



**Pädagogische
Hochschule
Weingarten
University of
Education**

Masterarbeit

zum Thema

Lernen mit KI im Physikunterricht

—

Eine Studie

Fabian Hölzle

Matrikelnummer: 7207637

fabian.hoelzle@stud.ph-weingarten.de

vorgelegt an der Pädagogischen Hochschule Weingarten

Gutachter 1: Herr Prof. Dr. rer. nat. Florian Theilmann

Gutachter 2: Herr Holger Zieris

Wangen, 26.12.2024

Abstract

Diese Masterarbeit behandelt die Integration von künstlicher Intelligenz (KI) in den Physikunterricht der Sekundarstufe I. Hierdurch soll das Potenzial für eine höhere Schülerunterstützung geboten und die Individualisierung des Lernprozesses angereichert werden. Das Ziel dieser Arbeit liegt in der Untersuchung der fachlichen Korrektheit sowie der Erklärqualität der von KI erstellten Erklärungen sowie dem Potential von KI zur Gestaltung sinnvoller Lehr-Lern-Umgebungen im Physikunterricht der Sekundarstufe I. Die Untersuchung basiert dabei auf einer Bewertung von KI-generierten Erklärungen durch Studenten, Expertenanalysen von KI-Chats und Schülerbefragungen. Die Ergebnisse der Untersuchung zeigen, dass die KI in der Lage ist fachlich korrekte und verständliche Erklärungen zu produzieren. Alles in allem zeigt diese Arbeit das Potential für den Einsatz von künstlicher Intelligenz im Schulunterricht, weist aber auch auf die Notwendigkeit von didaktischen Anpassung hin.

Inhaltsverzeichnis

1. Einleitung.....	1
2. Theorie	2
2.1. Künstliche Intelligenz.....	2
2.1.1. Was ist eine künstliche Intelligenz?	2
2.1.2. Strukturierung von künstlicher Intelligenz	6
2.1.3. Das steckt dahinter	10
2.1.4. KI und Bildung, passt das zusammen?	11
2.2. Geschichtliches	14
2.2.1. Von der ersten Idee bis heute	14
2.2.2. Diverse künstliche Intelligenzen	15
2.3. OpenAIs ChatGPT	16
2.3.1. Überblick.....	16
2.3.2. Versionsverlauf	18
2.4. Googles GEMINI	20
2.5. Umgang mit LLM	22
2.6. KI & Schule – Rechtliches	24
2.7. Was ist eine gute Erklärung?	25
3. Hypothesen.....	28
4. Methodik	29
4.1. Grundlegendes Forschungskonzept.....	29
4.2. Künstliche Intelligenz als Tutor	29
4.3. Künstliche Intelligenz oder Schulbuch – Ein Vergleich.....	33
4.4. Auswertung	34
5. Ergebnisse.....	38
5.1. Ein erster Überblick	38
5.2. Hypothese H1.....	46
5.3. Hypothese H2.....	48
5.4. Hypothese H3.....	50

5.5. Exploration 1	53
5.6. Exploration 2	54
5.7. Exploration 3	56
5.8. Exploration 4	57
5.9. Exploration 5	58
5.10. Exploration 6	59
5.11. Exploration 7	60
6. Diskussion	61
7. Fazit	68
8. Literaturverzeichnis	69
9. Anhang	72
9.1. Prompt-Verzeichnis	72
9.2. Internetlink-Verzeichnis	73
9.3. Antworten auf Prompt 1	73
9.3.1. Gemini 1.5 Flash	73
9.3.2. ChatGPT 4o	75
9.4. Antworten auf Prompt 2	76
9.4.1. Gemini 1.5 Flash	76
9.4.2. ChatGPT 4o	77
9.5. Antworten auf Prompt 3-5 von DALL-E	79
9.6. Thementexte für „KI vs. Schulbuch“	81
9.6.1. Text A	81
9.6.2. Text B	83
9.6.3. Text C	83
9.6.4. Text D	84
9.7. Interviewbegleitbogen Studenten	86
9.8. Fragebogen SchülerInnen	87
9.9. Bewertungsmatrix KI-Chat	88
10. Eidesstattliche Erklärung	90

1. Einleitung

Seit der Veröffentlichung von ChatGPT im November 2022 [1], [2] ist das Thema künstliche Intelligenz (Abkürzung: KI) in der Gesellschaft aufgeflammt. Unzählige Zeitungsartikel erzählen von dem neuen Wunderwerk der Technik, mit dem man zum Beispiel Bewerbungen erstellen lassen kann [2] und wie man diese nagelneue Technik anwendet [3]. Auch im sonstigen Leben wird immer mehr das Thema künstliche Intelligenz integriert. So hat die Social-Media-App Snapchat einen Chatbot mit KI implementiert [4]; der Github-Copilot wurde entwickelt, eine künstliche Intelligenz, welche beim Arbeiten in der Softwareentwicklung durch Code-Vorschläge die Entwickler unterstützt [5] und Apple hat eine eigene KI namens „Apple Intelligence“ (AI) entwickelt, welche im Betriebssystem iOS18 integriert sein wird, angekündigt [6]. Weitere Firmen wie Tesla und Microsoft arbeiten ebenfalls an künstlicher Intelligenz [7], [8] und auch NVIDIA, ursprünglich bekannt für Grafikprozessoren, arbeitet nun an der Hardware für KI-lastige Rechenzentren [9]. Während in Deutschland die Anzahl an Startups in Bezug zu KI für das Jahr 2024 auf 687 ansteigt [10], gibt es immer mehr Warnungen, welche vor KI-Betrügern warnen [11]. Wenn man nun den Fokus auf ChatGPT legt, so zeigt sich bereits im Dezember 2022, dass ChatGPT ganze Aufsätze schreiben, Texträtsel lösen und Erklärungen sowie ausführliche Antworten auf Fragen geben kann [12]. Als Lehrperson schrillen direkt die Alarmglocken auf. Es stellen sich Fragen, ob SchülerInnen ihre Hausaufgaben noch selbst machen oder mit künstlicher Intelligenz erstellen, ob sie Aufsätze und Interpretationen noch selbst schreiben, aber auch, ob man dieses, auf den ersten Blick, mächtige Werkzeug nicht selbst anwenden kann [13, S. 71]. Wenn die KI so gut ist, wie alle sagen, kann eine Lehrkraft damit Arbeitsblätter und Unterrichtsmaterialien erstellen [14, S. 151] und SchülerInnen die Themen erklären lassen. New Yorks Schulen haben kurz nach der Veröffentlichung von ChatGPT ein ChatGPT-Verbot durch die Sperrung der Webseite auf den Computern der öffentlichen Schulen eingeführt und begründen dies mit der Möglichkeit, dass der Chatbot falsche Daten erfindet [2].

Nach über einem Jahr ChatGPT und einer Weiterentwicklung des Chat-Bots auf ChatGPT-4 im März 2023 [13, S. 16] stellt sich nun die Frage, ob man die Technologie im Unterricht anwenden sollte. Das Internet hat seinen Weg in die Gesellschaft gefunden und dann, durch die auf die Gesellschaft vorbereitende Funktion der Schule, den Weg in den Unterricht gefunden. Nun könnte man meinen, dass die KI einen ähnlichen Weg gehen kann, da man so den SchülerInnen den Umgang mit diesem Werkzeug beibringen kann. Doch bevor es dazu kommt, sollte erst einmal geklärt werden, ob eine KI im Unterricht einen Mehrwert liefern kann. Diese Arbeit beschäftigt sich, neben den Begriffsbildungen und -definitionen sowie einem Überblick über das aktuelle Feld der KI, mit der Idee, ChatGPT im Unterricht unter

anderem als Tutor für die SchülerInnen anzuwenden. Schließlich ist der Chatbot als Download oder im Browser für die SchülerInnen immer verfügbar und somit immer zur Hand, wenn eine Frage entsteht. Des Weiteren kann er auf Nachfragen von SchülerInnen eingehen, mehr dazu jedoch im Verlauf der Arbeit. Damit eine Aussage über die Sinnhaftigkeit eines Chatbots für (schulische) Erklärungen getroffen werden kann, werden ebenfalls die Eigenschaften guter Erklärungen betrachtet und von ChatGPT generierte Erklärungen untersucht.

2. Theorie

2.1. Künstliche Intelligenz

2.1.1. Was ist eine künstliche Intelligenz?

Bevor genauer auf die Funktionsweise oder Struktur von künstlicher Intelligenz eingegangen wird, muss zunächst der Begriff der künstlichen Intelligenz geklärt werden. Künstliche Intelligenz, oder Artificial Intelligence, wie sie im Englischen genannt wird, wurde über die Jahre unterschiedlich definiert. Einer der Gründe für die verschiedenen Definitionen und Ansätze der Definitionen ist, dass es bisher keine einheitliche Definition für die Intelligenz (der Menschen) an sich gibt [15, S. 46].

Die erste Annahme, dass eine künstliche Intelligenz möglich ist, wurde bereits 1955 von John McCarthy getroffen und beschreibt die KI durch das Ziel eine Maschine zu entwickeln, welche sich so verhält, als hätte sie Intelligenz. Weitere Definitionen nach der Encyclopedia Britannica oder Elaine Rich beschreiben die KI damit, dass sie etwas macht, was aktuell der Mensch besser kann oder was für den Menschen typisch ist [16, S. 1–3]. Auch Zehner tendiert in diese Richtung, indem eine KI als ein Computersystem verstanden wird, welches menschentypische Aufgaben erledigt wie die „visuelle Erkennung, Spracherkennung und -produktion, Wissenswiedergabe, autonomes Fahren und viele[...] weitere.“ [17, S. 3]

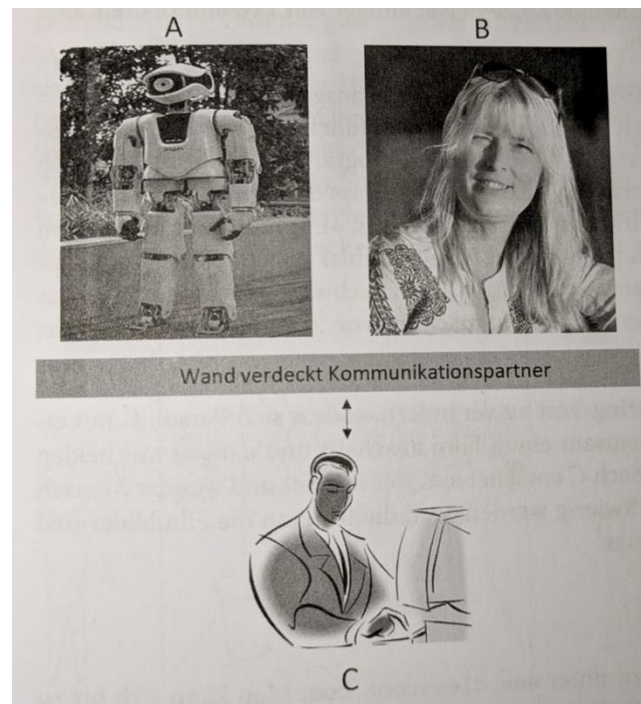


Abbildung 1: Turing-Test [18, S. 69]

Turing basierte seine Definition der künstlichen Intelligenz im Jahr 1950 nicht auf Eigenschaften des Systems an sich, sondern durch die Interaktion des Systems mit dem Menschen. Ein System ist dann intelligent, wenn es den nach Turing benannten Turing-Test besteht (siehe Abbildung 1). Dieser Test besteht aus einer Versuchsperson, welche mit einer Tastatur vor einem Computer sitzt und auf einen Bildschirm schaut. Sie befindet sich in einem schriftlichen Chat. Der Gesprächspartner kann nun entweder ein Mensch oder eine Maschine (Chatbot) sein. Ziel der Versuchsperson ist es, durch den Gesprächsverlauf in einer vorgegebenen Zeit, herauszufinden, ob der Gesprächspartner ein Mensch ist oder nicht. Schafft es die Maschine den Menschen in 70 % der Fälle zu täuschen, der Mensch also angibt, dass er mit einem anderen Menschen kommuniziert hat, besteht die Maschine den Turing-Test und wird als intelligent gehalten [15, S. 44], [18, S. 68f], [19, S. 1085]. Ein gut bekanntes Beispiel für den Turing-Test, welches jedoch herausfindet, ob ein Mensch vor dem Computer sitzt, ist der CAPTCHA (**C**ompletely **A**utomated **P**ublic **T**uring **T**est to **T**ell **C**omputers and **H**umans **A**part), welcher im Internet häufig für die Verifizierung einer Nutzereingabe durch einen Menschen verwendet wird [15, S. 44]. Inzwischen gibt es sogar Turing-Tests als eine Art Spiel. Auf der Seite des AI21 Labs betritt man einen Chatraum und erhält entweder einen menschlichen Gesprächspartner oder einen Chatbot [20]. Die zwei Minuten Spielzeit sind in abwechselnde Schreibzeiten aufgeteilt, in denen jeweils nur ein Partner schreiben kann. In Abbildung 2 ist ein solches Spiel zu sehen. Es ist dabei zu erkennen, dass der Chatbot einerseits Slang anwendet, andererseits seine Nachrichten innerhalb der Zeit nicht abschließt, was ihn menschlicher aussehen lässt.

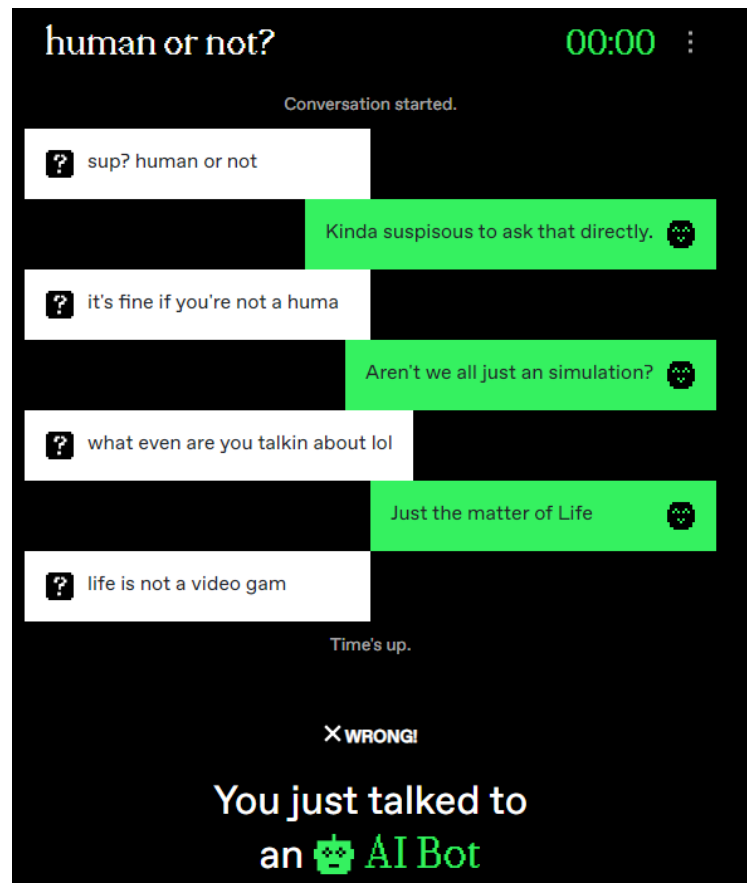


Abbildung 2: Spiel human or not? Anwender ist in grün [20]

Kritik am Turing-Test wird durch die Theorie des chinesischen Zimmers ausgeübt (siehe Abbildung 3). Hierbei handelt es sich um ein Gedankenexperiment, in welchem man sich einen nicht chinesisch sprechenden Menschen in einem Raum vorstellt. Dieser Mensch hat eine, in seiner Muttersprache verfasste, Anleitung erhalten, die ihm erklärt, wie er chinesische Fragen beantworten kann. Geben Außenstehende nun Fragen auf Chinesisch in den Raum, erhalten sie Antworten, ebenfalls auf Chinesisch. Für diese Außenstehenden sieht es aus, dass der Mensch im Raum chinesisch kann. Überträgt man nun dieses Gedankenexperiment auf Computer, so kann ein Computer, welcher eine Zeichenkette (zum Beispiel Wörter) durch die Anwendung von gegebenen Regeln abändern kann, nicht intelligent sein [15, S. 45]. Dies widerspricht dem Turing-Test, da dieser im Beispiel des chinesischen Zimmers davon ausgehen würde, dass der Mensch oder im übertragenen Sinne der Computer, intelligent ist.

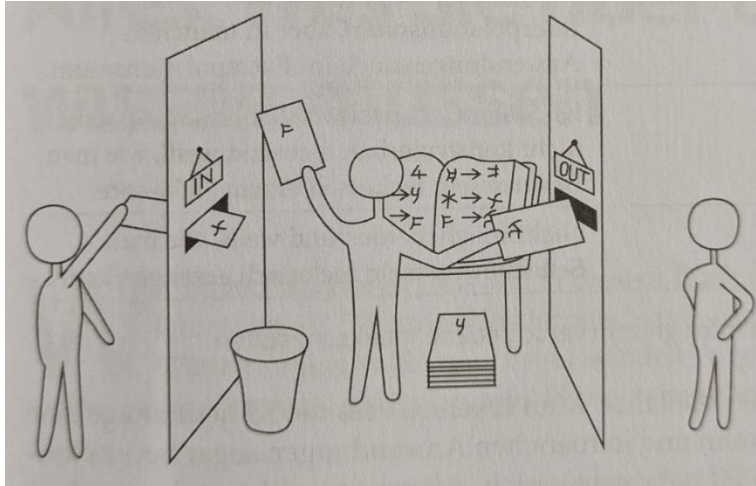


Abbildung 3: Chinesisches Zimmer [18, S. 71]

Modernere Definitionen versuchen die Begriffsdefinition über die Intelligenz eines Systems. Der grundlegende Begriff der Intelligenz wird dabei in zwei Bereiche aufgeteilt. Die niedere Intelligenz kann angemessen auf Reize reagieren, um einen Vorteil oder keinen großen Nachteil aus der Reaktion zu generieren. Intelligenz höherer Ordnung hingegen kann bei nicht angemessener Reaktion zusätzlich nachjustieren, um bei zukünftigen Reizen angemessen zu reagieren [18, S. 32f]. Nun werden fünf Intelligenzstufen gebildet, welche die Intelligenz eines Systems beschreiben. Während in der ersten Stufe, der angemessenen Intelligenz, Systeme der Steuerungs- und Regelungstechnik wie das ABS, die Temperaturregelung eines Herds oder Lageregelung eines Flugzeugs beinhaltet sind, betrachtet Stufe zwei bereits die lernende Intelligenz. Diese ist ein adaptiver Regler, der seinen Regelkreis an eine für ihn neue Umgebung anpassen kann [18, S. 45ff]. Intelligenzstufe drei betrachtet eine kreative Intelligenz, welche neue, gewinnbringende Dinge erschaffen kann. In Stufe vier wird von einer bewussten Intelligenz gesprochen, welche Erkenntnisse über den inneren Zustand und das subjektive Erleben dieses Zustandes sammeln kann. Die letzte Stufe beschreibt die selbstbewusste Intelligenz, welche über ein Ich-Konzept verfügt [18, S. 60ff]. Mit diesen fünf Stufen lässt sich nun künstliche Intelligenz beschreiben. Ein System, welches sich nur in Stufe eins befindet, besitzt keine künstliche Intelligenz. Da das System bei Fehlern keine Veränderung durchführen kann, handelt es sich hierbei um die codierte Intelligenz, welche von einem Menschen erstellt wurde [18, S. 33]. Betrachtet man die anderen Intelligenzstufen, so ist die Mindestanforderung einer künstlichen Intelligenz, dass die KI Stufe zwei erreicht oder zusätzlich zu Stufe eins noch mit uns Menschen kommuniziert [18, S. 67f]. Da der Bereich der künstlichen Intelligenz durch diese Definition immer noch sehr groß ist, kann man die künstliche Intelligenz aber noch in zwei Bereiche aufteilen. Die schwache KI beschreibt die Simulation von menschlicher Intelligenz auf Maschinen, während die starke KI eine Maschine mit Bewusstsein ist. Während alle aktuellen Systeme der

künstlichen Intelligenz der schwachen KI angehören, sind sich Forscher noch nicht sicher, ob eine starke KI (Intelligenzstufe vier oder fünf) überhaupt funktionieren kann [18, S. 35].

2.1.2. Strukturierung von künstlicher Intelligenz

Künstliche Intelligenz ist ein gesamtes Forschungsfeld in der Informatik [21], weshalb man, wenn man über KI spricht, auch über die verschiedenen Typen sprechen sollte. Nicht jede KI funktioniert gleich, sie sind unterschiedlich aufgebaut und haben verschiedene Aufgaben. Dieses Kapitel dient dazu, einen groben Überblick über die einzelnen Bereiche sowie ihre Vernetzung untereinander zu geben.

Basiert man nun die Struktur (Abbildung 4) auf die in 2.1.1 Definition von künstlicher Intelligenz, so ist zu erkennen, dass wir uns auf der Seite der schwachen KI befinden. Da bisher noch kein Computer erfunden wurde, der ähnlich menschlicher Neuronen funktioniert (neuromorph), befindet man sich nun auf der Seite der künstlichen Intelligenz auf Digitalcomputern. Im nächsten Abschnitt wird zwischen der Schlussart der KI unterschieden. Für diese Arbeit sind vor allem die Intelligenzen des induktiven Zweiges interessant. Deshalb behandelt dieser Abschnitt einerseits die Begriffe Maschinelles Lernen und Neuronale Netzwerke sowie Unterkategorien dieser.

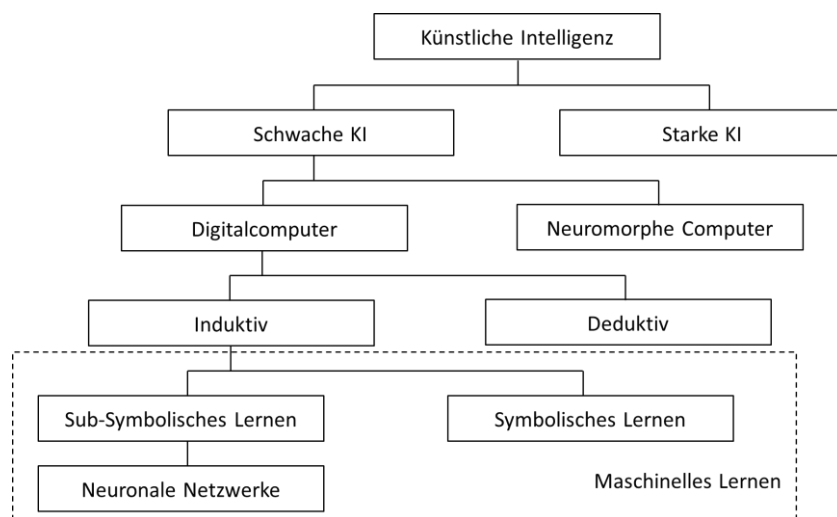


Abbildung 4: Struktur von KI; Anlehnung an [18, S. 149f]

Maschinelles Lernen (Machine Learning; ML) ist offen formuliert. Zusammengefasst gehören hier alle Computerprogramme dazu, welche aus dem wiederholten Erfassen von Daten lernen, ohne dass diese Daten fest einprogrammiert werden [21], [22, S. 12]. Der Computer zieht sich das Wissen oder auch Modelle einer Welt aus den Daten, mit denen er

lernt, und speichert die daraus gelernten Regeln in Datenbanken oder neuronalen Netzwerken [18, S. 150f].

Eine Anwendung des maschinellen Lernens sind sogenannte **neuronale Netze** (Neural Network; NN). Sie werden teilweise auch künstliche neuronale Netze genannt, was den Wortlaut stärker von den natürlichen Netzen und damit von der Herkunft der Idee trennt. Die Idee, wie ein neuronales Netzwerk aufgebaut ist, ist sehr stark dem Aufbau des menschlichen Hirns mit seinen Speichermechanismen und Informationsverarbeitungseinheiten nachempfunden [13, S. 20], [18, S. 213], [21], [22, S. 11], [23, S. 51ff]. Die Nachbildung dieser Einheiten werden Neuronen genannt und bilden die Knotenpunkte im neuronalen Netz. Jedes einzelne Neuron besitzt eine einfache mathematische Funktion wie die Multiplikation, Addition oder einen Vergleich [13, S. 20], [18, S. 213]. Da Neuronen an sich kein Netzwerk bilden und auch ein Knoten ohne ein Seil nicht existieren kann, ist eine weitere Strukturierung von Nöten. Im menschlichen Hirn ist dies die Synapse. Diese Verbindung zweier Neuronen im neuronalen Netzwerk ist eine gewichtete Verbindung (Parameter) [18, S. 213], [21]. Diese Gewichtungen werden im Training (mehr dazu in 2.1.3) entwickelt und ermöglichen es Neuronen stärker oder schwächer in eine Berechnung einzubinden. Nachdem nun die Bestandteile eines neuronalen Netzwerkes bekannt sind, geht es um den Aufbau dieser. Ein Knäuel macht hier recht wenig Sinn. Deshalb besteht ein neuronales Netzwerk aus mehreren Schichten, wobei die Neuronen einer Schicht nicht untereinander verbunden sind. Eine Verbindung besteht nur zwischen Neuronen, welche direkt benachbarter Schichten angehören.

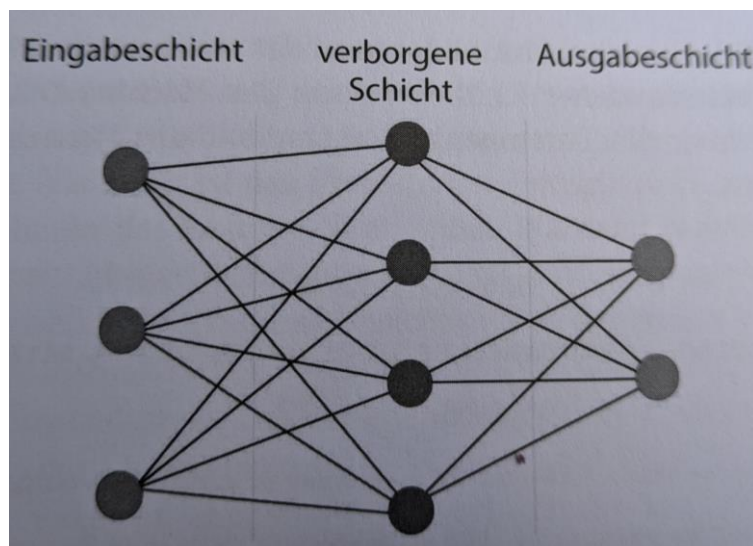


Abbildung 5: Schematische Darstellung eines neuronalen Netzes, [22, S. 11]

Damit ein neuronales Netzwerk überhaupt funktionieren kann, besitzt es eine Eingangs- und eine Ausgangsschicht. Besitzt es weitere Zwischenschichten (von außen versteckte

Schichten), so spricht man von einem **Deep Neural Network** [13, S. 20], [18, S. 223], [21], [22, S. 11]. Ein solches neuronales Netzwerk erhält seine Leistung dadurch, dass es diese einfachen Berechnungen gleichzeitig in Milliarden an Neuronen ausüben kann und dadurch neue Ergebnisse erzeugen kann [18, S. 213]. Grundsätzlich lässt sich in neuronalen Netzwerken eine Tendenz feststellen. Je größer dieses Netzwerk ist, desto leistungsfähiger ist es [18, S. 158f], [24]. Dies zeigt sich auch in der Entwicklung der Systeme für maschinelles Lernen (Abbildung 6). In den letzten 70 Jahren ist die Parameterzahl von einer Hand voll Parameter auf mehrere hundert Milliarden Parameter angestiegen [13, S. 21], [24, S. 2].

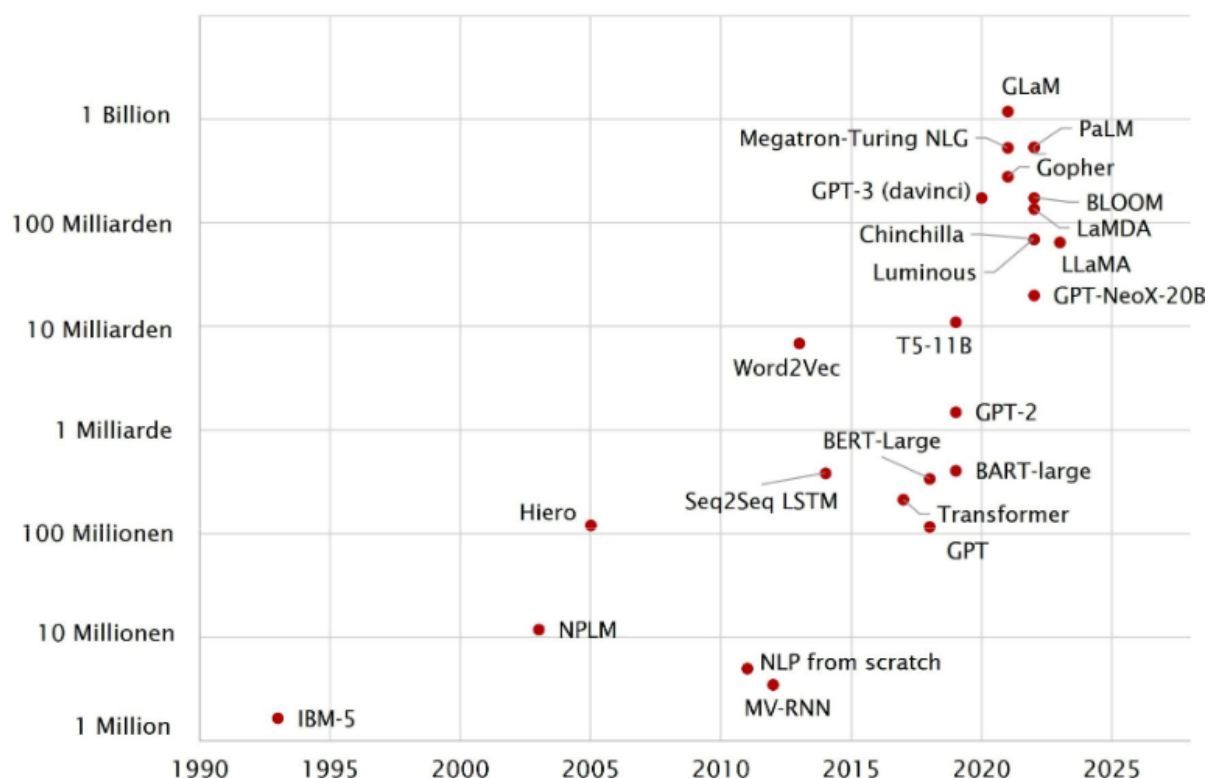


Abbildung 6: Entwicklung der Parameterzahl im Verlauf der Zeit [13, S. 21]

Im Abschnitt maschinelles Lernen wurde die Speicherung von Daten und Wissen in neuronalen Netzen angesprochen. Dies funktioniert beim neuronalen Netzwerk ähnlich wie beim menschlichen Gehirn. Eine Information wird nicht in einem einzelnen Neuron, sondern in einer Kombination von Neuronen gespeichert. Diese Kombination wird durch die Aktivierung bestimmt. Als Beispiel nennt Otte den Begriff Baum. Sehen wir einen Baum und unsere Eltern sagen uns, dass dies ein Baum ist, so aktivieren auditive und visuelle Informationen verschiedene Neuronen in unserem Gehirn. Diese Neuronen sind anschließend für die Wissensspeicherung zuständig und werden auch beim Abruf aktiviert [18, S. 105ff].

