabei eine Struktur für Diskussionen bieten, in Echtzeit Feedback geben und individuelle Unterstützung leisten. Der intensivere Austausch fördert die Problemlösefähigkeiten der Gruppe (Lo, 2023). Eine Gefahr ist jedoch, dass ChatGPT das kritische Denken und die eigene Arbeit ersetzt, vielmehr sollte es eine Unterstützung bei genannten Arbeitsprozessen darstellen (Rudolph et al., 2023).

Die KI hat trotz seines Erfolgs neue Herausforderungen für die Bildung geschaffen. Insbesondere durch die Möglichkeit, Aufgaben und Prüfungen für Schüler*innen zu erledigen, wird Schummeln mithilfe von KI hiermit zum Thema (Lo, 2023). Lehrkräfte könnten den Unterricht so gestalten, dass (bewertete) Lern- und Verstehensleistungen direkt im Unterricht erbracht werden müssen, während zuhause mit Unterstützung von KI-Systemen beispielsweise Texte vorbereitet oder Aufgaben gelöst werden. Dabei sollten fachdidaktische Überlegungen weiterhin die Grundlage für etwaige Prüfungsformate bilden, unabhängig davon welche potenziellen Hilfsmittel eingesetzt werden können. Analog zur Nutzung von Taschenrechnern im Mathematikunterricht könnte es Prüfungsteile geben, die ohne Hilfsmittel auskommen, sowie solche, in denen KI-Tools erlaubt sind (SWK, 2024). Lehrkräfte befürchten auch, dass Schüler*innen das Schreiben von Texten an die Sprachmodelle auslagern, da sie recht schnell akzeptable bis gute Inhalte generieren können (Mhlanga, 2023). Die verantwortungsvolle Nutzung von ChatGPT im Lernprozess ist daher dringend erforderlich, um seine Vorteile zu optimieren und Risiken zu minimieren (Lo, 2023).

Seit der Einführung des Internets hat die Menschheit eine stetige Beschleunigung digitaler Innovationen erlebt. In einer Zukunft, die zunehmend von generativer KI geprägt sein wird, stehen Lehrkräfte vor der Herausforderung, Bildungsprozesse neu zu gestalten. Dies erfordert von den betroffenen Personen insbesondere Anpassungsfähigkeit, Kreativität und eine schnelle Reaktion auf technologische Entwicklungen. Diese Dynamik wird Lehrer*innen fordern, aber auch deren eigene Weiterentwicklung vorantreiben (Weßels, 2023). Zusammenfassend kann gesagt werden, dass die gesellschaftliche Nutzung von KI und digitalen Medien einen erweiterten Bildungsauftrag mit sich bringt, da digitale Werkzeuge mittlerweile essenzielle Bestandteile des Alltags- und Berufslebens sind. Diese Veränderung eröffnet Potenziale für neue Lernkonzepte. KI bietet in diesem Kontext eine Möglichkeit, sowohl kognitive Prozesse zu fördern als auch das Lernen durch individualisierte Rückmeldungen zu unterstützen (Knaus et al., 2023).

3.2. KI in der österreichischen Bildungslandschaft

Das Thema KI ist auch in der österreichischen Bildungspolitik angekommen. Das Bundesministerium für Bildung, Wissenschaft und Forschung schreibt in einem Artikel von baldigen KI-bedingten Veränderungen des Alltags und der Arbeitswelt, die die Gesellschaft vor

Herausforderungen stellen werden. Im Bereich Schule wird die technologische Entwicklung aber durchaus als Chance gesehen und das Thema muss behandelt werden, um Schüler*innen auf ein verantwortungsbewusstes und eigenständiges Leben in einer Welt mit künstlicher Intelligenz vorzubereiten (BMBWF, o.D.). Auch den politischen Entscheidungsträgern ist die Relevanz des Themas bewusst. In einem Entschließungsantrag der Nationalratsabgeordneten Eva-Maria Himmelbauer und Sibylle Hamann vom 5. Juli 2023 wird der Bildungsminister Martin Polaschek um eine breite öffentliche Diskussionen mit Expert*innen aus Theorie und Praxis gebeten, um über die Chancen und Risiken von KI in der Schule zu sprechen. Außerdem soll auf Grundlage der bisherigen Maßnahmen eine KI-Strategie für den Bildungsbereich entwickelt werden. Ziel ist es, die Möglichkeiten von KI, wie zum Beispiel personalisiertes Lernen, zu nutzen und gleichzeitig die Risiken zu minimieren (Parlament der Republik Österreich, 2023).

Doch wie sieht es nun tatsächlich in den Klassenzimmern aus? Bei einer im März 2024 durchgeführten Studie vom Österreichischen Bundesverlag (ÖBV, 2024a) wurden 381 Lehrkräfte aus ganz Österreich zu ihrem KI-Einsatz im Unterricht und ihrer Einstellung gegenüber der Technologie befragt. Rund 30 % der Befragten sind an einer allgemeinbildenden höheren Schule tätig, ca. 27 % an einer Mittelschule und ca. 21 % an einer Volksschule. Die meistunterrichteten Fächer der befragten Lehrkräfte sind Deutsch, Naturwissenschaften, Fremdsprachen und Mathematik. 44 % der Lehrer*innen haben künstliche Intelligenz im Schulalltag schon verwendet. Auf die Frage, ab welchem Alter die Künstliche Intelligenz eine Rolle in der Bildung spielen sollte, antworteten rund 36 % mit „Sekundarstufe I“ und rund 56 % mit „Sekundarstufe II“. Die Tendenz bezüglich einer Verwendung in späterem Alter ist hiermit klar ersichtlich. Außerdem zeigt die Befragung, dass Lehrkräfte KI vor allem zur Unterstützung bei der Unterrichtsvorbereitung einsetzen (57 %). Viele nutzen sie auch zur Erstellung von Arbeitsblättern (54 %) oder von Materialien für unterschiedliche Lernniveaus (34 %). Eine Vielzahl an Lehrkräften beobachtet, dass sich Schüler*innen durch die KI verleitet fühlen, sinnvolle Arbeit abzugeben und nicht selbst durchzuführen. Rund 51 % bzw. 32 % stimmen dieser Aussage zu bzw. eher zu. Der Bedarf an Schulungen im Bereich ist hoch: 73 % der Lehrkräfte wünschen sich mehr Fortbildungen zum Umgang mit der KI. Außerdem sind 71 % der Lehrkräfte der Meinung, dass Schüler*innen den Umgang mit KI in der Schule lernen sollten. Unter die wichtigsten Kompetenzen im Umgang mit KI fallen laut den Befragten „kritische Denkfähigkeit“, „Ethik & soziale Verantwortung“ und „Grundverständnis für

Rechtliches & Datenschutz“. Die Meinungen in Bezug auf Bildungsgerechtigkeit sind zwiegespalten. Viele sehen die Technologie in diesem Zusammenhang eher als Gefahr (57 %), der Rest eher als Chance (43 %). Laut Experten wird sich die Bildungsschere mit den Jahren allerdings eher schließen. Der Zugang zu digitalen Hilfsmitteln und KI ist für Schüler*innen gegeben und sie können sich damit selbstständig Unterstützung holen, ganz unabhängig von der finanziellen Lage der Eltern (ÖBV, 2024b). Die Ö3-Jugendstudie, durchgeführt vom Sozialforschungsinstitut Foresight in Kooperation mit dem ORF, liefert unter anderem Einblicke, wie 16- bis 25-Jährige über das Thema KI denken. 49 % der über 30.000 befragten Jugendlichen gab an, KI als großartig zu erachten und die diversen Nutzungsmöglichkeiten ausschöpfen zu wollen. Auf der anderen Seite äußern 51 % Bedenken und geben an, dass ihnen KI Angst mache und sie befürchten, irgendwann ersetzt zu werden. Auch bei der Nutzung von ChatGPT oder anderen KI-Chatbots in Schule und Studium gehen die Meinungen auseinander. 39 % der Jugendlichen sagen, dass sie diese verwenden, da sie ihnen Arbeit abnehmen, 24 % nutzen KI-Chatbots zwar noch nicht, sind aber offen dafür und sehen Potenzial in der Anwendung. Eine Minderheit (11 %) steht der Technologie skeptisch gegenüber und verzichtet aus Gründen der Fehleranfälligkeit darauf. 26 % lehnen die Nutzung ab, da sie der Meinung sind, dass die Ausgaben keine eigene Leistung mehr darstellen. Die Ergebnisse verdeutlichen, wie unterschiedlich die Einstellungen der Jugendlichen gegenüber KI sind. Während viele die Vorteile erkennen, gibt es ebenso starke Vorbehalte und Bedenken (ORF, 2024).

Auch die Politik sieht bereits konkrete Maßnahmen vor. Der „8-Punkte-Plan“ zur Digitalisierung in den österreichischen Schulen soll um eine KI-Strategie ergänzt werden. Das Thema KI und Digitalisierung soll außerdem ein fester Bestandteil in der Aus-, Fort- und Weiterbildung von Lehrerinnen werden. Zusätzlich ist geplant, mehr Unterrichtsmaterial zu diesen Themen zu erstellen und online bereitzustellen. Prüfungen, Leistungsbeurteilungen und abschließende Arbeiten müssen weiterentwickelt werden, mit dem Ziel die Eigenleistungen der Schüler*innen besser zu fördern und sichtbar zu machen (Parlament der Republik Österreich, 2023).

3.3. Anwendungsmöglichkeiten und Kompetenzen von KI im Unterrichtsfach Mathematik

3.3.1. Lösungskompetenz

In einer Studie von Plevris et al. (2023) wurden die Leistungen von ChatGPT-3.5, ChatGPT-4 und Google Bard (Vorgänger von Google Gemini) beim Lösen von Mathematik- und Logikaufgaben verglichen. Das Ziel war das Verständnis und die Problemlösungsfähigkeiten der Chatbots bei unterschiedlichen Aufgaben zu bewerten. Es gab zwei Sets an Fragen, wobei ersteres (Set A) aus neu erdachten Fragen bestand, während zweiteres (Set B) Fragen enthielt, die bereits online verfügbar und bereits publiziert waren. Anzumerken ist, dass alle drei KIs jede Frage jeweils drei Mal beantworten mussten. Bei drei identen Eingaben kamen jedoch oft drei unterschiedliche Ausgaben zustande, die zum Teil korrekt und inkorrekt waren. Im Folgenden ist eine Beispielfrage aus Set A angeführt:

“The sum of three adults’ ages is 60. The oldest of them is 6 years older than the youngest. What is the age of each one of them? Assume that an adult is at least 18 years old” (Plevris et al., 2023, S. 957).

Bei dieser Frage gab es beispielsweise eine richtige Ausgabe von GPT-3.5, drei richtige Ausgaben von GPT-4 und keine richtige von Bard. Ein eigens durchgeführter Test mit GPT-4o und der exakt selben Eingabe, führte dreimal zum korrekten Ergebnis. Die Antworten bzw. Lösungswege unterschieden sich jedoch deutlich. Die Ergebnisse der restlichen Aufgaben zeigen, dass alle drei Chatbots in gewissem Maße Mathematik- und Logikprobleme verstehen und lösen können, jedoch mit Ausnahmen und Einschränkungen. Eine Studie von Dao und Le (2023) analysierte ChatGPTs Fähigkeiten die Mathematikaufgaben der vietnamesischen Schulabschlussprüfung (VNHSGE) zu beantworten. Es handelt sich hierbei um Multiple-Choice-Fragen, unterschiedliche Themen und Schwierigkeitsstufen wurden bei der Untersuchung berücksichtigt. Die Fragen wurden auf die vier Kategorien Wissen, Verständnis, Anwendung und höhere Anwendung aufgeteilt, sowie in zehn mathematische Themenbereiche. Die Ergebnisse zeigen, dass ChatGPTs Leistung je nach Schwierigkeitsgrad und Thema variiert. Am besten schnitt es in der Kategorie Wissen mit einer Genauigkeit von 83 % ab. In den Bereichen Verständnis, Anwendung und höhere Anwendung sank mit 62 %, 27 % und 10 % die Quote kontinuierlich.

Die Studie von Guler et al. (2024) zeigt vielversprechende Ergebnisse. ChatGPT konnte die meisten Aufgaben nationaler Mathematiktests der Türkei lösen, jedoch oft erst nach mehreren Versuchen und mit klaren Anweisungen. Wardat et al. (2023) konnten bei ihren Tests mit ChatGPT zu dem Schluss, dass das Modell kein tiefes Verständnis für Geometrie hat und Fehlvorstellungen oft nicht richtig korrigieren kann. Teils wurden bei einfachen Berechnungen von Seitenlängen mit dem Satz des Pythagoras oder den Winkelfunktion, Zahlen nicht korrekt quadriert oder Gleichungen falsch umgeformt. Bei Dao und Le (2023) war ChatGPT bei Themen wie Exponential- und Logarithmusfunktionen, sowie geometrischen und arithmetischen Reihen besonders erfolgreich. Schwierigkeiten hatte es hingegen bei Ableitungen und ihren Anwendungen und räumlicher Geometrie. Wichtig zu erwähnen ist hier, dass ChatGPT Probleme mit grafischer Interpretation hat und Bilder nicht lesen oder verstehen kann. Auch Guler et al. (2024) beobachteten schwache Leistungen von ChatGPT in den Bereichen Geometrie, Wahrscheinlichkeitsrechnung und Ungleichungen. In der Testung von Plevris et al. (2023) konnte bei einer einfachen Addition zweier neunstelliger Zahlen, nur die neueste Version GPT-4 dreimal die richtige Antwort liefern, Google Bard hingegen kein einziges Mal. Der Eigentest mit GPT-4o führte lediglich zu zwei korrekten Ausgaben. Bard brillierte stattdessen beim Lösen der bereits veröffentlichten Fragen aus Set B, was wahrscheinlich mit dessen Internetzugriff zu begründen ist. ChatGPT-4 übertraf GPT-3.5

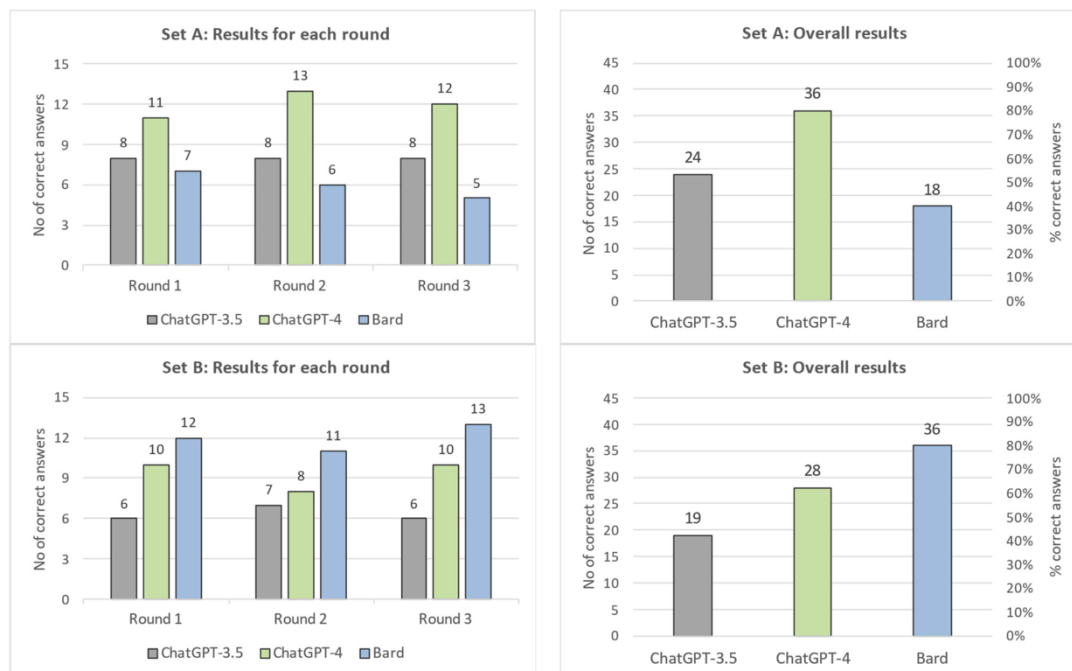


Abbildung 6: Vergleich der KI-Anwendungen (Plevris et al., 2023, S. 965)

deutlich in Genauigkeit und im Umgang mit komplexen Problemen (Plevris et al., 2023). Die genauen Ergebnisse sind in Abbildung 6 ersichtlich.

Eine Studie von Spreitzer et al. (2024) testete die mathematischen Modellierungsfähigkeiten von ChatGPT. Es stellte sich heraus, dass die getesteten GPT-Modelle zwar bei einfachen Aufgaben präzise Ergebnisse liefern können, jedoch bei mittel bis hoch komplexen Modellierungsaufgaben an ihre Grenzen stoßen. Selbst bei Aufgaben mit geringer bis mittlerer Komplexität lieferten alle GPT-Versionen häufig unzureichende Ergebnisse. Sie konnten Probleme mit authentischen Kontexten oder auf höherem mathematischen Niveau nur sehr rudimentär lösen und gaben teilweise falsche Antworten. OpenAI hat ChatGPTs mathematische Leistung unter anderem durch den Mathematik SAT-Test bewertet, eine standardisierte Prüfung für College-Anwärter*innen in den USA. Dabei erreichte GPT-3.5 das 70. und GPT-4.0 das 89. Perzentil unter geprüften Personen (Achiam et al., 2023). Ob SAT-Tests in den Trainingsdaten von GPT enthalten sind, ist unklar, weshalb diese Resultate mit Vorsicht zu betrachten sind (Spreitzer et al., 2024).

Als Illustration für die Lösungskompetenz von ChatGPT im Bereich der linearen Algebra soll folgendes Beispiel in Helfrich-Schkarbanenko (2023) herangezogen werden. Der Prompt lautet wie folgt:

„Gegeben sind drei Punkte $A(1|1|1)$, $B(-11|7|7)$ und $C(-3|13|-3)$.
a) Berechnen Sie den Punkt D , so dass das Viereck $ACDB$ ein Parallelogramm ist.
b) Der Punkt $T(-10|14|2)$ liegt in derselben Ebene wie das Parallelogramm $ACDB$.
Liegt der Punkt T im Parallelogramm oder außerhalb des Parallelogramms?
Begründen Sie Ihre Antwort durch eine Berechnung“ (Helfrich-Schkarbanenko, 2023, S. 41).

ChatGPT löst Unteraufgabe a) korrekt über die Berechnung des Vektors $\overrightarrow{AB}$ und anschließender Addition mit Punkt C für das Ergebnis des Punktes D . Bei Unteraufgabe b) hat die KI die richtige Lösungsidee, sie möchte den Vektor von A zu T als Linearkombinationen der Vektoren $\overrightarrow{AB}$ und $\overrightarrow{AC}$ darstellen und kontrollieren, ob die Koeffizienten zwischen 0 und 1 liegen. Die Berechnung der Koeffizienten ist allerdings fehlerhaft und auch die Zwischenschritte für die Berechnung werden nicht ausgegeben. Bei einer erneuten Prompt-Eingabe und

Aufforderungen zur Kontrolle der Koeffizienten, kommt die KI dann auf das richtige Ergebnis (Helfrich-Schkarbanenko, 2023). Auch Sachaufgaben aus dem Bereich der Kombinatorik löst ChatGPT durchaus zufriedenstellend. Durch mehrfaches Generieren von Antworten können verschiedene Lösungswege entstehen, die sich vergleichen lassen. Das ist besonders hilfreich, um alternative Ansätze zu entdecken oder Lösungen auf Effizienz und Eleganz zu prüfen (Spannagel, 2023).

Beim Lösen einer Aufgabe für die Sekundarstufe II aus dem Inhaltsbereich der Analysis scheitert ChatGPT allerdings. Es soll ein Polynom dritten Grades mit folgenden Eigenschaften gefunden werden.

$$p(-2) = 2 \quad p(2) = 2 \quad p'(-2) = 0 \quad p'(2) = 0$$

Das Aufstellen der Gleichungen wurde korrekt durchgeführt, bei der Berechnung scheitert ChatGPT allerdings (Helfrich-Schkarbanenko, 2023). Mit gängiger Mathematik-Software wie Computeralgebrasystemen (CAS) könnte die Berechnung recht einfach durchgeführt werden.

Was sind aber nun die Vorteile und Nachteile von gängigen KI-Systemen gegenüber bereits existierender CAS beim Lösen mathematischer Aufgaben? KI ist zwar flexibel in der Kommunikation mit natürlicher Sprache, macht jedoch mehr Fehler. CAS hingegen sind nahezu fehlerfrei in ihren Berechnungen, erfordern jedoch eine präzise Aufbereitung von Ein- und Ausgaben seitens der Nutzenden. In der Zukunft könnten verstärkt hybride Systeme entwickelt werden, die die Stärken beider Ansätze vereinen. Eine KI könnte ein CAS steuern, korrekte algebraische Berechnungen sicherstellen und zusätzlich die Ein- und Ausgabe in verständlicher Sprache übernehmen (Greefrath et al., 2024).