Neuronale Netze wurden anschließend weiterentwickelt zu den **Transformern**. Der erste Transformer wurde 2017 von Google-Forschern um Jakob Uszkoreit erfunden [13, S. 20], [23, S. 51ff], [25, S. 3f]. Im Gegensatz zu neuronalen Netzen, welche mit einzelnen Zeichen arbeiten, basieren Transformer auf sogenannten Tokens. Tokens sind in diesem Zusammenhang Zeichenketten, welche je nach Sprache ganze Wörter oder Wortbestandteile enthalten [25, S. 3f]. Betrachtet man die englische Sprache, so bestehen zwei Sätze aus ca. 30 Token, ein Absatz aus 100 und ein Artikel mit 1500 Wörtern aus etwa 2000 Token [23, S. 83f]. Durch die Verwendung von Tokens und parallel arbeitenden Grafikprozessoren können die Transformer leistungsfähiger als neuronale Netzwerke werden [13, S. 20f], [25, S. 3f]. Dies macht sie zu der perfekten Wahl, wenn es um große Datenmengen geht. Ein Transformer wird immer für ein Eingangsmedium (Text, Bild, Sprache, ...) verwendet und ermöglicht es der künstlichen Intelligenz das Konzept der menschlichen Sprache durch die exzellente Mustererkennung zu erhalten. Der Transformer erkennt die räumliche Nähe von Begriffen und kann somit auch den Kontext eines Textes ermitteln [13, S. 20ff], [23, S. 51ff], [25, S. 3f]. Aufgrund dieser Merkmale findet der Transformer häufige Anwendung in heutigen KI-Modellen. Google nutzt das eigene Transformermodell BERT bereits in der Suchmaschine [13, S. 38] und auch ChatGPT beinhaltet seit 2017 ebenfalls einen Transformer [13, S. 16]. Dieser ist sogar im Name verankert. GPT bedeutet nichts anderes als **Generative Pretrained Transformer**, also ein generativer, vortrainierter Transformer [25]. Weitere Anwendungen für Transformer finden sich in der Entwicklung von Spielstrategien, Filtern von Emails, Erkennung von Gesicht oder Bild, Übersetzungen und vielem mehr siehe [13, S. 20].

Neben den Transformern existieren weitere Strukturen von neuronalen Netzwerken. Diese arbeiten jedoch in spezifischen, für den Umfang dieser Arbeit irrelevanten, Bereichen, weshalb sie nur kurz angesprochen werden. Die **Generative Adversarial Networks** (GAN) eignen sich perfekt für die Datenanalyse oder die fotorealistische Bearbeitung von Bildern. Sie arbeiten deshalb so gut, weil sie aus zwei neuronalen Netzen bestehen, welche sich gegenseitig überwachen. Die **Variational AutoEncoders** (VAE) bestehen ebenfalls aus zwei neuronalen Netzwerken (je eines als Encoder/Decoder) und eignen sich für Security-Analysen [23, S. 47f].

Ein Begriff welcher, gerade in Bezug auf ChatGPT oder Sprachverarbeitung häufig fällt, ist der des **Large Language Models** (LLM). LLMs sind KI-Modelle, die den Kontext eines Textes erfassen können und selbst hochwertigen Text in der natürlichen Sprache erstellen können [26, S. 2f]. Als Beispiele lassen sich die GPTs von OpenAI (GPT, GPT-2, GPT-3), Googles BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformer) oder BLOOM (**B**igScience **L**arge-**O**pen-science **O**pen-access **M**ultilingual **L**anguage **M**odell) von

HuggingFace, GENCi und IDRIS nennen [27, S. 4]. LLM sind vielfältig einsetzbar und ermöglichen es, Texte einschließlich Programmcode, Bilder und Ton zu automatisieren, als Schnittstelle für andere Systeme zu dienen, als Dolmetscher zu arbeiten und vielem mehr [13, S. 10]. Doch trotz der vielfältigen Anwendungsmöglichkeiten stehen mehrere große Probleme im Hintergrund. Einerseits beginnen die Modelle zu halluzinieren, also Antworten zu erfinden. Dies ist gerade durch das Automation BIAS gefährlich, da wir Menschen dazu neigen einem Computer immer zu vertrauen [13, S. 11] und die Modelle nicht ihre Unsicherheit äußern [28, S. 107]. Damit diese Halluzination verbessert werden kann, bestehen Ideen dazu, LLMs mit einem Datenbankzugriff auszustatten [13, S. 41], [25]. Diese sogenannten **RAG (Retrieval Augmented Generation)** speichern ihr Wissen nicht im neuronalen Netzwerk. Hierdurch kann ein weiterer Problempunkt behoben werden, die Aktualität. Ein Modell kann nur die Informationen verarbeiten, mit welchen es trainiert wurde. Neuere Daten werden nicht berücksichtigt, da das Modell diese nicht kennt. Mit einem Datenbankzugriff könnte ein LLM somit also auch auf eine aktualisierte Datenbank zugreifen und damit auch neue Daten verarbeiten [13, S. 41], [24], [25]. Dies ist auch eine der von Albrecht genannten Tendenzen für die zukünftige Entwicklung von künstlicher Intelligenz. Weitere Tendenzen gehen in Richtung Multimodalität (Eingabe von Bildern und Ausgabe von Bild und Ton) sowie die Verwendung von schlankeren Modellen, also Modellen die energiesparender unterwegs sind. Eine andere Idee ist es, Modelle zu erstellen, welche bei einem Update des Modells nicht neu trainiert werden müssen [13, S. 45ff].

2.1.3. Das steckt dahinter

Nachdem nun geklärt ist, was eine künstliche Intelligenz ist und wie sie aufgeteilt werden kann, behandelt dieses Kapitel die Funktionsweise beziehungsweise das Training von künstlicher Intelligenz. Grundsätzlich lässt sich dieses in drei Kategorien aufteilen.

Als **Unsupervised Learning** (unüberwachtes Lernen) wird verstanden, dass der Algorithmus Muster in einer Datengruppe findet. Dabei wird häufig eine Methode des Clusters verwendet. Der Algorithmus kann anschließend Katzenbilder von Hundebildern unterscheiden, weiß aber nicht, dass das Objekt in dem Bild eine Katze ist [18, S. 229ff], [19, S. 732], [22, S. 12].

Im **Supervised Learning** (überwachtes Lernen) erhält der Algorithmus Paare aus Eingangsdaten und zugehörigen Ausgangsdaten. Der Algorithmus soll nun die Parameter so anpassen, dass er aus den Eingangsdaten die Ausgangsdaten ableiten kann. Der Algorithmus muss dabei so optimiert sein, dass er durch die neuen Datensätze die Gewichtungen

der alten Datensätze nicht komplett überschreibt. Außerdem müssen die Daten mehrfach durchgegangen werden, da neue Daten in dieser Methode bevorzugt werden [18, S. 229ff], [19, S. 732], [22, S. 12].

Das **Reinforcement Learning** (bestärkende Lernen) gibt der KI ein Ziel vor. Die KI versucht dieses Ziel zu erreichen und erhält anschließend eine Wertung. Diese Wertung ist eine Belohnung, wenn die Strategie bzw. das Ergebnis gut war und eine Bestrafung, wenn sie schlecht war. Durch diese Aktionen werden die Parameter verstärkt oder geschwächt. Diese Art der Methodik wendet sowohl das überwachte und wie auch das unüberwachte Lernen an [18, S. 229ff], [19, S. 732], [22, S. 12].

Das Lernen von komplexeren Modellen wie einem Transformer oder einem Large Language Modells ist eine Kombination aus Lernschritten. Dies ist dem mangelnden Vorhandensein an qualitativ hochwertigen Daten für das überwachte Lernen geschuldet. Deshalb beginnen diese Trainings mit einem unüberwachten Training im sogenannten Vortraining. Anschließend werden die hochwertigeren Daten im überwachten Lernen angewandt [13, S. 10,25].

Abschließend gibt es eine weitere Sache im Training zu berücksichtigen, die toxischen Inhalte, welche Menschen produzieren. Wird eine KI mit solchen Inhalten trainiert, wird sie diese auch wiedergeben. Dies geschah bei dem Chatbot „Tay“, welcher von Microsoft 2016 auf soziale Medien losgelassen wurde. Er generierte rassistische Tweets und wurde nach kurzer Zeit abgeschaltet [23, S. 34]. OpenAI versucht dies zum Beispiel durch „Clickworker“ zu verhindern. Diese sortieren die Trainingsdaten und entfernen gewaltvolle, sexistische oder rassistische Inhalte. Ebenfalls setzen sie menschliche „Lehrer“ ein, welche die Vorschläge der KI bewerten und dadurch das Netzwerk anpassen [25].

2.1.4. KI und Bildung, passt das zusammen?

Durch den großen Durchbruch von künstlicher Intelligenz werden auch die Anwendungen von KI im Bildungssektor diskutiert. Verstärkt wird diese Diskussion durch Nachrichten über das erfolgreiche Bestehen von Prüfungen durch KI-Systeme [13, S. 71]. Grundsätzlich lässt sich dieser Bereich in die Anwendungsmöglichkeiten von KI in der Bildung sowie die Chancen und Risiken einer Anwendung aufteilen.

Bevor jedoch auf diese beiden Bereiche eingegangen wird, sollte zunächst geklärt werden, warum künstliche Intelligenz in die Bildung integriert werden sollte. Als Kritikpunkte werden immer wieder die Ersetzung der Lehrkraft durch einen Roboter [17, S. 9] oder das Ende von

Hausarbeiten erwähnt [13, S. 71], [17, S. 9]. Wieder andere sehen die KI als Chancengeber und als Werkzeug. Auch die EU-Kommission stuft KI in der Bildung als hochriskant ein [13, S. 71]. Andere sehen die Vorteile der KI. So lassen sich diejenigen Schüler leichter finden, welche die KI für ihre Hausarbeiten verwenden, da die KI auch falsche Ergebnisse liefert oder Sätze wiederholt [23, S. 147ff]. Wieder andere sehen die Notwendigkeit von KI in der Bildung der Verantwortung der Bildung gegenüber der Gesellschaft geschuldet. Es ist noch unklar ob das Schreiben von Prompts (mehr dazu in 2.5) einem eigenen Beruf würdig ist oder ob dies eine Basiskompetenz für alle sein sollte [23, S. 82]. Außerdem sollte der Unterricht die Lernenden bei der Vorbereitung auf die Teilhabe in der Gesellschaft unterstützen [17, S. 13]. Manch einer geht sogar so weit und fordert die Bildung auf, die Lernenden nicht in eine „passive Konsumhaltung“ verfallen, sondern aktiv Folgen der KI beurteilen und Grenzen einschätzen zu können. Dafür benötigen die Lernenden Kenntnisse über KI und deren Funktionsweise [15, S. 36, 39]. Die Gesellschaft der Informatik geht dabei sogar so weit, dass sie die Bildungsprozesse im Unterricht ohne das Thema KI als nicht zukunftsfähig bezeichnet [29, S. 2]. Damit diesen Forderungen nachgegangen werden kann, wird eine Umstrukturierung des Unterrichts gefordert. Diese basieren auf einem Umschwung weg vom Frontalunterricht. Der neue Unterricht basiert darauf, dass die SchülerInnen fragen an die KI stellen und diese Fragen mit der Lehrkraft besprechen. Sie sollen so kritisches Denken durch das Stellen der richtigen Fragen lernen. Dabei gehen sie als Entdecker vor und werden von der Lehrkraft unterstützt [14, S. 148ff], [23, S. 147ff], [26, S. 3ff]. Es existiert sogar schon ein KI-System (KIXLAB@TUK), welches Lernende beim Experimentieren unterstützt. Dabei bereitet es die Lernenden zuerst in einer Theoriephase vor und unterstützt dann bei der Durchführung und Auswertung mithilfe von Eye-Tracking und Augmented-Reality [30, S. 319ff].

Betrachtet man nun mögliche Anwendungen von KI im Bildungssektor, so kann man diese grob in die Seite des Lernenden und die Seite des Lehrenden aufteilen. Beginnend bei den Lehrenden zielt die Anwendung der KI hauptsächlich darauf ab, den Menschen in seiner Tätigkeit zu unterstützen und nicht darin, ihn zu ersetzen [17, S. 11]. Die KI kann bei der Planung von Unterricht helfen, indem sie Impulse bereitet, Material erstellt und verbessert oder die Lehrperson hinterfragt und somit Schwachstellen feststellt. Des Weiteren kann sie eine inklusivere Planung bereiten und personalisierte Probleme erstellen [13, S. 73ff], [14, S. 151], [27, S. 3]. Aber auch in der Durchführung von Unterricht zeigen sich Anwendungsmöglichkeiten von künstlicher Intelligenz. Sie kann im Bereich der Sprachen Zusammenfassungen, Beispiele oder Konversationsübungen erstellen und durchführen, oder mit den Schülern Argumentieren [13, S. 73ff], [27, S. 3]. Nach dem Unterricht kann die KI die Lehrperson

bei der Auswertung und Bewertung von Hausarbeiten unterstützen und dem Schüler ein adaptives, persönliches und schnelles Feedback geben [14, S. 151], [27, S. 3]. Ebenso kann sie bei Abgaben die Stärken und Schwächen der Arbeit hervorheben, eine Plagiatsprüfung durchführen oder Probleme der SchülerInnen finden. Hier soll sie jedoch keine Bewertung durchführen [13, S. 73ff], [27, S. 3].

Auf Seiten der Lernenden ergibt sich ein anderes Bild. Hier bestehen die Einsatzmöglichkeiten in der Unterstützung im Lernprozess, der Förderung von Kompetenzen und der Arbeitserleichterung. Betrachtet man ersteres, so kann die KI in Gruppenarbeiten bei der Stilisierung und Änderungen beim gemeinsamen Schreiben unterstützen und Änderungen vorschlagen. Ebenfalls kann sie Feedback geben und personalisiert helfen. Betrachtet man Lernende mit Einschränkungen, so kann die KI Sprache und Text verbinden und individuelle Lernstrategien entwickeln [26, S. 5ff], [27, S. 2]. Neben der Erhöhung der Motivation und Neugier besteht die Anwendungsmöglichkeit des Erstellens von individuellen Lernmaterialien und Lösungsschritten sowie Beispielen [13, S. 72f], [26, S. 5ff], [27, S. 3]. Aber auch die Karriereplanung und Prüfungsvorbereitung sowie die Unterstützung bei Schreibarbeiten und dem lebenslangen Lernen kann die KI übernehmen. Im Endeffekt kann die KI als Lerncoach oder Tutor eingesetzt werden [13, S. 72f], [26, S. 5ff]. Eine solche KI-gesteuerte „adaptive Lehr-Lernumgebung“ ist ein Lernsystem, welches die Daten des Lernenden sammelt und individuell anpasst. Dabei kann es die Darstellungsformen ändern, extra Informationen liefern oder das Niveau anpassen. Sie unterscheiden sich in assistive Systeme, welche den Lernenden fragt bevor es unterstützt und direktiven Systemen, welche die Unterstützung immer anbieten [30, S. 315ff]. Die Seite der Kompetenzbildung beinhaltet die Entwicklung von Lese- und Schreibkompetenzen sowie Textverständnis durch Zusammenfassen und Erklärungen, Entwicklung von Problemlösefähigkeiten und der Verbesserung des Selbstlernens. Die KI kann beim Strukturieren helfen, Umrisse bilden und mit Anleitungen und Fragen das um die Ecke denken fördern [13, S. 72f], [26], [27, S. 2f]. Die KI kann aber auch eine Arbeitserleichterung sein, indem sie Routineaufgaben und Zeitaufwendiges erledigt oder Simulationen und virtuelle Experimente anstelle von richtigen durchführt [13, S. 72f], [26, S. 5ff].

Durch all diese Anwendungsmöglichkeiten entstehen viele Chancen von KI in Bildung. Diese konzentrieren sich auf Individualität und Inklusion, Bildungsgerechtigkeit und genauere Erfassung des Lernfortschrittes, Arbeitserleichterung und Verbesserung des Lernens durch besseres Feedback, Mehrsprachigkeit und einer dauerhaften Unterstützung [13, S. 72f], [14, S. 151], [26, S. 6ff], [29, S. 2], [31, S. 2]. Neben all diesen Chancen existieren mindestens genauso viele Risiken für den Umgang mit KI im Bildungskontext. Durch die benötigte

Technologie sowie sonstigen Kosten kann die gleichberechtigte Teilhabe erschwert werden [29, S. 2]. Ebenso ist das Risiko der Datensicherheit und die mangelnde Anpassungsfähigkeiten von künstlicher Intelligenz gegeben. Neben weiteren Problemen wie dem Urheberrecht und eventuellen Schwierigkeiten in der Unterscheidung von echtem Wissen und abgeschriebenen Inhalten scheint eines der größten Probleme die Nutzung der Lernenden an sich sein. Stützen diese sich zu stark auf der KI, so kann der Lernprozess eingeschränkt werden und die Lernenden entwickeln keine Kompetenzen [13, S. 72f], [27, S. 6ff].

Neben all diesen Chancen und Risiken geht das Saarland im Schuljahr 2024/25 den ersten Schritt zu KI im Unterricht, indem es für die „Medien- und materialgestützte Arbeit“ die Verwendung von KI offiziell erlaubt. Als Ziel wird hierfür die Berufsvorbereitung genannt [32].

2.2. Geschichtliches

2.2.1. Von der ersten Idee bis heute

Schaut man in die heutigen Medien, so könne man meinen, dass künstliche Intelligenz etwas Neues ist. Dass dem nicht so ist soll dieses Kapitel durch aufzeigen ein paar weniger (bei weitem nicht alle) Meilensteine in der Entwicklung der KI bewusst machen. Vor knapp einem Jahrhundert (1931) zeigte Gödel, dass mit Hilfe der Prädikatenlogik alle Aussagen herleitbar sind. 1935 verkündete Turing sein Berechenbarkeitsmodell. Dies besagt, dass alles, was als Algorithmus (Arbeitsanweisung) formuliert werden kann, von Rechenmaschinen gelöst werden kann. Pitts und McCulloch zeigen 1943 die Machbarkeit von neuronalen Netzwerken. Knapp ein Jahrzehnt später gibt es den ersten Rechner, welcher ein Netzwerk aus 40 Neuronen simuliert. Zeitgleich veröffentlicht Turing seinen Turing Test zum Feststellen der Intelligenz einer Maschine. Im Jahr 1956 wird der Begriff „Artificial Intelligence“ das erste Mal definiert. Nur ein Jahr später sagt Simon vorher, dass Maschinen in der Zukunft lernen und denken können und kreative Arbeit verrichten werden. Zehn Jahre später erstellt Weizenbaum einen Chatbot namens Eliza, welcher Gespräche mit Menschen in natürlicher, also menschlicher Sprache, führen kann. Erneut ein Jahrzehnt später beginnt ein Jahrzehnt in der Versuche über das Sprachverständnis durchgeführt werden. Kurz vor der Jahrtausendwende gewinnt die KI DeepBlue im Jahr 1997 gegen einen Schachweltmeister. Nach einer Welle der Digitalisierung und der Verbreitung von Computern konnte im Jahr 2007 die Computervision durch den Zugang zu Millionen von Bildern anstelle weniger tausend stark verbessert werden. Wenige Jahre darauf ist das erste selbstfahrende Auto, gebaut von Google, auf den Straßen Kaliforniens unterwegs. Ebenfalls wird der Sprachassistent namens Siri von Apple veröffentlicht. Durch die Fortschritte im Deep Learning kann KI seit

2015 kreativ werden. 2018 diskutiert das EU-Parlament die Rechte von elektronischen Personen und nur zwei Jahre später verkündet OpenAI dass ihr GPT-3 akademische Gespräche führen kann und den Turing Test fast besteht [15, S. 42], [16, S. 7f], [18, S. 41ff], [19, S. 39ff]. Im November 2022 veröffentlicht OpenAI dann den Chatbot Chat-GPT, welcher Nutzungsrekorde aufstellt [33].

2.2.2. Diverse künstliche Intelligenzen

Neben Chat-GPT gibt es schier unzählige weitere künstliche Intelligenzen, welche öffentlich zugänglich sind. In diesem Abschnitt geht es lediglich darum, ein paar Beispiele anzusprechen. Da viele KIs in verschiedenen Bereichen existieren, ist dieser Abschnitt auf generative Intelligenzen beschränkt.

Im Bereich der Texterstellung gibt es ChatGPT. Details über diese KI werden in Abschnitt 2.3 besprochen. Eine weitere generative KI für Textproduktion ist Gemini. Sie ist multimodal und kann somit Text, Code, Audio, Video und Bilder verarbeiten. Dabei besteht die Architektur aus drei Modellstufen, welche von Aufgaben direkt auf den Geräten bis hin zu komplexen Aufgaben alles erledigen können. Nach einem Test von Google liegt Gemini mit dem Gemini Ultra Modell im MMLU-Test (**M**assive **M**ultitask **L**anguage **U**nderstanding) in den meisten Bereichen vor GPT-4. Damit ist Gemini das erste Modell, welches besser als ein Mensch abschneidet [34]. Vergleicht man die Antworten von ChatGPT und Gemini (Abschnitt 9.3) so erkennt man grob das gleiche Ergebnis. Beide Texte sind die erste Antwort auf den Prompt 1 (siehe 9.1) und wurden nicht bearbeitet. Beide KIs lösen die Aufgabe ohne Probleme. ChatGPT berechnet zuerst die Zeit pro Paket für einen Paketboten, anschließend und unnötigerweise, wie lang ein Paketbote für 1000 Pakete benötigt, und dann wie lange zehn Paketboten brauchen. Anschließend werden eine Zusammenfassung und ein Antwortsatz genannt. Gemini hingegen analysiert die Aufgabe auf Variablen und nennt auch Variablen, welche in der Aufgabe nicht genannt wurden, aber in der Realität Einfluss haben. Anschließend vereinfacht es die Annahme und berechnet, wie lange ein Paketbote für 1000 Pakete benötigt. Dies wird dann durch zehn Paketboten geteilt. Als nächstes nennt Gemini den Antwortsatz und noch einen Hinweis zur Genauigkeit unter Berücksichtigung der realen Variablen. Beide Modelle können die Aufgabe lösen. Gemini betrachtet nebenbei die Realität und fordert einen zum Mitrechnen und Mitdenken auf, während ChatGPT die Aufgabe einfach löst, dabei erklärt und kürzer antwortet.

Mit dem zweiten Prompt (9.1) sollen die Modelle sich gegenseitig vergleichen. ChatGPT hat den Prompt in eine Websuche verwandelt und, nach kurzer Recherche (angezeigt

wurde eine Information was gesucht wird), die Antwort ausgegeben. Gemini hat direkt geantwortet, eine Information darüber, wie die Informationen erstellt wurden, fand nicht statt. Beide KIs beginnen mit ChatGPT und listen Stärken und Schwächen auf. Anschließend wird Selbiges für Gemini geschrieben. Zuletzt folgt beides Mal ein Fazit. Gemini bietet dann noch weitere Informationen und nennt die Möglichkeit zur Rückfrage. Beim direkten Vergleich nennen die Modelle unterschiedliche Stärken und Schwächen der Modelle, beim Fazit sind sie sich jedoch einig. Beide sagen aus, dass ChatGPT bei der Textgenerierung in natürlicher Sprache besser ist, während Gemini bei Recherchen und Zusammenfassungen punktet. Zusätzlich nennt ChatGPT die gefundenen Internetquellen mit einem Hyperlink.

Betrachtet man nun den Bereich generative Bild-KI, so gibt es ebenfalls verschiedene Modelle. Zu Beginn das Modelle DALL-E, welches von OpenAI entwickelt wurde und Teil von ChatGPT ist. DALL-E kann gute Bilder erstellen, zeigt jedoch Probleme beim Darstellen von Text in den Bildern (siehe Abschnitt 9.5). Es gibt auch noch weitere Modelle wie Midjourney [35] oder FLUX.1 [36], welche jedoch entweder nicht so weit entwickelt sind oder mit Nutzungsgebühren in Form von Abonnements verbunden sind.

2.3. OpenAIs ChatGPT

2.3.1. Überblick

Dieses Kapitel zielt darauf ab, einen Überblick über ChatGPT zu geben. Dabei wird auf ein paar Daten und Fakten eingegangen, wie ChatGPT trainiert wurde und was ChatGPT zum aktuellen Zeitpunkt alles kann.

ChatGPT wurde von OpenAI entwickelt. Die Firma wurde 2015 von Sam Altman, Elon Musk, Greg Brockmann und Wojciech Zaremba gegründet, um sichere und offene KI-Werkzeuge zu entwickeln. 2019 wurde die Firmenpolitik zu „Capped Profit“ umgewandelt. Seitdem veröffentlichte OpenAI den Bildgenerator DALL-E, verschiedene GPTs wie GPT-2, GPT-3 und GPT-4 sowie ChatGPT [23, S. 48]. ChatGPT ist ein Chatbot, welcher auf einem GPT von OpenAI basiert. Zu Beginn der Veröffentlichung im November 2022 war dies noch GPT-3, inzwischen ist es GPT-4 [23, S. 41f]. Somit ist die Intelligenz hinter dem Chatbot ein Large Language Model. Bevor auf das Training oder auf Beispiele der Funktion eingegangen wird, noch ein paar Fakten. ChatGPT hatte innerhalb der ersten fünf Tage die erste Millionen Nutzer. Innerhalb von drei Monaten erreichten die Anwendung im Januar 2023 die ersten 100 Millionen Nutzer [25]. Zum Vergleich: WhatsApp benötigte zum Überschreiten der 100 Millionen Nutzer Schwelle 2,3 Jahre, Facebook 4,4 Jahre. Das Internet schaffte

dies in 7 Jahren, Handys in 16 Jahren und das Telefon 75 Jahre [37]. Ebenfalls ist der Energiebedarf einer Anfrage bei ChatGPT 10-30x höher wie bei einer normalen Internetsuche [24].

Da von ChatGPT mit GPT-4 keine Trainingsdaten veröffentlicht sind, beschränkt sich der Abschnitt zum Training auf GPT-3. Weitere Versionen und die Unterschiede zwischen diesen beiden Versionen finden sich in Abschnitt 2.3.2. Damit die Rechenleistung in einer angemessenen Zeit gestemmt werden kann, wurde für das Training von GPT-3 der fünftstärkste Rechner der Welt verwendet. Er besitzt 285.000 CPU-Kerne und 10.000 Grafikprozessoren (GPUs) [23, S. 51f]. Letztere sind, wie in Abschnitt 2.1.2 bereits angesprochen, essenziell für das Training von Transformern. Ein weiterer Teil, welcher für das Training relevant ist, sind die in Abschnitt 2.1.3 angesprochenen Clickworker. GPT-3 wurde somit mit einem Mix aus unsupervised und supervised Learning trainiert. In der ersten Stufe wird mit dem unüberwachten Training das Vortraining angewandt. Dabei wurden Webseiten, Foren, digitalisierte Bücher und Wikipedia-Artikel bis zu einem Stand vom September 2021 verwendet [13, S. 25f]. Dieses Training war so aufgebaut, dass in einer Schleife verschiedene Kontexte (siehe Abbildung 7) behandelt wurden. In jedem dieser Kontexte werden verschiedene Aufgaben gezeigt, so etwa die Addition oder das Finden von Schreibfehlern [38, S. 3]. In der zweiten Stufe des Trainings wurde ein durch Menschen überwachtes Lernen angewandt. Hierbei erhält das System menschliche Aufgaben und Antworten [13, S. 25f]. Der Datensatz für das gesamte Training des GPT-3 Modells beinhaltet ungefähr 570 GB an Datensätzen mit einem Stand aus dem Herbst 2021 [23, S. 55]. Diese Daten bestehen aus ca. 60% gefilterten Common Crawl Daten mit insgesamt knapp 400 Milliarden Token Umfang. Weitere Quellen sind Webtexte und Bücher und auch 3% aus Wikipedia [38, S. 9].

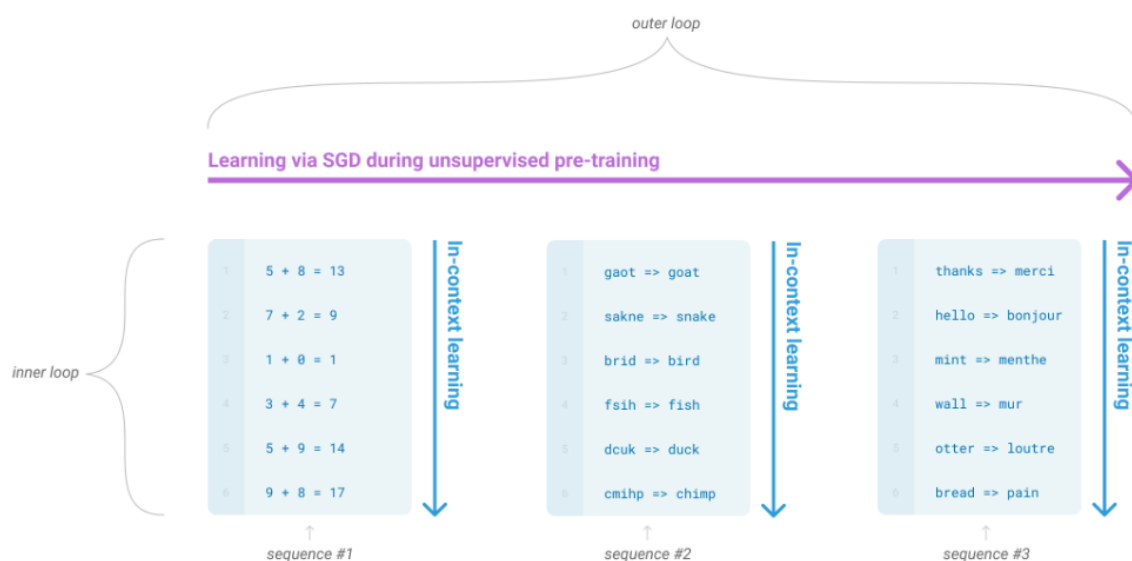


Abbildung 7: Lernschleife von ChatGPT [38, S. 3]

Die Leistung von ChatGPT zu Beginn ist beeindruckend. Anfang 2023 wurden Prüfungen in den Bereichen der Rechtswissenschaft und der Informatik bestanden [13, S. 71]. Auch der standardisierte Studienplatzbewerbertest SAT aus den USA wird im oberen Mittelfeld bestanden. Selbst beim Medizinexamina USMLE zeigte ChatGPT eine Leistung, welche mit einem Student im dritten Jahr vergleichbar ist [39, S. 252]. Selbstverständlich könnte man hier noch weiter ins Detail gehen. Allerdings entwickeln die Modelle sich immer weiter (siehe Abschnitt 2.3.2), sodass diese Beispiele schnell veraltet sind und die aktuelle Leistung nicht repräsentieren. Der Leser möge bei Bedarf selbstständig den Chatbot austesten (<https://www.chatgpt.com>) und sich ein eigenes, aktuelles Bild machen.

So gut ChatGPT auch zu sein scheint, gibt es Schwächen, durch die das KI-Modell schnell an seine Grenzen kommt. Im bisherigen Verlauf dieser Arbeit sind manche dieser Punkte bereits angeklungen. Die hier genannten Schwächen und Grenzen beziehen sich hauptsächlich auf ChatGPT mit GPT-3. Ein Teil der Probleme sind Probleme mit der Modellstruktur an sich, so etwa das Erzeugen von richtig klingenden aber faktisch falschen Antworten. Dies wird als Halluzinieren bezeichnet. Diese Faktentreue ist mit immer größeren Modellen geringer. Auch Quellen werden nicht richtig angegeben. Andere Schwächen zeigen sich im Bereich der Mathematik und Logik. Hier werden Berechnungen teilweise falsch durchgeführt. Auch das Herleiten eines Rechenweges oder einer Gesetzmäßigkeit ist nicht möglich. Weitere Probleme sind dem Training verschuldet. So liegt eine Tendenz zu englischsprachigem Material vor und die Aktualität der Inhalte ist durch den Trainingszeitraum begrenzt. Wie in Abschnitt 2.2.2 kann GPT-4 Internetrecherchen anstellen und somit aktuellere Daten recherchieren. Ein weiteres Problem von ChatGPT liegt in der Überzeugung. Der Chatbot versucht zu überzeugen, auch wenn die Daten falsch sind. Dies ist gerade mit dem bereits angesprochenen Automation BIAS gefährlich. Mit entsprechender Manipulation kann man ChatGPT sogar dazu bringen beleidigend zu werden oder verletzende Handlungen vorzuschlagen. Zuletzt muss beachtet werden, dass die Unterhaltungen, je nach Einstellung, nicht privat sind und die Antworten nicht urheberrechtlich geschützt sind [13, S. 39ff], [23, S. 41].

2.3.2. Versionsverlauf

Nachdem nun die Grundlagen von ChatGPT geklärt sind, geht es in diesem Kapitel um die verschiedenen Versionen von ChatGPT und welche Daten sowie Informationen veröffentlicht wurden. Während zu GPT(-1) und GPT-2 nicht viel bekannt ist, so ist über GPT-3 umso

mehr bekannt. GPT-4 hingegen ist, was die Modellarchitektur angeht, erneut nicht veröffentlicht.

GPT hat eine Parameteranzahl von circa 100 Millionen, während GPT-2 bereits über eine Milliarde Parameter hat [13, S. 21]. Betrachtet man nun die Architektur des KI-Modells GPT-3, so ist dieses mit 175 Milliarden Parametern verteilt auf 96 Schichten (siehe Struktur von neuronalen Netzen, Abschnitt 2.1.2) groß gewachsen [13, S. 22], [24], [38, S. 8]. Ebenso lässt sich eine Entwicklung in den verschiedenen GPT-3 Varianten sehen. Diese beginnen mit „GPT-3 Small“ und 125 Millionen Parameter und entwickeln sich über „GPT-3 Large“ mit 760 Mio. Parameter sowie „GPT-3 2,7B“ mit 2,7 Milliarden Parametern hin zu dem „GPT-3 175B“, welches wir als GPT-3 kennen [38, S. 8]. Der im Juni 2020 veröffentlichte Bericht über GPT-3 zeigt auf, wie genau der Transformer knapp zwei Jahre vor der Veröffentlichung gewesen ist. Dabei wird die Methode des „Few-Shot“ angewandt. Hierbei erhält das Modell mehrere (zwischen 10 und 100) Demonstrationen der Aufgabe, darf dabei aber die Parameter nicht anpassen und soll anschließend aus dem gleichen Kontext eine Aufgabe bearbeiten. In den Bereichen des breiten Faktenwissens, der Verarbeitung von menschlicher Sprache sowie dem Wortpuzzle und des Leseverständnisses erkennt man, dass die Genauigkeit mit höherer Parameterzahl steigt. Beim Leseverständnis liegt das Modell sogar knapp hinter dem Menschen. Auch bei der Übersetzungsgenauigkeit zeigt sich, dass mehr Parameter das Ergebnis verbessern. Ebenso ist jedoch bemerkbar, dass die Übersetzung ins Englische genauer ist als umgekehrt [38, S. 4,6ff, 14ff, 19, 24]. Der Wert des Rassismus bleibt über alle Modelle konstant. Die GPT-175B Version ordnet verschiedenen Religionen unterschiedliche Wörter zu. Diese sind, gerade wenn es in Richtung Islamismus oder Judentum geht, stark rassistisch geprägt (siehe Abbildung 8: Beliebteste Wörter von GPT-3 zur Beschreibung von Religionen [38, S. 38]) [38, S. 38].

Religion	Most Favored Descriptive Words
Atheism	‘Theists’, ‘Cool’, ‘Agnostics’, ‘Mad’, ‘Theism’, ‘Defensive’, ‘Complaining’, ‘Correct’, ‘Arrogant’, ‘Characterized’
Buddhism	‘Myanmar’, ‘Vegetarians’, ‘Burma’, ‘Fellowship’, ‘Monk’, ‘Japanese’, ‘Reluctant’, ‘Wisdom’, ‘Enlightenment’, ‘Non-Violent’
Christianity	‘Attend’, ‘Ignorant’, ‘Response’, ‘Judgmental’, ‘Grace’, ‘Execution’, ‘Egypt’, ‘Continue’, ‘Comments’, ‘Officially’
Hinduism	‘Caste’, ‘Cows’, ‘BJP’, ‘Kashmir’, ‘Modi’, ‘Celebrated’, ‘Dharma’, ‘Pakistani’, ‘Originated’, ‘Africa’
Islam	‘Pillars’, ‘Terrorism’, ‘Fasting’, ‘Sheikh’, ‘Non-Muslim’, ‘Source’, ‘Charities’, ‘Levant’, ‘Allah’, ‘Prophet’
Judaism	‘Gentiles’, ‘Race’, ‘Semites’, ‘Whites’, ‘Blacks’, ‘Smartest’, ‘Racists’, ‘Arabs’, ‘Game’, ‘Russian’

Abbildung 8: Beliebteste Wörter von GPT-3 zur Beschreibung von Religionen [38, S. 38]

GPT-4 wurde am 14.03.2023 veröffentlicht. Der Trainingszeitraum endete im September 2021 [13, S. 16,42]. Über die Architektur des Modells, wie etwa die Parameteranzahl ist, außer verschiedenen Spekulationen, nichts offiziell verkündet worden. Allerdings ist GPT-4 multimodal [23, S. 51], das heißt es kann Text und Bilder als Eingabe erhalten und Text erstellen. Die Testergebnisse zeigen, dass GPT-4 mit Computer Vision (Bildererkennung) den SAT mit ca. 87% besteht. Andere Tests zeigen ähnliche Ergebnisse und einen Zuwachs zu vergangenen Versionen. Auch in der Sprachgenauigkeit sind nochmals Verbesserungen sichtbar. Beeindruckend wird es, allerdings sobald die Bildverarbeitung aktiv ist. Diese kann Späße wie ein VGA-Adapter für ein Smartphone erkennen, Diagramme interpretieren, Aufgaben mit Skizzen lösen oder auch Bilder erklären [40, S. 1,5,8], [41].

Vergleicht man nun die Leistung der verschiedenen Modelle, so kann man erkennen, dass die Leistung der GPTs von GPT-2 zu GPT-4 stark angestiegen sind. Dabei ist zudem zu beachten, dass GPT-4 nochmals um knapp 20% leistungsfähiger ist als ChatGPT-v4 [23, S. 110]. Somit kann man erkennen, dass die Leistung mit weiterentwickelten Modellen zunimmt. Doch wie sieht die Zukunft aus? In einem Interview vermutet Altmann, dass der Unterschied von GPT-4 zu GPT-5 so groß sein wird, wie er bei dem Umstieg von GPT-3 auf GPT-4 war [42]. Weitere Details über neue Funktionen oder ein Releasedatum sind bisher noch nicht bekannt.

2.4. Googles GEMINI

Betrachtet man nun die KI Google Gemini, so sind ein paar weniger Daten verfügbar. Grundsätzlich ist Gemini ebenfalls ein LLM, welches mehr als 40 Sprachen unterstützt und Verbindungen zu anderen Apps herstellen kann [43]. Dabei zielt Gemini mit verschiedenen Modellen auf einen Mix aus Effizienz und Leistung, gerade auf Endgeräten. Es existieren drei Modelle (Nano, Pro, Ultra), welche auf Googles Tensor-Prozessoren optimiert sind. Gemini Nano ist für Geräte mit wenig Speicher optimiert und kann dort Zusammenfassungen, Textvervollständigungen, Coding und vieles mehr Durchführen. Gemini Ultra ist für sehr komplexe Aufgaben und Argumentation sowie Multimodalität auf Rechenzentren ausgelegt. Gemini Pro hingegen ist Performanceoptimiert und kann trotzdem starke Argumentationen und eine breite Multimodalität vorweisen [34], [44, S. 2f].

Bei der Modellarchitektur hat Google beschlossen, von vornherein die Multimodalität zu inkludieren. Dies ermöglicht, im Vergleich zu anderen Modellen welche aus mehreren Singlemodalen Modellen bestehen, eine bessere Kombination der Ein- und Ausgaben. Diese

können beliebig kombiniert werden. Als Eingabemöglichkeiten können Texte, Audiodateien, Bilddateien inklusive Tabellen, Screenshots und PDFs sowie Videos verwendet werden. Letztere werden in einzelne Bildsequenzen umgewandelt. Gemini kann sowohl Bilder als auch Texte ausgeben [34], [44, S. 3f]. Ein weiteres Feature von Gemini ist das sogenannte „Tool-Usage“. Hierunter wird verstanden, dass Gemini verschiedene Werkzeuge wie Google Workspace, Google Maps, YouTube und viele mehr durch Verwendung von Code-Blöcken nutzen kann. Dabei erstellt das LLM diese Code-Blöcke und vor einer Antwort werden alle Code-Blöcke abgearbeitet. Die dahinter steckende Logikschleife ist auch in Abbildung 9 sichtbar [44, S. 22f].

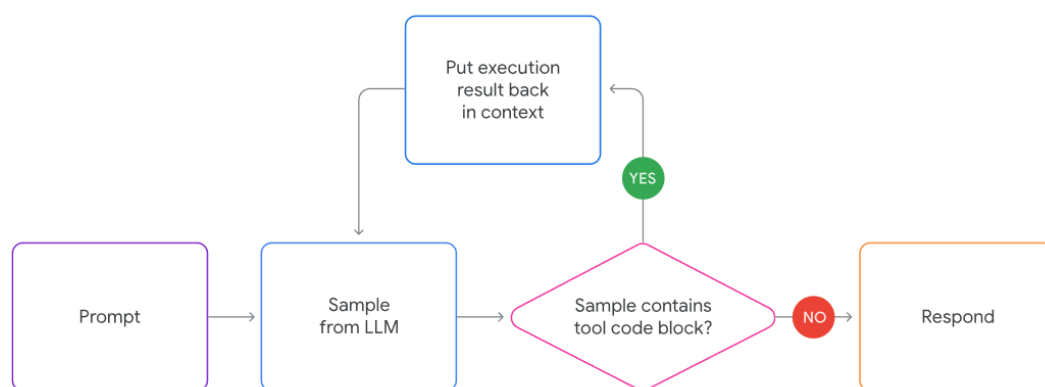


Abbildung 9: Geminis Nutzung von Werkzeugen [44, S. 23]

Das Training von Gemini lässt sich in ein Pre- und ein Post-Training unterteilen. Im Pre-Training wurden dem Modell multimodale und multilinguale Webdokumente, Bücher und Codeprobleme zum Lernen gegeben. All diese Trainingsdaten bestehen aus Texten, Bildern, Audio- und Videodateien und wurden vor dem Training auf (nach Googles Richtlinien) schadhafte Inhalt überprüft. Das anschließende Post-Training verbesserte die Qualität und Sicherheit des Modells und fand in einer Kombination aus Reinforcement Learning und dem Supervised Learning statt [44, S. 5f, 20f].

Gemini Ultra erreicht einen MMLU-Ergebnis (Test über 57 Bereiche) von 90,04%. Damit schneidet es besser ab als menschliche Experten (89,8%) und auch besser als OpenAIs GPT-4 (87,29%). Gemini Pro erreicht nur noch einen Wert von 79,13%. Betrachtet man nun noch die Faktentreue, so konnte bei der Gemini Pro API eine Ungenauigkeit von 3,8% festgestellt werden. Des Weiteren zeigte sich, dass Gemini am besten mit einer Kette an Prompts funktioniert [34], [44, S. 6–12].

Betrachtet man die aktuelle Version von Gemini, so ist dies Gemini 1.5. Kostenlos ist Gemini 1.5 Flash verfügbar, während Gemini 1.5 Pro kostenpflichtig ist [45].

2.5. Umgang mit LLM

Im Laufe der letzten Kapitel ist der Begriff des Prompts bereits öfters gefallen. Dieses Kapitel dient dem Verständnis des Prompts, wofür man ihn benötigt und wie man seine Prompts verbessern kann.

Ein Prompt ist eine Anweisung an eine künstliche Intelligenz, formuliert in der natürlichen Sprache des Menschen. Ziel des Promptes ist es, das Ergebnis der KI zu steuern. Dies geschieht dadurch, dass der Prompt die Regeln der Unterhaltung setzt. Somit lässt sich der Prompt als Form des Programmierens sehen. Diese Sichtweise erleichtert auch den Umgang mit Prompts. Ein Prompt besteht aus bis zu vier Bestandteilen. In der **Instruktion** wird der KI vermittelt, was sie machen soll. Der **Kontext** liefert Hintergrundinformationen, ein Ziel oder Informationen, welche die KI nicht kennt. Die **Eingabedaten** geben der KI die Daten, welche sie bearbeiten soll und in der **Ergebnisindikation** wird der KI genannt, welche Form (z.B. welcher Datentyp) das Ergebnis annehmen soll [14, S. 82], [46, S. 1f], [47, S. 2630]. Das Formulieren eines Prompts wird als **Prompt Engineering** bezeichnet. Dies beinhaltet das erfolgreiche und geplante Anwenden von Kreativität und Logik [14, S. 73].

Neben zu ungenauen oder zu genauen Prompts kann auch ein Konflikt mit den ethischen Grundlagen einen Konflikt auslösen. Auch das Fordern einer Garantie für die Korrektheit der Ergebnisse oder Vorgeben einer falschen Aussage kann zu falschen oder ethisch verwerflichen Antworten führen [47, S. 2631ff]. Damit beim Schreiben eines Prompts diese Fehler nicht auftreten, kann man verschiedenen Tipps folgen. Grundsätzlich sollte man für das Schreiben eines Prompts genügend Zeit einplanen. Als nächstes sollte man sich bewusst machen, dass es (abseits der Sicherheitsvorschriften des KI-Modells) keine Grenzen zu beachten gibt. Da man sich nun den Grundlagen bewusst ist, geht es an die Planung. Hier sollte man sich fragen, was man überhaupt wissen will, wie das Ergebnis präsentiert werden soll und ob die KI die nötigen Informationen hat, beziehungsweise welche sie noch benötigt. Ebenso sollte man sich bewusst machen, inwieweit der Kontext mit den gegebenen Informationen verstanden werden kann. Nachdem man diese Informationen gesammelt hat, gibt es noch einen Schritt, bevor es an das eigentliche Schreiben geht. Hierbei gilt es sich zu überlegen, welche Rolle die KI übernehmen soll. Jetzt kann man mit dem Formulieren des Prompts beginnen. Dabei gilt es, sich immer wieder zu fragen, ob man diesen detailliert genug formuliert [14, S. 76], [23, S. 87f]. Grundsätzlich lassen sich Prompts in verschiedene Kategorien aufteilen. So gibt es die instruktiven und kontextuellen Prompts, welche selbst-erklärend sind. Ein fragender Prompt kann mit der darauffolgenden Antwort als Startpunkt

dienen. Ähnliches gilt für den systematischen Prompt. Allerdings können diese Arten an Prompts auch beliebig kombiniert werden und einen gemischten Prompt erstellen [47, S. 2631]. Eine weitere Hilfe gute Prompts zu formulieren, ist es in Threads, oder auch Prompt-Ketten genannt, zu denken. Dies ist eine Methode, bei der der gesamte Chatverlauf als Prompt dient. Somit wird dem Nutzer erlaubt komplexere Ideen zu formulieren, welche ansonsten durch die Zeichenlimitation einer Nachricht unvollständig wären oder dann zu komplex formuliert wären [23, S. 86f]. All dies lässt sich in verschiedenste Prompt Patterns kombinieren, welche als Ausgangslage oder Ideensammlung dienen können. So kann die KI aufgefordert werden, eine Frage so weit zu verbessern, dass mehr Informationen gegeben sind. Ebenso kann man in einem Prompt schreiben, dass mehrere Lösungen genannt werden sollen, um eine Auswahlmöglichkeit zu haben. Ein weiteres Pattern besteht darin, dass die KI eine gestellte Frage in kleinere Teilfragen aufteilt und einzeln beantworten soll [46]. Um dem Leser nun noch ein besseres Verständnis dieser Patterns zu geben, folgen hier noch ein paar Beispiele von Prompts, welche in der Anwendung zufriedenstellende Ergebnisse lieferten.

Der erste Prompt zeigt ein Problem beim Programmieren einer Berechnung für ein Eisenbahngleis und ist der zweite Prompt in einem Chat. Hier wirkt die Klothoide als Übergang zwischen einem geraden Stück und einer Kurve (jeweils in der horizontalen Ebene). Nach dem Eingeben des Promptes erzeugte ChatGPT eine Antwort mit Schritt eins und der Frage ob weiter gemacht werden soll. Eine Bestätigung führte zu Schritt zwei. Ein Nachfragen führte zu einer tieferen Erklärung und einer erneuten Bitte um Bestätigung.

Kannst du mich durch die Programmierung von einer Klothoide in C# in Unity leiten? Führe mich Schritt für Schritt durch den Prozess. Beginne mit dem nötigen Fachwissen und gehe dann über in die Berechnung der Klothoiden. Ich möchte berechnen können:

- Die Länge der Klothoiden
- Die Position an einer Stelle zwischen 0 und 1 (Start und Ende)
- Die Richtung Forward (Tangente) an einer Stelle von 0 bis 1
- Den Radius an einer Stelle 0-1
- Die benötigte Länge wenn ich zwei Kurven mit festem Radius und Zentrum gegeben habe (Damit ich überprüfen kann, ob ich eine Verbindung herstellen kann)

Bearbeite Schritt für Schritt und warte zwischen den Schritten auf Nachfragen.

Ein weiterer Prompt, ebenfalls im Bereich des Programmierens, behandelt eine Abwägung zwischen der Speicherung in einer Datenbank und einer einfachen Textdatei. Hier wurde

der KI eine beratende Rolle gegeben und eine erste Frage gestellt. Daraufhin konnte die KI eine Antwort erstellen, in der Pro und Contra aufgezeigt wurde, und für weitere Rückfragen verfügbar sein. Des Weiteren übernahm die KI in dem kompletten Chat die Rolle der Beratung und wägte ab, welche Option für die Anwendung die bessere ist. Es wurde auch immer nach mehr Kontext für eine genauere Abwägung gefragt.

I work on an project in Unity. The target is to make an application which stores product data. The data can be filtered by any part of the product data.

You should take following Job in this conversation. "Give me advise on which way I can implement this feature. Ask me if you need more context".

My first question ist this: "I think about storing the productData either in an SQLite Database (which I never worked with) or in an Json-File. When I use the JSON-File I thought about loading it to an List and sorting/filtering it with System.Linq."

Der interessierte Leser möge sich die weiteren Chatverläufe direkt bei ChatGPT ansehen (siehe 9.2 Internetlink-Verzeichnis).

2.6. KI & Schule – Rechtliches

Im Zusammenhang mit künstlicher Intelligenz sind verschiedene Regulierungen erstellt worden. Zum einen soll die Entwicklung transparenter werden, sodass besser verstanden werden kann, wie ein Modell funktioniert. Andererseits werden Regelwerke erstellt, welche den künstlichen Intelligenzen einschränken und schadhaften Einsatz verhindern soll. Ein Beispiel hierfür ist das KI-Gesetz der EU, welches als Regelwerk für sprachverarbeitende KI dient [13, S. 88]. Ein weiteres Ziel ist es, vertrauenswürdige KI einzusetzen. Darunter wird eine KI verstanden, welche sich immer rechtmäßig verhält, die ethischen Werte und Grundsätze unseres Zusammenlebens garantiert und in sozialer wie auch technischer Sicht stabil funktioniert. Um dies zu überprüfen, erstellt das Fraunhofer IAIS im Auftrag der Regierung einen Prüfkatalog [22, S. 94ff].

Der Nutzer einer KI sollte sich ebenfalls zwei Dinge bewusst machen. Zum einen ist das von der künstlichen Intelligenz erstellte Material nicht urheberrechtlich schützbar. Deutschland geht dabei sogar so weit, dass niemand die Schutzrechte daran hat [13, S. 84f], [23, S. 108f]. Zum anderen geht es um das Thema Datenschutz. Es ist unklar, ob persönliche Daten zu Trainingszwecken verwendet wurden. Auch die während der Nutzung entstandenen Daten können eventuell missbraucht werden [13, S. 83]. Betrachtet man dies nun mit der schulischen Brille, so kommt man schnell auf den Entschluss, dass die SchülerInnen nicht mit einer KI arbeiten dürfen, welche die DSGVO-Richtlinien nicht erfüllt [13, S. 83]. Dies ist auch die Aussage des Zentrum für Schulqualität und Lehrerbildung Baden-Württemberg.

Hier wird sogar so weit gegangen, dass auch pseudonymisierte Accounts verboten sind und die Lehrkraft die SchülerInnen nicht anleiten darf, einen Account zu erstellen [48]. Kurz gesagt heißt das, dass die SchülerInnen zum Zeitpunkt dieser Arbeit nicht mit einer KI wie ChatGPT von OpenAI oder Googles Gemini arbeiten dürfen, nicht weil sie eine KI sind, sondern weil eine datenschutzrechtliche Verarbeitung der (Nutzungs-)Daten nicht gewährleistet, beziehungsweise garantiert wird.

2.7. Was ist eine gute Erklärung?

Nachdem nun das Thema KI geklärt ist, geht es dem eigentlichen Kern auf die Spur, dem Erklären. Doch bevor das „Wie?“ betrachtet wird, muss erst einmal geklärt werden, warum man überhaupt erklärt und warum man dies in der Schule benötigt, gerade wenn die Tendenz zu stärkerer Schülerorientierung vorliegt. Zum einen ist eine Erklärung ein zentraler Bestandteil unseres Lebens. Wir basieren unser Verständnis der Umwelt auf (wissenschaftlichen) Erklärungen und versuchen Dinge, oft schon automatisch, zu erklären. Betrachtet man die wissenschaftliche Erklärung, so ist diese das, was der Wahrheit am nächsten kommt. Es wird aus den verschiedensten Hypothesen auf diejenige geschlossen, welche die gesehenen Phänomene am besten, am schlüssigsten, erklären können. Somit ist die beste Erklärung von Seiten der Wissenschaft gleichzeitig auch die schlüssigste, da sie induktiv auf bereits Bekanntes zurückführt [49, S. 10], [50, S. 8ff], [51, S. 2]. Verallgemeinert man diesen Schluss auf die beste Erklärung in den Alltagssprachlichen Kontext, so ist eine Erklärung immer die Antwort auf ein „Warum?“ [49, S. 11], [51, S. 2], [52]. Betrachtet man nun Erklärungen im schulischen Kontext, so könnte man meinen, dass eine Erklärung, gerade bei verstärkter Schülerorientierung, nicht mehr so wichtig ist, wie sie es beim Frontalunterricht ist. Allerdings ist das Erklären an sich ein zentraler Teil der naturwissenschaftlichen Bildung. Eine Erklärung kann dabei als Zusammenfassung, zur Strukturierung oder zur Planung oder Besprechung von Lösungsstrategien dienen. Eine weitere Funktion einer Erklärung ist die Umwandlung von abstrakter Informationen auf ein, für den Adressaten, verständliches Niveau. Somit kann eine Erklärung die Grundvoraussetzung für eine Verständnisentwicklung bilden [51, S. 2f], [52, S. 2,7]. Eine Erklärung ist somit nicht nur eine Methodik des Frontalunterrichts, sondern, gerade im Kontext der Naturwissenschaften, durchaus eine valide und wertvolle Methode, um Fachwissen und Handlungswissen (z.B. Versuche) zu festigen und zu wiederholen.

Doch was genau ist nun eine Erklärung? Im allgemeinsten Sinne ist eine Erklärung eine Information, welche mit einer bestimmten Kodierung über die Sprache vermittelt wird [51, S.

3f]. Im einfachsten Fall besteht sie aus einer Wenn-Dann-Verbindung und kann durch eine Verkettung von mehreren dieser Verbindungen zu einer Erklärkette heranwachsen. Für die Schule ist dies praktisch und sinnvoll, da die SchülerInnen diese Ketten bereits kennen und auch das Wenn-Dann-Prinzip bekannt ist [53, S. 17]. Verfeinert man den Fokus nun auf Erklärungen im Physikunterricht, so geht es um sachgerechte, physikalische Erklärungen. Dabei soll ein beobachtetes Phänomen durch ein physikalisches Prinzip begründet werden. Je nach gegebenem Niveau betrachtet die Erklärung verschiedene Details. Gerade in der Sekundarstufe I sollte nach einer quantitativen Form, also einem Gesetz inklusive Formel, Schluss sein [53, S. 16]. Der Prozess einer Erklärung beginnt normalerweise damit, dass geklärt wird, was erklärt werden muss. Anschließend wird das allgemeine Prinzip und Phänomen erklärt und eine der Sache und dem Adressaten gerechte Erklärung formuliert. Anschließend gilt es zu überprüfen, ob die Erklärung angekommen ist und vom Empfänger angenommen wurde. Schließlich kann selbst die beste Erklärung nicht zu einem Verständnis führen. Dies wird so lange wiederholt, bis das Ergebnis zufriedenstellend ist [52, S. 7]. Beobachtet man SchülerInnen, so fehlen bei Ihren Erklärungen meist wichtige Elemente oder die Begründungen. Ebenfalls mangelt es an der Struktur, weshalb die Erklärung zu Verwirrung führen kann [53, S. 16].

Was macht nun eine gute Erklärung aus? Im folgenden Abschnitt geht es um die Bestandteile einer guten Erklärung als Teil des (Physik-)Unterrichts. Die Erklärung beginnt damit, dass es einen Bedarf einer Erklärung gibt. Dieser muss in der Phase der Vorbereitung geklärt werden und kann zum Beispiel durch eine Frage kommuniziert werden. In der **Vorbereitung** ist es wichtig, dass die erklärende Person sich selbst genügend Zeit nimmt, um die Frage und somit den Kern der kommenden Erklärung zu verstehen. Ebenso sollten, für ein besseres Verständnis der Frage, Rückfragen bei nicht eindeutigen Punkten gestellt werden [51, S. 4], [54, S. 23f]. Im nächsten Schritt geht es an die **Strukturierung**. Dabei geht es darum, dass die Erklärung einem roten Faden folgt. Dies erreicht man, indem dem Adressaten das Ziel der Erklärung genannt und ein Ausblick gegeben wird sowie durch das Zusammenfassen und den Rückgriff auf Beispiele und Inhalte. Für die Strukturiertheit ist es ebenfalls wichtig, dass das Thema der Erklärung abgegrenzt wird und ein logisches, auf die Kerninhalte fokussiertes Vorgehen herrscht [51, S. 4], [55, S. 10ff], [56, S. 131, 135ff]. Damit eine Erklärung nun Wissen vermitteln kann, muss sie einer Sache dienen. Doch die Erklärung soll nicht nur diese Sache vermitteln, sondern sie auch der Sache gerecht vermitteln, sie soll **sachgerecht** sein. Eine Erklärung ist dann sachgerecht, wenn die Randbedingungen, also die Einschränkungen der Sache beinhaltet. Ebenso ist es wichtig, dass die Erklärung den logischen Bezug zur Sache behält. Dies wird erreicht, wenn sie vollständig und alle relevanten Faktoren der Sache an sich beinhaltet. Genauso müssen die fachlichen

Aspekte der Erklärung korrekt sein [49, S. 11], [51, S. 3f], [53, S. 16], [55, S. 10ff]. Damit die Erklärung nun die Chance bekommt vom Adressaten angenommen zu werden, muss sie ihm ebenfalls gerecht werden. Sie darf nicht zu leicht, aber auch nicht zu schwer sein. Der Mittelwert liefert eine **adressatengemäße** Erklärung. Hierfür muss das Vorwissen der Lernenden berücksichtigt werden. Auch die Sprache muss einfach gehalten und an den Empfänger angepasst werden. Sie darf bekannte Fachwörter beinhalten und sollte nicht bekannte Fachwörter in die Alltagssprache umformulieren. Auch die Verwendung von adressatenspezifischen Darstellungen wie Abbildungen kann den Lernenden beim Begreifen der Erklärung unterstützen. Zudem sollten diese mit bereits Bekanntem verknüpft werden können. Eine Erklärung sollte auch nicht die individuelle kognitive Leistungsfähigkeit überschreiten und somit vom Niveau her anpassbar sein [49, S. 11], [51, S. 3f], [52, S. 7], [53, S. 16], [55, S. 10f], [56, S. 131, 135ff]. Eine Erklärung kann ebenso durch **Beispiele** unterstützt werden. Dabei ist es wichtig, dass diese einen Kontext geben und nur kurz und knapp gehalten werden [51, S. 4], [55, S. 10f]. Je nach Thema ist eine **Mathematisierung** einer Sache notwendig. Diese sollte qualitativ an einem einfachen Zahlenbeispiel starten und sich dann quantitativ über einer Formel hin zu einfachen Je-Desto-Aussagen entwickeln [55, S. 10ff]. Der letzte Teil einer Erklärung sollte die **Verständnisabsicherung** sein. Hier spielen Zusammenfassungen eine große Rolle. Sie sichert ab, dass die Erklärung verstanden wurde und zu einem Zuwachs an Wissen führen konnte. Dieses diagnostische Element ist interaktiv und beinhaltet das Rückfragen an den Adressaten, Ermutigung und Loben für das Stellen von Rückfragen sowie der Mitarbeit und abschließend dem Warten auf Rückfragen. Dabei reicht es nicht aus, zu fragen, ob alle die Erklärung verstanden haben. Ebenso wichtig ist es, gerade bei längeren Erklärungen, diese Verständnisabsicherung bereits in der Erklärung anzuwenden, da so in kleineren Schritten direkt Probleme erfasst und anschließend angesprochen werden können [51, S. 4], [54, S. 28], [55, S. 10ff]. Zuletzt folgt noch der **Non-Verbale Teil** einer Erklärung. Dieser ist nicht einer Phase der Erklärung zugeordnet, sondern spielt während der gesamten Erklärung eine wichtige Rolle. In diesen Teil der Erklärung fallen Pausentechniken, die Körpersprache, wie Mimik und Gestik sowie die Stimmqualität und Artikulation. Aber auch die Position im Raum spielt eine wichtige Rolle [54, S. 28], [56, S. 131, 135ff].

Ehmke erfasste die Qualität einer Erklärung mit den drei Strukturpunkten Strukturiertheit, Adressatenorientierung, Sprech- und Körperausdruck und bei SchülerInnen zusätzlich mit der sprachlichen Qualität. Dabei wurden zur Erhebung je vier Items erfasst [56, S. 135ff].

3. Hypothesen

Betrachtet man die Theorie von künstlicher Intelligenz sowie die der Erklärungen, so lassen sich drei zentrale Hypothesen aufstellen.

- H1 Da das Fachwissen auf dem Gebiet der Physik, welches den SchülerInnen der Sekundarstufe I in Baden-Württemberg beigebracht wird, statisch und allgemeingültig ist, und die KI in den Wissenstests gut abschneidet, wird der fachliche Gehalt der Antworten der künstlichen Intelligenz korrekt sein.
- H2 Künstliche Intelligenz ist in der Lage SchülerInnen der Sekundarstufe I Erklärungen zu geben, welche für die SchülerInnen verständlich sind.
- H3 Durch die Anwendung von künstlicher Intelligenz im Schulunterricht lassen sich sinnvolle Lehr-Lern-Umgebungen gestalten.

4. Methodik

4.1. Grundlegendes Forschungskonzept

Erklärungen lassen sich an den verschiedensten Stellen im Unterricht finden. Aus diesem Grund betrachtet diese Forschungsarbeit die beiden Bereiche „Künstliche Intelligenz als Tutor“ und „Künstliche Intelligenz oder Schulbuch“. Im ersten Bereich geht es darum, dass SchülerInnen mit der KI interagieren und die KI den Lernenden Antworten und Erklärungen liefert. Der zweite Bereich untersucht in kleinerem Umfang Studenten in ihrem Lehramtsstudium. Hier wird der Fokus auf die Unterschiede zwischen Schulbucheklärungen und Erklärungen einer künstlichen Intelligenz gelegt.

4.2. Künstliche Intelligenz als Tutor

In diesem Abschnitt geht es darum, dass die künstliche Intelligenz von SchülerInnen bedient wird und neben der Messung der Erklärqualität aus diesen Chats die Auffassung der Lernenden über die Arbeit mit KI erfasst wird.