Im Hinterkopf zu behalten ist, dass KI-Systeme wie ChatGPT auf Sprachmodellen basieren und keine mathematischen Expertensysteme sind. Pepin et al. (2024) meinen, dass ChatGPT über kein umfassendes Verständnis fortgeschrittener mathematischer Konzepte verfügt und mitunter ungenaue Lösungen liefert. Relevant ist allerdings, mit welcher Version die Aufgaben getestet werden. Spannagel (2023) merkt hier an, dass bei Tests von mathematischen Sachaufgaben mit dem älteren GPT-3.5-Modell nur 13 von 20 Versuchen korrekt waren. Das neuere Modell GPT-4 lieferte hingegen in allen 20 Versuchen das richtige Ergebnis. Auch die neueste Version

GPT-4o verspricht Verbesserungen im Vergleich zu den Vorgängerversionen. Wie bereits im Kapitel zu Prompt-Engineering erwähnt, kann mit geeigneten Prompts eine wesentliche Leistungssteigerung beim Lösen mathematischer Aufgaben erzielt werden. Liefert das Modell nach der ersten Eingabe keine richtige bzw. zufriedenstellende Lösung, schlagen Schorcht et al. (2023) eine schrittweise Optimierung des Prompts vor:

1. Zero-Shot-Prompt: Ohne Angabe von Beispielen, sollte die Eingabe erstmal sprachlich klar und deutlich formuliert werden.
2. Zero-Shot- & Chain-of-Thought-Prompt: Bringt Schritt 1 kein zufriedenstellendes Ergebnis, kann die Eingabe mit der Formulierung „Lass uns Schritt für Schritt vorgehen“ ergänzt werden.
3. Zero-Shot- & Chain-of-Thought-Prompt mit Abfrageschleife: Die Eingabe von Schritt 2 wird mit der Phrase „Stelle bei Bedarf Fragen, die dir bei der Lösung des Problems helfen“ ergänzt.
4. Few-Shot- & Chain-of-Thought-Prompt: Die Eingabe wird mit kurzen Lösungsbeispielen, die die Denkstruktur für ähnliche Probleme darstellen, versehen. Wichtig ist eine logische Gedankenkette bei den Beispielen, um dem Modell eine Chain-of-Thought-Struktur zu bieten. Auch die Lösung muss angegeben werden (Schorcht et al., 2023).

Eine mögliche weitere Abhilfe bei Lösungsschwierigkeiten stellt das Wolfram Alpha-Plugin dar, das in der kostenpflichtigen Plus-Version von ChatGPT aktiviert werden kann und auf die Verarbeitung mathematischer Aufgaben spezialisiert ist. Mit dem aktivierten Wolfram-Plugin lassen sich nicht nur Berechnungen präzise durchführen, sondern auch visuelle Darstellungen wie Diagramme oder detaillierte Lösungsschritte anzeigen. Diese Integration macht ChatGPT besonders hilfreich für mathematische Fragestellungen und wissenschaftliche Anwendungen (Spannagel, 2023).

Was bedeutet dies nun alles für die beteiligten Personen am Mathematikunterricht? Eine Studie von Wardat et al. (2023) konnte mithilfe von Interviews die Wahrnehmung von Schüler*innen und Lehrkräften zur Nutzung von ChatGPT im Mathematikunterricht untersuchen. Die Antworten der Teilnehmer*innen zeigen, dass ChatGPT für seine verbesserten mathematischen Fähigkeiten und seinen Beitrag zum Bildungserfolg geschätzt wird, indem es grundlegendes

Wissen in Mathematik und anderen Themen vermittelt. ChatGPT dient zudem als hilfreiches Werkzeug, um Begriffsdefinitionen zu lernen und sich auf den Unterricht vorzubereiten. Schüler*innen können mit der KI Lösungen von Aufgaben auf grundlegendem Niveau überprüfen und direktes Feedback erhalten. Dadurch haben sie schnellen Zugang zu Antworten auf einfache Fragen, die grundlegende Konzepte und Operationen betreffen, was Lehrkräften mehr Zeit für komplexere Fragen verschaffen könnte (Guler et al., 2024). Dies deutet darauf hin, dass die digitalen Fähigkeiten von ChatGPT aktuelle Lehrmethoden verändern und Reformen in der Mathematikbildung fördern könnten (Wardat et al., 2023).

Zusammenfassend kann gesagt werden, dass KI-Systeme wie ChatGPT, insbesondere GPT-4, in der Lage sind, mathematische Aufgaben mit unterschiedlichem Schwierigkeitsgrad zu lösen, wobei die Genauigkeit je nach Problemtyp variiert. Während GPT-4 deutliche Verbesserungen gegenüber älteren Modellen zeigt und teilweise sogar komplexe Probleme korrekt löst, bestehen Schwächen bei geometrischen Aufgaben, grafischen Interpretationen und fortgeschrittenen mathematischen Konzepten. Die Studien zeigen, dass die Leistung durch optimierte Eingaben mithilfe von Prompt-Engineering gesteigert werden kann. Im Vergleich zu klassischen Computeralgebrasystemen punkten KI-Modelle durch die natürliche Sprachverarbeitung, zeigen jedoch häufiger Rechenfehler. Insgesamt bietet ChatGPT Potenzial als unterstützendes Werkzeug im Mathematikunterricht, insbesondere zur Vermittlung grundlegender Inhalte und zur Bereitstellung von Lösungsansätzen.

3.3.2. Planungskompetenz

ChatGPT zeigte sich laut Baytak (2024) besonders nützlich bei der Erstellung detaillierter Unterrichtspläne. Es konnte Aktivitäten, Übungen und Beispiele zur Veranschaulichung mathematischer Konzepte vorschlagen, interaktive Elemente einbauen und die Pläne mit den Lehrplanstandards abstimmen. Zudem bot ChatGPT die Möglichkeit, die Unterrichtspläne individuell anzupassen, um auf unterschiedliche Lerngeschwindigkeiten der Schüler einzugehen. KI-Tools können die Zeit, die Lehrkräfte für die Erstellung und Überarbeitung von Unterrichtsmaterialien aufwenden, erheblich reduzieren. Durch die Unterstützung bei der Erstellung von Unterrichtsplanungen, Arbeitsblättern, Hausaufgaben und Prüfungen gewinnen Lehrkräfte mehr Zeit, die es ihnen ermöglicht, sich stärker auf die Umsetzung der Pläne im Unterricht zu konzentrieren. Die KI reagiert auf die gewünschten Eingaben und kann dadurch

angepasste Unterrichtsplanungen erstellen. Auch wenn der erste Entwurf nicht perfekt ist, können Änderungen vorgenommen und die Planung verfeinert werden (Shah, 2023).

Als Nachteil bei KI-generierten Stundenplanungen sehen Pepin et al. (2024) die Qualität der generierten Inhalte. Diese variiert von Planung zu Planung und eine Tendenz in Richtung lehrerzentrierter Unterrichtsformen konnte festgestellt werden. Auch die von Cameron und Mesiti (2024) generierten Pläne waren grundsätzlich geeignet, um festzulegen, was unterrichtet werden soll. Sie boten jedoch keine ausreichenden Details zu Methoden und wie diese sinnvoll umgesetzt werden können. Dies unterstreicht die Notwendigkeit von Lehrkraft-Expertise, um KI-generierte Unterrichtspläne zu verfeinern und effektiv umzusetzen.

Für die Formulierung von Prompts sollte beachtet werden, dass die gewünschte Notation der Ausgabe explizit an die KI kommuniziert wird. Präzise Angaben sind essenziell, um optimale Ergebnisse zu erzielen (Helfrich-Schkarbanenko, 2023). Auch Shah (2023) fordert dazu auf, diverse Unterrichtselemente, wie einen kreativen Stundeneinstieg, spezifische Fragen zum Thema, Gruppenaktivitäten, die Unterrichtsform oder einen passenden Stundenabschluss explizit im Prompt anzufragen. Auch die Angabe der Dauer der Einheit und die Schulstufe ist natürlich essenziell, damit die KI die Schwierigkeit und das Ausmaß der Einheit miteinbeziehen kann.

Stundenplanungen in Mathematik oder auch Schulbücher enthalten meist Erklärungen diverser mathematischer Konzepte und Zusammenhänge. Bei der Formulierung solcher Sachverhalte kann die KI unterstützend wirken (Pepin et al., 2024). Prompts wie „Erkläre den Satz des Pythagoras auf dem Niveau eines 10-Jährigen“ oder „Erkläre die reellen Zahlen in einfachen Worten“ sind äußerst hilfreich. Die Sprachmodelle greifen für die Erklärungen auf passende Analogien und didaktische Reduktion zurück (Helfrich-Schkarbanenko, 2023). Weiters kann KI in der Planung von Mathematikunterricht vor allem bei der Aufgabenerstellung hilfreich sein. Insbesondere wenn es darum geht, Übungen mit aufsteigendem Schwierigkeitsgrad zu erstellen (Shah, 2023). Die Zone der proximalen Entwicklung (ZPD) ist ein von Lev Wygotski entwickeltes pädagogisch-psychologisches Konzept, das den Bereich zwischen dem aktuellen Entwicklungsstand von Lernenden und dem nächsthöheren Entwicklungsstand beschreibt. Der aktuelle Stand bezieht sich auf Fähigkeiten, die Lernende selbstständig lösen können, während der nächsthöhere Entwicklungsstand Aufgaben umfasst, die alleine nicht bewältigt werden

können. Dazwischen befindet sich die ZPD, die Aufgaben umfasst, die die Lernenden mit Unterstützung von kompetenteren Personen wie Lehrkräften oder Peers bewältigen können. Die ZPD verdeutlicht, dass Lernen am effektivsten erfolgt, wenn Lernende in diesem Bereich unterstützt werden, da sie dadurch neue Fähigkeiten und Konzepte schrittweise verinnerlichen. Die Unterstützung erfolgt durch Anleitung, Zusammenarbeit oder den Einsatz von Werkzeugen, bis der Lernende eigenständig in der Lage ist, das Gelernte anzuwenden (Shabani et al., 2010). Ein Prompt, um solch eine Lernumgebung mit entsprechenden Aufgaben zu generieren, könnte wie folgt lauten:

„Design a series of increasingly challenging math problems where students start with problems they can solve independently and then gradually move onto problems that require peer or teacher assistance“ (Shah, 2023, S. 60).

In einer Untersuchung von Cameron und Mesiti (2024) wurde ein wissenschaftlich fundierter Prompt verwendet, um vier Unterrichtspläne zur Division von Brüchen für Schüler*innen der siebenten und achten Schulstufe zu erstellen. Die Pläne wurden auf ihre Struktur und didaktischen Praktiken analysiert. Trotz geringfügiger Unterschiede folgten alle Pläne einem ähnlichen Ansatz: Einführung von Schlüsselkonzepten und -verfahren, Demonstration mit Beispielen sowie Aufgaben zum eigenständigen Üben. Die KI bietet an dieser Stelle Potenzial und Lehrkräfte können Aufgaben mit relativ wenig Aufwand an die Lernvoraussetzungen der Schüler*innen anpassen, zum Beispiel in Bezug auf kognitive Anforderungen, dem Lebensweltbezug oder der Fachsprache (Buchholtz et al., 2024). Die Division von Brüchen in den Planungen von Cameron und Mesiti (2024) wurde durch die Methode „Kehrwert bilden und multiplizieren“ dargestellt. Dieser syntaktische Ansatz fokussiert jedoch auf das bloße Vermitteln des Rechenverfahrens. Die Autor*innen bemängeln, dass Praktiken zur Förderung des Verständnisses, wie zum Beispiel das Konzept von Balken-Modellen, fehlten.

ChatGPT kann laut Helfrich-Schkarbanenko (2023) auch mathematische Ausdrücke zu geometrischen Sachverhalten, wie zum Beispiel lineare Gleichungen, erstellen. Der in diesem Fall gewählte Zero-Shot-Prompt lautet wie folgt.

„Stelle beispielsweise ein konkretes Lineares Gleichungssystem für folgende Situation im $\mathbb{R}^2$ auf: Drei geraden [*sic*] schneiden sich in einem Punkt S mit Koordinaten $(2|1)$ “ (Helfrich-Schkarbanenko, 2023, S. 34).

Die KI liefert nach korrekten theoretischen Erklärungen, drei Gleichungen, die die gewünschten Anforderungen erfüllen. Auch die Verwendung von ganzen Zahlen anstelle von „komplexeren“ Dezimalzahlen ist hervorzuheben (Helfrich-Schkarbanenko, 2023).

KI ermöglicht es Lehrpersonen auch auf einfache Art und Weise Feedback der Schüler*innen direkt in ihre Unterrichtsplanungen miteinzubeziehen. Wenn Schüler*innen der Lehrperson Feedback geben, kann dies einen wichtigen Beitrag zu einem erfolgreichen und wertschätzenden Unterricht leisten. Die Lehrkraft erhält dadurch Einblicke in den Leistungsstand der Lernenden, ihre Verständnisprobleme und deren Eindruck vom Unterricht. Laut der bekannten Hattie-Studie besitzt der Faktor Feedback eine Effektstärke von $d = 0,73$ und damit einen großen Einfluss auf die Lernleistungen von Schüler*innen (Hattie & Clarke, 2018). Die Integration von Feedback in die Unterrichtsplanung mittels KI wird in folgendem Prompt deutlich.

„My students found my last classroom activity to be too static; they want something where they can move around. Generate a 45-minute lesson plan for one-digit multiplication for my 3rd-grade classroom, incorporating that feedback“ (Shah, 2023, S. 97).

ChatGPT integrierte das Feedback und bot daraufhin zwei unterschiedliche Aktivitäten an, die den geäußerten Bedürfnissen der Schüler*innen entsprachen. Sowohl der Ablauf in einzelnen Schritten, Lernziele, Dauer der einzelnen Phasen und benötigte Materialien wurden aufgelistet. Dies ermöglicht eine schnelle Anpassung des Unterrichts an die Schülerbedürfnisse, ohne stundenlangen Aufwand für die Überarbeitung der Planung (Shah, 2023).

Helfrich-Schkarbanenko (2023) testete auch, ob die KI Musterlösungen für einen gesamten Mathematik-Test, bestehend aus zehn Aufgaben für Maschinenbau-Studierende im zweiten Semester, erstellen kann. Themen wie uneigentliche Integrale, die Konvergenz von Reihen und Differentialgleichungen werden darin behandelt. Zu beachten ist allerdings, dass die Eingabe

hier in Form von LaTeX-Code erfolgte. ChatGPT kann alle Aufgaben korrekt lösen und gibt entsprechende Musterlösungen aus. Der Schwierigkeitsgrad der einzelnen Aufgaben wird auf Nachfrage ebenfalls eingeschätzt. Weiters kann ein äquivalenter, einfacherer oder schwierigerer Test mit anderen Angaben recht einfach erstellt werden.

„Stelle Dir vor, dass Du ein Mathematik-Professor bist. Für das kommende Semester benötige ich vom folgenden Test eine neue Version. Der Schwierigkeitsgrad einzelner Aufgaben sollte aber gleich bleiben. Hier ist der Test vom letzten Semester: [...] Erstelle bitte einen neuen Test. Gebe [*sic*] ihn als LaTeX-Code aus“ (Helfrich-Schkarbanenko, 2023, S. 162).

Auch der Autor war von der Qualität der Ausgabe überrascht. Es wird angenommen, dass die KI bei Aufgaben auf Schulniveau ähnliche, wenn nicht sogar bessere Leistungen zeigt. Insbesondere der Faktor der Zeitersparnis bei der Prüfungserstellung soll hier hervorgehoben werden.

Zusammenfassend kann gesagt werden, dass KI-Tools wie ChatGPT Lehrkräfte effektiv bei der Erstellung und Anpassung von Unterrichtsmaterialien unterstützen können, wodurch sie mehr Zeit für die eigentliche Unterrichtsgestaltung gewinnen. Dabei ist besonders die Möglichkeit hervorzuheben, Unterrichtspläne und Aufgaben individuell auf die Bedürfnisse der Schüler*innen anzupassen. Allerdings wird die Qualität der generierten Inhalte als uneinheitlich bewertet, und es zeigt sich eine Tendenz zu lehrerzentrierten Ansätzen, weshalb die Nachbearbeitung durch Lehrkräfte essenziell bleibt. Präzise und durchdachte Prompts sind entscheidend, um optimale Ergebnisse zu erzielen und die didaktischen Anforderungen zu erfüllen. Zudem bietet KI ein großes Potenzial für die Integration von Schüler*innenfeedback und die Erstellung von Materialien mit abgestuften Schwierigkeitsgrad.

3.3.3. Kontroll- und Bewertungskompetenz

KI-Technologien eröffnen neue Möglichkeiten in der Bewertung und Diagnostik mathematischer Lernprozesse. In Middendorf (2024) wird diskutiert, wie Lehrkräfte durch den Einsatz von KI bei der digital gestützten schulischen Leistungsbewertung entlastet werden können. Schon heute stehen Lehrkräften grundlegende Unterstützungssysteme zur Verfügung.

Erstens gibt es KI-Modelle wie ChatGPT oder Google Gemini, die frei zugänglich sind. Weiters existieren bereits E-Assessment-Tools, über die digitale Tests bzw. Prüfungen erstellt werden können. Beispiele hierfür sind Tests, die auf der Lernplattform Moodle oder anderen Online-Plattformen abgewickelt werden und automatisierte Prüfungs- und Bewertungsprozesse ermöglichen. Tests zeigen, dass ChatGPT in der Lage ist, Fehler von Schüler*innen in Mathematikaufgaben zu erkennen, wenn die zugrunde liegenden Aufgaben nicht zu komplex sind. Die Fehler- und Kompetenzanalysen funktionieren jedoch noch nicht immer zuverlässig und erfordern eine individuelle Nachprüfung (Schorcht et al., 2024). KI-gestützte Leistungsbewertung birgt laut Buchholtz et al. (2024) allerdings das Risiko, standardisierte und prozedurale Aufgaben zu stark in den Vordergrund zu rücken, da KI momentan vor allem diese zuverlässig überprüft. Das Überprüfen des konzeptuellen Verständnisses mathematischer Inhalte könnte durch den KI-Einsatz seltener ausfallen.

Ein Test von Schorcht et al. (2024) kommt hingegen zu einem anderen Schluss. Sie gaben ChatGPT eine handschriftlich nachgestellte Schülerantwort einer Aufgabe, bei der ein einfacher Term so weit wie möglich vereinfacht werden sollte. Weiters sollte die Schülerlösung hinsichtlich mathematischer Kompetenzen analysiert werden. Die Handschrift wird von der KI korrekt erkannt und auch der Fehler wird gefunden, sowie die richtige Lösung ausgegeben. Mehrere Kompetenzen, die von ChatGPT ohne weitere Quellenangabe herangezogen werden, werden aufgelistet und anhand der Aufgabe beurteilt. Bei einem weiteren Test von Buchholtz et al. (2024) ist ChatGPT nicht in der Lage, handschriftliche Schülerlösungen zuverlässig zu bewerten. In Tests wurden Schülerfehler, etwa bei der schriftlichen Multiplikation,



Ein Stromlieferant bietet zwei Tarife an:
 Tarif 1: 90,00 EUR Grundpreis und 35 Ct/kWh Arbeitspreis.
 Tarif 2: 60,00 EUR Grundpreis und 42 Ct/kWh Arbeitspreis.
 Beachte, dass bei beiden Tarifen noch zusätzlich 7 % Steuern auf den Grundpreis und 19 % Steuern auf den Arbeitspreis gezahlt werden müssen.
 Erstelle eine richtige und vier falsche Gleichungen, um zu berechnen, bei welchem Verbrauch der Tarif gewechselt werden sollte. Auch die falschen Gleichungen sollen möglichst plausibel wirken.