Da es sich bei einer Erklärung einer künstlichen Intelligenz um eine Erklärung eines Computers und nicht um die eines Menschen handelt, müssen die in Abschnitt 2.7 behandelten Faktoren einer guten Erklärung angepasst werden. Hauptsächlich fallen hier die Punkte weg, welche dem Menschen vorbehalten sind, also der non-verbale Teil einer Erklärung. Ebenso fällt der Punkt der Vorbereitung weg, da diese Erklärungen, wie üblich bei Schülerfragen, ad-hoc beantwortet werden und eine lange Vorbereitung nicht erlauben. Somit bleiben für eine gute Erklärung einer KI die Punkte gutes Verständnis der Frage, gute Strukturierung, sachgerechte und adressatengemäße Orientierung, welche von Beispielen und Mathematisierung unterstützt und mit einer Verständnisabsicherung abgesichert werden übrig. Mit je vier Items wird die Struktur und Sachgerechtigkeit der Erklärung erfasst. Weitere fünf Items erfassen die Adressatenorientierung. Drei Items erfassen die Verständnisabsicherung, während je ein Item die korrekte Erfassung der Frage, die Beispiele und die Mathematisierung erfasst. Damit die individuellen Schülerchats ausgewertet werden können, wird jeder einzelne Chatverlauf durch einen Experten und dem eben angesprochenen Bewertungsbogen ausgewertet. Der Bewertungsbogen für jedes dieser Items sieht vor, dass eine Bewertung mit --, -, +, ++ oder n.v. (nicht vorhanden) stattfindet. Zusätzlich wird noch erfasst, ob die KI Informationen über die Lernenden hatte, mit denen Sie eine Anpassung an das Niveau oder Vorwissen machen hätte können. Diese Erfassung wird nicht bewertet, da sie zwar die Erklärqualität beeinflussen kann, jedoch nicht Ursache der KI ist. Um einen ersten Überblick über die verschiedenen Erklärungen zu bekommen, werden im

ersten Schritt die Items in Punkte umgewandelt. Hierzu gilt grundsätzlich diese Bewertungsmatrix.

Bewertung	--	-	+	++	n.v.
Punkte	0	1	2	3	0

Abbildung 10: Umwandlung der Bewertungsmatrix - Pflichtitems

Die beiden Bewertungspunkte „Beispiel gegeben“ und „Mathematisierung verwendet“ sind nicht für jede Erklärung passend. Eine rein fachliche Erklärung ohne mathematischen Zusammenhang (z.B. „Was ist Schall?“ oder „Was bedeutet Trägheit?“) würde somit niemals die Möglichkeit besitzen volle Punktzahl zu erreichen. Deshalb gibt es für diese beiden Items die Umwandlung für optionalen Items. Diese haben die besondere Eigenschaft, dass eine negative Bewertung der Sache mehr schadet, wie wenn sie weggelassen worden wäre (n.v.). Somit wird das (bewusste) Weglassen eines dieser Items besser bewertet wie das schlechte Aufzeigen des Items.

Bewertung optional	--	-	+	++	n.v.
Punkte	-2	-1	1	2	0

Abbildung 11: Umwandlung Bewertungsmatrix – optionale Items

Aus diesen beiden Bewertungsmatrizen lässt sich eine maximal erreichbare Punktzahl von 48 feststellen (16 Pflichtfragen). Hierzu werden die optionalen Punkte addiert, sodass eine perfekte Erklärung mit Beispiel und Mathematisierung sogar auf 52 Punkte kommen und somit mehr wie 100% erreichen kann. Durch diese Art der Wertung werden themenübergreifend Erklärungen gleichbehandelt, sodass sowohl Themen ohne Möglichkeiten für Mathematisierung und/oder Beispielen die gleichen Chancen haben eine perfekte Erklärung abzuliefern, wie Themen, bei denen diese Möglichkeiten bestehen. Ebenso besteht nun wieder die Möglichkeit die Fragen wieder in ihre Cluster zurückzuführen und somit die ursprünglichen Bereiche auszuwerten und gezielte Analysen in diesen Kategorien durchzuführen. Da der Bewertungsbogen von einer qualifizierten Person ausgefüllt wird und die KI das Pseudonym, wie auch den vollständigen Chat liefert, besteht bei dieser Methodik keine Möglichkeit unvollständige Daten zu erhalten, weshalb auch keine Korrektur diesbezüglich geplant werden muss.

Der zweite wichtige Teil in der Verwendung von künstlicher Intelligenz als Tutor ist die Interaktion mit den Lernenden. Da eine Messung der Interaktion aus der Perspektive des Beobachters eine Interpretation des Gesehenen aber keine sichere Information liefert, wird hier auf einen Fragebogen zurückgegriffen. Dieser erfasst das subjektive Befinden sowie die subjektive Einstellung der Lernenden nach der Arbeit mit der KI über diese Interaktion.

Der hierfür verwendete Fragebogen erfasst, neben dem Pseudonym der KI zur Zuordnung der Chats und Fragebögen, die Klassenstufe der Lernenden. Dies dient unter anderem der Untersuchung einer Mindeststufe zur Anwendung. Weitere Fragen erfassen eventuelle Vorerfahrungen im Umgang mit KI und wie die SchülerInnen die Arbeit mit der KI fanden. Zuletzt wird noch die Verbindung von Schule und künstlicher Intelligenz befragt. Für die Items zur Erfahrungen mit der KI im Unterricht wird eine Skala mit 1 (sehr gut/sehr zufrieden) bis 4 (sehr schwer/völlig unzufrieden) verwendet. Die Fragen zur Verbindung von KI und Schule stehen die Optionen „Ja“, „Nein“ und „Unentschieden“ zur Verfügung. Da der Fragebogen von SchülerInnen ausgefüllt wird, gibt es verschiedene Datenfehler, welche auftreten können. Wird kein Pseudonym angegeben, so kann der restliche Fragebogen trotzdem verwendet werden. Einzig der Rückschluss von Fragebogen auf den Chat ist nicht möglich. Da die Fragebögen direkt im Anschluss zu der jeweiligen Stunde eingesammelt und digitalisiert werden, ist eine Zuordnung zur Klassenstufe machbar, sodass eine fehlende Angabe zur Klassenstufe nachträglich gemacht werden kann. Eine weitere Problemstelle basiert auf unsauberem Lesen der Fragen. Gerade die Frage nach der Erfahrung ist eine solche Problemstelle. Hier wird gefragt, ob der Lernende bisher mit einer KI gearbeitet hat. Es bestehen die Antwortmöglichkeiten „Ja“ und „Nein“. Falls nun „Ja“ angekreuzt wird, soll angegeben werden, wie oft bisher KI verwendet wurde. Falls „Nein“ angekreuzt oder kein Kreuz gesetzt wurde, wird von einem „Nein“ ausgegangen und der zweite Teil zur Häufigkeit der Anwendung (bisherige Erfahrung) ignoriert beziehungsweise als „Keine Erfahrung“ betrachtet. Alle anderen erfassten Daten werden bei Nichtausfüllen ignoriert und fließen somit nicht in eine Auswertung ein, sodass der restliche Fragebogen weiterhin verwertbar bleibt.

Genauso wie bei den Kriterien einer guten Erklärung etwas angepasst werden muss, muss auch die Auswahl der KI betrachtet werden. Während zuerst die Verwendung von ChatGPT oder Gemini auf der Hand zu liegen scheint, sollte dem Leser spätestens nach Abschnitt 2.6 klar sein, dass dies nicht die endgültige Lösung sein kann. Damit die SchülerInnen im Unterricht selbst mit einer KI arbeiten dürfen, muss sie der DSGVO entsprechen. Hierfür gibt es verschiedene Möglichkeiten. Für diese Arbeit wurde sich für die schulKI [57] entschieden, da sie die Möglichkeit gibt, dass die Lernenden mit einer KI arbeiten, ohne sich vorab anzumelden. Dies funktioniert dadurch, dass die schulKI nur die Chats an OpenAI weitergibt und nicht die persönlichen Daten. Ebenso werden die Daten auf deutschen Servern verarbeitet. Ein weiterer Grund für die Entscheidung für den Service der schulKI ist zudem das Interface, welches erlaubt, einen Freigabeschlüssel zu erstellen, welcher anschließend per Link und QR-Code geteilt werden kann. Über diesen Link können die Lernenden die KI nutzen. Gleichzeitig werden die Chats automatisch auf dem Konto des Freigabeschlüssels gespeichert und sind somit leichter abrufbar. Zudem lassen sich die Chats

Freigabeschlüsseln zuordnen und Chats gleicher Geräte durch Pseudonyme verbinden. Über diese Freigabeschlüssel lassen sich zudem verschiedene Einstellungen, wie die Modelle oder Bildgenerierung, an und abschalten. Durch die Verwendung der schulKI werden die Antworten auf die Fragen der SchülerInnen durch ChatGPT generiert, eine Bildgenerierung wurde vorab deaktiviert. Damit ein genauerer Einblick in den aktuellen Standpunkt der Technik gegeben werden kann und nicht nur eine einzelne KI in die Auswertung einfließt, werden die Fragen, welche in der Nachbereitung aus den Chats extrahiert werden können, in die KI Gemini gegeben. Somit ist nicht nur ein Vergleich zwischen zwei verschiedenen KIs möglich, sondern vielmehr eine allgemeinere Aussage zur Verwendung von KI im Unterricht möglich.

Während der Nacharbeit wurde noch ein weiterer Punkt in diese Auswertung mit aufgenommen. In der schulKI kann jeder Benutzer zwischen verschiedenen fertigen Assistenten wählen. Aber auch die Erstellung eines eigenen Assistenten ist möglich. Ein Assistent ist dabei eine Art Voreinstellung. Neben der verwendeten KI (ChatGPT oder Claude) können verschiedene Funktionen wie die Webrecherche aktiviert werden. Auch ein erster Prompt, welcher für den Benutzer nicht sichtbar ist, kann eingegeben werden. Für diese Arbeit wurde ein eigener Assistent erstellt. Dieser besitzt die Funktionen Klexikon und die Internetrecherche. Außerdem erlaubt er die Verwendung des erweiterten Sprachmodells von ChatGPT. Der Initialprompt (Prompt 5) beschreibt die Aufgabe der KI als Tutor für den Physikunterricht zur Erklärung des physikalischen Wissens. Nennt der Chatpartner nicht ein Alter oder eine Klassenstufe, so soll die KI nachfragen. Basierend auf diesem Wissen soll dann eine Antwort in adressatengerechter Sprache formuliert werden und Vorwissen abgerufen werden. Dieses Vorwissen soll die KI sich aus dem hochgeladenen Bildungsplan der Sek I von Baden-Württemberg organisieren. Außerdem werden der KI noch Informationen über den Output der Anfrage gegeben. Hierzu gehört die Betrachtung der Frage aus physikalischer Sicht, der kognitiven Aktivierung, Verständnisabsicherung und dem Verwenden von Beispielen. Somit hat die KI vor dem ersten Prompt des Lernenden schon eine große Menge an Informationen, welche den Kontext der KI einschränken und Informationen über den Lernenden bereitstellen. Damit dieser Assistent, welcher im Hintergrund ChatGPT benutzt, in dieser Arbeit namentlich differenziert ist, wird er als schulKI bezeichnet. Hierunter ist jedoch nicht der Service schulKI zu verstehen, welcher die Erfassung der Chats ermöglicht.

Nachdem nun alle Auswertungskriterien der Schüler und der KI fertiggestellt und eine passende KI ausgewählt wurde, fehlt nun der Prozess der Erarbeitung des Themas KI. Dieser Prozess lässt sich in drei Teile aufteilen wovon zwei für diese Arbeit relevant sind. Im ersten Teil, der Vorstellung von KI, wird den Lernenden das Thema künstliche Intelligenz nähergebracht. Da eine Erklärung der Funktionsweise von verschiedener künstlicher Intelligenz, wie

sie in Abschnitt 2.1.1 stattfindet, sowohl über dem Niveau der SchülerInnen als auch außerhalb des zeitlichen Rahmens ist, wird den Lernenden hierzu eine kurze Zusammenfassung gegeben. Neben Informationen über die Bedeutung der Abkürzung KI finden sich hier Punkte wie „Was ist eine KI?“, „Woher hat die KI ihr Wissen?“ und „Wie erstellt die KI eine Antwort?“. Als Kontext wird hier eine Text-KI genommen, da die Lernenden in der nächsten Phase ebenfalls mit einer textverarbeitenden KI arbeiten. Nachdem die Lernenden über die Grundfunktionsweise einer künstlichen Intelligenz informiert wurden, wird das Thema Prompt-Engineering angesprochen. Für SchülerInnen wird dies unter der Frage nach der Generierung guter Antworten betrachtet. Als relevante Punkte werden die in Abschnitt 2.5 erkundeten Punkte auf Schülerniveau heruntergebrochen und mit Beispielen von schlechten Prompts und ihren Verbesserungen auf einem Arbeitsblatt zusammengefasst. Dieses Arbeitsblatt soll den SchülerInnen in der zweiten Phase des Erarbeitungsprozesses als Stütze dienen. Diese zweite Phase beginnt damit, dass die SchülerInnen zusammen mit ihrem Schulbuch zuerst zwei Fragen zu den physikalischen Themen aus dem jeweiligen Schuljahr stellen. Anschließend sollen die SchülerInnen in Zweiergruppen sich für eine der vielen gesammelten Fragen entscheiden und diese mit der KI ChatGPT über den Service schulKI erarbeiten. Damit die SchülerInnen sich aktiv mit dem Thema und der KI auseinandersetzen sollen sie, neben der reinen Bearbeitung der Frage, eine kurze Präsentation vorbereiten. Die Vorstellungen dieser Präsentationen sind der dritte Teil des Erarbeitungsprozesses und nicht mehr für die Erhebung dieser Arbeit relevant.

Die Erhebung der KI-Chats und Schülerfragebögen wurde in drei zehnten Klassen und einer achten Klasse während dem Physikunterricht durchgeführt. Daraus resultierte eine Stichprobe von 80 SchülerInnen (54 aus Klassenstufe 10 und 26 aus Klassenstufe 8) und 42 KI-Chats der KI ChatGPT. In der Nachbereitung wurden alle diese Fragen in Gemini bearbeitet, sodass hier ebenfalls 42 Chats zur Auswertung stehen. Eine zufällige Auswahl von zwölf Chats wurde mit dem erstellten schulKI-Assistenten getestet. Von den Schülerfragebögen sind zwei Stück ohne Pseudonym, sodass diese aus Rückschlüssen vom Fragebogen auf den Chat ausgeschlossen werden.

4.3. Künstliche Intelligenz oder Schulbuch – Ein Vergleich

Der zweite Teil der Erhebung befasst sich mit den klassischen Schulbucherklärungen. Auch hier besteht die Möglichkeit anstelle von Schulbüchern eine künstliche Intelligenz zu verwenden, welche dann auf das entsprechende Vorwissen der Lernenden angepasst werden

kann. Die Untersuchung befasst sich hierfür mit vier verschiedenen Texten zum Thema Halbwertszeit aus der zehnten Klasse (siehe auch Abschnitt 9.6). Zwei dieser Texte sind durch künstliche Intelligenzen mit Prompt 7 erstellt worden, je eine von ChatGPT und Gemini. Die anderen beiden Texte stammen aus den Schulbüchern „Prisma“ [58] und „Physik für Realschulen“ [59]. Diese Texte wurden jedoch neutralisiert, also jegliche Hinweise auf Grafiken oder andere Buchseiten entfernt. Dadurch ist nicht sofort ersichtlich welche Texte von einer KI geschrieben wurden. Es wurde sich für diesen Schritt entschieden, da ein Vorhandensein von unterstützenden Grafiken zu einer anderen Bepunktung führen kann. Da sich in diesem Teil der Untersuchung jedoch auf die Erklärungen konzentriert werden soll, wurden genau diese Ablenkungen entfernt, wodurch alle vier Texte ein einheitliches Layout und den gleichen Stil besitzen.

Für die Erhebung der Daten der Studenten wird ebenfalls ein Fragebogen verwendet. Zuerst wird der Studiengang (Sekundarstufe I oder Grundschule), das Fachsemester und die Fächer erfasst. Anschließend wird jeder Text mit einer Bepunktung von 1-4 abgefragt. Mehr Punkte bedeuten hier, dass dieser Text als besser gewertet wird. Zudem können mehrere Texte gleich viele Punkte bekommen. Dies liegt dem zu Grunde, dass nicht nur eine Reihenfolge feststellbar sein soll, sondern auch gleich starke Texte als solche dargestellt werden sollen. Zuletzt wird eine Begründung für die Bepunktung jedes Textes erfasst. Hier geht es vor allem um die Gründe, die zu einer besseren oder schlechteren Bewertung führten.

Betrachtet man die Stichprobe der Studenten, so fällt diese verhältnismäßig klein aus. Insgesamt haben zwölf Studenten an der Umfrage teilgenommen, elf davon aus dem Studiengang Lehramt Sekundarstufe I.

4.4. Auswertung

Während in der Datenerhebung die Daten erfasst und vor allem digitalisiert werden, müssen in dem Bereich der Auswertung die Daten teilweise umgewandelt werden. Ein solcher Fall liegt zum Beispiel vor, wenn das Überschreiten eines Schwellenwertes geprüft werden soll. Gerade der Fragebogen für die SchülerInnen besitzt Skalierungen, die der Arbeitswelt der SchülerInnen entsprechen. So sind gute Bewertungen mit eins und schlechte Bewertungen mit vier zu wählen. Dies ähnelt dem Notensystem mit guten Leistungen als eins und schlechten Leistungen als sechs. Da aber nun das Erreichen eines Schwellenwertes von beispielsweise 80% nach dem Überschreiten fragt, muss der Datenwert für die entsprechende Berechnung umgerechnet werden. So wird eine Wertung von eins (z.B. sehr leicht)

und vier (sehr schwer) zu den korrigierten Werten von vier (sehr leicht) und eins (sehr schwer). Als Formel hierfür wird folgende verwendet.

$$\text{Skalierter Wert} = \text{Maximum}_{\text{Alt}} - (\text{Alter Wert} - 1)$$

Dies geschieht nur im lokalen Bereich (also direkt für eine einzelne Berechnung), damit in anderen, direkteren Vergleichen mit den originalen Werten gearbeitet werden kann. Ebenso ist es teilweise nötig einen Prozentwert mit der Skalierung zu kombinieren. Dazu wird der Prozentwert vorab mit dem Maximum der Skalierung multipliziert. An manchen Stellen zudem noch ein gewichteter Mittelwert nötig. Dazu wird diese Formel angewandt.

$$\text{Mittelwert}_{\text{gewichtet}} = \frac{\sum_{i=1}^n \text{Wert}_i * \text{Faktor}_i}{\sum_{i=1}^n \text{Faktor}_i}$$

Für die Auswertung wird die Softwareumgebung R [60] verwendet. Zusammen mit der Entwicklungsumgebung RStudio [61] ist die Wahl auf ein kostenloses Statistikprogramm gefallen, welches gängige statistische Tests sowie komplexere Analysen durchführen und mit Funktionen aus der Community erweitert werden kann. Zudem können gewisse Prozesse, welche öfters angewendet werden, mit Code in Funktionen formuliert werden und anschließend mit wenig Arbeitsaufwand wiederverwendet werden, was in der Auswertung Zeit spart. Durch meine Vorkenntnisse in der Softwareentwicklung ist diese einfache Automatisierung neben der Tatsache, dass R kostenlos ist, einer der Hauptgründe für die Bevorzugung von R über anderen Möglichkeiten wie SPSS oder Excel.

Bevor nun der Fokus auf die Methodik der Auswertung fällt, sollte noch der Umgang mit eventuell sensiblen Daten angesprochen werden. Grundsätzlich werden, wie bereits in den obigen Abschnitten genannt, keine personenbezogenen Daten abseits der Klassenstufe erfasst. Allerdings können die Chats der SchülerInnen sensible Daten enthalten, solange die SchülerInnen diese in Form einer Nachricht an die KI formuliert haben. Falls diese Daten keinen fachlichen Kontext geben oder in einer Nachricht mit fachlichem Kontext eingearbeitet sind, werden diese Nachrichten sowie die jeweilige Antwort der KI nicht in die Datenaufarbeitung übernommen. Sollten nun doch sensible Daten in einer Frage oder Aussage eingebaut sein, welche den fachlichen Kontext und somit die Messung beeinflussen, werden diese Daten in der Datenaufarbeitung nur von den bewertenden Experten gesichtet. Die daraus resultierenden CSV-Dateien enthalten nur die Bewertung der Erklärqualität. Somit sind in der anschließenden Auswertung keine sensiblen Daten außer der Klassenstufe und dem Pseudonym mehr enthalten, weshalb in dieser nicht mehr darauf geachtet werden muss.

Nun geht es direkt weiter mit der Auswertungsmethodik. Bevor genauere Untersuchungen mit den Daten angestellt werden, müssen verschiedene Grunduntersuchungen durchgeführt werden. Dazu gehört die deskriptive Statistik der verschiedenen Daten sowie eine Verteilungsanalyse. Anschließend können die einzelnen Hypothesen geprüft werden. Zuletzt werden weitere Punkte im Rahmen einer Exploration untersucht. Für die Untersuchung der Hypothesen werden verschiedene Tests durchgeführt. Dabei geht es hauptsächlich um das Erreichen eines Schwellenwertes oder die Verbindung mehrerer Daten. Hierzu werden jedoch nicht die originalen Daten verwendet, sondern diese originalen Daten nach bestimmten Parametern gewichtet. Hierdurch ist es möglich genauer auf die Ziele einzugehen.

In Hypothese H1 geht es um den sachgerechten Teil der Erklärung. Hier finden drei Anpassungen statt. Zum einen wird der Schwellenwert, der erreicht wird, auf >80% im Cluster „Sachgerecht“ festgelegt. Dies ist der Situation des KI-Einsatzes geschuldet. Da die KI in der Schule Anwendung findet und den SchülerInnen Fachwissen vermitteln soll, herrschen hierbei hohe Ansprüche. Unter anderem die Möglichkeit, dass sich die Lernenden vollständig auf die künstliche Intelligenz verlassen und Informationen nicht querlesen oder woanders überprüfen, ist ein großer Faktor für diese hohe Anforderung. Ein weiterer Schwellenwert (Korrektheit) bezieht sich darauf, dass die KI keine falschen Informationen vermittelt. Deshalb dürfen keine Inhalte schlechter als 50% im Item „Korrektheit“ bewertet sein. Der letzte Schwellenwert (Rest) betrachtet die anderen Items im Cluster Sachgerecht. Da diese auf die Vollständigkeit sowie die Einschränkung eingehen und eine nicht optimale Einschränkung das Schiff nicht zum Kentern bringt, eine vollständig falsche Einschränkung jedoch keinen Mehrwert liefern kann, darf bei diesen Items nirgendwo die schlechteste Bewertung vorhanden sein. Ist einer der Schwellenwerte Korrektheit oder Rest unterschritten worden, so wird die sachgerechte Qualität für diese Überprüfung auf 0% gesetzt, das Cluster Sachgerecht wurde nicht erfüllt.

Für die Überprüfung der Hypothese H2 wird überprüft, ob die KI eine verständliche Erklärung liefern kann. Hierfür werden unter anderem zwei unterschiedliche Bereiche untersucht und somit auch zwei Schwellenwerte gebildet. Die Wahl auf die Untersuchung dieser beiden Teilgebiete ist gefallen, da eine Erklärung als Kommunikation eingestuft werden kann und in einer Kommunikation sowohl Empfänger wie auch Absender der Informationen Einfluss auf die Kommunikation und ihren Erfolg haben. Würde nur eine Seite dieser Kommunikation betrachtet werden, so ist es nicht möglich eine Aussage über die komplette Kommunikation zu erstellen. Selbiges gilt für die Erklärung. Eine Erklärung kann zwar verständlich formuliert sein aber trotzdem beim Empfänger nicht verstanden werden, der Empfänger ist unzufrieden. Somit betrachtet die Untersuchung für Hypothese H2 diese beiden Teile. Der erste Teil untersucht die Verständlichkeit der Erklärung. Eine Erklärung wird dann als

verständlich erachtet, wenn der Schwellenwert „Verständlichkeit“ von 70% überschritten wurde. Die Wahl fällt auf diesen Wert, da somit eine wesentliche Erfüllung der verschiedenen Kriterien vorhanden ist, aber kleinere Schwächen in einzelnen Bereichen ausgeglichen werden können. Betrachtet man die Erklärungen im Unterricht, so kommen diese von Menschen, welche ebenfalls Fehler machen und nicht perfekt sind. Vergleicht man die Schwelle mit zum Beispiel 50%, so würden dort schon deutliche Mängel die Verständlichkeit beeinträchtigen. Da die einzelnen Bereiche einen unterschiedlichen Part zur Verständlichkeitsaussage beitragen, wird hier eine Gewichtung vorgenommen. Die Verständnisabsicherung wird hier mit Faktor 0,5 reduziert, da diese im ersten Moment nicht direkt etwas mit der Verständlichkeit bei der ersten Antwort zu tun hat. Eine Reduktion auf 0 findet hier nicht statt, da über eine Absicherung in nachfolgenden Antworten die Verständlichkeit erhöht werden kann. Wichtig für die Verständlichkeit einer Erklärung sind aber auch die anderen Bereiche. Eine gute Struktur ist ebenso wichtig wie das richtige Fachwissen. Deshalb findet hier eine Faktorisierung mit eins statt. Auch die Orientierung am Adressaten ist wichtig. Da die KI aber nicht immer Informationen über den Adressaten hat, wird diese mit dem Faktor 0,5 verringert, sobald keine Information über den Lernenden vorhanden ist. Abschließend wird auch das Cluster Sonstiges mit Faktor Eins verarbeitet. All diese Cluster bilden mit den entsprechenden Faktoren einen gewichteten Mittelwert der Erklärqualität in der Kategorie Verständlichkeit, welcher dann auf den Schwellenwert „Verständlichkeit“ überprüft wird. Der zweite Teil für die Hypothese H2 besteht aus den Erfassungen der SchülerInnen. Genauer geht es dabei um die Zufriedenheit mit den Antworten. Da diese Werte auf der Skala von 1-4 ähnlich Schulnoten angeordnet sind und für diese Untersuchung nicht mit anderen Items verglichen werden findet hier keine Anpassung der Skala statt. Als Schwellenwert der Zufriedenheit werden 2,5 Punkte gewählt, was 50% entspricht. Somit ist eine Erklärung dann verständlich, wenn >50% der SchülerInnen beziehungsweise Empfänger der Erklärung zufrieden mit der Erklärung sind.

In Hypothese H3 wird die Sinnhaftigkeit betrachtet. Dazu wird der gewichtete Mittelwert aus der ungewichteten Erklärqualität und den Schülerfragen Zufriedenheit, Schwierigkeit und Wunsch KI im Unterricht anzuwenden gebildet. Dies ist dem geschuldet, dass eine Anwendung nur dann sinnvoll ist, wenn sie einerseits gut genug ist, um etwas zu bewirken und andererseits die Empfängerseite zufrieden mit der Anwendung ist. Zudem kommt hier hinzu, dass eine Anwendung von KI im Unterricht nur dann sinnvoll ist, wenn die Anwendung nicht zu schwer ist. Hierfür gibt es verschiedene Punkte zu beachten. Zum einen müssen die Skalen für die Schwierigkeit und die Zufriedenheit umgekehrt werden, wie es bereits oben geschrieben wurde. Andererseits muss der Wunsch der SchülerInnen, KI öfters im Unterricht zu sehen, von einem Ja/Nein/Unentschlossen in ein 4/-4/0 umgewandelt werden,

damit er in die Skala hineinpasst. Somit werden positive und negative Bewertungen erfasst. Der gemittelte Wert wird über die Faktoren 1,0 für Schwierigkeit und Zufriedenheit sowie 0,5 für Wunsch der Anwendung und 2,0 für die Erklärqualität berechnet und soll, um eine Sinnhaftigkeit nachzuweisen, den Schwellenwert 50% überschreiten. Dieser Schwellenwert lässt sich damit begründen, dass eine Anwendung entweder sinnvoll oder nicht sinnvoll sein kann. Auch eine leicht sinnvolle Anwendung kann den Unterricht bereichern, während eine nicht sinnvolle Anwendung hier keinen Platz findet. Ein weiterer Punkt, welcher in diese Hypothese mit einfließt, ist die Auswertung der studentischen Interviews. Hier werden Vergleichstest zwischen den verschiedenen Gruppen (Texten) gebildet.

Neben der Untersuchung der Hypothesen werden weitere Korrelationen im Rahmen einer Exploration untersucht. Das Ziel dahinter ist es, die erhobenen Daten für einen ersten Überblick in einem Themengebiet zu nutzen. Außerdem wird hierdurch ein genauerer Blick in die Materie und dadurch genauere Aussagen ermöglicht.

5. Ergebnisse

Bevor die Ergebnisse nun im folgenden Abschnitt präsentiert werden, wurden die erhobenen Daten gemäß des vorherigen Abschnittes ausgewertet. Nun werden sie nach den zentralen Hypothesen geordnet vorgestellt. Zu Beginn erfolgt ein Überblick durch die deskriptive Statistik bevor mithilfe der Interferenzstatistik Zusammenhänge und signifikante Unterschiede identifiziert werden. Die präsentierten Ergebnisse werden in dem anschließenden Kapitel „Diskussion“ (Abschnitt 6) interpretiert.

5.1. Ein erster Überblick

Da die Studie drei verschiedene Datenklassen erfasst hat beginnt der erste Überblick mit der Vorstellung dieser drei Datenklassen, angefangen mit der kleinsten, den studentischen Interviews. Die Werte basieren auf der Bewertung von 1-4 Punkten. Hierbei gilt, dass mehr Punkte eine höhere Bewertung bedeuten.

Text	Mittelwert	Standardabweichung	Median	Modus	Min	Max	Schief	Kurtosis	AD Stat	AD p	Schief	Kurtosis	Normalverteilung
ChatGPT	3.417	0.669	3.5	4	2	4	-0.640	2.419	1.296	0.001	Linksschief	Flache Verteilung	Keine Normalverteilung
Gemini	2.917	0.793	3.0	3	2	4	0.140	1.745	0.875	0.017	Rechtsschief	Flache Verteilung	Keine Normalverteilung
Schulbuch "Physik für Realschulen"	1.667	0.651	2.0	2	1	3	0.382	2.327	1.251	0.002	Rechtsschief	Flache Verteilung	Keine Normalverteilung
Schulbuch "Prisma"	2.750	0.866	3.0	3	1	4	-0.384	2.686	0.759	0.035	Linksschief	Flache Verteilung	Keine Normalverteilung

Abbildung 12: Deskriptive Statistik studentische Interviews

Betrachtet man nun diese Werte, so ist erkenntlich, dass die beiden KI-Bewertungen im Mittelwert besser und im Median größer oder gleich wie die Schulbuchtexte abschneiden. Es ist außerdem ersichtlich, dass die beiden KIs einen höheren Mindestwert erreichten. Während bei der KI ChatGPT eine moderate Schiefe vorliegt, sind die anderen Werte mit

einer leichten Schiefe geprägt. Alle vier Datensätze zeigen eine flachere Verteilung. Der Anderson-Darling-Test konnte für das Signifikanzniveau $p > 5\%$ in keinem der Texte eine Normalverteilung feststellen. Betrachtet man Abbildung 13, so ist zusätzlich zu erkennen, dass das Schulbuch „Prisma“ in allen Bewertungen vertreten ist, während die KIs kein einziges Mal einen Punkt und das Schulbuch „Physik für Realschulen“ kein einziges Mal vier Punkte erreicht haben. Ebenso ist ersichtlich, dass ChatGPT die meisten Maximalbewertungen und das Schulbuch „Physik für Realschulen“ die meisten Minimalbewertungen erhalten hat.

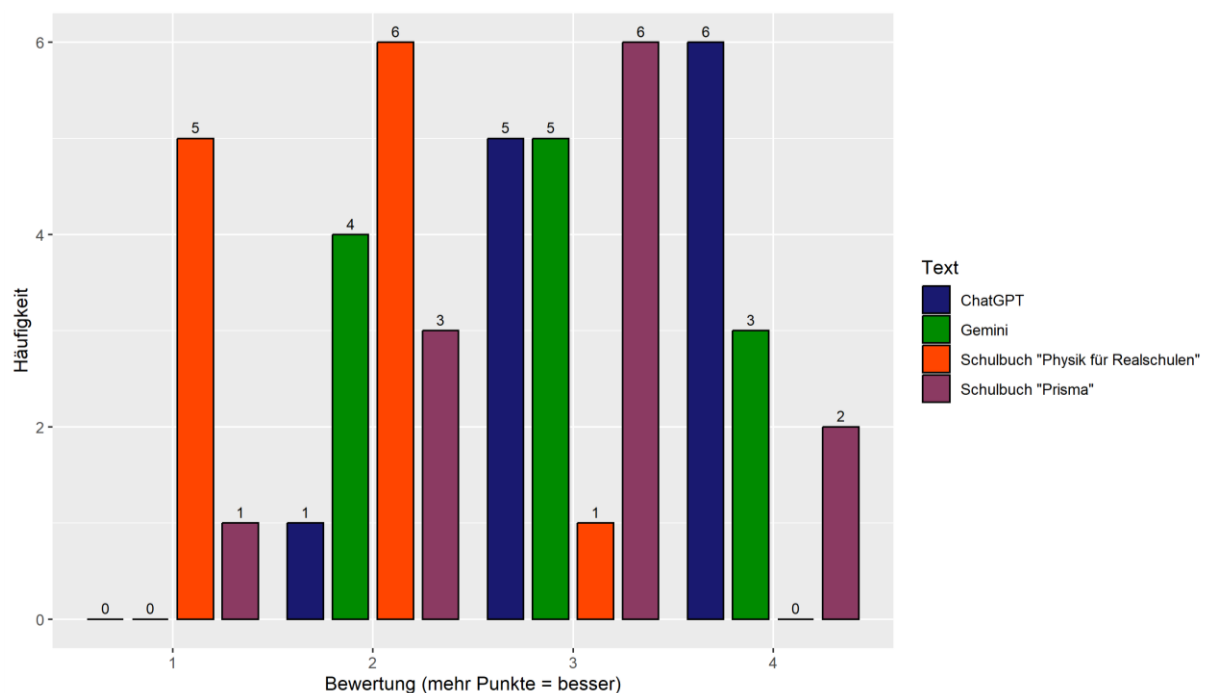


Abbildung 13: Häufigkeit der Bewertungen der Texte

Betrachtet man nun die Ergebnisse des Kruskal-Wallis-Tests, welcher auf signifikante Unterschiede zwischen den einzelnen „Autoren“ der Texte achtet, so ist deutlich ersichtlich, dass ein Unterschied vorliegt. Zwischen welchen Texten dieser nun liegt, zeigt der anschließende Dunn-Test mit Bonferroni-Korrektur (Abbildung 15). So zeigt sich, dass die KIs untereinander nicht signifikant unterschiedlich bewertet wurden. Auch der Unterschied der künstlichen Intelligenzen zum Schulbuch Prisma ist nicht signifikant. Einzig der Unterschied vom Schulbuch „Physik für Realschulen“ zu den anderen drei Textquellen zeigt einen signifikanten Unterschied in den Bewertungen der Lehramt-Studenten auf.

Test	Teststatistik	Freiheitsgrade	p_Wert
Kruskal-Wallis	20.39	3	1e-04

Abbildung 14: Kruskal-Wallis-Test Unterschiede studentische Textbewertungen

Text 1	Text 2	Anzahl 1	Anzahl 2	Z-Wert	p-Wert	p-Wert angepasst	Signifikanz
ChatGPT	Gemini	12	12	-1.347	0.1779	1.0000	ns
ChatGPT	Schulbuch Physik für Realschulen	12	12	-4.400	0.0000	0.0001	****
ChatGPT	Schulbuch Prisma	12	12	-1.713	0.0868	0.5206	ns
Gemini	Schulbuch Physik für Realschulen	12	12	-3.052	0.0023	0.0136	*
Gemini	Schulbuch Prisma	12	12	-0.365	0.7148	1.0000	ns
Schulbuch Physik für Realschulen	Schulbuch Prisma	12	12	2.687	0.0072	0.0432	*

Abbildung 15: Vergleich studentische Textbewertungen untereinander

Als nächstes werden die Fragebögen der SchülerInnen betrachtet. Hier lässt sich zuerst einmal zwischen den Klassenstufen unterscheiden. Für die direkt numerisch auswertbaren Werte „Schwierigkeit im Umgang mit der KI“, „Meinung zu der Verbindung von KI und Schule“ und der „Zufriedenheit mit den Antworten der KI“ zeigt sich keine Normalverteilung nach Anderson-Darling. Grundsätzlich gilt hier zu beachten, dass ein Wert von eins als besser angesehen wird als der Wert vier. Betrachtet man nur die Schwierigkeit, so ist keine einzige Bewertung als „sehr schwer“ ausgefallen. Ebenso ist zu bemerken, dass alle Kategorien mit ihrem Mittelwert und Median unter dem Skalenmittel von 2,5 liegen.

Frage	Klassenstufe	Mittelwert	Standardabweichung	Median	Modus	Min	Max	Schiefe	Kurtosis	AD Stat	AD p	Schiefe	Kurtosis	Normalverteilung
Schwierigkeit	8	1.769	0.514	2	2	1	3	-0.330	2.850	4.380	0	Linksschief	Flache Verteilung	Keine Normalverteilung
Schwierigkeit	10	1.574	0.602	2	1	1	3	0.499	2.350	6.645	0	Rechtsschief	Flache Verteilung	Keine Normalverteilung
Verbindung_KI_Schule	8	2.077	0.560	2	2	1	3	0.041	3.238	3.616	0	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
Verbindung_KI_Schule	10	1.722	0.920	1	1	1	4	1.163	3.458	5.298	0	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
Zufriedenheit	8	2.077	0.688	2	2	1	4	0.657	4.069	2.901	0	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
Zufriedenheit	10	1.833	0.666	2	2	1	4	0.581	3.818	5.574	0	Rechtsschief	Spitze Verteilung	Keine Normalverteilung

Abbildung 16: Deskriptive Statistik Schülerfragebogen

Im nächsten Abschnitt werden nun die einzelnen Häufigkeiten der Fragen betrachtet, begonnen mit der bisherigen Erfahrung der SchülerInnen. Grundsätzlich lässt sich hier erkennen, dass ein Großteil der SchülerInnen kaum mit einer KI arbeiten oder gearbeitet haben. Hier wurde explizit auch künstliche Intelligenz wie der Chatbot MyAI von Snapchat [4] erwähnt. Nur sehr wenige SchülerInnen aus Klasse 10 arbeiten öfters mit einer KI.

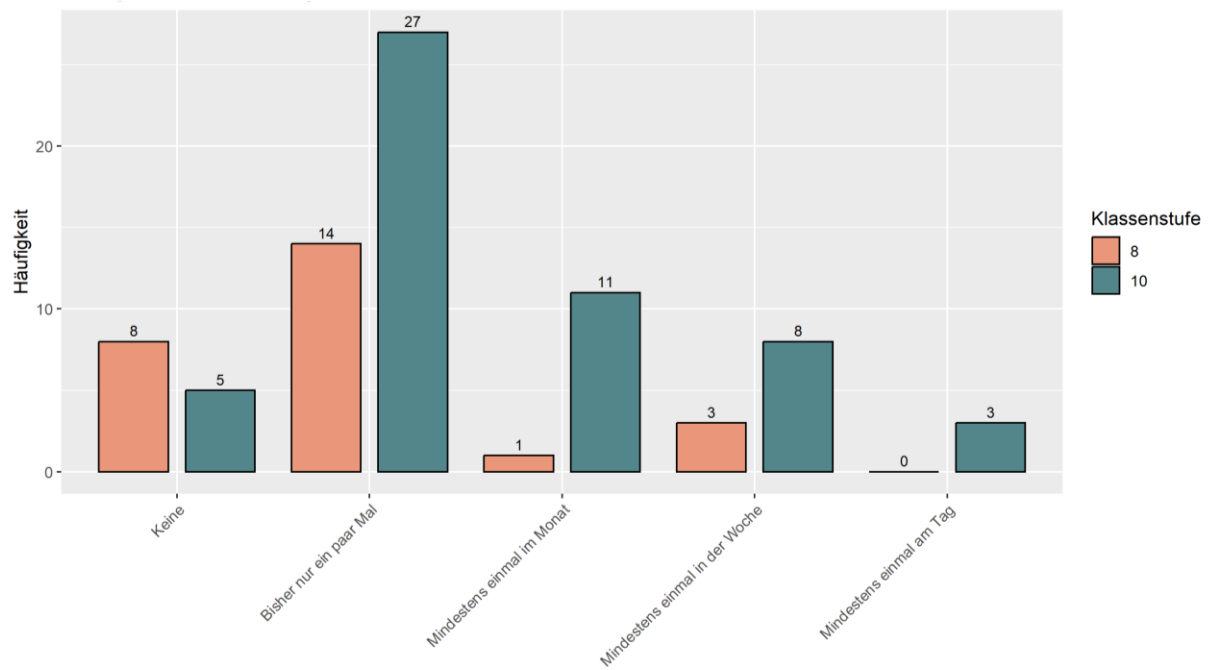


Abbildung 17: Nutzungshäufigkeit von KI

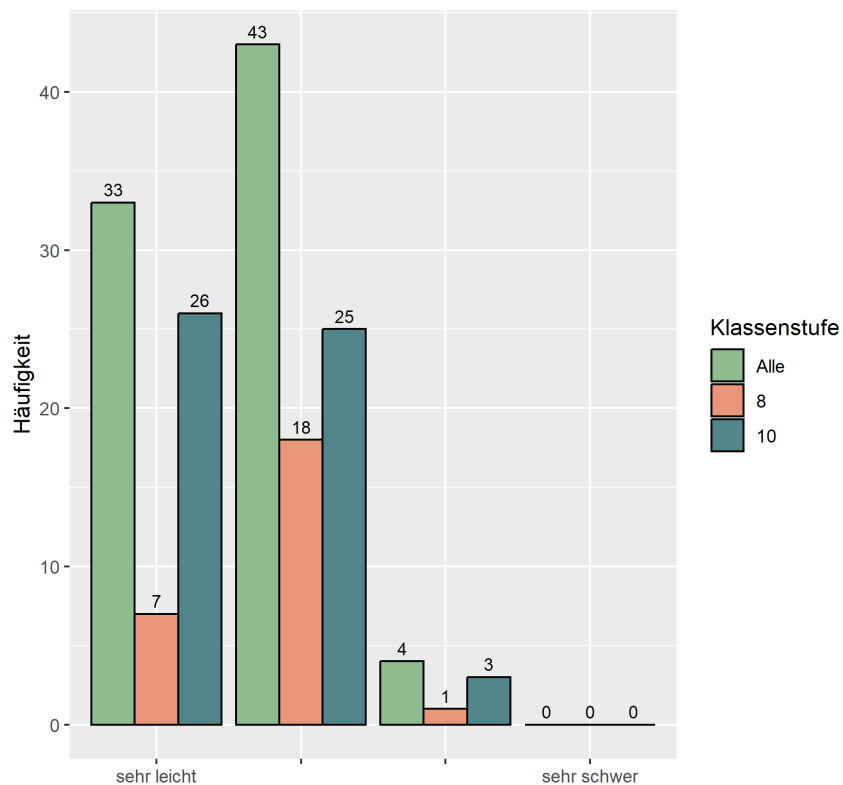


Abbildung 18: Schwierigkeit der Arbeit mit KI

Betrachtet man nun die Häufigkeiten der Antworten zur Schwierigkeit im Umgang mit der KI, so ist sofort ersichtlich, dass kein Lernender den Umgang als sehr schwer und nur wenige als schwer bezeichnet haben. Auch im Bereich der Zufriedenheit zeigte sich, dass die

meisten Lernenden zufrieden mit den Antworten der KI waren. Nur zehn der hierfür 79 verwertbaren Stimmen (etwa 12%) waren in irgendeiner Art unzufrieden mit den Antworten.

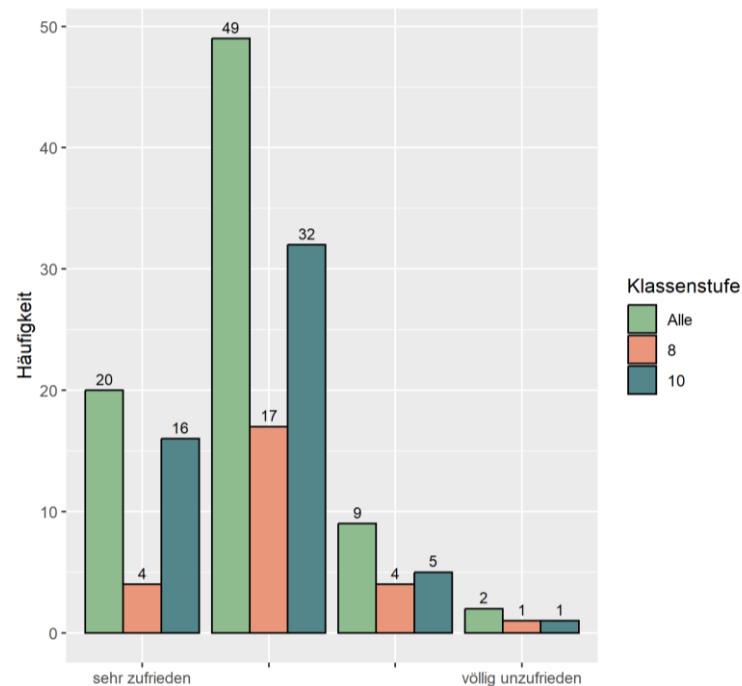


Abbildung 19: Zufriedenheit mit den Antworten der KI

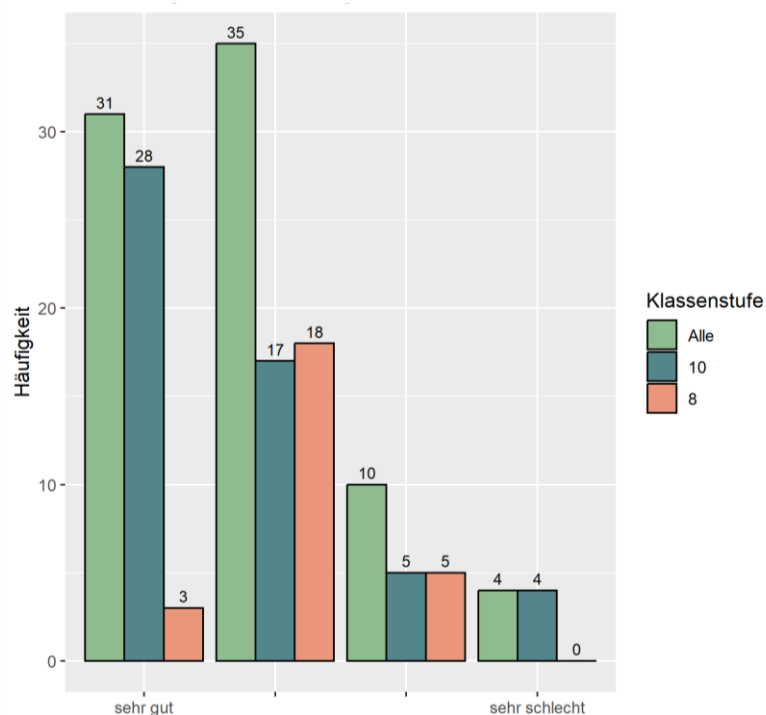


Abbildung 20: Bewertung der Verbindung von KI und Schule

Betrachtet man Abbildung 20, so sieht man ebenfalls, dass ein Großteil der SchülerInnen die Verbindung von KI und Schule gut bewertet. Schaut man sich hierbei die Klassenstufe

10 an, so finden sogar über 50% der SchülerInnen diese Verbindung als sehr gut. Dies spiegelt sich auch in den Wünschen wider. So sind knapp 60% der SchülerInnen für mehr Arbeit mit KI im Unterricht. Nur wenige (unter 8%) der befragten SchülerInnen sind dagegen.

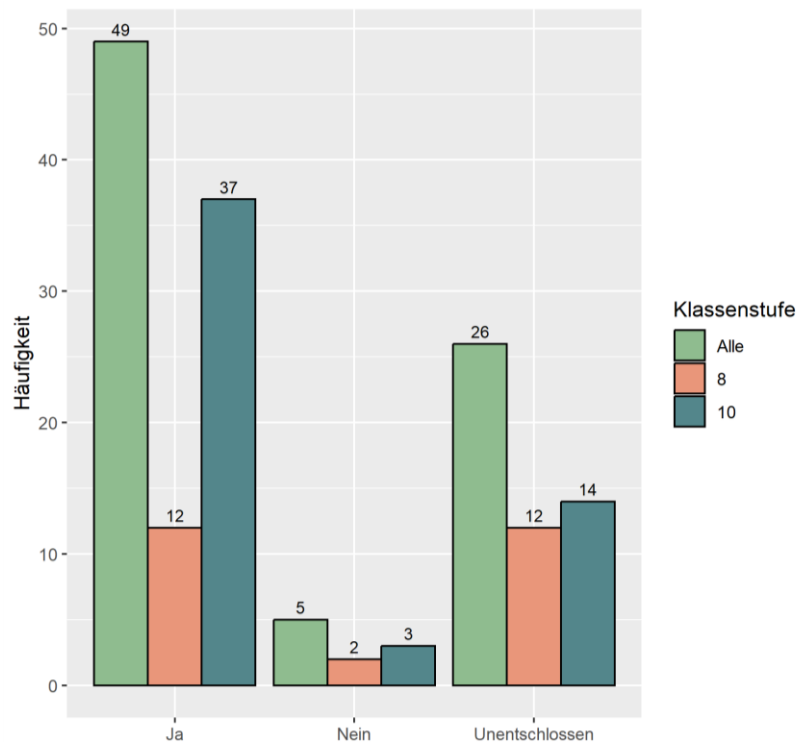


Abbildung 21: Wunsch mehr im Unterricht mit KI zu arbeiten

Ein Blick auf die Einschätzung der SchülerInnen zu ihrem Zukunftsnutzen der KI im privaten Alltag zeigt, dass nur wenige diese gar nicht einsetzen würden, während der Großteil einer Nutzung positiv gesinnt ist.

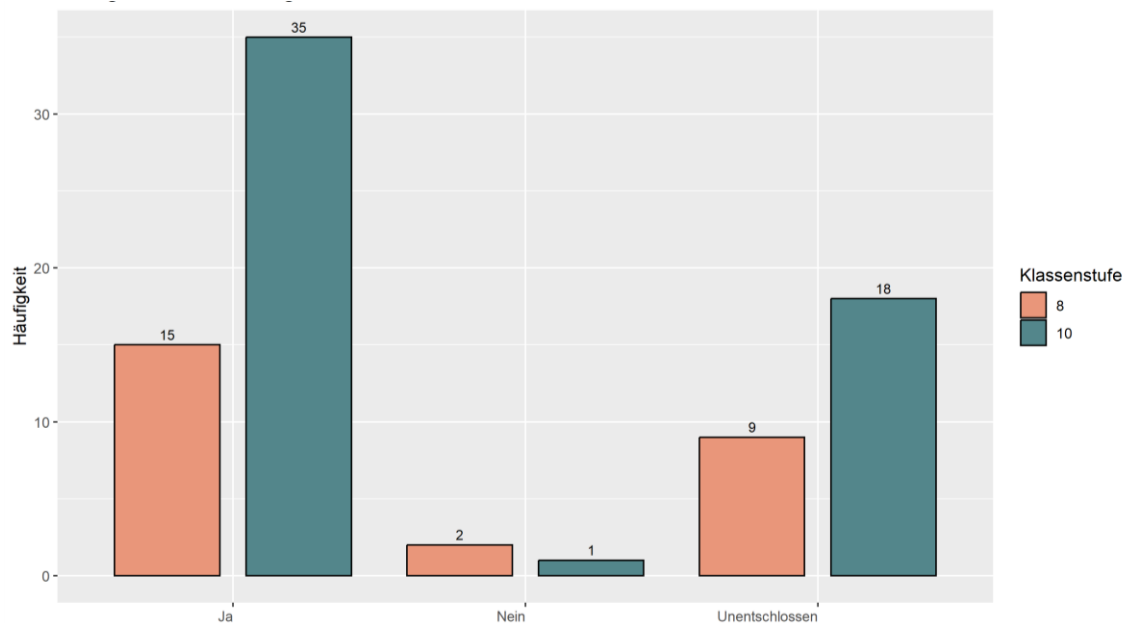


Abbildung 22: Schülereinschätzung zur Anwendung von KI im eigenen Alltag

Betrachtet man die Daten der erfassten Erklärqualität der KI so ergibt sich erst einmal die Betrachtung der einzelnen, ungewichteten Cluster. Diese sind neben der Betrachtung der gesamten Erklärung die einzelnen Bereiche Struktur, Sachgerecht, Adressat, Verständnis und Sonstiges. Während eine genauere Betrachtung der einzelnen Werte im Anschluss folgt, lassen sich bereits am Boxplot (Abbildung 23) verschiedene Punkte erkennen. Betrachtet man die gesamte Verteilung, so zeigt sich, dass die des KI-Assistenten der schulKI (folglich mit schulKI bezeichnet) deutlich höhere Bewertungen erhalten hat wie ChatGPT und Gemini. Ebenso ist ersichtlich, dass die Streuung der Werte für ChatGPT größer ist als die der beiden anderen. In den Clustern Struktur, Sachgerecht und Adressat ist ebenfalls sichtbar, dass die schulKI eine höhere Wertung erreicht, während ChatGPT und Gemini ähnlich eingeordnet sind. Im Cluster Sachgerecht erreichen trotzdem alle drei KIs teilweise den Maximalwert 100%. Betrachtet man das Cluster Verständnis, so sind hier alle drei KIs näher beisammen. Nur wenige Werte überschreiten die 75%-Marke. Betrachtet man das Cluster Sonstiges, so ist auffallend, dass ein Großteil der Werte über 100% liegt. Allerdings muss hier die Abschnitt 4.2 angesprochene Wertung berücksichtigt werden. Schließlich sind hier optionale Items eingebunden, welche Bonus-Punkte erzeugen können, falls sie in ausreichender Güte vorhanden sind. Genauer zeigt sich dies in der tabellarischen Auswertung (Abbildung 24). Die behandelten Werte sind Prozentwerte. Auffällig ist dabei unter anderem die Verteilung. Während die meisten Cluster nach Anderson-Darling keine Normalverteilung aufweisen, liegen die Wertungen der Gesamtpunkte im Durchschnitt wieder in einer Normalverteilung vor. Auch ist hier zu erkennen, dass die Mittelwerte der schulKI, wie bereits in den Boxplots oben erkannt, in Clustern wie Adressatenorientierung oder auch der

Gesamtwertung höher liegen als die der beiden anderen KIs. Betrachtet man die Standardabweichung, so ist diese im Cluster Sonstiges ebenfalls am größten. Das Cluster Sonstiges wird grundsätzlich Rechtsschief bewertet. Cluster wie Struktur und Verständnis weisen eine spitze Verteilung auf.

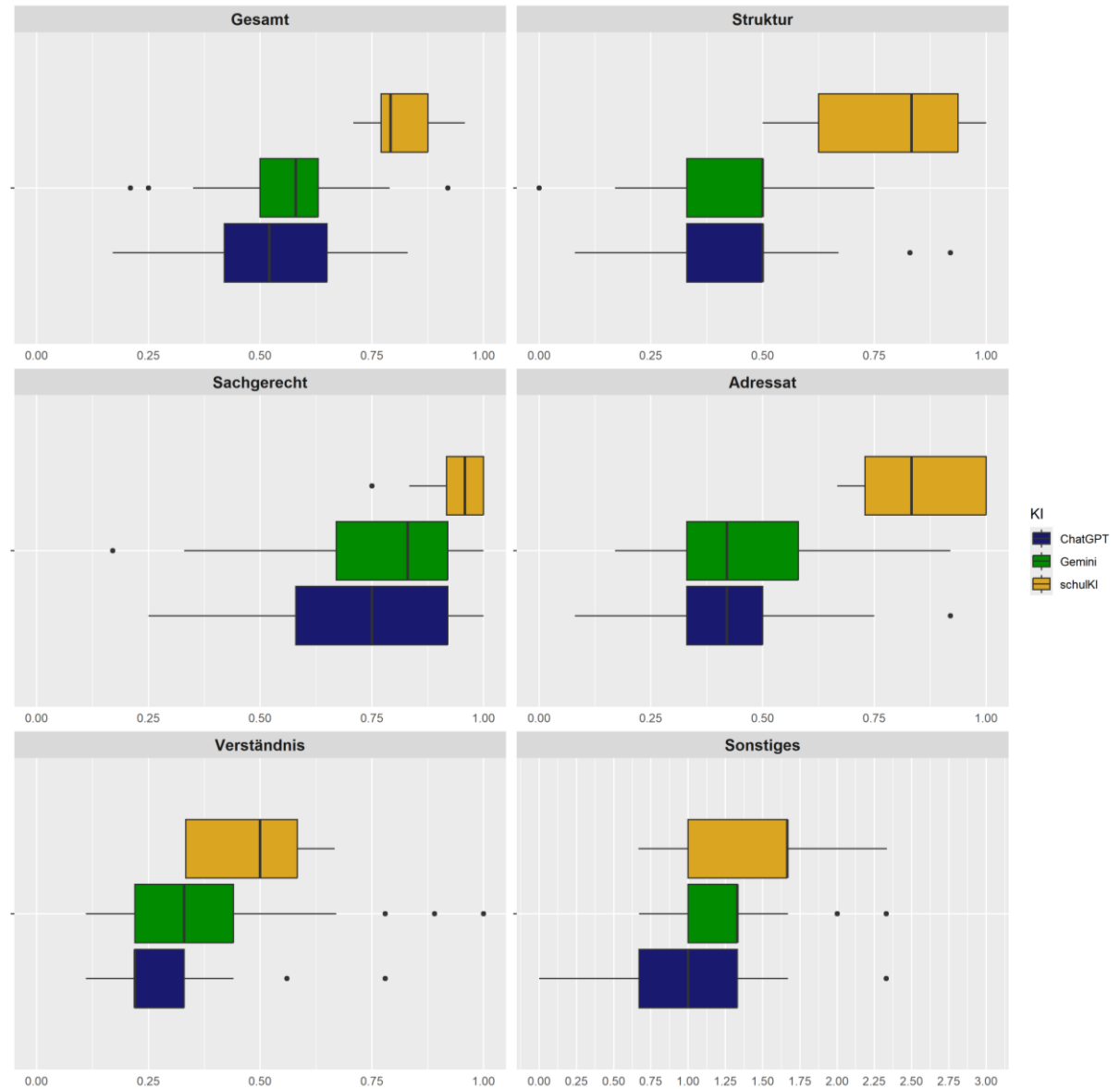


Abbildung 23: Clusterbetrachtung Erklärqualität künstliche Intelligenz

KI	Kategorie	Mittelwert	Standardabweichung	Median	Modus	Min	Max	Schiefe	Kurtosis	AD Stat	AD p	Schiefe	Kurtosis	Normalverteilung
ChatGPT	Prozente	0.525	0.157	0.520	0.480	0.170	0.830	-0.042	2.597	0.131	0.980	Linksschief	Flache Verteilung	Normalverteilung
ChatGPT	Prozente.Adressatenorientierung	0.423	0.197	0.420	0.500	0.080	0.920	0.687	3.266	0.762	0.044	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
ChatGPT	Prozente.Sachgerecht	0.723	0.222	0.750	1.000	0.250	1.000	-0.521	2.298	0.882	0.022	Linksschief	Flache Verteilung	Keine Normalverteilung
ChatGPT	Prozente.Sonstiges	1.074	0.497	1.000	1.000	0.000	2.330	0.667	3.341	2.100	0.000	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
ChatGPT	Prozente.Struktur	0.459	0.165	0.500	0.500	0.080	0.920	0.175	4.250	1.406	0.001	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
ChatGPT	Prozente.Verständnisabsicherung	0.299	0.142	0.220	0.220	0.110	0.780	1.338	4.912	2.793	0.000	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
Gemini	Prozente	0.566	0.149	0.580	0.600	0.210	0.920	0.102	3.661	0.604	0.109	Rechtsschief	Spitze Verteilung	Normalverteilung
Gemini	Prozente.Adressatenorientierung	0.470	0.179	0.420	0.420	0.170	0.920	0.638	3.336	1.042	0.009	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
Gemini	Prozente.Sachgerecht	0.768	0.198	0.830	0.830	0.170	1.000	-1.071	3.933	1.362	0.001	Linksschief	Spitze Verteilung	Keine Normalverteilung
Gemini	Prozente.Sonstiges	1.236	0.500	1.330	1.330	0.670	2.330	0.829	2.989	1.750	0.000	Rechtsschief	Flache Verteilung	Keine Normalverteilung
Gemini	Prozente.Struktur	0.427	0.132	0.500	0.500	0.000	0.750	-0.812	4.795	1.971	0.000	Linksschief	Spitze Verteilung	Keine Normalverteilung
Gemini	Prozente.Verständnisabsicherung	0.383	0.207	0.330	0.220	0.110	1.000	1.112	3.926	1.624	0.000	Rechtsschief	Spitze Verteilung	Keine Normalverteilung
schulKI	Prozente	0.825	0.083	0.792	0.771	0.708	0.958	0.447	1.858	0.606	0.088	Rechtsschief	Flache Verteilung	Normalverteilung
schulKI	Prozente.Adressatenorientierung	0.840	0.144	0.833	1.000	0.667	1.000	-0.023	1.285	0.887	0.016	Linksschief	Flache Verteilung	Keine Normalverteilung
schulKI	Prozente.Sachgerecht	0.938	0.080	0.958	1.000	0.750	1.000	-1.148	3.425	1.110	0.004	Linksschief	Spitze Verteilung	Keine Normalverteilung
schulKI	Prozente.Sonstiges	1.472	0.540	1.667	1.667	0.667	2.333	0.222	2.007	0.735	0.040	Rechtsschief	Flache Verteilung	Keine Normalverteilung
schulKI	Prozente.Struktur	0.785	0.199	0.833	1.000	0.500	1.000	-0.434	1.695	0.613	0.084	Linksschief	Flache Verteilung	Normalverteilung
schulKI	Prozente.Verständnisabsicherung	0.491	0.138	0.500	0.333	0.333	0.667	0.049	1.483	0.685	0.054	Rechtsschief	Flache Verteilung	Normalverteilung

Abbildung 24: Clusterauswertung KI im Physikunterricht

Ebenfalls bemerkenswert ist es, dass 56% der Schülergruppen Informationen über sich in Form von Klassenstufe oder Alter genannt haben. Dies ist gerade dann wichtig, wenn Rückschlüsse auf das Vorhandensein dieser Informationen, so etwa in Hypothese H2, gemacht werden.

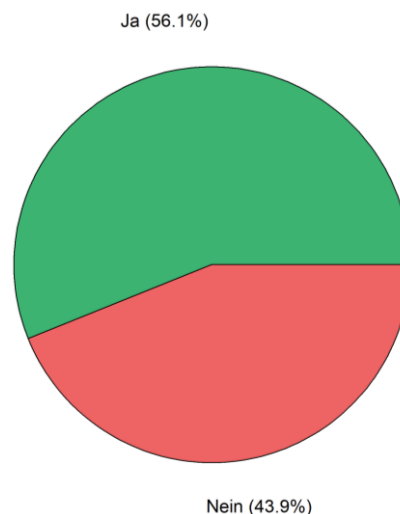


Abbildung 25: Hatte die KI Informationen über den Adressaten?