Richtige Gleichung:

$$90 + 0,07 \cdot 90 + (35 \cdot x) + 0,19 \cdot (35 \cdot x) = 60 + 0,07 \cdot 60 + (42 \cdot x) + 0,19 \cdot (42 \cdot x)$$

Falsche Gleichungen:

- $90 + 0,07 \cdot 60 + (35 \cdot x) + 0,19 \cdot (35 \cdot x) = 60 + 0,07 \cdot 90 + (42 \cdot x) + 0,19 \cdot (42 \cdot x)$
- $90 + 0,07 \cdot 90 + (35 \cdot x) + 0,19 \cdot (35 \cdot x) = 60 + 0,07 \cdot 60 + (35 \cdot x) + 0,19 \cdot (35 \cdot x)$
- $90 + 0,07 \cdot 60 + (35 \cdot x) + 0,19 \cdot (35 \cdot x) = 60 + 0,07 \cdot 90 + (35 \cdot x) + 0,19 \cdot (42 \cdot x)$
- $90 + 0,07 \cdot 60 + (35 \cdot x) + 0,19 \cdot (35 \cdot x) = 60 + 0,07 \cdot 60 + (42 \cdot x) + 0,19 \cdot (35 \cdot x)$

Abbildung 7: Generierung einer Multiple-Choice-Aufgabe mittels ChatGPT (Middendorf, 2024, S. 3)

häufig falsch beurteilt. Solche Fehler in der Bewertung könnten den Lernprozess eher behindern als fördern. Ob und in welchem Umfang diese Bildinterpretationen tatsächlich für diagnostische Zwecke eingesetzt werden können und ob Lehrkräfte diese in ihrer Praxis nutzen würden, ist bislang wenig erforscht. Die Potenziale für den Einsatz sind grundsätzlich vorhanden (Schorcht et al., 2024). Prüfungsaufgaben können in offener, halboffener oder geschlossener Form gestellt werden. Soll die elektronische Auswertung und Bewertung mit E-Assessment-Software erfolgen, sind jedoch nur geschlossene Aufgabenformate geeignet. Beispiele dafür sind Single-Choice-Aufgaben, Multiple-Choice-Aufgaben, sowie Zuordnungsaufgaben und Anordnungsaufgaben (Middendorf, 2024). Ein Beispiel für die Erstellung solch einer Aufgabe mittels KI wird in Abbildung 7 gezeigt. KI-Modelle können für geschlossene Prüfungsaufgaben sowohl korrekte Antwortmöglichkeiten als auch plausible Falschantworten, sogenannte Distraktoren, schnell generieren, was für die meisten Unterrichtsfächer nützlich ist. Sie minimieren das Risiko, dass Schüler*innen richtige oder falsche Antworten aufgrund formaler Strukturen oder Signalwörter erraten (Middendorf, 2024). Damit übernehmen KI-Modelle die zeitintensive Aufgabe, plausible Distraktoren zu entwerfen, die mitunter die größte und zeitaufwendigste Komponente bei der Erstellung von Multiple-Choice-Fragen sind (Lindner et al., 2015). Auch im Bereich der Individualdiagnostik könnten spezialisierte Systeme Entlastungen für Lehrer*innen bringen. Intelligente Tutorsysteme (ITS) ermöglichen eine automatisierte Analyse von Schüler*innenlösungen. Sie identifizieren spezifische Lernschwierigkeiten und liefern Lehrkräften detaillierte Lernprozessdaten, etwa zur Bearbeitungsdauer von Aufgaben. Diese Einblicke ermöglichen eine gezielte Unterstützung der Schülerinnen und Schüler sowie individuell angepasstes Feedback (Buchholtz et al., 2024). Zusammenfassend kann gesagt werden, dass die Integration von KI-Technologien zukünftig dazu beitragen könnte, zeitaufwendige Aufgaben wie die Erstellung und Bewertung von Prüfungen zu automatisieren und Lehrkräften mehr Raum für pädagogische Kernaufgaben zu schaffen. Voraussetzung dafür bleibt jedoch die Weiterentwicklung dieser Technologien sowie die Anpassung an die spezifischen Anforderungen des schulischen Umfelds (Middendorf, 2024).

3.3.4. Weitere Kompetenzen

An dieser Stelle soll nicht unerwähnt bleiben, dass ChatGPT auch LaTeX beherrscht, ein Textsatzsystem, das speziell für mathematische Formeln und Dokumente geeignet ist. Damit lassen sich komplexe mathematische Ausdrücke präzise und elegant darstellen. Als Beispiel

wird hier ChatGPT um die Erstellung einer 5 x 5 Einheitsmatrix gebeten, der generierte LaTeX-Code, sowie die kompilierte Version sind in Abbildung 8 ersichtlich.

Nicht nur die Ausgabe von LaTeX beherrscht ChatGPT. Experimente von Frieder et al. (2024) zeigen, dass auch die Eingaben von mathematischen Problemen in Form von LaTeX-Code erfolgreich verarbeitet und gelöst werden. Auch bei Tippfehlern der Bedienenden war ChatGPT in der Lage die beabsichtigte Bedeutung zu erkennen. Beispielsweise wurde $\$cup\$$ korrekt als $\$\cup\$$ interpretiert. Neben LaTeX beherrscht die KI auch TikZ, eine auf LaTeX basierende Sprache zur Erstellung von Vektorgrafiken. TikZ ermöglicht es, Diagramme, Plots oder geometrische Figuren flexibel und detailgenau zu entwerfen. Um ein möglichst präzises Ergebnis zu erzielen, sollten im Prompt jedoch Einschränkungen für den Definitions- und Wertebereich, die Linienstärken des Koordinatensystems sowie die Abstände der Gitterlinien geliefert werden. Es ist zu beachten, dass ChatGPT lediglich den TikZ-Code generiert. Um die Grafik zu erstellen, muss dieser Code in ein LaTeX-Programm kopiert und dort kompiliert werden (Helfrich-Schkarbanenko, 2023). In einem Eigentest konnte festgestellt werden, dass Diagramme oder Plots jedoch standardmäßig auch direkt in ChatGPT angezeigt werden können. Die Darstellung basiert dabei auf der Programmiersprache Python, dessen Quellcode ebenfalls angezeigt werden kann.

$$\begin{array}{l} \backslash[\\ \backslash\textbf{begin}\{\textbf{bmatrix}\} \\ 1 \& 0 \& 0 \& 0 \& 0 \backslash\backslash \\ 0 \& 1 \& 0 \& 0 \& 0 \backslash\backslash \\ 0 \& 0 \& 1 \& 0 \& 0 \backslash\backslash \\ 0 \& 0 \& 0 \& 1 \& 0 \backslash\backslash \\ 0 \& 0 \& 0 \& 0 \& 1 \\ \backslash\textbf{end}\{\textbf{bmatrix}\} \\ \backslash] \end{array} \longrightarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Abbildung 8: Ausgegebener und kompilierter LaTeX-Code einer 5x5 Einheitsmatrix
(Helfrich-Schkarbanenko, 2023, S. 28)

4. Testdurchführung mit ChatGPT

Im Rahmen dieser Arbeit wurden eigene Tests mit der Software ChatGPT durchgeführt, um ihre Leistungsfähigkeit in verschiedenen mathematisch-didaktischen Kontexten zu untersuchen. Ziel war es, die Stärken und Schwächen des Modells diesbezüglich zu analysieren und seine potenziellen Anwendungsmöglichkeiten zu bewerten.

4.1. Lösen von Mathematikaufgaben mit ChatGPT

Um die Lösungsfähigkeiten von ChatGPT in einem bildungsrelevanten Kontext zu untersuchen, werden die Aufgaben der österreichischen Mathematik-Matura herangezogen. Diese Aufgaben umfassen zum Teil einfache Rechenoperationen, aber auch komplexe Problemstellungen und mathematische Modellierungen. Die Prüfungen beinhalten aufgrund ihres „Abschlusscharakters“ diverse Themengebiete der Mathematik und eignen sich daher besonders gut, um die Fähigkeiten der KI breit zu testen. Weiters sind wie bereits diskutiert, die Leistungen von ChatGPT bei weniger komplexen Aufgaben durchaus beachtlich, weswegen hier bewusst auf eine Erprobung mit „klassischen Schulbuchaufgaben“ der Sekundarstufe I oder II nach dem Schema „Löse die Gleichung...“ verzichtet wird. In den weiteren Kapiteln erfolgt eine quantitative Analyse der gelösten Aufgaben, sowie Darstellungen ausgewählter Antworten und Lösungswege.

4.1.1. Methode und Durchführung

Für die Testung werden die Aufgaben der jüngsten Mathematik-Maturen herangezogen. Konkret handelt es sich um die Haupt- und Herbsttermine 2023/24 und die Haupt-, Herbst- und Wintertermine 2022/23. Prüfungstermine von September, Mai und Jänner 2024, sowie September und Mai 2023 der allgemeinbildenden höheren Schulen (AHS). Die Prüfung gliedert sich in zwei Teile, wobei die sogenannten Typ-1-Aufgaben meist aus einem kurzen Rechenweg bestehen und auf die gefragten Grundkompetenzen abzielen. Teil-2-Aufgaben sind meist umfangreichere und kontextbezogene Aufgaben, die komplexere Lösungsmethoden und eine Vernetzung der Grundkompetenzen erfordern. Alle Prüfungen bestehen aus 24 Teil-1-Aufgaben und 4 Teil-2-Aufgaben, die wiederum 1 bis 4 Teilaufgaben beinhalten (BMBWF, 2023). Insgesamt werden im Zuge der vorliegenden Arbeit fünf Maturen und damit 200 (Teil)-Aufgaben getestet.

In den Prüfungen werden vier große Inhaltsbereiche behandelt: Algebra und Geometrie (AG), Funktionale Abhängigkeiten (FA), Analysis (AN) und Wahrscheinlichkeit und Statistik (WS). Die Algebra dient als Sprache der Mathematik und ermöglicht den Umgang mit Variablen, Termen, Gleichungen, Ungleichungen und Gleichungssystemen. Neben der regelhaften Umformung stehen die Interpretation und Anwendung dieser Konzepte im Vordergrund. Algebra verbindet sich mit der Geometrie, indem geometrische Probleme algebraisch und algebraische Sachverhalte geometrisch dargestellt werden können. Funktionen beschreiben Beziehungen zwischen Größen und bilden ein zentrales Werkzeug der Mathematik. Der Fokus liegt auf dem Erkennen und Interpretieren wichtiger Funktionstypen und deren Eigenschaften. Funktionen dienen sowohl zur Modellierung als auch zur Kommunikation mathematischer Zusammenhänge in verschiedenen Kontexten. In der Analysis wird das Änderungsverhalten von Funktionen mithilfe von Differenzial- und Integralrechnung behandelt. Zentral sind das Verständnis und die Deutung dieser Konzepte in mathematischen und realen Kontexten, weniger deren rechnerische Umsetzung. Im Inhaltsbereich der Statistik werden Daten analysiert, Tabellen und Grafiken erstellt, sowie Kennzahlen wie Mittelwert oder Standardabweichung berechnet. In der Stochastik stehen Wahrscheinlichkeiten, Zufallsgrößen und Wahrscheinlichkeitsverteilungen im Fokus. Ziel ist die Interpretation und Bewertung von Aussagen, anhand durchgeführter Berechnungen. Diese Bereiche fördern die Fähigkeit, mathematische Begriffe, Modelle und Darstellungen verständlich anzuwenden und in verschiedenen Kontexten zu kommunizieren (BMBWF, 2024a).

Weiters treten in den Mathematik-Maturen sechs unterschiedliche Antwortformate auf: offenes Antwortformat, halboffenes Antwortformat, Konstruktionsformat, Multiple-Choice-Format, Lückentext und Zuordnungsformat. Beim offenen Antwortformat erfolgt die Bearbeitung frei und ohne Einschränkungen, wobei die Antworten je nach Aufgabenstellung flexibel gestaltet sein können, etwa durch eigene Erklärungen, Herleitungen oder Darstellungen. Das halboffene Antwortformat verlangt, dass die richtige Antwort in eine vorgegebene Struktur wie eine Gleichung, Funktion oder ein Diagramm eingefügt wird. Im Multiple-Choice-Format wird ein Fragenstamm mit vorgegebenen Antwortmöglichkeiten bereitgestellt, von denen die korrekten Antworten angekreuzt werden müssen. Hier stehen entweder fünf Möglichkeiten mit zwei korrekten Antworten (2 aus 5) oder sechs Möglichkeiten mit einer korrekten Antwort (1 aus 6) zur Verfügung. Beim Lückentext-Format sind im Aufgabentext zwei Lücken ausgewiesen, die mit jeweils drei vorgegebenen Antwortmöglichkeiten ergänzt werden können.

Die Bearbeitung erfolgt durch das Ankreuzen der zutreffenden Optionen für beide Lücken. Das Zuordnungsformat besteht aus sechs Auswahlmöglichkeiten, beispielsweise Aussagen, Tabellen oder Abbildungen, die vier Antwortmöglichkeiten zugeordnet werden müssen. Hier wird durch Eintragung eines Buchstabens (A bis F) jede Antwortmöglichkeit mit der passenden Auswahlmöglichkeit verbunden (BMBWF, 2019). Im ersten Teil wird jede Aufgabe mit 0 oder 1 Punkt bzw. 0, 0.5 oder 1 Punkt bewertet. Aufgaben mit halben Punkten sind gekennzeichnet. Der Nachweis der Grundkompetenz ist hier für die Punktevergabe entscheidend. Der zweite Teil besteht aus einer Aufgabe mit reduziertem Kontext, bei der maximal vier Punkte erzielt werden können. Die Aufgabenstellungen sind dabei unabhängig voneinander lösbar. Es folgen drei weitere Aufgaben, die auch jeweils mit bis zu vier Punkten bewertet werden können. Bei den übrigen drei Aufgaben kommt die sogenannte Best-of-Wertung zum Einsatz. Dies bedeutet, dass nur die beiden Aufgaben berücksichtigt werden, bei denen die meisten Punkte erzielt wurden, und die Aufgabe mit der geringeren Punktzahl nicht in die Bewertung einfließt. Die Aufgaben werden mit 0 oder 1 Punkt bzw. mit 0, 0.5 oder 1 Punkt bewertet. Die Bewertung orientiert sich an der Fähigkeit der Kandidat*innen, die Aufgaben inhaltlich und methodisch korrekt zu lösen und dabei gegebenenfalls mehrere Kompetenzen miteinander zu verknüpfen (BMBWF, 2023). Der Beurteilungsschlüssel ist in Tabelle 1 zu sehen.

Erreichte Punkte	Note
32 – 36 Punkte	Sehr gut
27 – 31,5 Punkte	Gut
22 – 26,5 Punkte	Befriedigend
17 – 21,5 Punkte	Genügend
0 – 16,5 Punkte	Nicht genügend

Tabelle 1: Beurteilungsschlüssel SRP Mathematik (BMBWF, 2023, S. 2)

Die Testung wird in der Desktopanwendung von ChatGPT mit der neuesten verfügbaren Version GPT-4o (Stand November 2024) durchgeführt. Verwendet wurde die kostenpflichtige Plus-Version, um eine höhere Anzahl an Eingaben für das 4o-Modell zur Verfügung zu haben. Um alle Arten von Aufgaben miteinzubeziehen zu können, werden die Eingaben in ChatGPT mittels Bilddateien getätigt. Eine im PDF-Format geöffnete Matura-Aufgabe wird mittels Screenshot-Funktion in die Zwischenablage kopiert und in der Eingabezeile von

ChatGPT eingefügt. Das Absenden des Bildes erfolgt via Enter-Taste. Eine reine Texteingabe ist aufgrund gegebener Antwortformate oder grafischen Darstellungen bei manchen Aufgaben nicht möglich. Aufgaben, wo zum Beispiel Vektoren aus Koordinatensystemen oder Steigungen von Funktionen ausgelesen werden müssen, können somit miteinbezogen werden. Dasselbe gilt für Aufgaben mit Zuordnungen, Konstruktionen oder Lückentexten als Antwortformat.

Wie bereits in Kapitel 2.10. erläutert, können durch die richtige Eingabe und die Optimierung der Prompts häufiger korrekte Lösungen forciert werden. Um die Wirksamkeit der verschiedenen Promptarten zu testen, wurden diese bei den eingefügten Aufgaben variiert. Jede Aufgabe wurde einmal ohne Prompt (oP), einmal mit einem Chain-of-Thought-Prompt (CoT), einmal mit einem Ask-me-Anything-Prompt (AmA) und einmal mit einer Kombination aus CoT und AmA (CoT & AmA) versehen. Die Prompts wurden nach Hochladen des Aufgaben-Screenshots in das Textfeld des Chatbots eingefügt, was in Abbildung 9 am Beispiel der CoT-Eingabe zu sehen ist. Folgende Prompts wurden jeweils unabhängig voneinander in separaten Chats verwendet.

1. oP: keine zusätzliche Texteingabe
2. CoT: „Lass uns Schritt für Schritt vorgehen.“
3. AmA: „Stelle bei Bedarf Fragen, die dir bei der Lösung des Problems helfen.“
4. CoT & AmA: „Lass uns Schritt für Schritt vorgehen. Stelle bei Bedarf Fragen, die dir bei der Lösung des Problems helfen.“

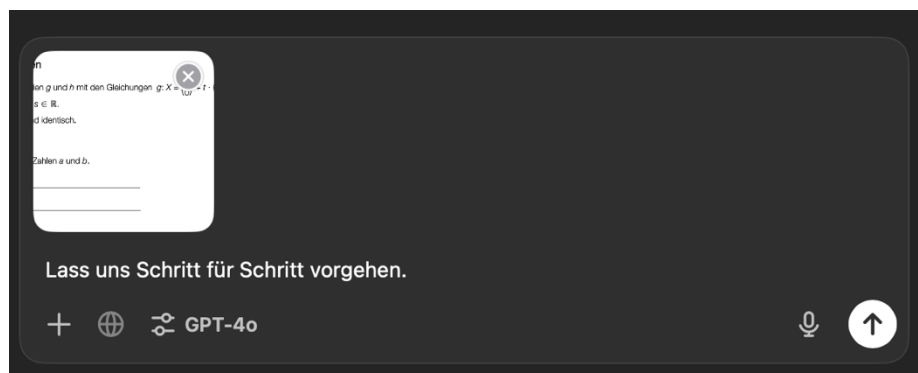


Abbildung 9: Eingabefeld von ChatGPT mit Aufgaben-Screenshot und CoT-Prompt (OpenAI, 2024a)

Die jeweiligen Ausgaben der KI wurden mit dem Korrekturheft der entsprechenden Prüfung abgeglichen und anhand des bereits vorgestellten Punkteschemas bewertet. Für jeden der vier

Prompts wurde getrennt entschieden, ob die Ausgabe korrekt oder inkorrekt ist. Die Konstruktionsaufgaben wurden als richtig gewertet, wenn von ChatGPT ein neues Bild generiert wurde, welches der korrekten Konstruktion entspricht. Der Grund dafür ist, dass eine Manipulation von hochgeladenen Bilder für die KI nicht möglich ist und damit beispielsweise auch die Fähigkeit zum Einzeichnen in gegebene Koordinatensysteme. Die Aufzeichnung und Auswertung der Ergebnisse erfolgten in einer Excel-Tabelle.

4.1.2. Ergebnisse

Von insgesamt 200 getesteten (Teil-)Aufgaben wurden gemittelt über die vier verschiedenen Promptarten 70 % der Aufgabenstellungen korrekt gelöst, wobei die Eingaben ohne Prompt mit 73 % die höchste Erfolgsquote aufweisen. Es folgen der CoT- & AmA-Prompt mit 70 %, der CoT-Prompt mit 69 % und der AmA-Prompt mit 68 %. Beim Prüfungstermin vom September 2023 konnten durchschnittlich 31,875 von 40 (Teil-)Aufgaben korrekt gelöst werden, er stellt damit die am besten gemeisterte Prüfung dar. Am wenigsten erfolgreich war die KI beim Termin vom Mai 2024 mit gemittelt 25,375 von 40 richtig gelösten (Teil-)Aufgaben. Aufgrund der Best-of-Wertung im Teil 2, ergeben sich die in Tabelle 2 ersichtliche und letztendlich mit dem Beurteilungsschlüssel aus Tabelle 1 zu bewertende Punktezahlen und Beurteilungen.