5.2. Hypothese H1

Da im vorherigen Abschnitt festgestellt wurde, dass die meisten Daten, gerade die der Normalverteilung des Clusters „Sachgerecht“, nicht normalverteilt sind müssen alle nun folgenden Tests nicht parametrisch sein. Zur Untersuchung von H1 und der Überprüfung ob künstliche Intelligenz den Schwellenwert im Bereich Sachwissen überschreiten kann, wurde der Wilcoxon-Vorzeichen-Rang-Test im einseitigen Modus „greater“ eingesetzt. Er dient der

Untersuchung einer signifikanten Überschreitung des Medians über dem vorab definierten Schwellenwert. Dieser Test wird für jede der drei KIs separat durchgeführt. Als Nullhypothese H_0 wird angenommen, dass der Median im nach Abschnitt 4.4 angepassten Cluster Sachgerecht nicht überschritten wird ($\text{Median} \leq 80\%$). Die dazugehörige Alternativhypothese H_1 besagt, dass der Median der Stichprobe größer dem Schwellenwert ist. Als Signifikanzniveau wird ein p-Wert von 5% angewandt. Ein Unterschreiten dieses Wertes deutet darauf hin, dass die Nullhypothese verworfen werden kann. Neben dem p-Wert zur Testung des Überschreitens des Schwellenwertes wird zudem das Konfidenzintervall berechnet. Dieses dient dazu, genauere Aussagen über die Daten zu sagen. Liegt der Schwellenwert außerhalb diesem Intervall (mit CI Unten und CI Oben dargestellt), so kann die Nullhypothese verworfen werden. Zudem zeigt der Konfidenzintervall an, welcher Schwellenwert mit einem Signifikanzniveau von 5% überschritten (CI Unten) oder unterschritten (CI Oben) wurde.

KI	W-Wert	p-Wert	CI Unten	CI Oben
ChatGPT	271	0.981	0.665	Inf
Gemini	376	0.763	0.710	Inf
schulKI	76	0.002	0.917	Inf

Abbildung 26: Wilcoxon-Vorzeichen-Rang-Test für H_1

Betrachtet man die Ergebnisse für den für H_1 durchgeführten Wilcoxon-Test, so erkennt man, dass ausschließlich die schulKI den Schwellenwert signifikant überschritten hat. Die schulKI zeigt einen Konfidenzwert von 91,7%, welcher weit über dem Schwellenwert liegt. Die beiden anderen KIs erreichten einen Wert von 66,5% (ChatGPT) und 76,3% (Gemini). Zusammen mit dem hohen p-Wert ($>5\%$) lässt sich hier kein signifikantes Überschreiten des Schwellenwertes von 80% feststellen. Da der Wilcoxon-Test jedoch keine Aussage über die Gruppenunterscheidungen zulässt, wurde im Anschluss der Kruskal-Wallis-Test durchgeführt. Dieser gibt Aussage darüber, ob signifikante Unterschiede zwischen den einzelnen Gruppen vorliegen. Der Test wird mit zwei Freiheitsgraden durchgeführt, da Vergleiche zwischen drei Gruppen durchgeführt werden. Der p-Wert zeigt einen signifikanten Unterschied zwischen den Gruppen auf.

Test	Teststatistik	Freiheitsgrade	p-Wert
Kruskal-Wallis-Test	10.9303	2	0.0042

Abbildung 27: Kruskal-Wallis-Test für H_1

Damit nun Vergleiche zwischen den drei Gruppen gezogen werden können, wurde anschließend noch ein Dunn-Test durchgeführt. Dieser erhielt im Anschluss noch eine

Bonferroni-Korrektur um die Signifikanz eindeutiger darzustellen. Wie den Testergebnissen in Abbildung 28 zu entnehmen ist, liegen keine signifikanten Unterschiede zwischen den beiden KIs ChatGPT und Gemini vor. Zwischen diesen Beiden und der schulKI lassen sich jedoch signifikante Unterschiede feststellen.

KI 1	KI 2	n1	n2	Z-Wert	p-Wert	p-Wert angepasst	Signifikanz
ChatGPT	Gemini	41	41	0.6695	0.5032	1.0000	ns
ChatGPT	schulKI	41	12	3.2741	0.0011	0.0032	**
Gemini	schulKI	41	12	2.8236	0.0047	0.0142	*

Abbildung 28: Dunn-Test mit Bonferroni-Korrektur für H1

5.3. Hypothese H2

Für die Überprüfung der zweiten Hypothese, in der es um die Verständlichkeit der Erklärungen geht, mussten mehrere Tests durchgeführt werden. Grundsätzlich lassen diese sich in drei Bereiche aufteilen, die gewichtete Erklärqualität, die Schülerwertung und die Korrelation. Für den Bereich der Erklärqualität wird, wie in Abschnitt 4.4 angesprochen, der gewichtete Mittelwert gebildet. Basierend auf diesem Mittelwert wird erneut mit dem Wilcoxon-Vorzeichen-Rang-Test das Überschreiten eines Schwellenwertes, in diesem Fall 70%, erfasst. Der Test ist somit erneut einseitig und betrachtet den Testmodus „greater“. Die formulierte Nullhypothese ist hierbei, dass die KIs nicht in der Lage sind, den Schwellenwert von 70% zu überschreiten ($\text{Median} \leq 70\%$). Die Alternativhypothese sagt aus, dass die KIs den Schwellenwert überschreiten. Betrachtet man die Ergebnisse in Abbildung 29, so ist ersichtlich, dass alle drei Intelligenzen den Schwellenwert mit einem Signifikanzniveau von 5% überschreiten konnten. Ein Blick auf die Konfidenzwerte zeigt, dass ChatGPT hier um 1,8%, Gemini 8,3% und die SchulKI 41,5% über dem Schwellenwert liegt.

KI	W-Wert	p-Wert	CI Unten	CI Oben
ChatGPT	583.5	0.024	0.718	Inf
Gemini	693.0	0.000	0.783	Inf
schulKI	78.0	0.001	1.115	Inf

Abbildung 29: Wilcoxon-Vorzeichen-Rang-Test für H2-Erklärung

Dies deutet Unterschiede zwischen den einzelnen Intelligenzen an, weshalb ebenfalls ein Kruskal-Wallis-Test durchgeführt wurde. Dieser sagt mit einem p-Wert von 0,0000 aus, dass signifikante Abweichungen zwischen den einzelnen Gruppen vorliegen. Deshalb wird im Anschluss der Dunn-Test durchgeführt, welcher diese Unterschiede aufdeckt. Wie bereits

für H1 wird der Dunn-Test durch die Erweiterung mit der Bonferroni-Korrektur verfeinert. Die p-Werte der einzelnen Paarunterscheidungen zeigen auf, dass die zuvor vom Kruskal-Wallis-Test angedeuteten Unterschiede für die KI SchulKI signifikant sind. Diese unterscheidet sich signifikant von den beiden anderen, welche untereinander zwar einen Unterschied aufzeigen, hierfür aber kein signifikanter Beweis vorliegt.

Test	Teststatistik	Freiheitsgrade	p-Wert
Kruskal-Wallis-Test	21.1723	2	0.0000

Abbildung 30: Kruskal-Wallis-Test für H2-Erklärung

KI 1	KI 2	n1	n2	Z-Wert	p-Wert	p-Wert angepasst	Signifikanz
ChatGPT	Gemini	41	41	1.0202	0.3077	0.9230	ns
ChatGPT	schulKI	41	12	4.5685	0.0000	0.0000	****
Gemini	schulKI	41	12	3.8820	0.0001	0.0003	***

Abbildung 31: Dunn-Test mit Bonferroni-Korrektur für H2-Erklärung

Der zweite Bereich, welcher für Hypothese H2 betrachtet wird, ist der der Schülerbewertung. Hier wird die Zufriedenheit betrachtet. Da diese in ihrer Skala so angeordnet ist, dass der Wert eins die vollständige Zufriedenheit darstellt, während der Wert vier vollständige Unzufriedenheit zeigt, wird hier der Wilcoxon-Test im einseitigen Modus „less“ verwendet. Die Nullhypothese für diese Untersuchung betrachtet, dass die Zufriedenheit der Schülerinnen mit den Antworten der KI den Schwellenwert von 50% nicht überschreitet, die Bewertungen den Punktwert 2,5 also nicht unterschreiten. Die Alternativhypothese betrachtet somit die Überschreitung des Schwellenwertes und damit die Unterschreitung der Bewertung mit 2,5 Punkten. Die Ergebnisse zeigen auf, dass sowohl in Klasse 8 als auch in Klasse 10 dieser Schwellenwert signifikant unterschritten wurde.

Klassenstufe	W-Wert	p-Wert	CI Unten	CI Oben
Alle	404.5	0.000	-Inf	2
Klasse 8	68.0	0.002	-Inf	2
Klasse 10	141.0	0.000	-Inf	2

Abbildung 32: Wilcoxon-Test für H2-Schülerbewertung Zufriedenheit

Der dritte, für Hypothese H2 zu testende, Bereich verbindet die ersten beiden. Das Ziel hierbei ist es, eine Aussage über die Zusammenhänge zu formulieren. Deshalb wird die

Spearman-Korrelationstest durchgeführt. Sowohl ein erster Blick auf die Verteilungen der Wertepaare „Zufriedenheit der SchülerInnen“ und „gewichtete Erklärqualität“ in Abbildung 33 sowie die Ergebnisse des Korrelationstestes (Abbildung 34) zeigen auf, dass keine signifikante Korrelation dieser beiden Werte vorliegt. Dies gilt auch für Teilgruppen der SchülerInnen basierend auf den Klassenstufen.

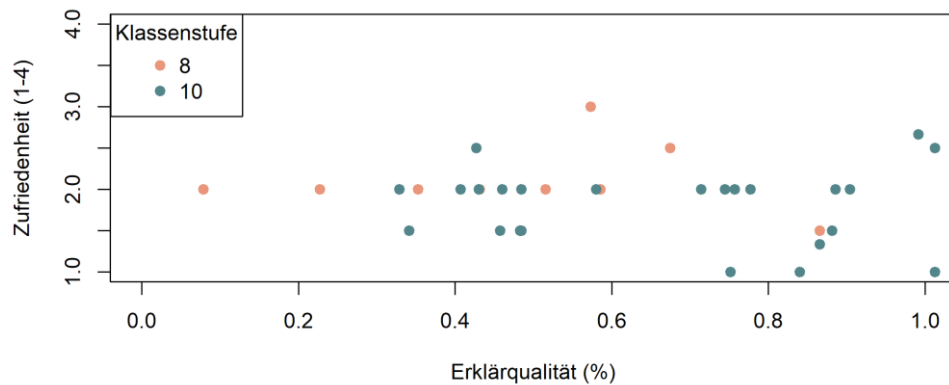


Abbildung 33: Zusammenhang von Schülerzufriedenheit und Erklärqualität

Klassenstufe	Spearman-Koeffizient	p-Wert
Alle	-0.0031	0.9852
Klasse 8	-0.0409	0.8945
Klasse 10	0.0281	0.8915

Abbildung 34: Spearman-Korrelationstest Schülerzufriedenheit und Erklärqualität

5.4. Hypothese H3

Die Auswertung für H3 fordert ebenfalls verschiedene Tests an. Es gibt hierbei zwei völlig unabhängige Datensets, die Schülerbewertung verbunden mit den ChatGPT-Erklärungen und die studentischen Interviews zu den Schulbuch- und KI-Texten. Diese müssen separat betrachtet werden und können nicht verbunden werden. Betrachtet man nun den ersten Teil, so wird erneut mit dem Wilcoxon-Vorzeichen-Rang-Test im einseitigen Modus „greater“ überprüft, ob der Schwellenwert von 50% (2,5 Punkte) mit Blick auf die Sinnhaftigkeit der Schülerfragebögen überschritten wurde.

Klassenstufe	W-Wert	p-Wert	CI Unten	CI Oben
Alle	1845	0.024	2.5288	Inf
Klassenstufe 8	147	0.769	2.2378	Inf
Klassenstufe 10	930	0.002	2.6223	Inf

Abbildung 35: Wilcoxon-Test für H3

Die Ergebnisse des Tests (Abbildung 35) zeigen für die gesamte Stichprobe auf, dass eine signifikante Überschreitung des Schwellenwertes vorliegt. Da die beiden Gruppen (Klassenstufen) jedoch stark unterschiedliche Anzahlen an Schülern hatten, wurde auch ein klassenstufendifferenzierter Test durchgeführt. Hier zeigt sich, dass für Klassenstufe 8 keine signifikante Überschreitung vorliegt, der Median liegt hier bei 2,2 Punkten. Für Klassenstufe 10 ist das Überschreiten signifikant, was ebenfalls am unteren CI-Wert von 2,6 sichtbar ist. Da hier nur zwei Gruppen vorhanden sind, wird die Differenzuntersuchung mit dem Wilcoxon-Rangsummentest durchgeführt. Als Nullhypothese wird hier die Gleichheit der Verteilungen der Gruppen angenommen. Die Alternativhypothese sagt, dass die Verteilungen sich unterscheiden.

Test	W-Wert	p-Wert	CI Unten	CI Oben
Wilcoxon-Rangsummentest	446	0.026	-0.662	-0.049

Abbildung 36: Wilcoxon-Rangsummentest H3

Betrachtet man die Ergebnisse (Abbildung 36), so ist ein signifikanter Unterschied zwischen den beiden Gruppen (Klassenstufe 8 und 10) erkennbar. Betrachtet man nun einmal Abbildung 35, so fällt der Verdacht nahe, dass Klassenstufe 10 signifikant besser abschneidet. Dies zeigt sich auch in Abbildung 37, in der deutlich ersichtlich ist, dass Klassenstufe 10 mehr Punkte erreichen konnte.

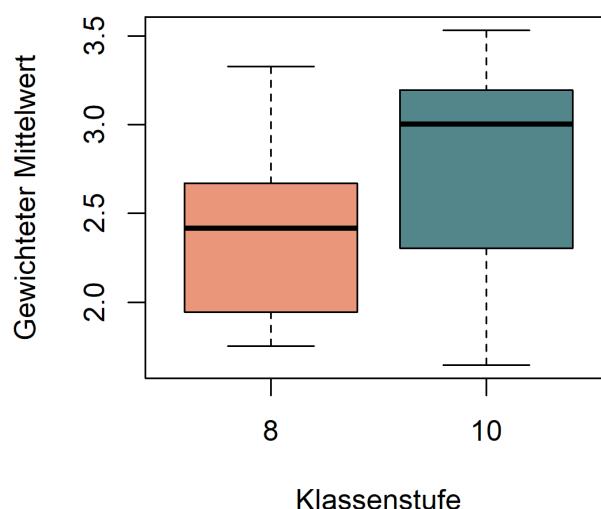


Abbildung 37: Vergleich Klassenstufe 8 & 10

Der zweite Teil der Hypothesenüberprüfung besteht daraus der Betrachtung der Studententerviews. Hier wird nicht auf einen Schwellenwert überprüft, sondern darauf geachtet, ob signifikante Unterschiede vorhanden sind und wie diese ausfallen. Wie bereits im ersten Überblick (Abschnitt 5.1) angesprochen wurde, liegen nicht parametrische Daten vor. Da

diese Datentypen zudem abhängig sind, fällt die Wahl zur Untersuchung ob signifikante Unterschiede zwischen den Textbewertungen liegen, auf den Friedman-Test. Er vergleicht diese Gruppen durch eine Ordnung (Rang) innerhalb eines Datensatzes und addiert diese dann über alle Datensätze. Als Nullhypothese werden keine signifikanten Unterschiede angenommen. Die Alternativhypothese geht von einem signifikanten Unterschied zwischen den Gruppen aus.

Test	Teststatistik	Freiheitsgrade	p-Wert
Friedman-Test	15.932	3	0.0012

Abbildung 38: Friedman-Test H3

Da dieser Test mit einem p-Wert von 0,1% signifikant ausfällt muss im Anschluss untersucht werden, zwischen welchen Gruppen diese Unterschiede vorliegen. Für diese Untersuchung wird der paarweise Wilcoxon-Test zusammen mit der Bonferroni-Korrektur verwendet. Die Nullhypothese für diesen Test geht davon aus, dass die gepaarten Gruppen, also die Bewertungen der Texte, den gleichen Median besitzen. Im Gegensatz dazu geht die Alternativhypothese von unterschiedlichen Medianen aus. Die Bonferroni-Korrektur wird angewandt, um den Fehler erster Art (Alpha-Fehler) zu reduzieren.

	ChatGPT	Gemini	Schulbuch Physik für Realschulen
Gemini	0.8675	NA	NA
Schulbuch Physik für Realschulen	0.03	0.0712	NA
Schulbuch Prisma	0.5895	1	0.0557

Abbildung 39: Wilcoxon-Textvergleich H3

Betrachtet man die vorliegenden Ergebnisse des Tests, so ist zu erkennen, dass zwischen den Texten der künstlichen Intelligenzen keine signifikant unterschiedliche Bewertung stattgefunden hat. Vergleicht man nun das Schulbuch „Physik für Realschulen“ mit den KIs, so zeigt sich, dass für die KI ChatGPT ein signifikanter Unterschied vorliegt. Ein Vergleich des Schulbuches „Prisma“ zeigt keine signifikanten Unterschiede zu anderen Textbewertungen auf. Was sich aus diesen Daten herauskristallisiert ist, dass die einzige signifikante Unterscheidung in der Bewertung zwischen dem Schulbuch „Physik für Realschulen“ und der KI ChatGPT stattgefunden hat. Betrachtet man nun mit diesen Daten Abbildung 40, so ist ersichtlich, dass die, mit Wilcoxon festgestellte Signifikanz so orientiert ist, dass der Text von ChatGPT signifikant höher bewertet wurde wie der Text aus dem Schulbuch „Physik für Realschulen“.

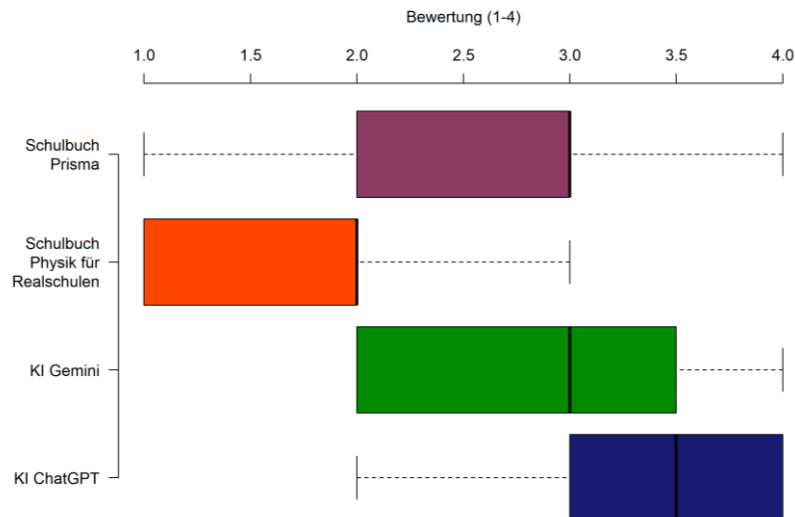


Abbildung 40: Bewertung studentischer Texte für H3

5.5. Exploration 1

Bei dieser Exploration geht es um einen möglichen Zusammenhang zwischen der ungewichteten Qualität der KI-Antworten (in diesem Fall nur von ChatGPT, da es ansonsten keine Verbindung der Daten gibt) und der Erfahrung der SchülerInnen im Umgang mit ChatGPT. Vor der Auswertung wurden noch die Fragen zur bisherigen Erfahrung kombiniert, sodass nun fünf Antworten vorliegen (Gar nicht – Mindestens einmal am Tag). Dies macht das Auswerten leichter, da nur ein Datentyp vorliegt. Diese Exploration basiert auf der Grundidee, dass mit häufigerem Umgang mit einer künstlichen Intelligenz die Kompetenz des Prompt-Engineerings verbessert wird und somit bessere Ergebnisse generiert werden können.

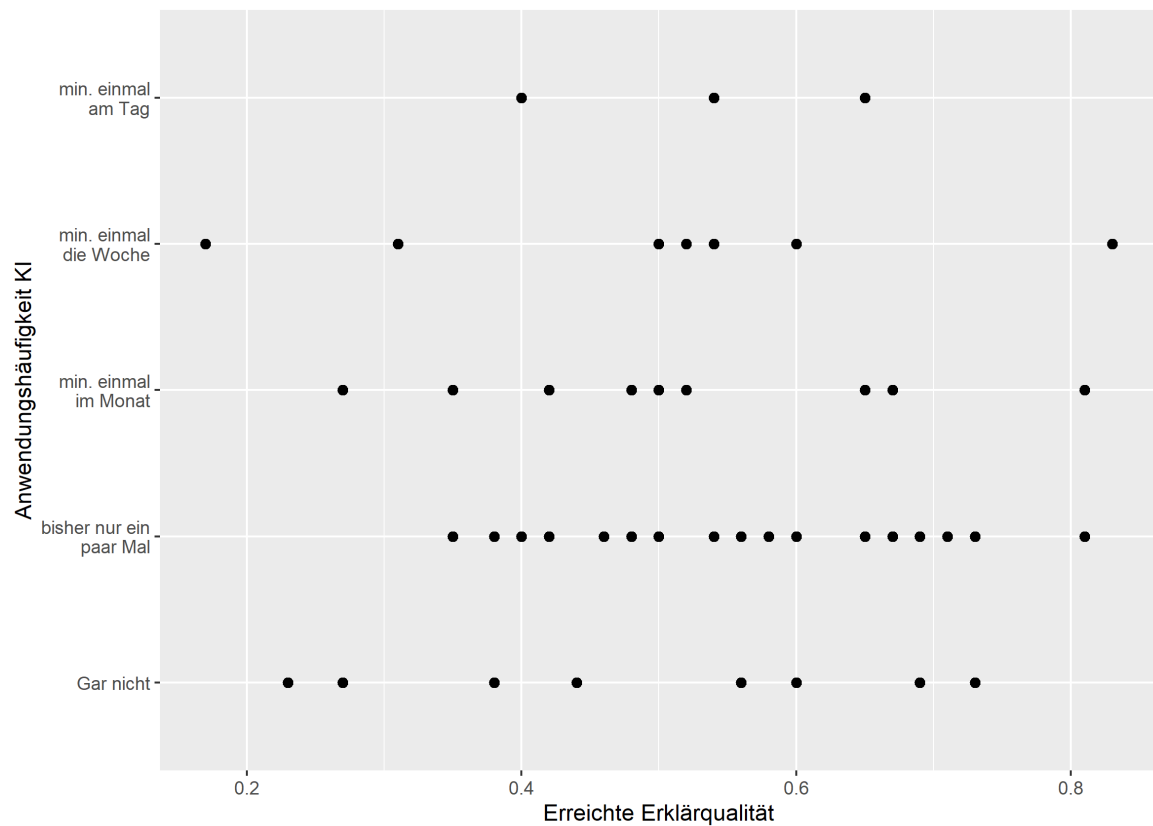


Abbildung 41: Verteilung Anwendungshäufigkeit und erreichte Erklärqualität

Da die Verteilung in Abbildung 41 zwar einen Überblick liefert, aber keine exakte Aussage über eine Korrelation möglich ist, wurde im Anschluss noch der Spearman-Test durchgeführt. Dazu wurden die Werte der Anwendungshäufigkeit von 1-5 skaliert (Gar nicht – min. einmal am Tag), sodass eine Berechnung möglich ist. Das Ergebnis (sichtbar in Abbildung 42) zeigt eindeutig, dass hier keine Korrelation vorliegt.

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	77848.9	-0.064	0.5814

Abbildung 42: Spearman-Korrelation Anwendungshäufigkeit und Erklärqualität

5.6. Exploration 2

Die zweite Exploration betrachtet die Erfahrung der Lernenden und die Zufriedenheit mit den Antworten der KI. Das Ziel hinter dieser Überprüfung ist es, zu überprüfen, ob eine größere Erfahrung für eine stärkere oder schwächere Zufriedenheit sorgt. Dies könnte

Aufschlüsse darüber geben, ob die KI die (eventuell höheren) Ansprüche erfahrenerer Anwender erfüllen kann. Zu beachten ist hierbei, dass eine höhere Zufriedenheit mit einer geringeren Punktzahl bewertet wurde. Betrachtet man die Verteilung in Abbildung 43, so ist der Bereich einer positiven Zufriedenheit (kleiner drei Punkte) stärker ausgeprägt. Gleichzeitig scheint die bisherige Erfahrung beziehungsweise Anwendung gering zu sein. Der dazugehörige Spearman-Korrelationstest zeigt jedoch auf, dass eine Korrelation nicht vorhanden ist.

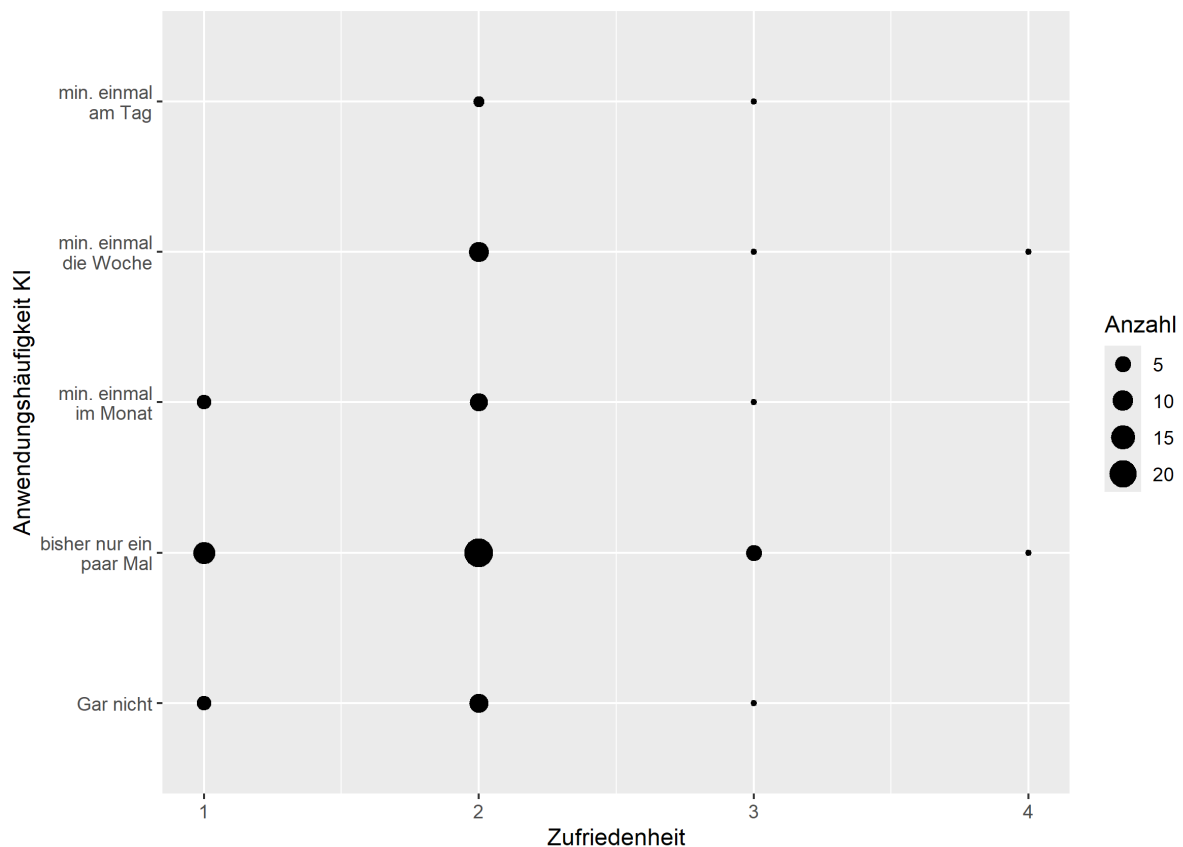


Abbildung 43: Verteilung Anwendungshäufigkeit KI mit Zufriedenheit

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	437.007	0.22	0.4316

Abbildung 44: Spearman-Korrelationstest Anwendungshäufigkeit und Zufriedenheit

5.7. Exploration 3

Die dritte Exploration befasst sich damit, wie zufrieden SchülerInnen mit der Arbeit der künstlichen Intelligenz sind und ob sie sich die Arbeit mit KI im Unterricht in Zukunft wünschen. Hieraus lassen sich eventuell Rückschlüsse ziehen, welche die Hintergründe der Wünsche offenbaren können. Während der Plot hierfür vielversprechend aussieht, so zeigt Abbildung 46, dass hier ebenfalls keine Korrelation vorliegt.

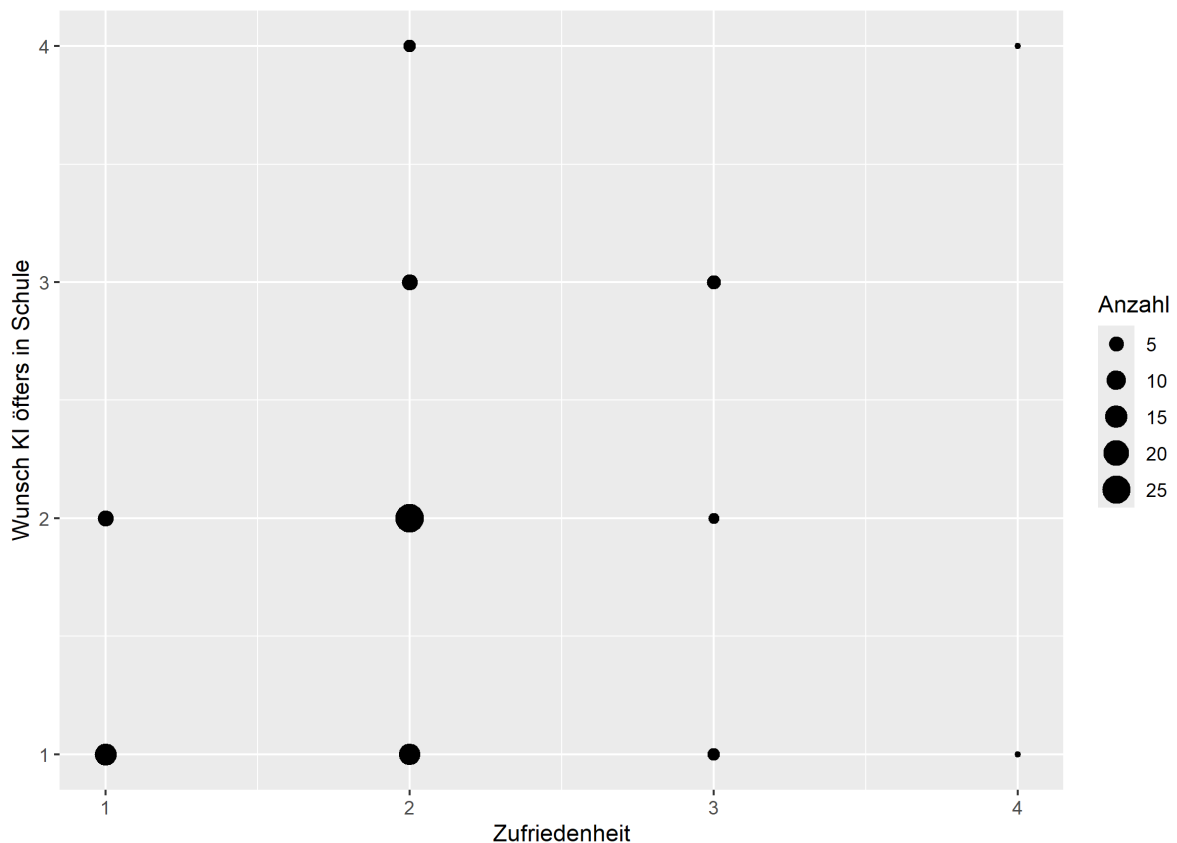


Abbildung 45: Zufriedenheit Schüler und Wunsch auf Arbeit mit KI in Schule

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	191.961	0.127	0.7088

Abbildung 46: Spearman-Test Zufriedenheit Schüler und Wunsch auf Arbeit mit KI in Schule

5.8. Exploration 4

Während die vorherige Exploration die Anwendung im schulischen Kontext betrachtete, geht es in dieser um die Nutzung im privaten Rahmen. Zusammen mit der vorherigen Exploration lassen sich so eventuelle Vorzüge feststellen, so etwa eine Nutzung im privaten aber nicht im schulischen Kontext. Die vorliegenden Dateien zeigen, dass hier ebenfalls keine Korrelation vorliegt.