	Ø Punktezahl	Note
Herbsttermin 2023/24: Teil 1	19,375	
Herbsttermin 2023/24: Teil 2	4,75	
Herbsttermin 2023/24: Gesamt	24,125	Befriedigend
Haupttermin 2023/24: Teil 1	18,625	
Haupttermin 2023/24: Teil 2	4,5	
Haupttermin 2023/24: Gesamt	23,125	Befriedigend
Wintertermin 2022/23: Teil 1	18,625	
Wintertermin 2022/23: Teil 2	7,625	
Wintertermin 2022/23: Gesamt	26,25	Befriedigend
Herbsttermin 2022/23: Teil 1	22,25	
Herbsttermin 2022/23: Teil 2	7	
Herbsttermin 2022/23: Gesamt	29,25	Gut
Haupttermin 2022/23: Teil 1	21,25	
Haupttermin 2022/23: Teil 2	6,375	
Haupttermin 2022/23: Gesamt	27,625	Gut

Tabelle 2: Über die Promptarten gemittelte Punktezahlen und Notengrade der einzelnen Prüfungstermine

Betrachtet man die einzelnen Prüfungstermine und die erreichten Punktezahlen der unterschiedlichen Promptarten, kann man in den meisten Fällen geringe Punkteabweichungen unter den verschiedenen Prompts erkennen, wie in Tabelle 3 zu sehen ist.

	oP	CoT	AmA	CoT & AmA	Ø Punkte
Herbsttermin 2023/24	25	23,5	23,5	24,5	24,125
Haupttermin 2023/24	26	21,5	22	23	23,125
Wintertermin 2022/23	26,5	26,5	24,5	27,5	26,25
Herbsttermin 2022/23	30	28,5	29	29,5	29,25
Haupttermin 2022/23	28	27,5	26,5	28,5	27,625

Tabelle 3: Gesamtpunktezahlen bei unterschiedlichen Promptarten

Auch die erreichten Notengrade sind bei Variation der Prompts meist ident. Bei zwei Prüfungsterminen erhält man bei allen vier Promptarten denselben Notengrad als Endergebnis, bei den drei übrigen Prüfungsterminen bekommt man jeweils bei einer Promptart einen anderen Notengrad als Ergebnis, wie in Tabelle 4 ersichtlich.

	oP	CoT	AmA	CoT & AmA	Ø Note
Herbsttermin 2023/24	Befriedigend	Befriedigend	Befriedigend	Befriedigend	Befriedigend
Haupttermin 2023/24	Befriedigend	Genügend	Befriedigend	Befriedigend	Befriedigend
Wintertermin 2022/23	Befriedigend	Befriedigend	Befriedigend	Gut	Befriedigend
Herbsttermin 2022/23	Gut	Gut	Gut	Gut	Gut
Haupttermin 2022/23	Gut	Gut	Befriedigend	Gut	Gut

Tabelle 4: Notengrade bei unterschiedlichen Promptarten

Von den 200 (Teil-)Aufgaben entfielen 74 auf das offene und 52 auf das halboffene Antwortformat. 34 Stück waren per Multiple-Choice, 18 per Lückentext und 12 per Zuordnung zu beantworten. Lediglich 10 Aufgaben wiesen das Konstruktionsformat auf. Die prozentuellen Häufigkeiten der richtigen Antworten bei den jeweiligen Antwortformaten werden in Tabelle 5 dargestellt.

	oP	CoT	AmA	CoT & AmA	Ø Prozent
offen	78 %	73 %	74 %	76 %	75 %
halboffen	83 %	78 %	73 %	73 %	77 %
Lückentext	72 %	75 %	69 %	78 %	74 %
Zuordnung	63 %	58 %	50 %	54 %	56 %
Multiple-Choice	62 %	56 %	62 %	68 %	62 %
Konstruktion	30 %	40 %	30 %	30 %	33 %

Tabelle 5: Prozentuelle Häufigkeiten richtiger Antworten innerhalb des jeweiligen Antwortformats

Während die KI beim offenen und halboffenen Format mit 75 % bzw. 77 % Erfolgsquote am besten abschnitt, war beim Konstruktionsformat das Gegenteil der Fall. Lediglich 33 % der Aufgaben mit Konstruktion wurden korrekt beantwortet. Beim Inhaltsbereich entfielen 41 (Teil-)Aufgaben auf das Gebiet „Algebra & Geometrie“, 53 auf „Funktionale Abhängigkeiten“, 55 auf „Analysis“ und 51 auf „Wahrscheinlichkeit & Statistik“. Die prozentuellen Häufigkeiten der richtigen Antworten bei den jeweiligen Inhaltsbereichen sind in Tabelle 6 ersichtlich.

	oP	CoT	AmA	CoT & AmA	Ø Prozent
AG	66 %	62 %	60 %	63 %	63 %
FA	77 %	73 %	73 %	74 %	74 %
AN	74 %	65 %	61 %	68 %	67 %
WS	73 %	75 %	76 %	75 %	75 %

Tabelle 6: Prozentuelle Häufigkeiten richtiger Antworten innerhalb des jeweiligen Inhaltsbereichs

Analysiert man die Häufigkeiten richtiger Antworten aus den beiden Teilen der Prüfungen separat, erhält man Tabelle 7.

	oP	CoT	AmA	CoT & AmA	Ø Prozent
Teil 1	73 %	70 %	67 %	73 %	70 %
Teil 2	73 %	68 %	69 %	67 %	69 %

Tabelle 7: Prozentuelle Häufigkeiten richtiger Antworten beim jeweiligen Prüfungsteil

4.2. Fallbeispiele gelöster Aufgaben

In folgendem Kapitel sollen ausgewählte Fallbeispiele der von ChatGPT gelösten Aufgaben dargestellt werden. Die Schwächen und Stärken der KI sollen damit an konkreten Sachverhalten aufgezeigt werden und zusätzliche Einblicke zu den quantitativen Daten bieten. Bewusst wurden Aufgaben aus allen Inhaltsbereichen und allen Antwortformaten gewählt, um ein breites Spektrum abzubilden. Auch auf die Richtigkeit bei den unterschiedlichen Promptarten wurde geachtet, um etwaige Unterschiede aufzuzeigen. In Tabelle 8 sind alle Fallbeispiele und deren Ergebnisse aufgelistet.

	Teil	Inhaltsbereich	Format	oP	CoT	AmA	CoT & AmA
Fallbeispiel 1	1	AG	Lückentext	0,5	0,5	0,5	0,5
Fallbeispiel 2	1	FA	halboffen	0,5	0,5	0	0
Fallbeispiel 3	2	AG	Konstruktion	0	0	0	0
Fallbeispiel 4	1	AG	Konstruktion	1	1	1	1
Fallbeispiel 5	1	FA	Multiple-Ch.	1	0	0	0
Fallbeispiel 6	1	WS	halboffen	1	1	0	1
Fallbeispiel 7	2	WS	offen	0	0	0	0
Fallbeispiel 8	1	AN	Zuordnung	0,5	0,5	0	0,5
Fallbeispiel 9	2	AG	halboffen	1	1	1	1
Fallbeispiel 10	2	WS	Konstruktion	0	1	1	0

Tabelle 8: Übersicht der Fallbeispiele

4.2.1. Fallbeispiel 1: Zahlenmengen

Gegeben sind zwei natürliche Zahlen, a und b , mit $b > a$.

Aufgabenstellung:

Ergänzen Sie die Textlücken im nachstehenden Satz durch Ankreuzen des jeweils zutreffenden Satzteils so, dass eine richtige Aussage entsteht.

In jedem Fall ist $a - b$ eine ① und $b - a$ eine ②.

①		②	
natürliche Zahl	☐	natürliche Zahl	☒
rationale, aber keine natürliche Zahl	☒	ganze, aber keine natürliche Zahl	☐
rationale, aber keine ganze Zahl	☐	rationale, aber keine natürliche Zahl	☐

Abbildung 10: Fallbeispiel 1- Prüfungsaufgabe der AHS-Matura des Herbsttermins 2023/24 (BMBWF, 2024b)

In gegebener Aufgabe in Abbildung 10 soll eine Aussage über die Differenz zweier natürlicher Zahlen vervollständigt werden. Die korrekte Lösung ist farblich gekennzeichnet. In Abbildung 11 ist die Ausgabe von ChatGPT ohne zusätzlichem Prompt gezeigt. Die Antwort für die zweite Lücke wird mit „natürliche Zahl“ korrekt beantwortet, die Begründung ist ebenfalls valide. Wenn $b > a$ ist und beides natürliche Zahlen laut Voraussetzung sind, muss die Differenz $b - a$ folglich auch eine natürliche Zahl sein. Da die übrigen Antwortmöglichkeiten für die zweite Lücke die Formulierung „keine natürliche Zahl“ enthalten, ist die Antwort für ChatGPT

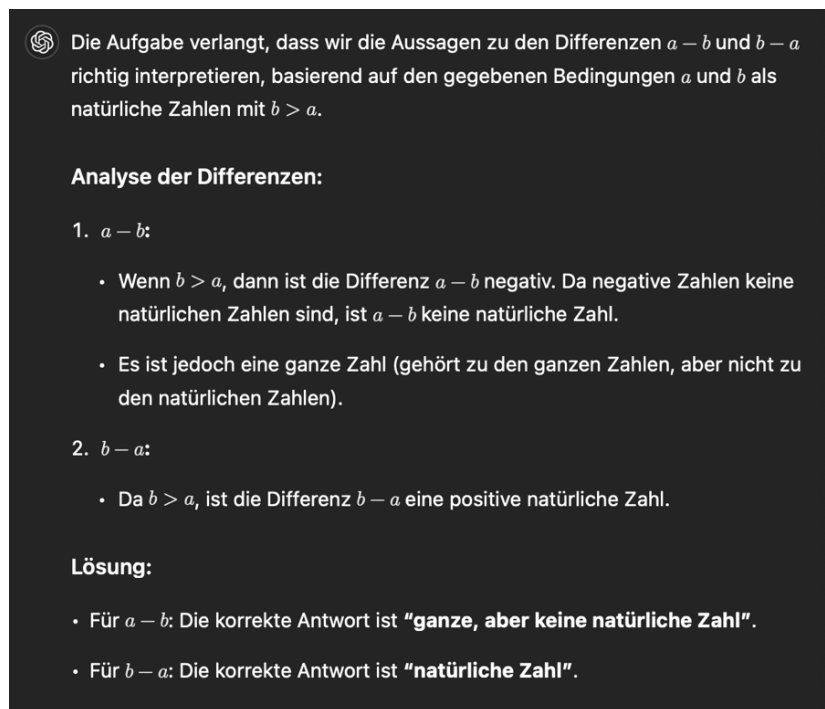


Abbildung 11: Fallbeispiel 1 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024aa)

offensichtlich eindeutig. Bei der ersten Lücke geht es um die Differenz $a - b$, die aufgrund der Voraussetzungen eine negative und damit ganze Zahl sein muss. Die KI deutet den Ausdruck ebenfalls als negative Zahl, und schließt daraus, dass es keine natürliche Zahl sein kann. Eine weitere Eingrenzung und Argumentation werden nicht geliefert und die Entscheidung für die Antwort „ganze, aber keine natürliche Zahl“ erfolgt scheinbar willkürlich. Diese Antwort ist allerdings für die zweite Lücke vorgesehen und damit ohnehin inkorrekt. Die Aufgabe wird für die Promptart „ohne Prompt“ wegen einer korrekten Teilantwort mit 0,5 Punkten bewertet. Das gleiche Ergebnis erzielen auch die anderen Promptversionen und auch die gegebenen Antworten sind vollständig ident. Einzig beim AmA-Prompt wird eine Rückfrage über den relevanten Definitionsbereich gestellt, welche mit „Der Definitionsbereich sind die reellen Zahlen.“ beantwortet wurde.

4.2.2. Fallbeispiel 2: Kosten, Erlös und Gewinn

In gegebener Aufgabe in Abbildung 12 sollen eine Erlös- und eine Kostenfunktion analysiert werden. Das Antwortformat ist halboffen und es handelt sich um den Inhaltsbereich „Funktionale Abhängigkeiten“. Die korrekte Lösung ist farblich gekennzeichnet. In Abbildung 13 ist die Ausgabe von ChatGPT mit CoT-Prompt gezeigt. ChatGPT geht, wie vom Prompt intendiert, schrittweise vor. Die Gewinnfunktion $G(x)$ wird als Differenz der Erlösfunktion E und der Kostenfunktion K definiert und die Lagebeziehungen zueinander werden erläutert. Ein positiver Gewinn wird in dem Bereich erzielt, wo die Erlösfunktion über der Kostenfunktion liegt. Der Wert x_1 wird als jene gefragte Stelle erkannt, bei der sich Erlös- und Kostenfunktion schneiden. Das Ablesen ist mit $x_1 = 100$ korrekt. Ebenso richtig erkannt wird die Tatsache, dass der größte Gewinn erreicht wird, wenn der senkrechte Abstand zwischen den Graphen maximal ist. Die KI lokalisiert diese Stelle x_2 bei 200 Mengeneinheiten, was außerhalb des Toleranzbereichs [225; 275] liegt. Diese Ausgabe des CoT-Prompts wurde aufgrund einer von zwei richtigen Antworten mit 0,5 Punkten bewertet. Bei der Ausgabe ohne Prompt war die Argumentation ähnlich, jedoch wurde die erste Stelle falsch ausgelesen. Dafür lag die zweite Stelle im Toleranzbereich und die Ausgabe bekommt ebenso 0,5 Punkte. Auch bei den AmA-

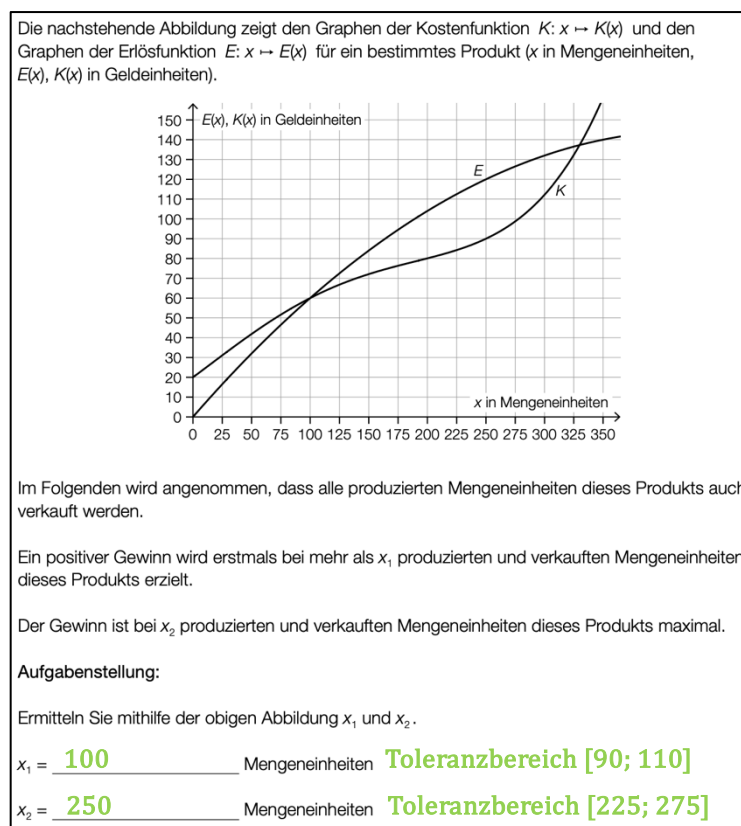


Abbildung 12: Fallbeispiel 2 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2023/24 (BMBWF, 2024b)

Schritt-für-Schritt-Lösung:

- Analyse der Graphen:**
 - Der Gewinn ergibt sich aus der Differenz zwischen der Erlösfunktion $E(x)$ und der Kostenfunktion $K(x)$:
 $G(x) = E(x) - K(x)$.
 - Ein positiver Gewinn wird erzielt, wenn $E(x) > K(x)$, also wenn der Graph von $E(x)$ oberhalb des Graphen von $K(x)$ liegt.
 - Der maximale Gewinn wird erreicht, wenn der vertikale Abstand zwischen den beiden Graphen maximal ist.
- Bestimmung von x_1 :**
 - x_1 ist die erste Stelle, bei der $E(x) > K(x)$, also der Punkt, an dem sich die beiden Graphen schneiden und $E(x)$ beginnt, oberhalb von $K(x)$ zu liegen.
 - Aus der Abbildung erkennt man, dass dies etwa bei:
 $x_1 \approx 100$ Mengeneinheiten
- Bestimmung von x_2 :**
 - x_2 ist die Stelle, an der der Gewinn maximal ist, also der Punkt, an dem der Abstand zwischen $E(x)$ und $K(x)$ am größten ist.
 - Aus der Abbildung erkennt man, dass dies etwa bei:
 $x_2 \approx 200$ Mengeneinheiten
- Ergebnisse:**
 - $x_1 = 100$ Mengeneinheiten
 - $x_2 = 200$ Mengeneinheiten

Abbildung 13: Fallbeispiel 2 - Lösung der Prüfungsaufgabe mit CoT-Prompt (OpenAI, 2024a)

und CoT- & AmA-Prompts war die theoretische Herleitung korrekt, das Ermitteln der Stellen im Koordinatensystem war mit $x_1 = 75$ und $x_2 = 200$ allerdings fehlerhaft. Beide Promptarten wurden mit 0 Punkten bewertet. Trotz fachlicher Korrektheit zeigt sich hier das erste Mal

a) Der Kurs eines Containerschiffs wird modellhaft durch die Geschwindigkeitsvektoren $\vec{k}$, $\vec{a}$ und $\vec{s}$ beschrieben. Der Zielkurs $\vec{k}$ ergibt sich aus dem angesteuerten Kurs $\vec{s}$ und der sogenannten Abdrift $\vec{a}$. Es gilt: $\vec{k} = \vec{s} + \vec{a}$. Für einen bestimmten Zeitpunkt sind die Geschwindigkeitsvektoren $\vec{k}$ und $\vec{a}$ in der nachstehenden Abbildung dargestellt.

1) Zeichnen Sie in der obigen Abbildung den Vektor $\vec{s}$ als Pfeil ausgehend vom Koordinatenursprung ein. [0/1 P.]

Abbildung 14: Fallbeispiel 3 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2023/24 (BMBWF, 2024b)

deutlich, welche Probleme die KI hat, Bilder zu analysieren und zu interpretieren. Diese Auffälligkeit soll auch bei weiteren Fallbeispielen deutlich werden.

4.2.3. Fallbeispiel 3: Vektoren

In gegebener Teil 2-Aufgabe soll ein Vektor $\vec{s}$ eingezeichnet werden, der die Gleichung in der Angabe erfüllt. Sie weist das Konstruktionsformat auf und zählt zum Inhaltsbereich „Algebra & Geometrie“. Die Lösung ist in Abbildung 14 farblich hervorgehoben. Bei der Lösung ohne Prompt formt ChatGPT die Gleichung aus der Angabe richtig um und man erhält $\vec{s} = \vec{k} - \vec{a}$. Anschließend werden die Komponenten der Vektoren mit $\vec{k} = (10, 15)$ und $\vec{a} = (-5, 0)$ ausgelesen. Betrachtet man die Vektorpfeile, erkennt man schnell, dass die Komponenten nicht korrekt sind. Die x -Koordinate des Vektors $\vec{a}$ stimmt, die anderen Koordinaten scheinen beliebig aus der Beschriftung des Koordinatensystems genommen worden zu sein. Die Berechnung des Vektors $\vec{s}$ unterliegt also einem Folgefehler und führt damit zu einem nicht korrekten Endergebnis. Bei der Ausgabe des CoT- und CoT & AmA-Prompts werden die gegebenen Vektoren ebenfalls falsch ausgelesen, beim AmA-Prompt liefert die KI lediglich eine schriftliche Anleitung zur Vorgangsweise, ohne konkrete Zahlen und Ergebnisse. Zum Teil generiert ChatGPT eigenständig Bilder von Koordinatensystemen, in denen die Vektoren eingezeichnet werden. Bei manchen Prompts mussten die Grafiken allerdings explizit angefordert werden. Die Verwendung der Formulierung „Erstelle mir zu gegebener Situation ein Bild“ führte zu einer kreativen Interpretation eines Koordinatensystems, die mathematisch

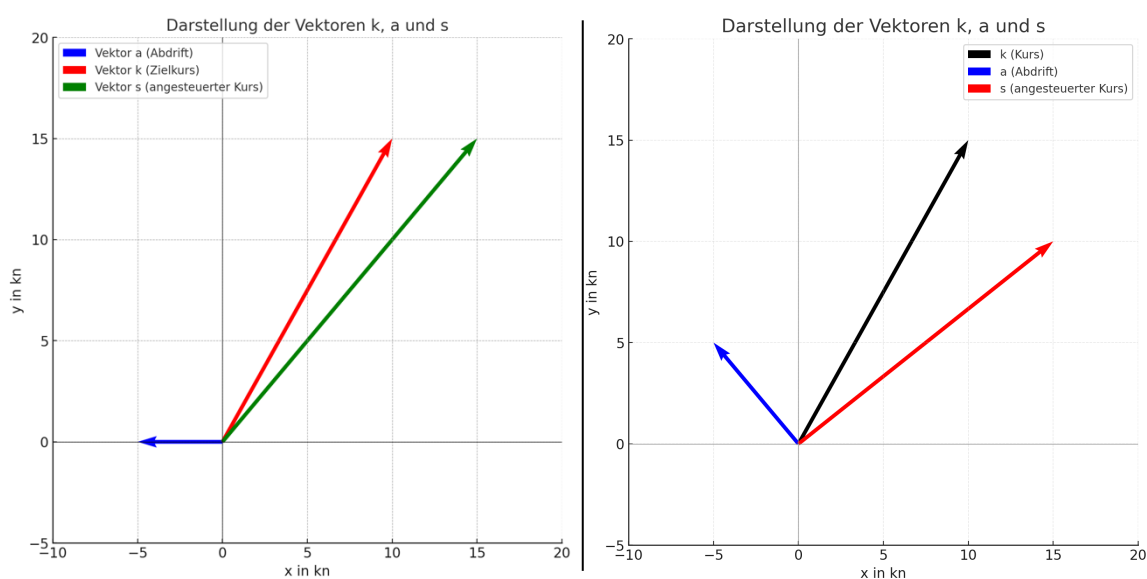


Abbildung 15: Fallbeispiel 3 - Generierte Bilder zur Prüfungsaufgabe ohne Prompt und mit CoT-Prompt (OpenAI, 2024a)

nicht korrekt war. Der Begriff „Koordinatensystem“ war hier erfolgreicher, die Ergebnisse zweier Promptarten sind in Abbildung 15 zu sehen. Es wird deutlich, dass die Vektorpfeile den falschen ausgelesenen Koordinaten entsprechen. Die Achsenbeschriftungen sind korrekt, sowie auch der Koordinatenursprung als Ausgangspunkt. Alle vier Promptarten werden mit 0 Punkten bewertet.

4.2.4. Fallbeispiel 4: Weg, Geschwindigkeit und Beschleunigung

Ein Körper bewegt sich im Zeitintervall $[0; 5]$ geradlinig mit einer konstanten Beschleunigung und kommt zum Zeitpunkt $t = 5$ zum Stillstand.