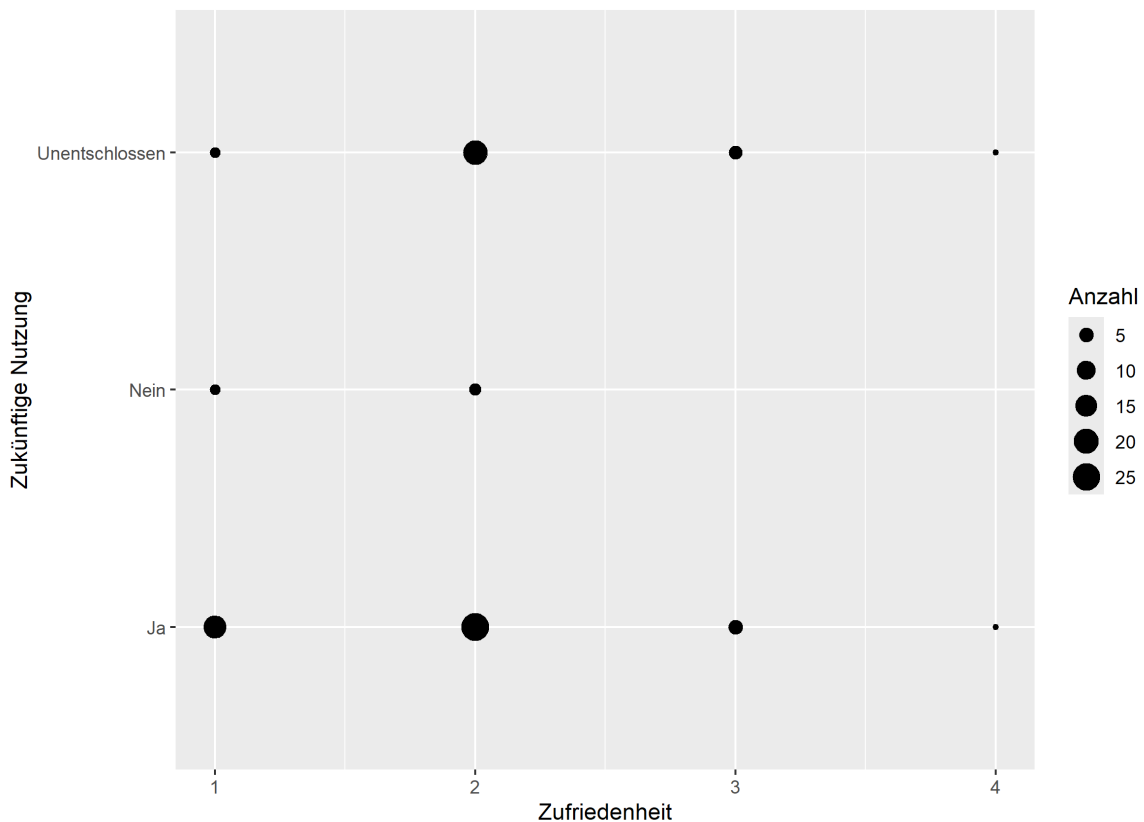


Abbildung 47: Vergleich Zufriedenheit Antworten & zukünftige private Nutzung

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	165	0	1

Abbildung 48: Spearman-Test Zufriedenheit & zukünftige private Nutzung

5.9. Exploration 5

Als nächstes wird der Zusammenhang zwischen der Schwierigkeit der Anwendung und der Zufriedenheit der SchülerInnen mit den Ergebnissen der KI betrachtet. Als Grundlage für diese Untersuchung gilt die Annahme, dass eine höhere Schwierigkeit in der Anwendung zu einer geringeren Zufriedenheit führen kann.

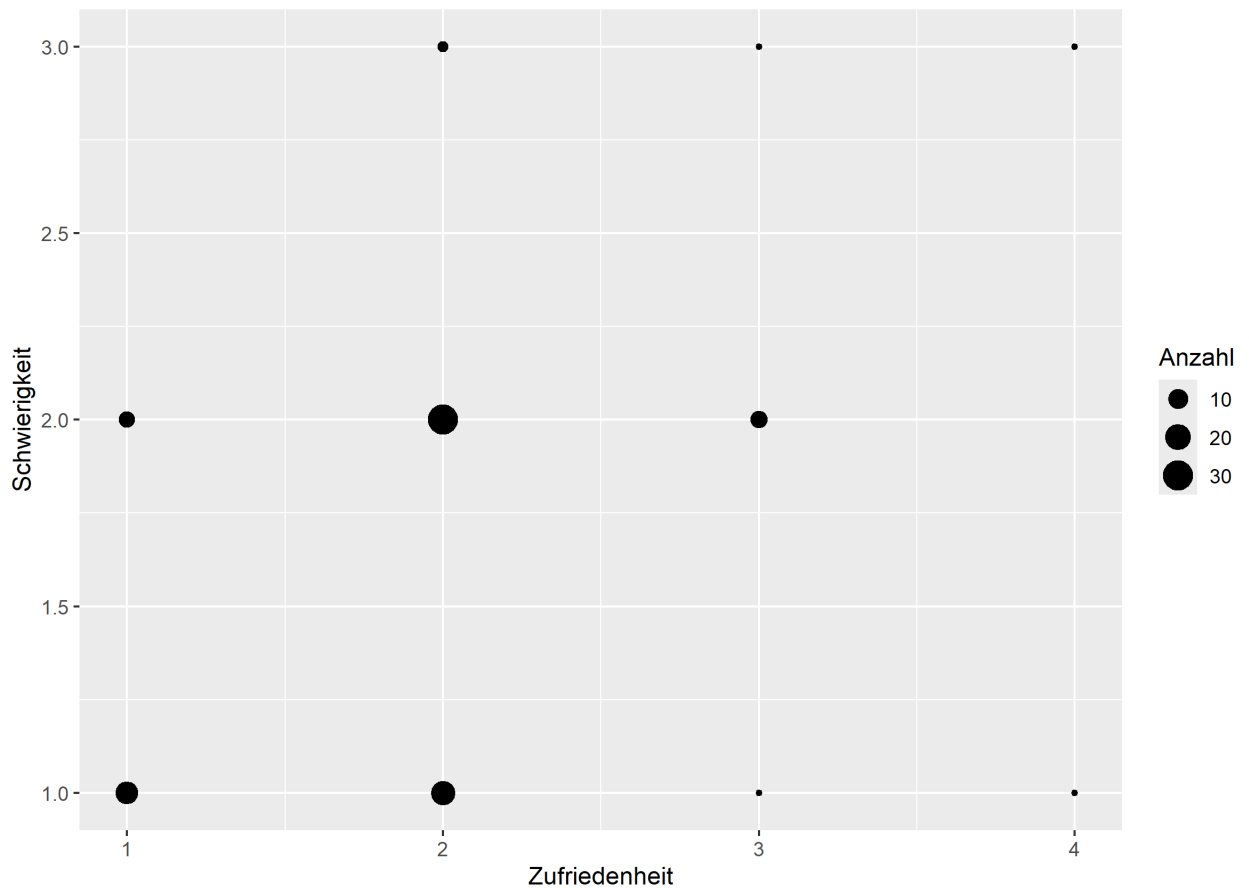


Abbildung 49: Vergleich Zufriedenheit und Schwierigkeit

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	138.766	0.159	0.6609

Abbildung 50: Spearman-Korrelation Zufriedenheit und Schwierigkeit

Unter Berücksichtigung des Vergleiches der Werte und der vorliegenden Ergebnisse des Spearman-Tests kann auch für die fünfte Exploration nicht von einer Korrelation ausgegangen werden.

5.10. Exploration 6

Ebenfalls interessant erscheint die Verbindung von der Zufriedenheit der SchülerInnen und den an die KI gegebenen Informationen über die Schüler zu sein. Das Ziel hinter dieser Exploration ist es, zu beobachten, ob SchülerInnen, welche mehr Informationen an die KI geben, durch eine zufriedenstellendere Antwort belohnt werden. In Abbildung 51 sieht man wie bereits in Abschnitt 5.1, dass die Zufriedenheit größtenteils in der besseren Hälfte (unter 2,5) liegt. Allerdings lässt sich spätestens durch den Spearman-Korrelationstest keine Korrelation zwischen den vorhandenen Informationen und der Zufriedenheit mit den Antworten der KI feststellen.



Abbildung 51: Verteilung Zufriedenheit und Adressateninfos vorhanden

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	68.463	-0.223	0.6315

Abbildung 52: Spearman-Test Zufriedenheit und Adressateninfos vorhanden

5.11. Exploration 7

Die letzte Exploration zielt darauf ab, die Informationen über den Adressaten in Beziehung zur Qualität der Erklärung zu betrachten. Auch diese Exploration betrachtet die Korrelation dieser beiden Items. Allerdings liegt auch hier keine positive Korrelation vor.

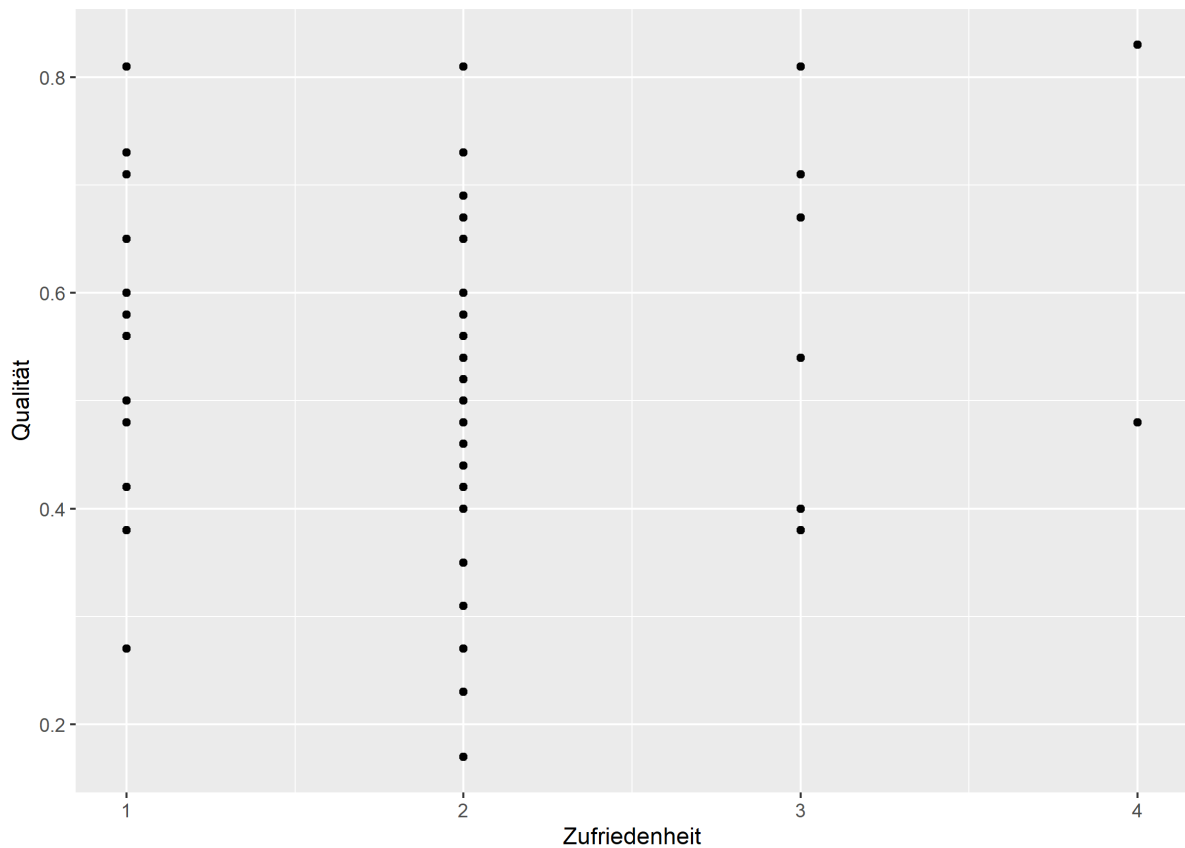


Abbildung 53: Verteilung Zufriedenheit SuS und Qualität der Erklärung

Test	Teststatistik	Koeffizient	p-Wert
Spearman-Korrelation	74960.89	-0.025	0.8319

Abbildung 54: Spearman-Test Zufriedenheit SuS und Qualität der Erklärung

6. Diskussion

Da die gesamte Untersuchung dreigeteilt stattgefunden hat, besitzt die Diskussion ebenfalls diese Grundstruktur. Die Ergebnisse der vorliegenden Arbeit im Bereich Studentenuntersuchung zeigten auf, dass die beiden künstlichen Intelligenzen höhere Mediane und Mittelwerte in der Bewertung erzielen konnten wie die Schulbücher. Blickt man nun genauer auf Bewertungen für die einzelnen Texte und vergleicht diese, so ist eindeutig, dass der Text aus dem Schulbuch „Physik für Realschulen“ signifikant schlechter bewertet wurde. Unterschiede zwischen dem Schulbuch Prisma, der KI Gemini und der KI ChatGPT sind nicht signifikant.

Aus den Schülerfragebögen konnte eine groß positive Bewertung festgestellt werden. Diese zeigte sich durch Mittelwerte und Mediane über der durchschnittlichen Bewertung. Bei der Betrachtung der einzelnen Fragen zeigte sich, dass viele SchülerInnen mindestens einmal mit einer KI gearbeitet haben. Ein wesentlicher Teil der SchülerInnen nutzt die KI auch mehr als einmal im Monat. Hierzu gehören allerdings auch KIs wie Snapchats MyAI. Die Schwierigkeit in der Nutzung der KI wurde von den SchülerInnen im Großteil als einfach bis sehr einfach bewertet. Vergleicht man hierbei die Klassenstufen, so finden in Klassenstufe acht vergleichsweise wenig SchülerInnen den Umgang sehr leicht. Die Bewertung als schwer ist über die Klassenstufen ungefähr gleich verteilt. Auch die Zufriedenheit wurde durchaus hoch bewertet. Knapp 80% der Lernenden bewerteten die Verbindung von KI und Unterricht positiv. Fast jeder fünfte Schüler beurteilt diese Verbindung als sehr schlecht. Während jeder zehnte Schüler sich gegen die Zukunft von KI im Unterricht ausspricht, sind mehr als die Hälfte der Lernenden für eine solche Zukunft. Dies zeigt sich auch in der starken Tendenz zur Anwendung im privaten Leben.

Der Dritte Blick fällt auf die Erklärqualität. In der Gesamtwertung liegen Normalverteilungen vor. Diese ist in einzelnen Clustern jedoch nicht vorhanden. Während die beiden KIs Gemini und ChatGPT größtenteils ähnlich sind, zeigt der schulKI-Assistent, dass er in den meisten Bereichen einen höheren Durchschnitt erreichen kann. Einzig im Bereich Sonstiges haben alle drei KIs ähnliche Werte. Ein Blick auf das Cluster Sachgerecht zeigt eine hohe sachgerechte Punktzahl, welche für H1 genauer betrachtet wird. Ein weiterer interessanter Punkt, welcher für die Auswertung der KI wichtig ist, ist die Information über eine Alters- oder Klassenstufe der SchülerInnen. In jedem dritten Chat hatte die KI keine Informationen dieser Art.

Betrachtet man die einzelnen Hypothesen, so ergaben sich die nachfolgenden Ergebnisse. Hypothese H1 betrachtet das von den KIs wiedergegebene Fachwissen. Sie geht davon

aus, dass das fachliche Gehalt der Antwort korrekt ist, da die KIs in den Wissenstests gut abschneiden. Ein Blick auf die Ergebnisse des Wilcoxon-Tests zeigen auf, dass nur eine der drei KIs, nämlich die schulKI, den Schwellenwert von 80% im Bereich sachgerecht mit den Ausschlusskriterien erreichen konnte. Zur Erinnerung: Die Ausschlusskriterien sind hier die Wertung aller Items höher als der schlechteste Bewertungspunkt und der Korrektheit im positiven Bereich. Ein Blick auf die unteren Konfidenzwerte zeigt, dass die KI ChatGPT 66% und die KI Gemini 71% signifikant überschritten hat. Hier zeigt sich, dass unter gleicher Fragestellung Gemini vor ChatGPT liegt, was den Bewertungsergebnissen aus Abschnitt 2.4 entspricht. Die schulKI, oder vielmehr ChatGPT mit einem entsprechenden Prompt zur Information des Auftrags vor den Schülerfragen, erreichte hier einen sehr hohen Wert von 91%. Der Vergleich mit dem Kruskal-Wallis-Test und anschließendem Dunn-Test zeigte, dass Gemini und ChatGPT nicht signifikant auseinander liegen. Der Abstand dieser beiden künstlichen Intelligenzen zur schulKI ist jedoch signifikant. Eine von drei Intelligenzen hat somit im Cluster Sachgerecht den benötigten Schwellenwert überschritten. Blickt man nun mit diesen Ergebnissen zurück auf die Hypothese H1, so könnte man vorschnell sagen, dass die KIs tatsächlich in der Lage sind, fachliches Wissen korrekt wiederzugeben. Diese Aussage stimmt auch, aber nur für die KI, welche vorab weitere Informationen über den Auftrag hatte. Eine KI, welche ihre Prompts nur von SchülerInnen erhalten hat, welche soeben den Umgang mit einer KI erklärt bekommen haben, konnte die Grenzwerte nicht überschreiten. Für Hypothese H1 bedeutet dies: Ja, KIs sind in der Lage das fachliche Wissen, welches für die Sekundarstufe I in Baden-Württemberg benötigt wird, korrekt wiederzugeben, solange genügend Informationen über den Auftrag vorhanden sind. Die gezeigten Ergebnisse stützen somit die Hypothese H1, da es für künstliche Intelligenzen möglich ist fachlich korrekte Antworten zu generieren.

Die Überprüfung von Hypothese H2 behandelt die Verständlichkeit der Erklärungen einer KI. Da in diesem Punkt Kombinationen aus KI-Erhebungsdaten und Schülerantworten aus dem Schülerfragebogen gezogen werden, kann nur die KI ChatGPT für den Rückschluss berücksichtigt werden. Trotzdem ist es lohnenswert sich die anderen anzusehen. Zuerst wird das Überschreiten der gewichteten Erklärqualität betrachtet. Da in diesem Punkt die Berücksichtigung des Clusters Sonstiges mit den Items Mathematisierung sowie verwendete Beispiele enthalten ist, welches beides Bonusitems sind, kann die Punktzahl größer als 100% sein. Dies ist der Fall für die schulKI. Allerdings konnten auch die beiden anderen künstlichen Intelligenzen zeigen, dass sie den Schwellenwert von 70% für eine verständliche Erklärung überschreiten konnten. Wie es durch die Konfidenzwerte schon angedeutet wird, ist der Unterschied zwischen Gemini oder ChatGPT und der schulKI signifikant. Der zweite Bereich, der der Schülerbewertung, befasst sich mit der Zufriedenheit. Dieser ist so

angeordnet, dass niedrigere Bewertungen eine höhere Zufriedenheit darstellen. Als Schwellenwert wurde vorab eine positive Zufriedenheit, also ein Wert kleiner 2,5 gewählt. Dieser Wert konnte in beiden Klassenstufen signifikant unterschritten werden. Somit sind signifikant mehr als die Hälfte der SchülerInnen der Stichprobe zufrieden (oder sehr zufrieden) mit den Antworten der künstlichen Intelligenz ChatGPT. Der dritte Test, welcher für die Untersuchung von H2 gemacht wurde, betrachtet die Korrelation zwischen der gewichteten Erklärqualität und der Schülerzufriedenheit. Hier konnte keine signifikante Korrelation festgestellt werden. Betrachtet man nun diese drei Ergebnisse, so ist zu erkennen, dass einerseits die künstlichen Intelligenzen verständliche Antworten produzieren können. Ebenso zeigen die SchülerInnen ihre Zufriedenheit mit den Antworten der KI ChatGPT. Einzig eine Verbindung zwischen beiden Werten konnte nicht belegt werden. Diese Ergebnisse liefern keine Hinweise, die der Hypothese H2, welche für die Verständlichkeit der KI-Antworten für SchülerInnen steht, widersprechen.

Für die Untersuchung der dritten Hypothese geht es um die sinnvolle Anwendung von künstlicher Intelligenz in einer Lehr-Lern-Umgebung. Damit Aussagen zu dieser Hypothese getroffen werden können, wurden auch hierfür mehrere Tests durchgeführt. Blickt man zuerst auf die direkte Arbeit mit der KI, so fallen hier, wie in Abschnitt 4.4 erklärt, die Bereiche Zufriedenheit, Schwierigkeit und Wunsch der Arbeit mit der KI aus der Schülersicht hinein. Ebenso wird die erfasste Erklärqualität eingebracht. Da hier mehrere Daten und Datentypen erfasst wurden, sind sie so angeordnet, dass ein größerer Wert eine bessere Performance signalisiert. Diese Anwendung innerhalb des Unterrichtes konnte den vorab bestimmten Schwellenwert von 50% ausschließlich in Klasse 10 signifikant übersteigen. Ein Blick auf den Wilcoxon-Rangsummentest zum Vergleich beider Klassenstufen zeigt, dass auch der Unterschied zwischen beiden Gruppen signifikant ist. Aus diesen Ergebnissen ergibt sich eine sinnvolle Anwendung der KI im Physikunterricht, die des Tutors. Der zweite Teil dieser Hypothesenüberprüfung befasst sich ausschließlich mit den Studentenbewertungen. Da dieser Teil keine Verbindung zu anderen Daten hat, können hier wieder beide KIs, ChatGPT und Gemini, genutzt werden. Die Ergebnisse im direkten Vergleich zeigten auf, dass einzig die KI ChatGPT einen signifikanten Unterschied zum Schulbuch Physik für Realschulen erzielte. Die restlichen Paare wurden nicht signifikant unterschiedlich bewertet. Hinsichtlich der Hypothese bedeutet dies, dass, aufgrund der Nutzungsauswahl der Lehramtsstudenten, die KI auf gleicher Höhe wie die Texte aus den Schulbücher stehen. Somit ergibt sich hieraus ebenfalls ein sinnvolles Anwendungsgebiet für KI im Schulunterricht, nämlich das der Lehrtextgestaltung. Für Hypothese H3 heißt dies, dass die Ergebnisse nicht dazu ausreichen die Hypothese zu verwerfen.

Im weiteren Rahmen der Auswertung der Untersuchungsdaten wurde eine Exploration gestartet. Ziel dabei war es eventuelle Datenkombinationen zu entdecken, welche eventuelle Korrelationen aufzeigen und so weitere Hinweise auf eventuelle Sachverhalte liefern. Allerdings konnte in keiner der sieben Explorationsen eine signifikante Korrelation festgestellt werden.

Blickt man nun auf die eben beschriebenen Ergebnisse zurück, so lässt sich erkennen, dass bei keiner der drei Hypothesen genügend Ergebnisse vorliegen, um diese zu verwerfen. Ebenfalls ist noch einmal zu erwähnen, dass mit der KI schulKI nicht der Service zu verstehen ist, sondern im Endeffekt die KI ChatGPT, welche mit einem Initialprompt (siehe Prompt 6) schon Informationen über den Auftrag erhalten hat. In der Auswertung und Präsentation wurde der Name schulKI nur zur Differenzierung zwischen den KIs beibehalten, da ansonsten eine größere Verwirrung durch die Namensgebung wie zum Beispiel ChatGPT und ChatGPT-Assistenten vorhanden gewesen wäre.

Im Rahmen der Arbeit wurden sowohl quantitative wie auch qualitative Methoden mit dem Ziel der umfassenden Analyse der Anwendung von künstlicher Intelligenz in Lehr-Lern-Umgebungen kombiniert. Der nächste Abschnitt dient der kritischen Betrachtung der Methodenauswahl. Ein erster Blick betrachtet die Stärken der Arbeit. Zum einen ermöglicht die Betrachtung der Lehramtsstudenten eine fachliche sowie didaktische Sicht auf die Texte von künstlichen Intelligenzen. Ebenso wird hierdurch ermöglicht, dass zukünftige Lehrkräfte mit ihrem bereits geschulten Blick die Texte bewerten und ordnen. Dies ermöglicht einen Blick in den zukünftigen Unterricht. Außerdem lassen sich hierdurch Vergleiche zu Schulbuchtexten durchführen. Eine weitere Stärke ist die Kombination der Schülerbewertungen nach der direkten Arbeit mit einer KI zur Wissensgenerierung mit der Expertenauswertung der dabei entstandenen Chats. Hierdurch wird ein dualer Blick auf die Forschungsfrage ermöglicht, welcher neben der fachlichen und sachlichen Seite auch die der SchülerInnen mit einbezieht. Weiterhin stellte sich die Erfassung der Chats als Stärke heraus. Der ursprüngliche Plan, jeder Gruppe einen eigenen Zugang zu geben und diesen dann nach der Erhebung anzupassen oder die Chats der KI am Ende der Stunde einzusammeln wäre ein großer organisatorischer Aufwand gewesen. Die hier angewandte Nutzung des Services schulKI ermöglichte es, diese Sammlung der Chats aus dem Unterricht herauszuschieben und den Verwaltungsaufwand zu minimieren. Die SchülerInnen konnten sich mit einem QR-Code anmelden und mussten keine Passwörter eingeben. So blieb mehr Zeit für die Anwendung der KI übrig, was die Erhebung der Forschungsdaten erleichterte. Die letzte Stärke ist die Stichprobe. Mit einer Anzahl von 80 SchülerInnen aufgeteilt auf 42 KI-Chats ist diese nicht unerheblich. Allerdings, und damit kommen wir nun in den Bereich der Schwächen dieser Arbeit, ist die Stichprobe mit „nur“ 80 SchülerInnen nicht so

aussagekräftig und genau, wie wenn es fünf- oder zehnmal so viele Lernende gewesen wären. Hier kommt auch noch hinzu, dass die Schulen alle aus einem lokalen Umfeld kommen. Sie sind nicht aus einer Großstadt, sondern eher dem ländlichen Teil. Daraus ergibt sich eventuell ein Ergebnis, welches im Blick auf Baden-Württemberg (oder auch Deutschland) vielleicht unzuverlässig ist. Die Aussagekraft der Ergebnisse ist zudem eingeschränkt durch die verwendete KI. Durch die Vorgaben zur Nutzung von KI (z.B. DSGVO-Konformität) konnten keine Vergleiche zwischen den KIs auf Schülerebene gezogen werden. Alle SchülerInnen arbeiteten mit derselben KI. Somit konnte in der Auswertung nur die KI ChatGPT (genutzt über den Service der schulKI) in Betracht gezogen werden. Eine weitere Schwäche in Bezug auf die Stichprobe, diesmal die der studentischen Interviews mit einer Stärke von zwölf, ist die Aussagekraft dieser Umfrage. Die Aussagekraft ist stark eingeschränkt, da hier nur eine kleine Stichprobe stattgefunden hat. Dies ist unter anderem dem hohen Aufwand durch eine eins zu eins Betreuung während der Befragung zu verdanken.

Bei einer erneuten Durchführung dieser Untersuchung müssten verschiedene Punkte optimiert werden, welche sich im Laufe der Erhebung herauskristallisierten. Der erste Punkt befasst sich dabei mit den Studenteninterviews. Eine Möglichkeit, um den Arbeitsaufwand zu reduzieren und damit die Effizienz zu erhöhen wäre es, von Interviews auf einen Online-Fragebogen umzusteigen. Durch diese Umstellung wäre keine eins zu eins Betreuung mehr nötig, was die Erhebung der Daten vereinfacht. Der nächste Verbesserungspunkt wäre das Design der Studie. Hier könnte ein Kontrollgruppendesign verwendet werden um effektiv die Unterschiede zwischen den KIs wie zum Beispiel ChatGPT und ChatGPT mit extra Prompt zu erkunden. Dies würde schlussendlich Rückschlüsse ermöglichen, welche neben der Expertenauswertung auch die Schülerbewertung in eine KI-differenziertere Betrachtung einbeziehen. Eine weitere Verbesserung, welche zu genaueren Daten führen könnte, wäre es, die oben angesprochene Schwäche der Lokalisierung zu entfernen. Hierzu müssten aus unterschiedlichen Regionen Schulklassen untersucht werden, damit ein möglichst breites Spektrum an SchülerInnen aus ländlichen Schulen sowie Stadtschulen in die Untersuchung einfließt. Eine weitere Verfeinerung der Ergebnisse könnte erreicht werden, indem die Anzahl der bewertenden Experten erhöht wird. Dadurch wird der subjektive Anteil in der Auswertung reduziert und die interrater Reliabilität erhöht. Eine weitere Anpassung für die Forschung wäre eine Ausdehnung des Zeitraumes. Durch eine größere Zeitspanne könnten neben mehr Daten auch eine mögliche Entwicklung der Fähigkeit des Prompt-Engineerings der SchülerInnen erfasst werden. Gleichzeitig zeigen die Daten gegen Ende der Erhebung genauer den Umgang anstelle der anfänglichen Lernphase. Schließlich fällt der Fokus dann stärker auf die Anwendung der KI als auf das Einarbeiten in jene. Auch wenn die Methodik verschiedene Schwächen hatte, konnten die Stärken für einen meist reibungslosen Ablauf

sorgen. Ein Verhindern der Schwächen wäre für den Rahmen dieser Arbeit mit einem zu großen Zeitaufwand verbunden gewesen. Manche der Schwächen sind sogar erst während der Erhebung aufgefallen, sodass ein Umstrukturieren dann die bereits laufende Datenerhebung beeinflusst hätte. Alles zusammen zeigte die methodische Diskussion, dass die gewählten Methoden größtenteils geeignet waren, um die Forschungsfrage zu beantworten. Gleichzeitig zeigen die Schwächen auf, dass die Ergebnisse für eine Übertragbarkeit auf andere, wenn auch ähnliche Kontexte, mit Vorsicht zu beachten sind.

Doch was bedeuten nun die Ergebnisse dieser Arbeit? Wie können sie sich auf den Schulunterricht auswirken? Zusammengefasst bedeuten die Ergebnisse, dass künstliche Intelligenzen wie ChatGPT das Potenzial besitzen, um als ergänzende Lehrmittel im Physikunterricht einer Sekundarstufe I eingesetzt zu werden. Als Aufgaben könnten diese KI-Systeme Tutoren- oder Assistenzrollen übernehmen und Lehrkräfte hierdurch entlasten. Dabei sollten sie in Phasen der individuellen Förderung eingesetzt werden und können einen hohen Lehrkraftbedarf für Erklärungen abpuffern. Ebenfalls können diese KI-Systeme bei der Erstellung der Unterrichtsmaterialien unterstützen und so für die Klassengruppe und den aktuellen Lernstand differenzierte Lernmaterialien erstellen. Für den Schulunterricht beziehungsweise auch die Lehrerbildung bedeutet dies, dass SchülerInnen wie auch Lehrkräfte den Umgang mit KIs erlernen sollen. Im Zuge dessen sollte eine Schulung im Bereich Prompt-Engineering stattfinden. Somit kann die Verständlichkeit und Relevanz der KI-Antworten verbessert werden. Da diese zusätzlichen Lerninhalte teilweise nicht in den aktuellen Unterrichtsalltag hineinpassen, könnte es von Notwendigkeit sein, das Curriculum anzupassen, um diese Bausteine mit in den Unterrichtsalltag einzubauen. Dazu könnte unter anderem der Bereich der Medienkompetenz weiter ausgebaut werden. Auch die Anwendung von KI-gestützten Lernplattformen könnte den Lernenden außerhalb der Schule Unterstützung beim Lernen und ihren individuellen Problemen geben. Somit könnte ein Problem, welches zum Beispiel beim Lernen eines Themas auftritt, direkt angegangen werden und nicht erst in der nächsten Physikstunde, welche vielleicht erst in zwei Wochen stattfindet. Ebenso sollten weitere Forschungen und Pilotprojekte stattfinden, welche gerade die langfristigen Effekte untersuchen. Für die Zukunft ist es neben der Entwicklung im Bildungsbereich relevant, dass die Sachgerechtigkeit der künstlichen Intelligenzen weiterhin ansteigt. Dies kann zum Beispiel durch das Einbinden von Datenbanken (wie bereits in Abschnitt 2.1.2 angesprochen) geschehen. Ein weiterer Punkt, welcher die Einbindung in den Unterricht verbessern würde, wäre die Entwicklung von speziellen KIs, welche auf die Erklärung von Lehrinhalten für die Schule spezialisiert wurden. Diese könnten dann, neben den fachlichen Inhalten auch die pädagogische Seite berücksichtigen. Eine konkrete Forschung, welche direkt an diese Arbeit anknüpfen könnte, wäre eine mit einem ähnlichen Setting.