Für die Beschleunigung gilt: $a(t) = -0,4$

t ... Zeit in s
 $a(t)$... Beschleunigung zum Zeitpunkt t in m/s^2
 $v(t)$... Geschwindigkeit zum Zeitpunkt t in m/s
 $s(t)$... zurückgelegter Weg zum Zeitpunkt t in m

In zwei der unten stehenden Abbildungen wird die Bewegung des Körpers richtig dargestellt.

Aufgabenstellung:

Kreuzen Sie die beiden zutreffenden Abbildungen an. [2 aus 5]

	☒		☒
	☐		☐
	☐		

Abbildung 16: Fallbeispiel 4 - Prüfungsaufgabe der AHS-Matura des Haupttermins 2023/24 (BMBWF, 2024b)

Bei der gegebenen Teil 1-Aufgabe sollen zum beschriebenen Sachverhalt die passenden Funktionsgraphen gefunden werden, das Format ist Multiple-Choice. Es handelt sich um eine geradlinig und konstant beschleunigte Bewegung, die zum Zeitpunkt $t = 5$ zum Stillstand kommt. Folgedessen muss die Geschwindigkeitsfunktion v eine lineare Funktion mit dem Beschleunigungswert als Steigung sein und die Beschleunigungsfunktion a eine konstante Funktion sein. Die Aufgabe ist in Abbildung 16 dargestellt und die Lösung farblich markiert.

Bei der Ausgabe ohne Prompt analysiert die KI zuerst die Angabe und dann schrittweise die einzelnen Funktionsgraphen, was in Abbildung 17 zu sehen ist. Sowohl die Beschreibungen der Graphen als auch die damit verbundene Mathematik wird präzise wiedergegeben. Hier zeigt das System sein doch vorhandenes „Verständnis“ von Grafiken. Bei den anderen Promptarten sind die Ausgaben allerdings teils fehlerhaft. Beim CoT-Prompt stellt ChatGPT zunächst eine Beschleunigungsfunktion auf und ermittelt über Integration eine Geschwindigkeitsfunktion v mit $v(t) = -0,4t + 2$ und eine Wegfunktion s mit $s(t) = -0,2t^2 + 2t$. Der zweite Funktionsgraph in der linken Spalte wird als korrekter Graph zur Wegfunktion identifiziert, was offensichtlich falsch ist. Die Wegzuwächse müssten bei einer negativen Beschleunigung kleiner und nicht größer werden. Weiters ist der negative Koeffizient von t^2 ein Hinweis auf eine nach unten geöffnete Parabel, die hier mit keinem der Funktionsgraphen gezeigt wird. Der andere

Analyse der Abbildungen:

- 1. Erste Abbildung (Geschwindigkeit $v(t)$):**
 - Die Geschwindigkeit nimmt linear ab und erreicht bei $t = 5\text{ s}$ den Wert 0. Dies entspricht der Bewegung mit konstanter negativer Beschleunigung.
 - Richtig.**
- 2. Zweite Abbildung (Beschleunigung $a(t)$):**
 - Die Beschleunigung ist konstant bei $-0,4\text{ m/s}^2$. Diese Darstellung zeigt die Beschleunigung als konstant negativ.
 - Richtig.**
- 3. Dritte Abbildung (Weg $s(t)$):**
 - Der Weg nimmt quadratisch zu (parabolische Kurve), was nicht mit der abnehmenden Geschwindigkeit und der konstant negativen Beschleunigung übereinstimmt. Der Weg sollte linear abnehmend oder weniger stark ansteigend sein.
 - Falsch.**
- 4. Vierte Abbildung (Geschwindigkeit $v(t)$):**
 - Die Geschwindigkeit bleibt konstant, was der gegebenen Beschleunigung nicht entspricht.
 - Falsch.**
- 5. Fünfte Abbildung (Beschleunigung $a(t)$):**
 - Die Beschleunigung ist positiv, was nicht mit den Angaben übereinstimmt.
 - Falsch.**

Abbildung 17: Fallbeispiel 4 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a)

korrekte Funktionsgraph wird richtig erkannt. Beim AmA-Prompt werden Erwartungen für die Beschleunigungs-, Geschwindigkeits- und Wegfunktion formuliert und ohne weitere Argumentationen die zwei Funktionsgraphen als Lösung deklariert. Der Geschwindigkeitsgraph wird auch hier wieder korrekt gewählt. Beim CoT- & AmA-Prompt wird aus dem ersten Graphen in der rechten Spalte eine konstante Beschleunigung von 0 herausgelesen, was zur Nichtbeantwortung der Aufgabe führt. Zusammengefasst wird nur die Eingabe ohne Prompt mit einem Punkt bewertet, alle anderen erhalten null Punkte.

4.2.5. Fallbeispiel 5: Einheitskreis

In vorliegender Teil 1-Aufgabe in Abbildung 18 soll ein Punkt entsprechend den Bedingungen am Einheitskreis eingezeichnet werden, sie stammt aus dem Inhaltsbereich „Algebra & Geometrie“. Die korrekte Lösung ist farblich gekennzeichnet. Gegeben ist ein Winkel α im Intervall $[0^\circ; 360^\circ]$ mit $\sin(\alpha) = -0,5$ und $\cos(\alpha) < 0$. Der AmA-Prompt führt zu einer Ausgabe, bei der mithilfe der Sinus-Bedingung der mögliche Bereich des Punktes zunächst auf den dritten und vierten Quadranten eingeschränkt wird. Die Cosinus-Bedingung schränkt die Situation nochmals enger auf den dritten Quadranten ein. Dann wird mithilfe der

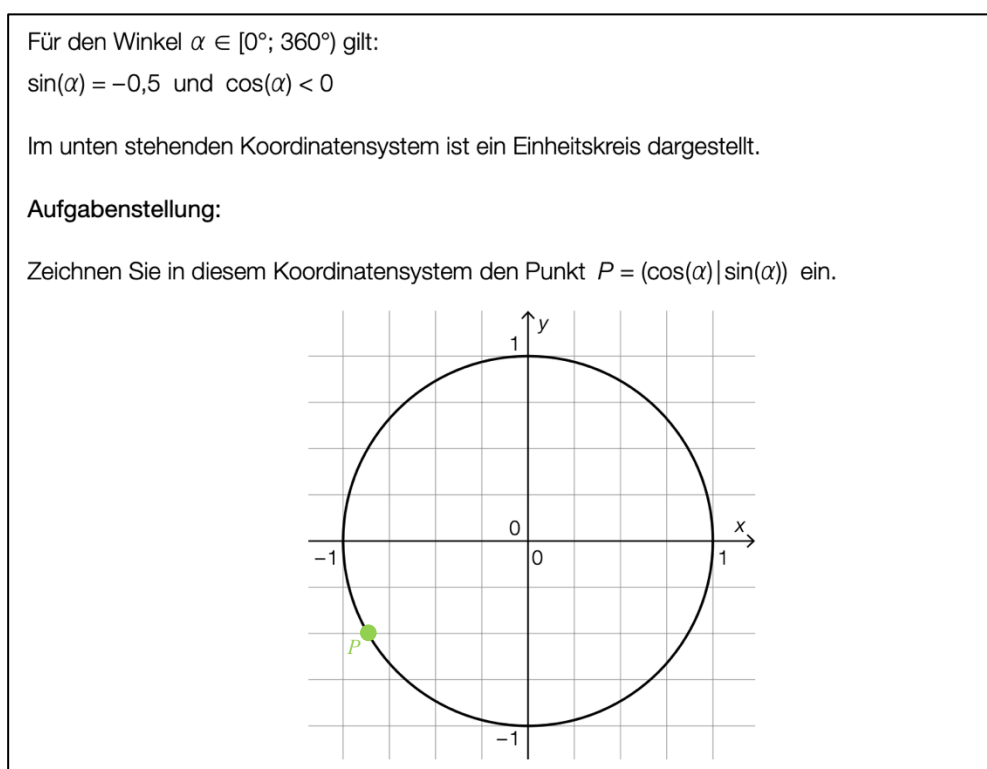


Abbildung 18: Fallbeispiel 5 - Prüfungsaufgabe der AHS-Matura des Haupttermins 2023/24 (BMBWF, 2024b)

Beziehung $\sin^2 \alpha + \cos^2 \alpha = 1$ und dem Einsetzen von α der Cosinus-Wert berechnet. Es ergibt sich der Punkt $P = (-0,866; -0,5)$ und das Koordinatensystem in Abbildung 19. Beim CoT-Prompt ist die Ausgabe ähnlich, es wird jedoch noch mehr auf die schrittweise Vorgangsweise eingegangen und auch, wie der Punkt im Koordinatensystem eingezeichnet werden soll wird erklärt. Bei allen Promptarten wird der Punkt richtig berechnet und es werden auch jeweils akkurate Koordinatensysteme generiert. Im Kontrast zu Fallbeispiel 4 geht der Bildgenerierung hier eine Berechnung voraus, deren Ergebnis dann im Bild verwertet wird. Bei genanntem anderen Beispiel musste zuerst aus der gegebenen Grafik ausgelesen und dann wieder eingezeichnet werden, was die KI vor Schwierigkeiten stellte. Alle Promptarten wurden bei diesem Beispiel mit einem Punkt bewertet. Im Vergleich zu vorherigen Szenarien, liefert das System unabhängig vom Prompt-Typ die richtige Lösung.

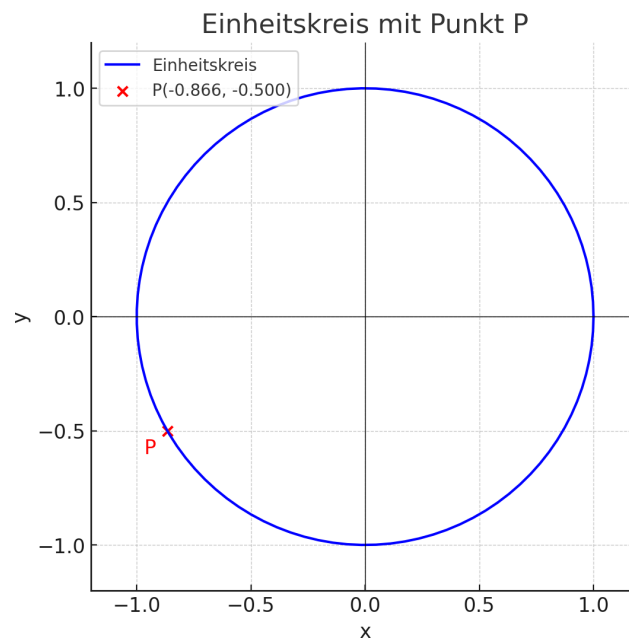


Abbildung 19: Fallbeispiel 5 - Lösung der Prüfungsaufgabe mit CoT- & AmA-Prompt (OpenAI, 2024a)

4.2.6. Fallbeispiel 6: Ereignisse eines Zufallsversuchs

Eine Münze hat eine Kopfseite K und eine Zahlseite Z .

Diese Münze wird 3-mal geworfen. Dabei ist zum Beispiel ZKK ein möglicher Ausgang dieses Zufallsversuchs.

Dabei bedeutet der 1. Buchstabe das Ergebnis des 1. Wurfes, der 2. Buchstabe das Ergebnis des 2. Wurfes und der 3. Buchstabe das Ergebnis des 3. Wurfes.

Mit E wird das Ereignis bezeichnet, dass der 2. Wurf das Ergebnis Z hat.

Aufgabenstellung:

Geben Sie das Ereignis E als Teilmenge des zugehörigen Grundraums dieses Zufallsversuchs an.

$E = \{ \text{KZK, KZZ, ZZK, ZZZ} \}$

Abbildung 20: Fallbeispiel 6 - Prüfungsaufgabe der AHS-Matura des Wintertermins 2022/23 (BMBWF, 2024b)

Die vorliegende Aufgabe des ersten Prüfungsteils kommt aus dem Bereich „Wahrscheinlichkeit & Statistik“ und ist im halboffenen Format in Abbildung 20 gegeben. Es soll die Ereignismenge eines Zufallsversuchs mit einer klassischen Münze ermittelt werden. Der Grundraum dieses Zufallsversuchs umfasst alle möglichen Kombinationen der drei Würfe mit den Ausgängen Kopf (K) und Zahl (Z). Da E das Ereignis bezeichnet, dass der zweite Wurf das Ergebnis Z hat, umfasst E alle Kombinationen, bei denen der zweite Buchstabe Z ist. Wie in Abbildung 21 zu sehen, wird dies bei der Eingabe ohne Prompt auch genauso von ChatGPT erläutert und

🌀 Das Ereignis E beschreibt, dass der **2. Wurf** der Münze ein Z ergibt. Um das Ereignis als Teilmenge des Grundraums zu formulieren, müssen wir alle möglichen Ergebnisse betrachten, bei denen der zweite Buchstabe Z ist.

Der Grundraum Ω umfasst alle möglichen Kombinationen der drei Würfe, also:

$\Omega = \{KKK, KKZ, KZK, KZZ, ZKK, ZKZ, ZZK, ZZZ\}$

Das Ereignis E , dass der zweite Wurf Z ist, enthält die Ergebnisse, bei denen der zweite Buchstabe Z ist:

$E = \{KZK, KZZ, ZZK, ZZZ\}$

Antwort:

$E = \{KZK, KZZ, ZZK, ZZZ\}$

Falls du weitere Schritte benötigst, lass es mich wissen!

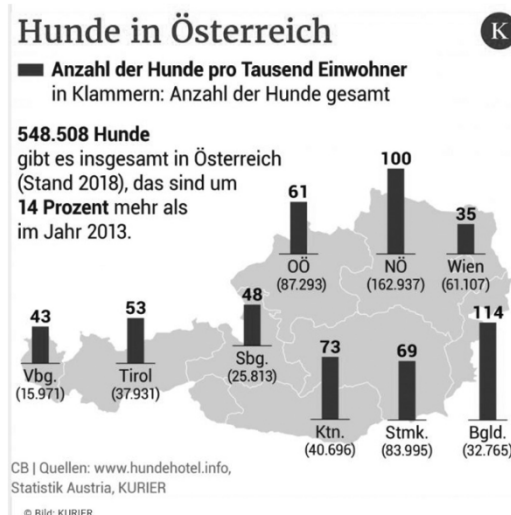
Abbildung 21: Fallbeispiel 6 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a)

korrekt gelöst. Ähnliche Varianten der Lösung treten beim CoT- und beim CoT- & AmA-Prompt auf und erhalten damit jeweils einen Punkt. Einzig der AmA-Prompt liefert die Ausgabe $E = \{KZK, KZZ, ZZK, ZZZ, KZZ, ZZZ\}$, wo die Ereignisse KZZ und ZZZ doppelt vorkommen. Der Grund hierfür ist unklar, bei einer vergleichsweise einfachen Aufgabe wie dieser, ist diese Ungereimtheit verwunderlich. Es wird deutlich, dass eine eigenständige Kontrolle der Ausgaben unabdingbar ist und man sich nie zur Gänze auf die KI verlassen sollte.

4.2.7. Fallbeispiel 7: Median

In vorliegender Teil 2-Aufgabe in Abbildung 22 aus dem Inhaltsbereich „Wahrscheinlichkeit & Statistik“ soll anhand eines Diagramms das Bundesland ermittelt werden, welches der gegebenen statistischen Aussage entspricht. Die korrekte Lösung ist farblich gekennzeichnet. Die Grafik zeigt sowohl die Gesamtanzahl der Hunde, als auch die Anzahl der Hunde pro Tausend Einwohner*innen im jeweiligen Bundesland. Gefordert ist zunächst der Median der Anzahl der Hunde pro Tausend Einwohner*innen, der durch Sortierung der entsprechenden Werte ermittelt werden kann. Dieser Wert ist aufgrund der ungeraden Anzahl der Werte auch ein tatsächlich vorkommender Wert in einem der Bundesländer. Dies wird auch in der

- a) Die nachstehende Abbildung zeigt die Verteilung der Anzahl der Hunde in Österreich im Jahr 2018 nach Bundesländern.



Bildquelle: <https://kurier.at/chronik/oesterreich/plus-14-prozent-hunde-liegen-voll-im-trend/400573877> [16.03.2021].

Der Median der Anzahl der Hunde pro tausend Einwohner/innen in den 9 Bundesländern ist gleich der Anzahl der Hunde pro tausend Einwohner/innen in einem bestimmten Bundesland.

- 1) Ermitteln Sie dieses Bundesland. **Oberösterreich**

[0/1 P.]

Abbildung 22: Fallbeispiel 7 - Prüfungsaufgabe der AHS-Matura des Wintertermins 2022/23 (BMBWF, 2024b)

Aufgabenstellung nochmals verdeutlicht, die allerdings sprachlich anspruchsvoll formuliert ist. Trotz der sprachlichen Komplexität muss lediglich der Median ermittelt werden. In der Ausgabe ohne Prompt liest die KI das Diagramm nur zum Teil richtig aus, wie in Abbildung 23 zu sehen ist. Die Werte für Vorarlberg und die Steiermark sind mit 53 statt 43 bzw. 61 statt 69 nicht korrekt. Auffällig ist, dass eine Ziffer in beiden Fällen richtig ausgelesen wurde. Die Fehler führen jedoch zu keinem richtigen Ergebnis, die KI deklariert sowohl Oberösterreich als auch die Steiermark als Lösung, da beide mit dem Wert 61 als Median gewertet werden. Die Ausgabe des AmA-Prompts kommt zum gleichen Ergebnis, während der AmA- & CoT-Prompt nur acht der neun Werte ausliest, aber dennoch den fünften Wert der sortierten Liste (Kärnten) als Antwort festlegt. Grundsätzlich werden die Werte gut erkannt, obwohl die Bundesländer zum Teil mit ihren Abkürzungen angeführt sind. Im vorliegenden Fall wurde aber zumindest immer einer der Werte falsch ausgelesen. Alle vier Prompt-Typen wurden mit 0 Punkten bewertet.

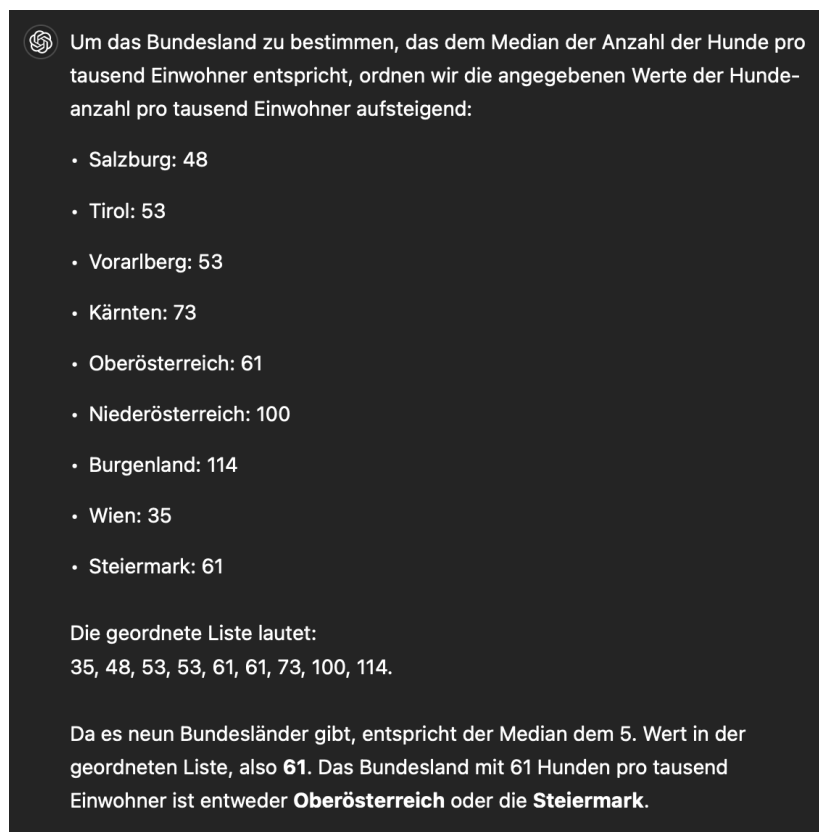


Abbildung 23: Fallbeispiel 7 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a):

4.2.8. Fallbeispiel 8: Integral und Flächeninhalt

Die in Abbildung 24 vorliegende Teil 1-Aufgabe stammt aus dem Inhaltsbereich der Analysis, für deren Lösung die Verknüpfung der mathematischen Objekte Integral und Flächeninhalt

gegeben sein muss. Flächeninhalte sind stets positive Werte, während Integrale, deren entsprechende Flächen vollständig unterhalb der x -Achse liegen, negativ sind. Auch die Größen der grau markierten Flächeninhalte geben Ausschluss über einen positiven bzw. negativen Integralwert, wenn beispielsweise über die Grenzen 0 und c integriert wird. Mithilfe der angegebenen Flächeninhalte kann die Aufgabe gelöst werden. Die Ausgabe von ChatGPT beim AmA-Prompt ist in Abbildung 25 zu sehen. Der erste Ausdruck wird korrekt ausgelesen und die Flächen werden addiert, wobei der Flächeninhalt A_2 negativ in die Summe einfließt. Die KI erhält mit -31 das richtige Ergebnis und die Zuordnung wird korrekt vorgenommen. Beim zweiten Ausdruck werden die Integralgrenzen falsch ausgelesen, statt a und b werden b und c als Stellen verwendet. Dies führt zu einer falschen Berechnung und folglich auch Zuordnung. Beim dritten Ausdruck schreibt Chat GPT „Die Summe von positiver Fläche A_1 , negativer Fläche A_2 und A_3 “. Die Vorzeichen der einzelnen Werte werden also richtig erkannt, das Minus im Ausdruck

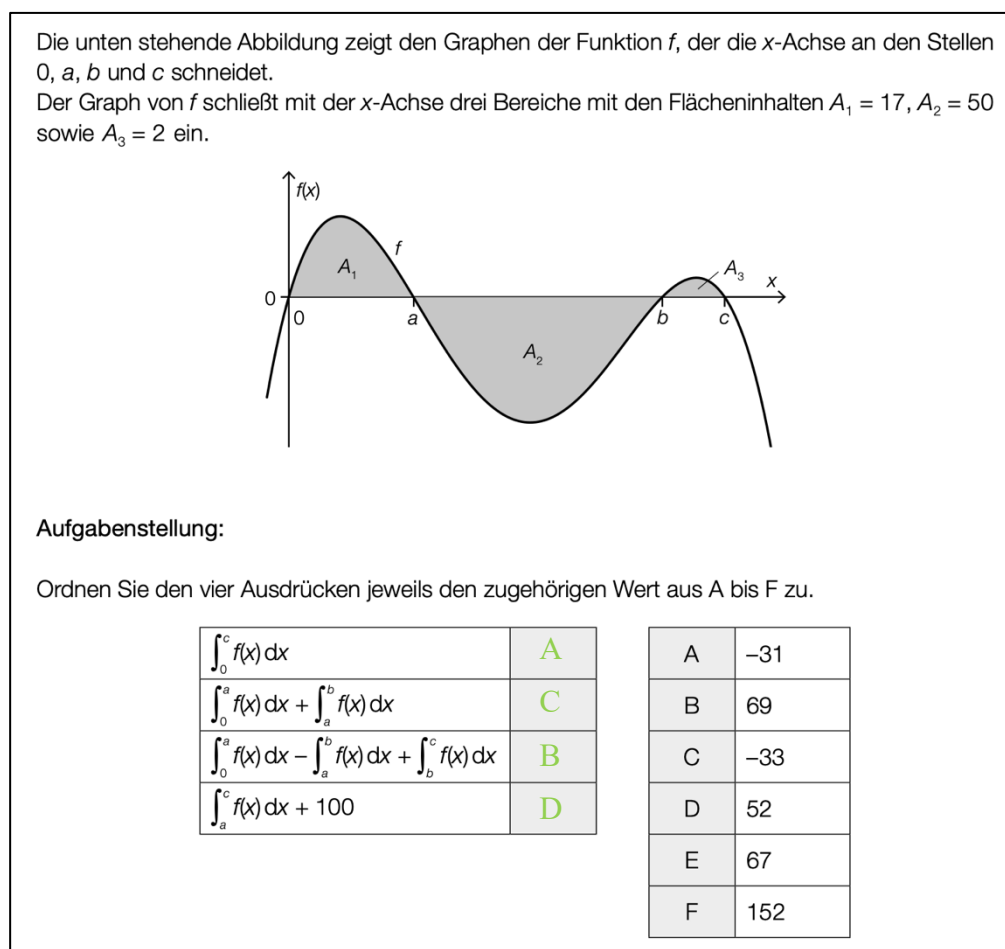


Abbildung 24: Fallbeispiel 8 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2022/23 (BMBWF, 2024b)

Die vollständige Zuordnung der gegebenen Werte ist wie folgt:

- $\int_0^c f(x)dx$: -31
 - Dies ist die Gesamtfläche von 0 bis c , wie oben berechnet.
 - Zuordnung: A = -31.**
- $\int_0^a f(x)dx + \int_b^c f(x)dx$: 69
 - Die Teilflächen von 0 bis a und von b bis c sind positiv addiert.
 - Zuordnung: B = 69.**
- $\int_0^a f(x)dx - \int_a^b f(x)dx + \int_b^c f(x)dx$: -33
 - Die Summe von positiver Fläche A_1 , negativer Fläche $-A_2$, und A_3 .
 - Zuordnung: C = -33.**
- $\int_a^c f(x)dx + 100$: 152
 - Gesamtfläche zwischen a und c , plus 100.
 - Zuordnung: F = 152.**

Abbildung 25: Fallbeispiel 8 - Lösung der Prüfungsaufgabe mit AmA-Prompt (OpenAI, 2024a)

wird allerdings ignoriert, sowie das Integral für A_3 vollständig weggelassen. Das führt zu einer falschen Berechnung und Zuordnung. Beim vierten Ausdruck wird das Integral fälschlich als positiv aufgefasst. Die AmA-Prompt-Ausgabe konnte daher nur einen von vier Werten richtig zuordnen. Auch bei den anderen Promptarten werden die Ausdrücke großteils richtig erkannt. Die Ausgabe ohne Prompt ordnet drei von vier Werte richtig zu, eine Auswahlmöglichkeit wird zweimal verwendet. Allgemein sind die Berechnungen detaillierter. Der erste Ausdruck

a) Der Verlauf der Schwimmstrecke eines bestimmten Triathlons ist in der nachstehenden Abbildung modellhaft dargestellt. Der Schwimmbewerb startet im Punkt S und endet im Punkt Z, dazwischen müssen die Kontrollpunkte B_1 , B_2 , B_3 , S, B_1 und B_2 in genau dieser Reihenfolge erreicht werden.

Die Entfernung vom Punkt $S = (600|0)$ zum Punkt B_1 beträgt 700 m.

1) Berechnen Sie die y-Koordinate von B_1 .

$B_1 = (0 | \boxed{360,5\dots})$

[0/1 P.]

Abbildung 26: Fallbeispiel 9 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2022/23 (BMBWF, 2024b)

wird wie folgt erläutert: $\int_0^c f(x)dx = A_1 - A_2 + A_3 = 17 - 50 + 2 = -31$. Der CoT-Prompt generiert drei korrekte Zuordnungen. Beim vierten Ausdruck erhält die KI einen Wert, der nicht zur Auswahl steht. Zusammengefasst wird der AmA-Prompt mit 0 Punkten, die übrigen Typen aufgrund von zwei bis drei korrekten Zuordnungen mit 0,5 Punkten bewertet.

4.2.9. Fallbeispiel 9: Rechtwinkeliges Dreieck

Die in Abbildung 26 gegebene Teil 2-Aufgabe ist im Inhaltsbereich „Algebra & Geometrie“ angesiedelt, die Lösung ist im Bild farblich markiert. Gegeben ist ein Koordinatensystem mit eingezeichneten Punkten, die die Schwimmstrecke eines Triathlon-Bewerbs abstecken. Die Punkte S und B_1 befinden sich auf der x - bzw. y -Achse. Weiters sind die Koordinaten von S sowie die Entfernung von S zu B_1 gegeben. Wird das rechtwinkelige Dreieck erkannt, erfordert die Aufgabe lediglich die Berechnung einer Seitenlänge mit dem Satz des Pythagoras. Trotz der Angabe in Form eines Koordinatensystems stellt dies für die KI kein Problem dar. Die Ausgabe

Um die y -Koordinate des Punkts B_1 zu berechnen, benötigen wir mehr Informationen zu der Strecke und den gegebenen Koordinaten. Aus der Abbildung und den gegebenen Werten können wir den Verlauf der Schwimmstrecke und die Positionen der Kontrollpunkte nachzeichnen.

Die Entfernung vom Punkt $S = (600|0)$ zum Punkt B_1 beträgt 700 Meter. Dies ist die Entfernung entlang der Strecke S zu B_1 , die sich als Hypotenuse eines rechtwinkligen Dreiecks mit den Koordinaten von S und B_1 darstellen lässt.

Vorgehensweise:

1. Die Entfernung von $S = (600|0)$ zum Punkt B_1 beträgt 700 m.
2. Die Koordinate von B_1 auf der x -Achse ist gegeben als $x = 0$.
3. Die Entfernung in der y -Richtung ist also die Höhe des Dreiecks, die wir berechnen müssen.

Verwenden wir den Satz des Pythagoras für das Dreieck $\triangle SB_1O$:

$$\text{Entfernung}^2 = \text{Abstand in } x^2 + \text{Abstand in } y^2$$

Da der x -Abstand von S nach B_1 600 m beträgt, können wir die Formel umstellen, um die y -Koordinate zu berechnen:

$$700^2 = 600^2 + y^2$$

Nun berechnen wir y :

$$490000 = 360000 + y^2$$

$$y^2 = 490000 - 360000 = 130000$$

$$y = \sqrt{130000} \approx 360,6$$

Ergebnis:

Die y -Koordinate von B_1 beträgt etwa **360,6 m**.

Abbildung 27: Fallbeispiel 9 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a)

ohne Prompt legt die Strecke von S zu B_1 sogleich als Hypotenuse fest. Weiters wird die y -Koordinate von B_1 als Höhe des Dreiecks und zu berechnende Größe erkannt. Der Abstand entlang der x -Achse von B_1 zu S wird korrekt als 600 angenommen. Nach dem Einsetzen und Umformen des Satzes des Pythagoras wird das richtige Endergebnis ausgegeben. Der Lösungsweg der Eingabe ohne Prompt ist in Abbildung 27 nachzuverfolgen. Auch die anderen Promptarten meistern die Aufgabe und erkennen in der Grafik immer das rechtwinkelige Dreieck. Alle vier Typen werden daher mit einem Punkt bewertet. Beim CoT- & AmA-Prompt werden explizit die Koordinatenachsen und ihre senkrechte Lage zueinander erwähnt.

4.2.10. Fallbeispiel 10: Histogramm

Die in Abbildung 28 vorliegende Aufgabe aus dem Inhaltsbereich „Wahrscheinlichkeit & Statistik“ gehört zum zweiten Prüfungsteil und behandelt das Erstellen von Histogrammen. Die Lösung der Aufgabe ist farblich eingezeichnet. In einer Tabelle sind die absoluten Häufigkeiten der prozentuellen Kalorienabweichungen von Fitnessuhren dargestellt. Mit diesen Informationen soll im vorgegebenen Koordinatensystem ein entsprechendes Histogramm

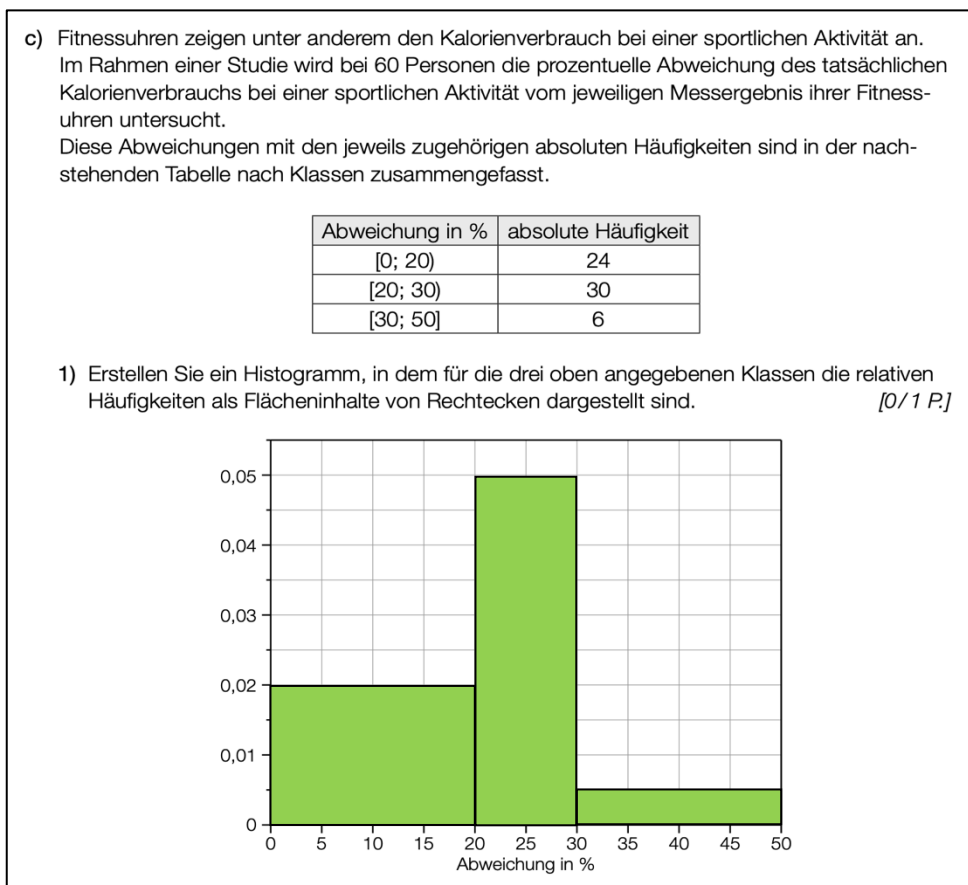


Abbildung 28: Fallbeispiel 10 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2022/23 (BMBWF, 2024b)

gezeichnet werden. Dabei sollen die relativen Häufigkeiten den Flächeninhalten der Rechtecke der drei angegebenen Klassen entsprechen. Die Ausgabe ohne Prompt liefert die berechneten relativen Häufigkeiten für die einzelnen Klassen und erklärt, dass die Höhen der Rechtecke entsprechend angepasst wurden. Weiters liefert die KI ein generiertes Bild, das die Anforderungen nur mäßig erfüllt. Es fehlt die Skalierung der x -Achse, weshalb nicht evaluiert werden kann, ob die Rechtecksbreiten korrekt sind. Weiters befindet sich zwischen y -Achse und erstem Balken ein Abstand, der den Flächeninhalt des ersten Rechtecks unerwünscht reduziert ist. Beim CoT- & AmA-Prompt werden alle Rechtecke gleich breit gezeichnet. Die beiden genannten Promptarten werden daher mit 0 Punkten bewertet. Die Ausgabe des CoT-Prompts ist in Abbildung 29 dargestellt. Es ist erkenntlich, dass sowohl die x - als auch die y -Achse korrekt beschriftet und die Rechtecke richtig eingezeichnet sind. Auch der AmA-Prompt liefert ein ähnliches Bild, beide Prompts wurden daher mit einem Punkt bewertet.

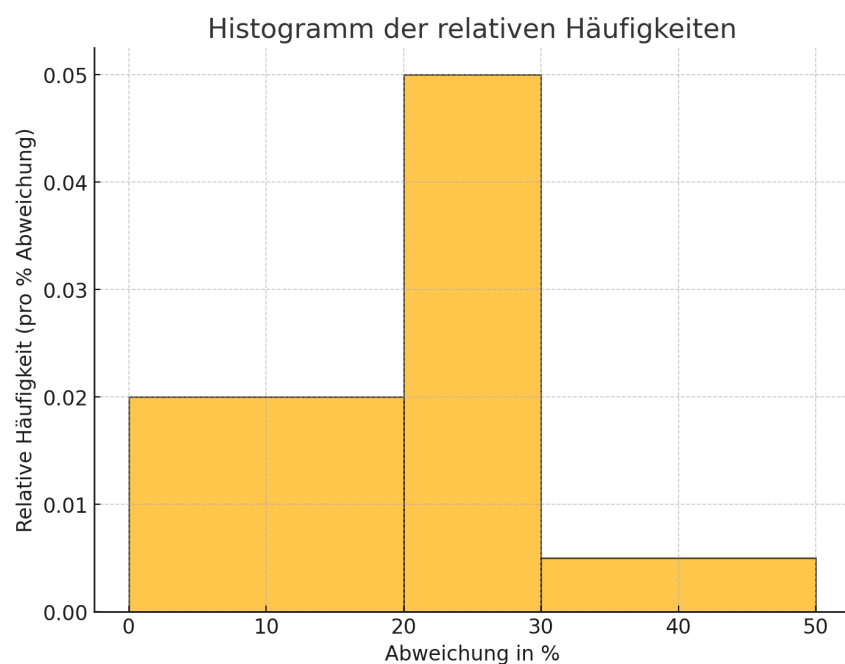


Abbildung 29: Fallbeispiel 10 - Lösung der Prüfungsaufgabe mit CoT-Prompt (OpenAI, 2024a)

5. Diskussion

5.1. Zusammenfassung der Erkenntnisse

Auf Basis der quantitativen Auswertung und der erreichten Punktezahlen und Notengrade kann gesagt werden, dass bei Verwendung von unterschiedlichen Promptarten lediglich marginale Unterschiede bestehen. Überraschenderweise waren über alle Aufgaben hinweg betrachtet, die Eingaben ohne zusätzlichen Prompt die erfolgsversprechendsten. Bei drei Prüfungsterminen konnte diese Art der Eingabe die meisten Punkte erzielen. Bei zwei weiteren Terminen landet die Eingabe ohne Prompt auf Platz 2. Betrachtet man die Lösungshäufigkeiten der unterschiedlichen Antwortformate, so stechen das offene, das halboffene und das Lückentext-Format als am einfachsten zu lösen hervor. Beim Zuordnungs- und Multiple-Choice-Format wurden mit 56 % bzw. 62 % Lösungsquote geringere Leistungen festgestellt. Die Trennung der verfügbaren Antworten mit Linien beim Multiple-Choice-Format wird von der KI erkannt, auch die Aussagen werden richtig abgelesen. Oftmals führt allerdings die Anzahl der richtigen Antworten zu Problemen, die Angaben „2 aus 5“ oder „1 aus 6“ werden ignoriert oder der Bezug zu den verfügbaren Antwortmöglichkeiten kann nicht hergestellt werden. Aufgrund von Fehlern in der eigenen Argumentation wählt ChatGPT folglich zu viele oder zu wenig korrekte Antworten aus. Diese Diskrepanz sollte dem System auffallen und nach einem etwaigen falschen Schluss, seine Antworten anhand dieser Infos revidieren und nochmal kontrollieren. Auch das Ausschlussprinzip, das von vielen Schüler*innen bei Multiple-Choice-Aufgaben angewendet wird, übernimmt die KI nicht. Jede Aussage wird stattdessen separat evaluiert. Auch beim Lückentext-Format tritt ein ähnliches Phänomen zu Tage. Antwortmöglichkeiten, die für die zweite Lücke vorgesehen sind, werden beispielsweise auch für die erste Lücke verwendet und umgekehrt. ChatGPT erkennt nicht, welche Antworten jeweils für Lücke 1 und 2 gedacht sind. Ein Fehler, der Schüler*innen nach dem Verstehen der Funktionsweise des Formats womöglich nur sehr selten passiert.

Das Konstruktionsformat weist die mit Abstand geringste Lösungshäufigkeit (33 %) auf und ist somit die Schwachstelle der KI. Bereits das Auslesen und Analysieren von Graphen gestaltet sich schwierig. Steigungen und Krümmungen werden an bestimmten Stellen als positiv oder negativ bewertet, obwohl oft genau das Gegenteil der Fall ist. Auch Seitenlängen oder Längen von Vektoren werden teils falsch bewertet. Diese Tätigkeiten stellen jedoch meist die Vorarbeit

bei Konstruktionsformaten dar, weiters muss dann oft noch etwas zur analysierten Grafik ergänzt werden. ChatGPT ist es nicht möglich hochgeladene Bilder zu manipulieren oder Einzeichnungen darin vorzunehmen. Daher muss das Programm ein eigenständiges Bild auf Basis des hochgeladenen Bildes reproduzieren. Die reine Reproduktion ist meist nicht korrekt und folglich auch nicht die ergänzte Konstruktion. Weiters sind keine wesentliche Leistungssteigerungen mit bestimmten Promptarten bei einem bestimmten Antwortformat festzustellen.

Die Erfolgsquoten variieren auch je nach Inhaltsbereich. Das beste durchschnittliche Ergebnis erzielt die KI im Bereich „Wahrscheinlichkeit & Statistik“ (75 %), gefolgt von „Funktionalen Abhängigkeiten“ (74 %). In den Bereichen „Analysis“ (67 %) sowie „Algebra & Geometrie“ (63 %) fällt die Leistung ein wenig ab. Dies deutet darauf hin, dass die KI mit geometrischen Problemstellungen größere Schwierigkeiten hat, was zum Teil auch in den Fallbeispielen dargestellt wurde. Vor allem beim AmA-Prompt waren die Differenzen zwischen den zwei besten und den zwei schlechtesten Inhaltsbereichen am größten. Hier verzeichnete die KI in den Inhaltsbereichen „Algebra & Geometrie“ und „Analysis“ lediglich 60 % bzw. 61 % Erfolgsquote, in „Funktionalen Abhängigkeiten“ und „Wahrscheinlichkeit & Statistik“ 73 % bzw. 76 %. Die Analyse der Prüfungsteile zeigt eine insgesamt gleichmäßige Leistung. Im ersten Prüfungsteil liegt die durchschnittliche Erfolgsquote bei 70 %, während im zweiten Teil ein leichter Rückgang auf 69 % zu beobachten ist. Dieser Unterschied ist jedoch minimal und deutet darauf hin, dass die Leistungsfähigkeit der KI nicht stark von der Struktur oder dem Schwierigkeitsgrad der Prüfungsabschnitte abhängt.

5.2. Interpretation

Wie bereits im vorangegangenen Kapitel erwähnt, kann ChatGPT Aufgaben im offenen bzw. halboffenen Format am besten lösen. Dies hat womöglich damit zu tun, dass diese Formate noch am ehesten einem klassischen Textprompt ähneln, der der KI im Eingabefeld geliefert wird und sie damit auch am besten hantieren kann. Die Angaben erfüllen bereits die in Kapitel 2.10.2. angeführten Eigenschaften eines guten Prompts wie Klarheit, Prägnanz und Kontext. Weiters führt bei offenen Aufgabenstellungen in der Form von „Berechne...“ oder „Löse...“ eine strikt formale Berechnung zur Lösung, die Computersysteme gut durchführen können.

Auch bei den halboffenen Aufgabenstellungen wird vor der Lösungszeile, auf der das Ergebnis festgehalten werden muss, eine klare Aufgabenstellung geliefert. Es ist damit bereits offensichtlich, was zu berechnen ist und die Lösungszeile in den meisten Fällen nicht unbedingt notwendig. Aus Sicht der KI sind das offene und halboffene Format damit als gleichwertig zu betrachten. Auch in vielen Schulbüchern sind prozedurale Aufgaben im offenen Format gängig. Wissen Schüler*innen über die Bildeingabe Bescheid, so ist es ein Leichtes, vorgegebene Aufgaben von ChatGPT lösen zu lassen. Die Aufgabe wird mit dem Smartphone fotografiert, in der Smartphone-App von ChatGPT hochgeladen und der Lösungsweg ausgegeben. Besteht die Mathematik-Hausübung lediglich aus solchen Aufgaben, ist diese für technikaffine Schüler*innen eine schnelle Angelegenheit mit wenig Eigenleistung. Es besteht die Gefahr, dass die im Unterricht erlernten Fertigkeiten nicht wiederholt werden. Bei offenen Interpretationsaufgaben existiert im diesem Sinne keine stringente Berechnung, dennoch kann ChatGPT aufgrund seiner guten Argumentationsfähigkeiten auch diese Aufgaben meist bravourös bewältigen. Auch diese Art von Aufgabe läuft Gefahr für Schüler*innen von der KI gelöst zu werden.

Bei durchgeführter Untersuchung war ChatGPT insbesondere beim Konstruktionsformat nur mäßig erfolgreich. Hierbei war nicht die Mathematik per se das Problem, da die Argumentationen im Bereich der Geometrie meist als korrekt identifiziert werden konnten. Schwierigkeiten treten bei der Analyse und Erfassung der Angabe in Form von Grafiken aus der Statistik, Koordinatensystemen, Funktionsgraphen und Vektoren auf. Die „Übersetzung“ in Zahlen und Variablen, mit denen ChatGPT dann wieder besser hantieren kann, läuft oftmals schief oder ist mehr oder weniger vom Zufall abhängig. Wird dieser Schritt gemeistert, ist der weitere Weg zum formalen Ergebnis selten ein Problem. Vor allem in Fallbeispiel 7, wo Zahlenwerte aus einer Statistik ausgelesen werden sollen, ist diese Problematik wunderbar zu erkennen. Es werden beispielsweise nicht alle Werte ausgelesen, Werte vertauscht oder schlichtweg falsche Werte ausgelesen. Auch in durchgeführten Untersuchungen von Johnson et al. (2023) und Wang et al. (2024), in denen die visuellen Analysefähigkeiten von ChatGPT mithilfe von diversen Darstellungen von Funktionen und Grafiken, klassischen Bildern und biomedizinischen Diagrammen untersucht wurden, konnten Limitationen festgestellt werden. Oberflächliche Merkmale werden gut erkannt, die KI hat jedoch Schwierigkeiten mit detaillierten Informationen, der farblichen Wahrnehmung und bei quantitativen Analysen. Abgesehen von der Analyse, ist auch die Ausgabe grafischer Inhalte fehleranfällig, wenn auch

nicht so stark. Eine Bildgenerierung zum Sachverhalt ist bei geringer Komplexität machbar, wie in Fallbeispiel 5 durch die Erstellung des Einheitskreises gezeigt wird. Auch das Zeichnen von Vektoren in Fallbeispiel 3 ist grundsätzlich machbar, gleiches gilt auch für Funktionen. Lehrpersonen könnten sich die Bildgenerierung zu Nutze machen und beispielsweise auf schnellem Wege Funktionsgraphen mit bestimmten Eigenschaften erstellen, die wiederum für Multiple-Choice- oder Zuordnungsaufgaben verwendet werden können. Aufgrund der geringen Lösungshäufigkeiten von Aufgaben im Konstruktionsformat, besteht hier eine geringere Gefahr, dass Schüler*innen den Lösungsprozess vollständig an ChatGPT auslagern und das Gelernte selbstständig geübt wird. Für Lehrpersonen, die dies stärker forcieren möchten, ist das Format der Schlüssel, solange sich die visuellen Analysefähigkeiten auf dem aktuellen Niveau befinden. Eine Aufgabe im Konstruktionsformat ist allerdings nicht in jedem inhaltlichen Themengebiet möglich, weshalb hier das Zuordnungsformat Abhilfe schaffen könnte. Dies war in durchgeführter Untersuchung das Format mit der zweitgeringsten Lösungshäufigkeit. ChatGPT hat hier aufgrund der grafischen Anordnung der Auswahlmöglichkeiten ähnliche Probleme bei der visuellen Analyse wie beim Konstruktionsformat. Das Format ist vielfältig einsetzbar und für Inhaltsbereiche wie Analysis, Wahrscheinlichkeit, Statistik und Funktionen geeignet.

Auch bei umfangreicheren Aufgaben, wie jene im zweiten Teil, kann ChatGPT punkten. Teil 2 gilt allgemein hin als komplexer, kontextbezogener und kompetenzübergreifender als Teil 1, die Lösungshäufigkeit ist in der durchgeführten Untersuchung aber nahezu ident. ChatGPT kann damit auch eine gute Lösungsfähigkeit bei umfangreicheren Mathematikproblemen zugeschrieben werden. Bei genauerer Betrachtung steht dies allerdings im Widerspruch zur geforderten Klarheit und Prägnanz von Eingaben, wie in Kapitel 2.10.2 diskutiert. Die KI arbeitet laut Literatur besser und effizienter, wenn die Eingaben kurz, prägnant und frei von überflüssigen Informationen sind. In Teil-2-Aufgaben sind Schüler*innen zum Teil mit einer Menge an Text in der Angabe konfrontiert, was für die KI hier offensichtlich kein Problem darstellt. Möglicherweise wirkt hier die Klarheit der Angabentexte entgegen, da diese zweifelsfrei genau darstellen, was bei gegebener Aufgabe zu tun ist. Für das Sprachmodell ist die Absicht der Eingabe offensichtlich deutlich genug.

Weiters geht aus den Ergebnissen der durchgeführten Testung hervor, dass die Verwendung der unterschiedlichen Promptarten nicht nutzbringend ist, da die Aufgaben ohne Prompt jene mit

der höchsten Lösungshäufigkeit waren. Im getesteten Fall ist es sogar umgekehrt, ein zusätzlicher Prompt zur Bildeingabe verringert die Erfolgsquote, wenn auch nur minimal. Dies ist aufgrund der bereits diskutierten und von Schorcht et al. (2023) und Wei et al. (2023) erwähnten Leistungssteigerungen durch Verwendung von CoT- oder AmA-Prompts mehr als verwunderlich. Es stellt sich die Frage nach der Ursache dieser Diskrepanz. Während in dieser Testung die GPT-Version 4o verwendet wurde, waren bei den Studien von Schorcht et al. (2023) und Wei et al. (2023) GPT-4 bzw. GPT-3 und weitere Sprachmodelle wie LaMDA und PaLM im Einsatz. Das CoT-Prompting ist laut Wei et al. (2023) bei größeren Modellen, die mit einer höheren Anzahl an Parametern trainiert wurden, nochmals spürbar effektiver. Diese Tatsache geht aus folgendem Zitat hervor:

„Chain-of-thought prompting enables large language models to solve challenging math problems. Notably, chain-of-thought reasoning is an emergent ability of increasing model scale“ (Wei et al., 2023, S. 5).

Auch in Abbildung 30(A) ist klar zu erkennen, dass bei einem Parameteranstieg von 10 auf 100 Milliarden beim CoT-Prompting von Textaufgaben eine Genauigkeitssteigerung von etwa 14 Prozentpunkten erzielt wird. Bei den weiteren angeführten Bereichen, wie dem Befolgen von Anweisungen oder der Addition achtschelliger Zahlen, ist der Anstieg noch wesentlich größer.

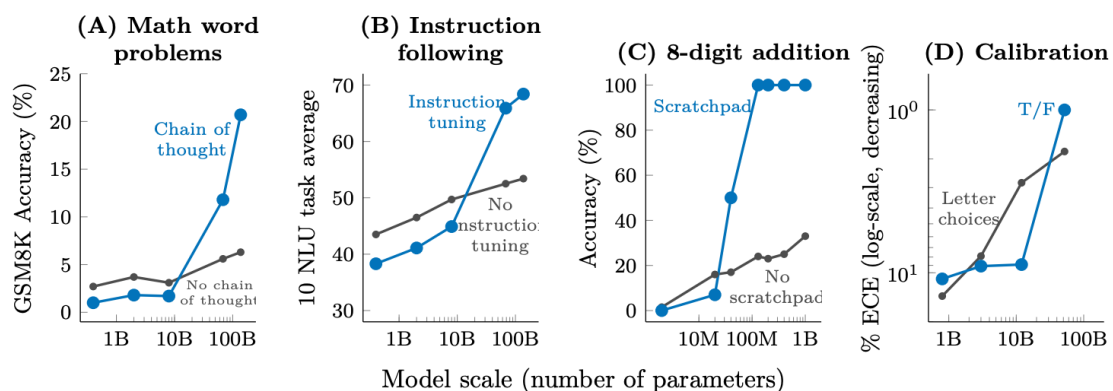


Abbildung 30: Leistungssteigerung beim CoT-Prompting mit wachsender Parameterzahl (Wei et al., 2022, S. 28)

Die Parameteranzahl von GPT-4o ist ebenso wenig offiziell bekannt, wie jene von GPT-4. Folgt man dennoch den geschätzten Parameterzahlen, die online für GPT-4 und GPT-4o kursieren, befindet man sich über den Größenordnungen, der in Abbildung 30 gezeigten Parameterzahlen. Damit gibt es keine Erklärungs- und Argumentationsgrundlage für die nicht nachgewiesene Leistungssteigerung bei der Testung und auch keine Erkenntnisse bezüglich der Effektivität des

CoT-Promptings bei Modellen mit solch großer Parameterzahl. Auf Basis eines Textes von OpenAI (2024b) zur Veröffentlichung der GPT-Version „o1“ lässt sich dennoch eine Vermutung aufstellen. Dort wird von einer direkten Implementierung von Chain-of-Thought in das Modell o1 gesprochen, was laut OpenAI zu einer Verfeinerung der Denkprozesse und Strategien der KI führt. Mithilfe dieser automatisierten Gedankenketten werden Fehler selbst erkannt und korrigiert, komplexe Schritte vereinfacht und damit die Fähigkeit zur logischen Argumentation gesteigert. Die Gedankenkette, die ursprünglich durch den CoT-Prompt eingeleitet werden sollte, passiert in diesem Modell wohl automatisch. Der IBM-Mitarbeiter Hay (2024) vermutet, dass frühe Versionen des o1-Modells das vorangegangene 4o-Modell trainiert haben. Daher scheint es plausibel, dass GPT-4o bereits gewisse CoT-Eigenschaften besitzt. Zur Untermauerung stellte Hay GPT-4o eine komplexe mathematische Rätselfrage und forderte im Prompt, dass die KI keine Zwischenschritte ausführen und so schnell wie möglich antworten solle. Als er dann die gleiche Frage ohne diese zusätzlichen Anweisungen an die KI formulierte, zeigte das Modell automatisch einen schrittweisen Denkprozess und gab sogar wortwörtlich an, dies zu tun. Eine ähnliche CoT-Schrittfolge zeigte auch das o1-Modell bei gleicher Eingabe.

5.3. Folgen und Nutzen für Schüler*innen und Lehrpersonen

Was bedeuten die aufgezeigten Erkenntnisse nun für die Beteiligten des Mathematikunterrichts? Die mathematischen Fähigkeiten von ChatGPT auf Schulniveau sind unbestritten und werden sich mit dem technologischen Fortschritt noch weiter verbessern. Schüler*innen können mit der KI Mathematikaufgaben in kurzer Zeit und Schritt für Schritt lösen lassen. Dies birgt sowohl neue Möglichkeiten als auch Gefahren. Die Leistungsbeurteilung anhand von Aufgaben, die nicht unter Aufsicht einer Lehrperson bearbeitet wurden, ist damit zu hinterfragen. Ausgaben der KI können von Schüler*innen ohne viel Aufwand übernommen und als deren Leistungen präsentiert werden. Hausübungen, die aus rein prozeduralen Aufgaben bestehen und auch noch bewertet werden, sind unter diesen Gesichtspunkten wenig sinnvoll. Bewertete Leistungen sollten in den Unterricht verlagert werden und dort punktuell stattfinden. Andererseits kann die KI Schüler*innen als virtueller Tutor dienen, der den Lernprozess unterstützt. Gerade unter Abwesenheit einer Lehrperson, können die Lernenden dann Fragen zu bestimmten Sachverhalten stellen und sich Lösungswege von Aufgaben erklären lassen. Dies ist eine bisher einzigartige Möglichkeit in der Geschichte

des Mathematiklernens und bietet ein Menge Potential. Diese Chance sollte den Schüler*innen im Unterricht bewusst gemacht und die beispielhafte Anwendung vorgeführt werden. So können die Lernenden motiviert werden, die KI sinnvoll und zu Hause zu ihrem eigenen Vorteil zu nutzen. Auch hier könnte im Unterricht sensibilisiert werden, für welche Arten von Aufgaben die KI problemlos einsetzbar ist und welche Ausgaben mit Vorsicht zu genießen sind. Somit lernen die Schüler*innen die Technologie zu benutzen, sie aber auch kritisch zu hinterfragen, was wiederum den Lernprozess ankurbelt. Laut der diskutierten Umfrage unter Jugendlichen in Kapitel 2.3. ist die Bereitschaft die KI zu nutzen, ohnehin beim Großteil der Befragten vorhanden. Diese Motivation sollte in der Schule aufgegriffen werden. Die Jugendlichen, die Skepsis hegen und Angst haben von der KI ersetzt zu werden, sollen sich folglich erst recht mit der Technologie auseinandersetzen.

Bei Lehrpersonen sind die Anwendungsmöglichkeiten und der potenzielle Nutzen vielfältiger. Das getestete Lösen von Aufgaben kann nutzbringend sein, wenn detaillierte Musterlösungen zu gegebenen Aufgaben erstellt werden sollen. Der Lösungsweg sollte den fachlich geschulten Lehrpersonen ohnehin klar sein, eine Zeitersparnis wäre es natürlich dennoch. Weiters kann die KI auch verschiedene Lösungswege einer Aufgabe aufzeigen, an die auch Lehrer*innen möglicherweise im ersten Moment nicht gedacht haben. Verschiedene Lösungswege können damit schnell gefunden und beispielsweise im Unterricht diskutiert werden. Somit können verschiedene Herangehensweisen für unterschiedliche Schüler*innen angeboten werden. Insbesondere im Bereich der Planung von Mathematikunterricht und der Bewertung von Schüler*innenleistungen könnte die KI in Zukunft effektiv genutzt werden. Gerade diese Bereiche sind besonders zeitaufwendig und bieten ein großes Potenzial für Zeiteinsparungen. Lehrpersonen könnten sich in Folge mehr auf weitere wichtige Aspekte des eigenen Unterrichts fokussieren. Die KI kann Erklärungen diverser mathematischer Konzepte und Zusammenhänge liefern, Sachverhalte umformulieren und für Schüler*innen verständlicher gestalten. Die Modelle beherrschen bereits jetzt didaktische Reduktion und verwenden kindgerechte Analogien. Durch die passenden Eingaben können ganze Stundenplanungen, mit den einzelnen Phasen der Einheit, den jeweiligen Tätigkeiten und passenden Aufgaben generiert werden. Die Ideenfindung und das Konzept einer Stunde, die den Lehrpersonen obliegt, kann ausgelagert und der Planungsprozess beschleunigt werden. Eine Kontrolle und geeignete Adaptierung ist trotzdem unabdingbar. Insbesondere beim Erstellen von Aufgaben ist eine einfache Differenzierung möglich. Wenn gewünscht, kann man sich Aufgaben mit einem bestimmten

Schwierigkeitsgrad oder nach Schwierigkeit gestufte Aufgaben erstellen lassen. Die Chance zur Differenzierung unterschiedlicher Schüler*innengruppen und personalisierter Unterstützung ist gegeben. Den Lernenden könnte einfacher denn je, das geboten werden, was sie aufgrund ihres Lernstands gerade benötigen. Auch im Bereich der Bewertung ist Potential zur Anwendung vorhanden. Eine Korrektur von Schüler*innenantworten muss jedoch vorab rechtlich geregelt sein, sodass bei Eingabe in die KI der Datenschutz gegeben ist. Tests zeigen, dass die KI Fehler in Mathematikaufgaben erkennen kann, Fehleranalysen sind jedoch noch nicht immer zuverlässig und erfordern eine Nachprüfung. Die Integration von KI-Technologien könnte zukünftig helfen, zeitaufwendige Aufgaben wie die Unterrichtsplanung und Prüfungserstellung und -bewertung zu automatisieren und Lehrkräften mehr Zeit für andere pädagogische Aufgaben zu verschaffen. Voraussetzung ist jedoch die Testung und passende Implementierung von Planungs- und Bewertungskompetenzen der KI in der Praxis.

5.4. Ausblick und Desiderata

In möglichen weiteren Arbeiten könnte ChatGPT mit komplexeren Aufgaben auf Universitätsniveau geprüft werden, um noch genauere Aussagen zur Leistungsfähigkeit in gewissen Themenbereichen treffen zu können. Möglicherweise können so weitere Schwachstellen und Lücken identifiziert werden, die bei einer Testung dieses Ausmaßes nicht gefunden werden konnten. Der Fortschritt von KI-Systemen muss definitiv im Auge behalten werden. Aufgaben, die der KI beispielsweise in behandelter Literatur des Jahres 2022 noch Probleme bereitet haben, konnten zum jetzigen Zeitpunkt einwandfrei gelöst werden. Auch die Aufgaben, die in der eigens durchgeführten Testung mit 0 Punkten bewertet wurden, konnten mit der neuesten Version GPT-o1 zum Teil richtig beantwortet werden. Verbesserungen eröffnen neue Möglichkeiten, können aber auch Gefahren mit sich bringen. Eine ständige Evaluierung neuer Versionen wird daher notwendig sein. Insbesondere die Leistung der KI bei den noch schwer zu lösenden Aufgaben im Konstruktionsformat wird mit Spannung zu verfolgen sein. Möglicherweise werden sich die visuellen Analysefähigkeiten verbessern und damit auch dieses Format unter der KI-Anwendung von Schüler*innen zu lösen sein. Im Unterricht sollte eine geeignete Implementierung von KI erfolgen, sowie eine Sensibilisierung bezüglich der Technologie. Ähnlich wie bei Taschenrechnern und Computeralgebrasystemen muss eine passende Verwendung etabliert werden. Das reine Berechnen und Lösen von

mathematischen Aufgaben könnte weiter in den Hintergrund und das Begründen und das Finden von kreativen Lösungsansätzen mehr in den Vordergrund rücken.

Es gilt in einer weiteren Untersuchung die hier nicht untersuchten Fähigkeiten der Bewertungs- und Planungskompetenz zu testen. Es muss geklärt werden, auf welche Art und Weise Eingaben erfolgen müssen, um verwendbare Stundenplanungen und Aufgabenstellungen zu erhalten. Kann die KI bald auch Antworten bewerten und Schülerfeedback generieren, wäre dies eine zeitliche Entlastung für Lehrpersonen. Auch die Entwicklungen im Bereich des Prompt Engineerings müssen von KI-Anwender*innen verfolgt werden. Bereits in der vorliegenden Untersuchung zeigte sich, dass die Eingabe mit bestimmten Promptarten keine besonderen Vorteile bringt und die Techniken bereits in die Prozesse der KI implementiert wurden. Dieser Trend könnte sich fortsetzen und das Prompt Engineering als hochgelobte Schlüsselkompetenz der Zukunft weitaus weniger bedeutsam sein als gedacht. Vermutlich müssen die Eingaben künftig nicht mehr so präzise sein, und die KI wird dennoch das ausgeben, was intendiert wurde. Die Zukunft der KI im Bildungsbereich verspricht, Lernen und Lehren effizienter und kreativer zu gestalten, indem sie Lehrkräften und Schüler*innen neue Möglichkeiten eröffnet. Der Bildungsweg von morgen kann ein harmonisches Zusammenspiel von Mensch und Technologie werden, das neue Horizonte in der Mathematikdidaktik erschließt.

6. Literaturverzeichnis

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L. & McGrew, B. (2023). GPT-4 Technical Report.

arXiv, 2303.08774. <https://doi.org/10.48550/arXiv.2303.08774>

Al-Worafi, Y. M., Hermansyah, A., Goh, K. W., & Ming, L. C. (2023). Artificial Intelligence Use in University: Should

We Ban ChatGPT ?. *Preprints*. <https://doi.org/10.20944/preprints202302.0400.v1>

Alpaydin, E. (2019). *Maschinelles Lernen* (2. Aufl.). De Gruyter.

Arora, S., Narayan, A., Chen, M. F., Orr, L., Guha, N., Bhatia, K., Chami, I., Sala, F., & Ré, C. (2022). Ask Me Anything:

A simple strategy for prompting language models. *arXiv*, 2210.02441.

<https://doi.org/10.48550/arXiv.2210.02441>

Balasubramaniam, S., Kadry, S., Prasanth, A. & Kumar Dhanaraj, R. (2024). *Generative AI and LLMs: Natural*

Language Processing and Generative Adversarial Networks. De Gruyter.

Baytak, A. (2024). The Content Analysis of the Lesson Plans Created by ChatGPT and Google Gemini. *Research in*

Social Sciences and Technology, 9(1), 329-350. <https://doi.org/10.46303/ressat.2024.19>

BMBWF. (2019). *Antwortformate SRP Mathematik (AHS)*. Bundesministerium für Bildung, Wissenschaft und

Forschung. [https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=3036&token=](https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=3036&token=64c1ec19a7a13b762de0a9ab8ec104c690c4cb73)

64c1ec19a7a13b762de0a9ab8ec104c690c4cb73 (Abgerufen am 30.11.2024)

BMBWF. (2023). *Korrektur- und Beurteilungsanleitung zur SRP Mathematik (AHS)*.

Bundesministerium für Bildung, Wissenschaft und Forschung.

[https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=4874&token=2fe884a7184ba50bfc3643873e72](https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=4874&token=2fe884a7184ba50bfc3643873e7213cf775ef92f)

13cf775ef92f (Abgerufen am 11.12.2024)

- BMWF. (2024a). *Mathematische Grundkompetenzen für die SRP in Mathematik (AHS): Inhaltliche Grundlagen zur Sicherung mathematischer Grundkompetenzen*. Bundesministerium für Bildung, Wissenschaft und Forschung. https://aufgabenpool.at/amn/gk/srp_mat_grundkompetenzen_2024-10-25.pdf (Abgerufen am 30.11.2024)
- BMWF. (2024b). *Standardisierte Reife- und Diplomprüfung: Materialien & Publikationen*. Bundesministerium für Bildung, Wissenschaft und Forschung. <https://www.matura.gv.at/downloads> (Abgerufen am 17.11.2024)
- BMWF. (o.D.). *Künstliche Intelligenz im Bildungsbereich*. Bundesministerium für Bildung, Wissenschaft und Forschung. <https://www.bmbwf.gv.at/Themen/schule/zrp/ki.html> (Abgerufen am 15. November 2024)
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Buchholtz, N., Baumanns, L., Huget, J., Peters, F., Pohl, M. & Schorcht, S. (2023). Herausforderungen und Entwicklungsmöglichkeiten für die Mathematikdidaktik durch generative Sprachmodelle. *GDM Mitteilungen*, 114, 19–26.
- Buchholtz, N., Schorcht, S., Baumanns, L., Huget, J., Noster, N., Rott, B., Siller, H., Sommerhoff, D. (2024). Damit rechnet niemand!: Sechs Leitgedanken zu Implikationen und Forschungsbedarfen zu KI-Technologien im Mathematikunterricht. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, (117), 15–24.
- Cameron, S., & Mesiti, C. (2024). What kind of mathematics teacher is ChatGPT? Identifying the pedagogical practices preferred by generative AI tools when preparing lesson plans. *Proceedings of the 46th annual conference of the Mathematics Education Research Group of Australasia*
- Campeato, O. (2023). *Transformer, BERT, and GPT: Including ChatGPT and Prompt Engineering*. Mercury Learning and Information. <https://doi.org/10.1515/9781683928973>

- Copeland, M. (2016). *What's the difference between artificial intelligence, machine learning, and deep learning?*
NVIDIA Blog. <https://blogs.nvidia.com/blog/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/> (Abgerufen am 06.11.2024)
- Dao, X. Q., & Le, N. B. (2023). Investigating the effectiveness of ChatGPT in mathematical reasoning and problem solving: Evidence from the Vietnamese national high school graduation examination. *arXiv*, 2306.06331.
<https://doi.org/10.48550/arXiv.2306.06331>
- De Florio-Hansen, I. (2020). *Digitalisierung, Künstliche Intelligenz und Robotik: eine Einführung für Schule und Unterricht*. Waxmann.
- de Witt, C., Gloerfeld, C., & Wrede, S. E. (2023). *Künstliche Intelligenz in der Bildung* (1. Aufl.). Springer Fachmedien
Wiesbaden.
- Duranton, S. (2023). *Prompting progress: We'll all speak AI in the future workforce*. Forbes.
<https://www.forbes.com/sites/sylvainduranton/2023/06/15/prompting-progress-well-all-speak-ai-in-the-future-workforce/> (Abgerufen am 10.11.2024)
- Ertel, W., & Black, N. T. (2016). *Grundkurs Künstliche Intelligenz* (4. Aufl.). Wiesbaden: Springer Fachmedien
Wiesbaden.
- Europäisches Parlament. (2020). *Was ist künstliche Intelligenz und wie wird sie genutzt?*
<https://www.europarl.europa.eu/topics/de/article/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt> (Abgerufen am 24.10.2024)
- Field, H. (2024). *OpenAI launches new AI model and desktop version of ChatGPT*. CNBC.
<https://www.cnbc.com/2024/05/13/openai-launches-new-ai-model-and-desktop-version-of-chatgpt.html>
(Abgerufen am 12.11.2024)
- Fischer, J., Eckert, K., Wolf, I., & Wiley-VCH. (2024). *Maschinelles Lernen für Dummies* (1. Aufl.). Wiley.

Fraunhofer-Institut für Kognitive Systeme. (2022). *Künstliche Intelligenz (KI) und maschinelles Lernen*. Fraunhofer IKS. <https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html> (Abgerufen am 24.10.2024)

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical capabilities of ChatGPT. *Advances in neural information processing systems*, 36.

Google Cloud. (o. D.). *Was ist künstliche Intelligenz?*. <https://cloud.google.com/learn/what-is-artificial-intelligence?hl=de> (Abgerufen am 24.10.2024)

Greefrath, G., Oldenburg, R., Siller, H.-S., Ulm, V., & Weigand, H.-G. (2024). *Digitalisierung im Mathematikunterricht: Theorie und Praxis digitaler Medien in der Sekundarstufe I* (1. Aufl.). Springer Berlin Heidelberg.

Guler, N., Dertli, Z. G., Boran, E., & Yildiz, B., (2024). An artificial intelligence application in mathematics education: Evaluating ChatGPT's academic achievement in a mathematics exam. *Pedagogical Research*, 9(2), 1-12.

Hattie, J., & Clarke, S. (2018). *Visible Learning: Feedback* (1. Aufl.). Routledge.
<https://doi.org/10.4324/9780429485480>

Hay, C. (2024). *OpenAI O1 models probably trained GPT-4o and Turbo in chain of thought* [Video]. YouTube.
https://www.youtube.com/watch?v=3n_-1HxBiAM (Abgerufen am 13.01.2024)

Heinrichs, B., Heinrichs, J. & Rüther, M. (2022). *Künstliche Intelligenz*. De Gruyter.

Helfrich-Schkarbanenko, A. (2023). *Mathematik und ChatGPT: Ein Rendezvous am Fuße der technologischen Singularität* (1. Aufl.). Springer Berlin Heidelberg.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266.
<https://doi.org/10.1126/science.aaa8685>

Hischer, H. (2023). ChatGPT und Mathematikunterricht–eine didaktische Herausforderung?. *Mitteilungen der Deutschen Mathematiker-Vereinigung*, 31(3), 173-175.

- Hu, K. (2023). *ChatGPT sets record for fastest-growing user base - analyst note*. Reuters.
<https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (Abgerufen am 31.10.2024)
- Huang, Y. M., & Rocha, T. (2023). Innovative Technologies and Learning. *6th International Conference, ICITL 2023 Porto, Portugal, August 28–30, 2023 Proceedings*.
- Johnson, O. V., Alyasiri, O. M., & Johnson, O. E. (2023). Image Analysis through the lens of ChatGPT-4. *Journal of Applied Artificial Intelligence*, 4(2), 31-46. <https://doi.org/10.48185/jaai.v4i2.870>
- Khan, I. (2024). *The quick guide to prompt engineering: Generative AI tips and tricks for ChatGPT, Bard, Dall-E, and Midjourney*. John Wiley & Sons.
- Kitzmann, A. (2022). *Künstliche Intelligenz: Wie verändert sich unsere Zukunft?* (1. Aufl.). Springer Fachmedien Wiesbaden.
- Knaus, T., et al. (2023). Künstliche Intelligenz und Bildung: Was sollen wir wissen? Was können wir tun? Was dürfen wir hoffen? Und was ist diese KI? Ein kollaborativer Aufklärungsversuch. *Ludwigsburger Beiträge zur Medienpädagogik*, (23). doi.org/10.21240/lbzm/23/19
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information processing systems*, 35, 22199-22213.
- LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, 521(7553), 436–444.
<https://doi.org/10.1038/nature14539>
- Levy, S. (2024). *Eight Google employees invented modern AI with this paper*. Wired.
<https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper> (Abgerufen am 09.11.2024)
- Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen?: Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung. *Zeitschrift Für Pädagogische Psychologie*, 29(3–4), 133–149. <https://doi.org/10.1024/1010-0652/a000156>

- Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. *FinTech and artificial intelligence for sustainable development: The role of smart technologies in achieving development goals*, (387-409). Springer Nature Switzerland.
- Middendorf, W. (2024). Zur Möglichkeit digital gestützter Leistungsbewertung mithilfe Künstlicher Intelligenz in der Schulpraxis. 8 S. <https://doi.org/10.25656/01:28739>
- ÖBV. (2024a). *KI im Klassenzimmer: Umfrage des österreichischen Bundesverlags*. Österreichischer Bundesverlag Schulbuch GmbH & Co. KG. <https://www.oebv.at/images/product-images/oebv-umfrage-ki-im-klassenzimmer-2024.pdf> (Abgerufen am 02.12.2024)
- ÖBV. (2024b). *KI im Klassenzimmer: Wie künstliche Intelligenz Bildung verändert - Whitepaper*. Österreichischer Bundesverlag Schulbuch GmbH & Co. KG. <https://www.oebv.at/magazin/whitepaper-ki-im-klassenzimmer-wie-kuenstliche-intelligenz-bildung-veraendert> (Abgerufen am 02.12.2024)
- OpenAI. (2024). *ChatGPT (Version 4o)* [Large language model]. <https://chat.openai.com/chat> (Abgerufen am 17.12.2024)
- OpenAI. (2024). *Learning to reason with large language models*. <https://openai.com/index/learning-to-reason-with-llms> (Abgerufen am 13.01.2025)
- OpenAI. (o.D.). *How can I access GPT-4, GPT-4 Turbo, GPT-4o, and GPT-4o mini?*. OpenAI. <https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4-gpt-4-turbo-gpt-4o-and-gpt-4o-mini> (Abgerufen am 12.11.2024)
- ORF. (2024). *Alle sind gefragt - Die Ö3 Jugendstudie 2024*. https://www.oe3jugendstudie.at/ergebnisse.php?utm_source=int&utm_media=story (Abgerufen am 16.11.2024)

Otte, R., & Wiley-VCH. (2021). *Allgemeinbildung Künstliche Intelligenz - Risiko und Chance für Dummies* (1. Aufl.). John Wiley & Sons.

Pandey, M., Fernandez, M., Gentile, F., Isayev O., Tropsha A., Stern A., Cherkasov A. (2022). The transformational role of GPU computing and deep learning in drug discovery. *Nature Machine Intelligence*, 4, 211–221.
<https://doi.org/10.1038/s42256-022-00463-x>

Parlament der Republik Österreich. (2023). *Umgang mit Künstlicher Intelligenz (KI) in der Schule. Gegenstand XXVII/A/3490: Selbstständiger Entschließungsantrag*.
<https://www.parlament.gv.at/gegenstand/XXVII/A/3490> (Abgerufen am 16.11.2024)

Pepin, B., Buchholtz, N., & Salinas-Fernandez, U. (2024). A scoping survey of ChatGPT in mathematics education. *Research Square*. <https://doi.org/10.21203/rs.3.rs-4982227/v1>

Perry, T. S. (2018). Move over, Moore's law. Make way for Huang's law. *IEEE Spectrum*, 55(5), 7-7.

Phoenix, J., & Taylor, M. (2024). *Prompt engineering for generative AI: future-proof inputs for reliable AI outputs at scale*. O'Reilly Media.

Plevris, V., Papazafeiropoulos, G., & Jiménez Rios, A. (2023). Chatbots put to the test in math and logic problems: a comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google bard. *AI*, 4(4), 949-969.

Rüdian, S., & Pinkwart, N. (2024). Prompt Engineering zur Generierung von Sprachlerninhalten. *Informatik Spektrum*, 1-9.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?. *Journal of applied learning and teaching*, 6(1), 342-363.

Russell, S. J., Norvig, P., & Davis, E. (2016). *Artificial intelligence: a modern approach* (3. Aufl.). Pearson.

- Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F., & Pohl, M. (2023). Ask Smart to Get Smart: Mathematische Ausgaben generativer KI-Sprachmodelle verbessern durch gezieltes Prompt Engineering. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, (115).
- Schorcht, S., Baumanns, L., Buchholtz, N., Huget, J., Peters, F., & Pohl, M. (2024). Lernt die KI nun Sehen und Zeichnen?: Chancen und Herausforderungen der Bildgenerierung und Bildinterpretation mit ChatGPT für die mathematikdidaktische Forschung. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, (116).
- Shabani, K., Khatib, M., & Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English language teaching*, 3(4), 237-248.
- Shah, P. (2023). *AI and the Future of Education: Teaching in the Age of Artificial Intelligence*. John Wiley & Sons.
- Spannagel, C. (2023). Hat ChatGPT eine Zukunft in der Mathematik?. *Mitteilungen der Deutschen Mathematiker-Vereinigung*, 31(3), 168-172. <https://doi.org/10.1515/dmvm-2023-0055>
- Spreitzer, C., Straser, O., Zehetmeier, S., & Maaß, K. (2024). Mathematical Modelling Abilities of Artificial Intelligence Tools: The Case of ChatGPT. *Education Sciences*, 14(7), 698. <https://doi.org/10.3390/educsci14070698>
- Ständige Wissenschaftliche Kommission der Kultusministerkonferenz (SWK). (2024). Large Language Models und ihre Potenziale im Bildungssystem. *Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz*. <https://doi.org/10.25656/01:28303>
- Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart learning environments*, 10(1), 15.
- Wang, J., Ye, Q., Liu, L., Guo, N. L., & Hu, G. (2024). Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception. *NPJ Precision Oncology*, 8(1), 84.
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), em2286.

- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *arXiv*, 2206.07682. <https://doi.org/10.48550/arXiv.2206.07682>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter B., Xia, F., Chi, E. H., Le Q. V. & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- Weßels, D. (2023). ChatGPT – ein Meilenstein der KI-Entwicklung. *Mitteilungen der Deutschen Mathematiker-Vereinigung*, 31(1), 17-19. <https://doi.org/10.1515/dmvm-2023-0007>
- Wittpahl, V. (2019). *Künstliche Intelligenz: Technologien / Anwendung / Gesellschaft* (1. Aufl.). Springer Berlin Heidelberg.
- Ziegenbalg, J., Ziegenbalg, Oliver, & Ziegenbalg, Bernd. (2016). *Algorithmen von Hammurapi bis Gödel: Mit Beispielen aus den Computeralgebrasystemen Mathematica und Maxima* (4. Aufl.). Springer Spektrum.

7. Abbildungsverzeichnis

Abbildung 1: Traditionelle Programme und maschinelles Lernen (Wittpahl, 2019, S. 25) _____	14
Abbildung 2: Schematisches Modell eines künstlichen Neurons (Ziegenbalg et al., 2016, S. 324) _____	18
Abbildung 3: Neuronales Netz mit gegebener Anfangsgewichtung (Ziegenbalg et al., 2016, S. 329) _____	19
Abbildung 4: Vergleich neuronaler Netze (Ziegenbalg et al., 2016, S. 331f) _____	21
Abbildung 5: Ausgabegenauigkeit von Zero, One- und Few-Shot-Prompts (Brown et al., 2020, S. 4) _____	31
Abbildung 6: Vergleich der KI-Anwendungen (Plevris et al., 2023, S. 965) _____	40
Abbildung 7: Generierung einer Multiple-Choice-Aufgabe mittels ChatGPT (Middendorf, 2024, S. 3) _____	49
Abbildung 8: LaTeX-Code einer 5x5 Einheitsmatrix (Helfrich-Schkarbanenko, 2023, S. 28) _____	51
Abbildung 9: Eingabefeld von ChatGPT mit Aufgaben-Screenshot und CoT-Prompt (OpenAI, 2024a) _____	55
Abbildung 10: Fallbeispiel 1- Prüfungsaufgabe der AHS-Matura des Herbsttermins 2023/24 (BMBWF, 2024b) _	59
Abbildung 11: Fallbeispiel 1 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024aa) _____	60
Abbildung 12: Fallbeispiel 2 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2023/24 (BMBWF, 2024b) _	61
Abbildung 13: Fallbeispiel 2 - Lösung der Prüfungsaufgabe mit CoT-Prompt (OpenAI, 2024a) _____	62
Abbildung 14: Fallbeispiel 3 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2023/24 (BMBWF, 2024b) _	62
Abbildung 15: Fallbeispiel 3 – Lösung der Prüfungsaufgabe ohne und mit CoT-Prompt (OpenAI, 2024a) _____	63
Abbildung 16: Fallbeispiel 4 - Prüfungsaufgabe der AHS-Matura des Haupttermins 2023/24 (BMBWF, 2024b) _	64
Abbildung 17: Fallbeispiel 4 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a) _____	65
Abbildung 18: Fallbeispiel 5 - Prüfungsaufgabe der AHS-Matura des Haupttermins 2023/24 (BMBWF, 2024b) _	66
Abbildung 19: Fallbeispiel 5 - Lösung der Prüfungsaufgabe mit CoT- & AmA-Prompt (OpenAI, 2024a) _____	67
Abbildung 20: Fallbeispiel 6 - Prüfungsaufgabe der AHS-Matura des Wintertermins 2022/23 (BMBWF, 2024b) _	68
Abbildung 21: Fallbeispiel 6 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a) _____	68
Abbildung 22: Fallbeispiel 7 - Prüfungsaufgabe der AHS-Matura des Wintertermins 2022/23 (BMBWF, 2024b) _	69
Abbildung 23: Fallbeispiel 7 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a): _____	70
Abbildung 24: Fallbeispiel 8 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2022/23 (BMBWF, 2024b) _	71
Abbildung 25: Fallbeispiel 8 - Lösung der Prüfungsaufgabe mit AmA-Prompt (OpenAI, 2024a) _____	72
Abbildung 26: Fallbeispiel 9 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2022/23 (BMBWF, 2024b) _	72
Abbildung 27: Fallbeispiel 9 - Lösung der Prüfungsaufgabe ohne Prompt (OpenAI, 2024a) _____	73
Abbildung 28: Fallbeispiel 10 - Prüfungsaufgabe der AHS-Matura des Herbsttermins 2022/23 (BMBWF, 2024b) _	74
Abbildung 29: Fallbeispiel 10 - Lösung der Prüfungsaufgabe mit CoT-Prompt (OpenAI, 2024a) _____	75
Abbildung 30: Leistungssteigerung beim CoT-Prompting (Wei et al., 2022, S. 28) _____	80

8. Tabellenverzeichnis

Tabelle 1: Beurteilungsschlüssel SRP Mathematik (BMBWF, 2023, S. 2)	54
Tabelle 2: Über die Promptarten gemittelte Punktezahlen und Notengrade der einzelnen Prüfungstermine	56
Tabelle 3: Gesamtpunktezahlen bei unterschiedlichen Promptarten	57
Tabelle 4: Notengrade bei unterschiedlichen Promptarten	57
Tabelle 5: Prozentuelle Häufigkeiten richtiger Antworten innerhalb des jeweiligen Antwortformats	58
Tabelle 6: Prozentuelle Häufigkeiten richtiger Antworten innerhalb des jeweiligen Inhaltsbereichs	58
Tabelle 7: Prozentuelle Häufigkeiten richtiger Antworten beim jeweiligen Prüfungsteil	58
Tabelle 8: Übersicht der Fallbeispiele	59