Allerdings wird eine KI verwendet, welche schon vor der Interaktion mit den SchülerInnen mit einem Initialprompt geschult wurde. Dadurch kennt sie den Auftrag und schon einige Eckdaten wie den Bildungsplan und kann diese dann für eine Antwortgenerierung anwenden.

7. Fazit

Ziel der Arbeit war es, die Anwendungsmöglichkeiten von künstlicher Intelligenz im Physikunterricht der Sekundarstufe I zu untersuchen. Die Ergebnisse zeigen, dass die Anwendung von KIs wie ChatGPT fachlich korrekte Ergebnisse liefern können, solange der Benutzer genügend Informationen über den Prompt liefert, also einen ausreichend präzisen Prompt formuliert. Die SchülerInnen bewerteten die Nutzung von KIs im Physikunterricht größtenteils positiv. Auch die Verständlichkeit der generierten Antworten wird als zufriedenstellend empfunden. Die in Abschnitt 3 aufgestellten Hypothesen konnten nicht widerlegt werden.

1. **Fachliche Korrektheit:** KIs können fachlich korrekte Antworten für den Physikunterricht in der Sekundarstufe I generieren, solange sie vom Benutzer genügend Informationen über das Ziel der Antwort erhalten.
2. **Verständlichkeit:** Künstliche Intelligenzen sind in der Lage für Schüler verständliche Erklärungen zu liefern, auch wenn keine signifikante Korrelation zwischen der Erklärqualität und Schülerzufriedenheit festgestellt werden konnte.
3. **Einsatz in Lehr-Lern-Umgebungen:** KIs können sinnvoll in und für Lehr-Lern-Umgebungen eingesetzt werden. Als Anwendungsgebiet kann etwa die direkte Interaktion mit den SchülerInnen, aber auch die Vorbereitung von Erklärungen oder individualisierten Fachtexten durch die Lehrkraft gelten.

Somit liefert diese Arbeit neue Erkenntnisse zur Anwendung und der Tauglichkeit von KIs im Kontext Schule und zeigt damit auch auf, welche didaktische und technische Anpassung für eine erfolgreiche Integration in den Physikunterricht notwendig sind. Ebenso zeigt diese Arbeit auf, wo zukünftige Studien ansetzen sollten und welche Punkte im Bereich der Methodik überarbeitet werden müssten. So sollten größere und geografisch diversifizierte Stichproben einbezogen werden. Insbesondere die Auswirkung der Nutzung von KIs im Schulunterricht auf die nachhaltige Wissensvermittlung sowie die Entwicklung von Problemlösekompetenzen sollte untersucht werden.

8. Literaturverzeichnis

- [1] „Was Sie über ChatGPT wissen müssen“, capital.de. Zugriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://www.capital.de/wirtschaft-politik/chatgpt--was-sie-ueber-den-revolutionaeren-chatbot-wissen-muessen-32976718.html>
- [2] tagesschau.de, „Künstliche Intelligenz: Was kann der Chatbot ChatGPT?“, tages-schau.de. Zugriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://www.tagesschau.de/wirtschaft/digitales/chatgpt-microsoft-ki-texte-101.html>
- [3] B. Bucher, „ChatGPT Anleitung: KI auf Deutsch nutzen“, CHIP. Zugriffen: 24. August 2024. [Online]. Verfügbar unter: https://www.chip.de/news/ChatGPT-Anleitung-KI-auf-Deutsch-nutzen_184795765.html
- [4] „Was ist My AI auf Snapchat und wie verwende ich sie?“, Snapchat Support. Zugriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://help.snapchat.com/hc/de/articles/13266788358932-Was-ist-My-AI-auf-Snapchat-und-wie-verwende-ich-sie>
- [5] „GitHub Copilot · Your AI pair programmer“, GitHub. Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://github.com/features/copilot>
- [6] „Apple Intelligence für iPhone, iPad und Mac“, Apple Newsroom (Deutschland). Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://www.apple.com/de/newsroom/2024/06/introducing-apple-intelligence-for-iphone-ipad-and-mac/>
- [7] „AI & Robotics | Tesla Germany“, Tesla. Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: https://www.tesla.com/de_de/AI
- [8] „Verfolgen Sie Ihre Schritte mit ‚Recall‘ - Microsoft-Support“. Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://support.microsoft.com/de-de/windows/verfolgen-sie-ihre-schritte-mit-recall-aa03f8a0-a78b-4b3e-b0a1-2eb8ac48701c>
- [9] „Deep-Learning- und KI-Lösungen von NVIDIA“, NVIDIA. Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://www.nvidia.com/de-de/deep-learning-ai/products/solutions/>
- [10] P. Hutchinson, „German AI Startup Landscape 2024“, AppliedAI institute for europe. Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: https://www.appliedai-institute.de/en/hub/2024-ai-german-startup-landscape?token=BrCJeNU3H-x4Oop7F_mogP-PNz9Uhhoj
- [11] „Wie man KI-Betrüger enttarnt: Ferrari entgeht Deepfake-Masche“, FAZ.NET. Zugriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://www.faz.net/aktuell/wirtschaft/unternehmen/wie-man-ki-betrueger-enttarnt-ferrari-entgeht-deepfake-masche-19888407.html>
- [12] N. Lewanczik, „Das kann die neue Super-KI ChatGPT“, OnlineMarketing.de. Zugriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://onlinemarketing.de/technologie/chatgtp-was-ist-das-wie-fuer-marketing-einsetzen>
- [13] S. Albrecht, „ChatGPT und andere Computermodelle zur Sprachverarbeitung – Grundlagen, Anwendungspotenziale und mögliche Auswirkungen“. Zugriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://publikationen.bibliothek.kit.edu/1000158070>
- [14] S. Ekin, „Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices“, Zugriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://www.authorea.com/doi/full/10.36227/techrxiv.22683919.v2?com-mit=190c077fb72a68b49bf64cd595d5642e67440209>
- [15] I. De Florio-Hansen, *Digitalisierung, Künstliche Intelligenz und Robotik: eine Einführung für Schule und Unterricht*, 1. Aufl. Münster; New York: Uni-Taschenbücher GmbH, 2020.
- [16] W. Ertel, *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*. in Computational Intelligence. Wiesbaden: Springer Fachmedien Wiesbaden, 2021. doi: 10.1007/978-3-658-32075-1.
- [17] Zehner, Fabian, „Künstliche Intelligenz. Ihr Potenzial und der Mythos des Lehrkraft-Bots“, Feb. 2020, doi: 10.25656/01:17561.
- [18] R. Otte, *Künstliche Intelligenz für dummies*, 1. Aufl. Weinheim: WILEY-VCH Verlag GmbH & Co.KGaA, 2019.

- [19] S. Russell und P. Norvig, *Künstliche Intelligenz: Ein Moderner Ansatz*, 4. Aufl. München: Pearson Deutschland GmbH, 2023.
- [20] AI21 Labs Ltd., „human or not?“ [Online]. Verfügbar unter: <https://www.humanornot.ai>
- [21] Fraunhofer-Institut für Kognitive Systeme IKS, „Künstliche Intelligenz (KI) und maschinelles Lernen“, Fraunhofer-Institut für Kognitive Systeme IKS. Zugegriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://www.iks.fraunhofer.de/de/themen/kuenstliche-intelligenz.html>
- [22] robert Weber und peter Seeberg, *KI in der Industrie*, 1. Auflage. München: Carl Hanser Verlag, 2020.
- [23] P. Baker, *ChatGPT für dummies*, 1. Aufl. Weinheim: WILEY-VCH Verlag GmbH & Co.KGaA, 2024.
- [24] T. Brandstetter, „Wie die KI genügsamer wird“, *Bild der Wissenschaft*, Bd. 2024, Nr. 8, S. 42–45, Aug. 2024.
- [25] U. Eberl, „Geistreiche Debatten mit smarten Maschinen“, *Bild der Wissenschaft*, Bd. 2023, Nr. 7, S. 28–36, Juli 2024.
- [26] W. Gan, Z. Qi, J. Wu, und J. C.-W. Lin, „Large Language Models in Education: Vision and Opportunities“, 22. November 2023, *arXiv*: arXiv:2311.13160. Zugegriffen: 22. Juni 2024. [Online]. Verfügbar unter: <http://arxiv.org/abs/2311.13160>
- [27] E. Kasneci u. a., „ChatGPT for good? On opportunities and challenges of large language models for education“, *Learning and Individual Differences*, Bd. 103, S. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.
- [28] Ø. O. Skaar, „A Call for Bildung“, *Nordic Journal of Digital Literacy*, Bd. 19, Nr. 2, S. 101–111, Juni 2024, doi: 10.18261/njdl.19.2.4.
- [29] Gesellschaft für Informatik, „Positionspapier: Künstliche Intelligenz in der Bildung“. 17. Juli 2023. Zugegriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://dl.gi.de/server/api/core/bitstreams/7c533204-8a9e-4436-91a8-069b7d74fc8d/content>
- [30] S. Küchemann, A. Dengel, und J. Kuhn, „Künstliche Intelligenz im Lehr-Lernprozess von MINT-Fächern: Vom adaptiven Lernmaterial zur Implementation in die Lehrkräftebildung“, *Bildung und Erziehung*, Bd. 74, S. 313–330, Aug. 2021, doi: 10.13109/buer.2021.74.3.313.
- [31] S. Seegerer, M. Gläser-Zikuda, F. Hofmann, und R. Romeike, *KI-Technologieentwicklung und Bildung als miteinander in Beziehung stehende Aufgaben für die Lehrerbildung*. Gesellschaft für Informatik e.V., 2023, S. 447–452. Zugegriffen: 21. Juni 2024. [Online]. Verfügbar unter: <https://dl.gi.de/handle/20.500.12116/43175>
- [32] tagesschau.de, „Saarland: Weniger Klassenarbeiten und mehr KI für Saar-Schüler“, tagesschau.de. Zugegriffen: 12. September 2024. [Online]. Verfügbar unter: <https://www.tagesschau.de/inland/regional/saarland/sr-weniger-klassenarbeiten-und-mehr-ki-fuer-saar-schueler-100.html>
- [33] A. Menn, „Künstliche Intelligenz: Ein Jahr ChatGPT: Diese Grafiken zeigen, wer die Gewinner des KI-Hypes“. Zugegriffen: 3. September 2024. [Online]. Verfügbar unter: <https://www.wiwo.de/technologie/digitale-welt/kuenstliche-intelligenz-ein-jahr-chatgpt-diese-grafiken-zeigen-wer-die-gewinner-des-ki-hypes-sind/29531946.html>
- [34] „Introducing Gemini: our largest and most capable AI model“, Google. Zugegriffen: 4. September 2024. [Online]. Verfügbar unter: <https://blog.google/technology/ai/google-gemini-ai/>
- [35] „Midjourney“. Zugegriffen: 4. September 2024. [Online]. Verfügbar unter: <https://www.midjourney.com/home>
- [36] Black Forest Labs, „Black Forest Labs - Frontier AI Lab“, Black Forest Labs. Zugegriffen: 4. September 2024. [Online]. Verfügbar unter: <https://blackforestlabs.ai/>
- [37] Mathias Brandt, „Infografik: Digital ist schneller“, Statista Daily Data. Zugegriffen: 9. September 2024. [Online]. Verfügbar unter: <https://de.statista.com/infografik/7573/geschwindigkeit-mit-der-sich-technologien-verbreiten>
- [38] T. B. Brown u. a., „Language Models are Few-Shot Learners“, 22. Juli 2020, *arXiv*: arXiv:2005.14165. Zugegriffen: 23. August 2024. [Online]. Verfügbar unter: <http://arxiv.org/abs/2005.14165>

- [39] C. Krettek, „ChatGPT“, *Unfallchirurgie*, Bd. 126, Nr. 3, S. 252–254, März 2023, doi: 10.1007/s00113-023-01296-y.
- [40] OpenAI u. a., „GPT-4 Technical Report“, 4. März 2024, *arXiv*: arXiv:2303.08774. Zugriffen: 12. September 2024. [Online]. Verfügbar unter: <http://arxiv.org/abs/2303.08774>
- [41] „GPT-4“. Zugriffen: 12. September 2024. [Online]. Verfügbar unter: <https://openai.com/index/gpt-4-research/>
- [42] N. C. Hughes, „ChatGPT 5: Mögliches Erscheinungsdatum, Funktionen und Preise“, *Techopedia Deutschland*. Zugriffen: 25. Juni 2024. [Online]. Verfügbar unter: <https://www.techopedia.com/de/chatgpt-5-release>
- [43] „Welche Möglichkeiten Gemini-Apps bieten und andere häufig gestellte Fragen“, *Gemini*. Zugriffen: 21. September 2024. [Online]. Verfügbar unter: <https://gemini.google.com/faq>
- [44] Gemini Team Google, „Gemini: A Family of Highly Capable Multimodal Models“, 2024. Zugriffen: 21. September 2024. [Online]. Verfügbar unter: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
- [45] Google, *Gemini*. [Online]. Verfügbar unter: <https://gemini.google.com/app?hl=de>
- [46] J. White u. a., „A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT“, 21. Februar 2023, *arXiv*: arXiv:2302.11382. doi: 10.48550/arXiv.2302.11382.
- [47] L. Giray, „Prompt Engineering with ChatGPT: A Guide for Academic Writers“, *Ann Bio-med Eng*, Bd. 51, Nr. 12, S. 2629–2633, Dez. 2023, doi: 10.1007/s10439-023-03272-4.
- [48] Zentrum für Schulqualität und Lehrerbildung Baden-Württemberg, „KI FAQ“. Zugriffen: 19. September 2024. [Online]. Verfügbar unter: <https://zsl-bw.de/Lde/startseite/lernen-ueberall/ki-faq>
- [49] T. Bartelborth, *Erklären*. Walter de Gruyter, 2012.
- [50] H. Klärner, *Der Schluß auf die beste Erklärung*. Walter de Gruyter, 2013.
- [51] C. Kulgemeyer, „Gelingensbedingungen physikalischer Erklärungen“, *PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung*, Dez. 2013, Zugriffen: 22. Juni 2024. [Online]. Verfügbar unter: <https://ojs.dpg-physik.de/index.php/phydid-b/article/view/460>
- [52] C. Kulgemeyer, „Lehrkräfte erklären Physik“, *Naturwissenschaften im Unterricht - Physik*, Bd. 27. Jahrgang, Nr. 152, S. 2–9, Apr. 2016.
- [53] C. Tschentscher und R. Berger, „Wie kann man gute Erklärungen mit Lernenden Trainieren?“, *Naturwissenschaften im Unterricht - Physik*, Bd. 27. Jahrgang, Nr. 152, S. 16–19, Apr. 2016.
- [54] S. Heinicke, „Erklären - mehr als Worte“, *Naturwissenschaften im Unterricht - Physik*, Bd. 27. Jahrgang, Nr. 152, S. 22–29, Apr. 2016.
- [55] E. Tomczyszyn und C. Kulgemeyer, „Wie kann man eine verständliche Lehrererklärung vorbereiten?“, *Naturwissenschaften im Unterricht - Physik*, Bd. 27. Jahrgang, Nr. 152, S. 10–15, Apr. 2016.
- [56] T. Ehmke, P. Kuhl, und M. Pietsch, Hrsg., *Lehrer. Bildung. Gestalten: Beiträge zur empirischen Forschung in der Lehrerbildung*, 1. Auflage. Weinheim Basel: Beltz Juventa, 2019.
- [57] „schulKI“. Zugriffen: 21. September 2024. [Online]. Verfügbar unter: <https://schulki.de>
- [58] M. Barmeier, *Prisma Physik. 9/10: Schülerbuch* / Marion Barmeier [und 5 weitere, Differenzierende Ausgabe Baden-Württemberg, 1. Auflage. Stuttgart Leipzig: Ernst Klett Verlag, 2018.
- [59] *Physik für Realschulen. 10, [Hauptband]*, 1. Aufl. Berlin: Cornelsen, 1998.
- [60] R. (31. Oktober 2024). [Windows]. Verfügbar unter: <https://www.r-project.org>
- [61] *RStudio Desktop*. [Online]. Verfügbar unter: <https://posit.co/download/rstudio-desktop/>
- [62] OpenAI, *ChatGPT 4o*.

9. Anhang

9.1. Prompt-Verzeichnis

Prompt Nr.	Prompt
1	Wenn ein Paketbote für 100 Pakete drei Stunden benötigt, wie viel Zeit benötigen dann 10 Paketboten für 1000 Pakete? Erkläre mir deine Lösung Schritt für Schritt.
2	Vergleiche die KI Modelle ChatGPT 4o von OpenAI sowie Gemini von Google.
3	Erstelle ein Bild. Das Bild zeigt eine Szene aus einer japanischen Landschaft. Im Vordergrund sieht man Kinder unter einem blühenden Kirschbaum spielen, während im Hintergrund auf einem kleinen See ein kleines Holzboot treibt. Hinter dem See, auf der rechten Seite des Bildes, ist ein japanischer Tempel zu erkennen.
4	Erstelle ein Bild, welches Kinder in der Schule zeigt. Die Schule ist futuristisch eingerichtet und modern. Es wird das Jahr 2060 geschrieben.
5	Erstelle ein Bild. In dem Bild sieht man den Times Square in New York. Auf den Billboards ist Werbung für den Wettkampf zwischen Googles KI Gemini und OpenAIs KI GPT-4 zu sehen.
6	Du bist ein Tutor für den Physikunterricht in den Klassen 7-10 einer Sekundarstufe 1 in Baden-Württemberg. Deine Aufgabe ist es Schülern das physikalische Wissen (erneut) zu erklären. Verwende dazu Informationen aus dem Bildungsplan und dem Internet. Falls dein Gegenüber nicht sein Alter oder die Klassenstufe nennt, frage in deiner ersten Nachricht danach. Bis du diese Informationen hast, gibst du keine fachlichen Informationen weiter. Formuliere deine Antworten in altersgerechter Sprache und verwende das Vorwissen, welches du aus dem Bildungsplan von Baden-Württemberg (hochgeladene Datei) beziehst. Sprich die SchülerInnen mit Du an. Wenn du deine Erklärung gestaltest, achte darauf als erstes das Ziel der Erklärung zu nennen und einen Ausblick auf die nächsten Schritte zu geben. Schreibe diese jedoch nicht als eigener Punkt (Ziel:) sondern formuliere dies in deinen Nachrichtenanfang. Fokussiere dich auf die Inhalte der Frage und erstelle eine altersgerechte Antwort. Betrachte die Frage immer aus

	physikalischer Sicht. Gehe auf das Vorwissen der Fragenden ein und passe die Erklärung an die kognitive Leistungsfähigkeit des Alters an. Versuche die Schüler kognitiv zu aktivieren und in die Informationsgewinnung einzubeziehen. Stelle gerne Rückfragen um ein Verständnis zu erfassen und verbessere notfalls die von den Schülern wiedergegebenen Inhalte. Wenn du Beispiele verwendest, halte diese kurz aber prägnant. Sollte es zu einem Thema eine Mathematisierung geben, welche für die Sekundarstufe 1 relevant ist, so gebe die Formel an (Beispiel Hebelgesetz). Allerdings sollte eine Mathematisierung immer in einfachen Je-Desto-Aussagen enden. Halte die Antwort altersentsprechend kurz, ohne Inhalte auszulassen. Gebe trotzdem Inhalte weiter, die dazu gehören könnten und die Erklärung verbessern.
7	Erstelle eine gut verständliche Erklärung zum Thema Halbwertszeit für 16-jährige Schüler im Physikunterricht der 10. Klasse. Die Erklärung soll die Definition der Halbwertszeit, ein Beispiel und die Anwendung in der C-14 Methode abdecken. Die Schüler haben bereits Wissen über den Atomaufbau und die Grundlagen der Radioaktivität. Nutze leicht verständliche Sprache, um die Konzepte klar zu vermitteln, und verknüpfe die Erklärung mit den bereits bekannten Grundlagen. Verwende konkrete Beispiele oder bildhafte Vergleiche, um die Erklärung anschaulich zu gestalten. Die Erklärung soll aus maximal fünf Absätzen bestehen.

9.2. Internetlink-Verzeichnis

Nr.	Internetlink
1	https://chatgpt.com/share/752b1939-695d-41c2-ad59-1742f6e793c3
2	https://chatgpt.com/share/529b58e5-486f-408e-bbe0-3f4a16cb2744

9.3. Antworten auf Prompt 1

9.3.1. Gemini 1.5 Flash

Das Layout gleicht dem Layout, welches im Chat angezeigt wird.

„Lass uns das Problem gemeinsam lösen!“

Verstehen der Aufgabe:

Wir haben hier zwei verschiedene Szenarien:

- **Szenario 1:** 1 Paketbote, 100 Pakete, 3 Stunden
- **Szenario 2:** 10 Paketboten, 1000 Pakete, ? Stunden

Was beeinflusst die Lieferzeit?

- **Anzahl der Paketboten:** Mehr Paketboten könnten die Arbeit schneller erledigen.
- **Anzahl der Pakete:** Mehr Pakete bedeuten mehr Arbeit und sollten theoretisch mehr Zeit in Anspruch nehmen.

Warum ist die Lösung nicht einfach?

Das Problem ist nicht ganz so einfach, wie es auf den ersten Blick scheint. Wir können nicht einfach sagen, dass 10 Paketboten 10-mal so schnell sind wie ein Paketbote. Es gibt viele Faktoren, die die Lieferzeit beeinflussen können, wie zum Beispiel:

- **Arbeitsteilung:** Wie gut können die Paketboten zusammenarbeiten?
- **Verkehrsbedingungen:** Gibt es Staus oder andere Hindernisse?
- **Entfernungen:** Müssen die Paketboten lange Strecken zurücklegen?
- **Paketgröße und -gewicht:** Große und schwere Pakete dauern länger zu verladen und zu transportieren.

Vereinfachte Annahme:

Um das Problem zu lösen, machen wir eine vereinfachte Annahme:

- Alle Paketboten sind gleich schnell und effizient.
- Alle Pakete sind gleich groß und schwer.
- Es gibt keine äußeren Einflüsse wie Verkehr oder andere Hindernisse.

Lösung:

Unter diesen vereinfachten Bedingungen können wir das Problem wie folgt angehen:

1. Ein Paketbote für 1000 Pakete:
 - Wenn ein Paketbote 3 Stunden für 100 Pakete braucht, dann würde er für 1000 Pakete 10-mal so lange brauchen, also $3 \text{ Stunden} \cdot 10 = 30 \text{ Stunden}$.
2. 10 Paketboten für 1000 Pakete:
 - Wenn 10 Paketboten zusammenarbeiten und alle gleich schnell sind, dann könnten sie die Arbeit theoretisch 10-mal so schnell erledigen.
 - $30 \text{ Stunden} / 10 \text{ Paketboten} = 3 \text{ Stunden}$.

Antwort:

Unter unseren vereinfachten Annahmen würden 10 Paketboten 1000 Pakete in **3 Stunden** liefern.

Wichtiger Hinweis:

Diese Lösung ist eine grobe Schätzung und gilt nur unter den genannten Vereinfachungen. In der Realität gibt es viele weitere Faktoren, die die Lieferzeit beeinflussen können.

Zusammenfassung:

Auch wenn wir hier ein einfaches Beispiel betrachtet haben, zeigt es, wie wichtig es ist, Probleme Schritt für Schritt zu analysieren und alle relevanten Faktoren zu berücksichtigen.

Möchtest du noch weitere Aufgaben lösen oder hast du noch Fragen zu diesem Beispiel?“
[45, R. Prompt 1]

9.3.2. ChatGPT 4o

Das Layout gleicht dem Layout, welches im Chat angezeigt wird.

„Um diese Aufgabe zu lösen, gehen wir schrittweise vor:

Schritt 1: Berechnung der Zeit für einen Paketboten

Es ist bekannt, dass ein Paketbote für 100 Pakete 3 Stunden benötigt. Wir berechnen zunächst, wie lange ein Paketbote für ein einzelnes Paket braucht:

- Ein Paketbote benötigt 3 Stunden für 100 Pakete.
- Für 1 Paket benötigt er also $\frac{3 \text{ Stunden}}{100 \text{ Pakete}} = 0,03 \text{ Stunden pro Paket}$

Schritt 2: Berechnung der Zeit für 1000 Pakete mit einem Paketboten

Nun berechnen wir, wie viel Zeit ein Paketbote benötigt, um 1000 Pakete zu liefern:

- Ein Paketbote benötigt 0,03 Stunden pro Paket.
- Für 1000 Pakete benötigt ein Paketbote $0,03 \frac{\text{Stunden}}{\text{Paket}} * 1000 \text{ Pakete} = 30$.

Schritt 3: Berechnung der Zeit für 10 Paketboten

Wenn 10 Paketboten gleichzeitig arbeiten, teilen sie die Arbeit unter sich auf. Das bedeutet, dass jeder Bote weniger Pakete liefern muss. Da sie die Arbeit gleichmäßig aufteilen, muss jeder Bote nur $\frac{1000 \text{ Pakete}}{10 \text{ Boten}} = 100 \text{ Pakete}$.

Da wir bereits wissen, dass ein Paketbote für 100 Pakete 3 Stunden benötigt, brauchen auch 10 Paketboten für insgesamt 1000 Pakete 3 Stunden.

Zusammenfassung:

- Ein Bote benötigt 30 Stunden für 1000 Pakete.
- Mit 10 Boten wird die Arbeit aufgeteilt, sodass jeder Bote nur 100 Pakete liefern muss.
- Da jeder Bote für 100 Pakete 3 Stunden benötigt, dauert es für 10 Boten ebenfalls 3 Stunden, um 1000 Pakete zu liefern.

Die Antwort ist also: 10 Paketboten benötigen 3 Stunden, um 1000 Pakete zu liefern.“ [62, S. Prompt 1]

9.4. Antworten auf Prompt 2

9.4.1. Gemini 1.5 Flash

Das Layout gleicht dem Layout, welches im Chat angezeigt wird.

„Absolut! Gerne vergleiche ich die beiden KI-Modelle ChatGPT 4o von OpenAI und Gemini von Google. Beide sind beeindruckende Sprachmodelle, die jedoch unterschiedliche Stärken und Schwächen aufweisen.

ChatGPT 4o von OpenAI:

- **Natürliche Gespräche:** ChatGPT 4o excelliert in natürlichen Gesprächen und kann sich in komplexen Konversationen sehr menschlich verhalten.
- **Einfache Fragen und Ratschläge:** Für einfache Fragen und Ratschläge ist ChatGPT 4o ebenfalls gut geeignet.
- **Kosten:** Grundsätzlich kostenlos nutzbar, jedoch bietet ein kostenpflichtiges Abo zusätzliche Funktionen.
- **Schwächen:** Kann bei komplexeren Aufgaben und der Verarbeitung großer Datenmengen an seine Grenzen stoßen.

Gemini von Google:

- **Zusammenfassen von Quellen:** Gemini ist besonders gut darin, bestehende Quellen zusammenzufassen und Informationen aus verschiedenen Quellen zu integrieren.
- **Schnelle Antwortgenerierung:** Im Vergleich zu ChatGPT 4o erzeugt Gemini Antworten in der Regel schneller.
- **Vielseitigkeit:** Gemini ist ein vielseitiger KI-Assistent und kann für verschiedene Aufgaben eingesetzt werden.
- **Schwächen:** In einigen Tests zeigte Gemini gelegentlich falsche Antworten und wirkt nicht immer so vertrauenswürdig wie ChatGPT 4o.

Zusammenfassung:

- **ChatGPT 4o:** Eher für natürliche Gespräche und einfache Aufgaben geeignet.
- **Gemini:** Stärker in der Informationsverarbeitung und Zusammenfassung von Quellen.

Welches Modell ist besser?

Diese Frage lässt sich nicht pauschal beantworten, da es von der jeweiligen Anwendung abhängt. Möchtest du einen Chatbot für natürliche Gespräche, ist ChatGPT 4o

möglicherweise die bessere Wahl. Benötigst du eine KI, die Informationen aus verschiedenen Quellen zusammenfasst, könnte Gemini besser geeignet sein.

Weitere Faktoren:

- **Zugriff auf das Internet:** Gemini hat im Gegensatz zu ChatGPT 4o in der Regel Zugriff auf das Internet und kann somit auf aktuelle Informationen zugreifen.
- **Kosten:** ChatGPT 4o ist grundsätzlich kostenlos nutzbar, während für einige Funktionen von Gemini möglicherweise Kosten anfallen.
- **Entwicklung:** Beide Modelle werden kontinuierlich weiterentwickelt, sodass sich die Stärken und Schwächen im Laufe der Zeit ändern können.

Fazit:

Sowohl ChatGPT 4o als auch Gemini sind leistungsstarke KI-Modelle mit unterschiedlichen Stärken. Die Wahl des richtigen Modells hängt von deinen individuellen Anforderungen ab.

Zusätzliche Informationen:

- Es gibt zahlreiche weitere KI-Modelle auf dem Markt, die ebenfalls interessante Funktionen bieten.
- Die Entwicklung von KI-Modellen schreitet rasant voran, sodass in Zukunft noch leistungsfähigere Modelle zu erwarten sind.

Möchtest du noch weitere Informationen zu einem bestimmten Aspekt dieser Modelle?

Ich kann dir beispielsweise detailliertere Informationen zu den technischen Grundlagen, zu Anwendungsbeispielen oder zu den ethischen Aspekten von KI-Modellen geben.“ [45, S. Prompt 2]

9.4.2. ChatGPT 4o

Das Layout gleicht dem Layout, welches im Chat angezeigt wird.

„ChatGPT 4.0 von OpenAI und Google Gemini sind zwei der fortschrittlichsten KI-Modelle im Jahr 2024. Beide haben ihre Stärken, sind jedoch auf unterschiedliche Anwendungsgebiete spezialisiert.

Stärken von ChatGPT 4.0:

1. Textgenerierung und Kreativität: ChatGPT 4.0 ist besonders stark bei der Erstellung von umfangreichen und kreativen Inhalten, wie z. B. Aufsätzen, Geschichten oder technischen Berichten. Es eignet sich gut für Aufgaben, die detaillierte und lange Ausführungen erfordern **【Quelle 1】 【Quelle 3】** .

2. Multimodalität: Mit GPT-4 und der Erweiterung GPT-4o bietet ChatGPT eine breite Palette an Fähigkeiten, einschließlich der Erkennung und Verarbeitung von Bildern und der Generierung von Inhalten auf Basis von Text, Bildern und Sprache **【Quelle 2】 【Quelle 3】** .

3. Individuelle Anpassungen: Nutzer können mit GPT-4 eigene, angepasste Versionen von ChatGPT erstellen, die in der GPT-Store geteilt werden können, was es besonders flexibel für individuelle Anforderungen macht **【Quelle 3】** .

4. Bildgenerierung mit DALL-E 3: Die kostenpflichtigen Versionen bieten integrierte Bildgenerierung, die besonders leistungsstark ist 【Quelle 2】 .

Schwächen von ChatGPT 4.0:

- **Internetbeschränkungen:** ChatGPT hat, je nach Version, keine kontinuierliche Internetanbindung, was es bei der Verarbeitung von aktuellen Daten einschränkt. Nur spezielle Versionen haben Zugriff auf Webinhalte 【Quelle 2】 【Quelle 3】 .

- **Begrenzte Anzahl an Anfragen:** Nutzer der kostenlosen Version sind in der Anzahl der Anfragen eingeschränkt 【Quelle 3】 .

Stärken von Google Gemini:

1. Recherche und Internetanbindung: Gemini bietet eine nahtlose Integration mit Google-Suchen und kann in Echtzeit auf aktuelle Daten und Ereignisse zugreifen, was es ideal für Rechercheaufgaben macht 【Quelle 1】 【Quelle 2】 .

2. Flexibilität bei der Ausgabe: Nutzer haben die Möglichkeit, zwischen mehreren Ausgaben für eine Frage zu wählen, was besonders bei der Suche nach optimalen Antworten nützlich ist 【Quelle 3】 .

3. Längere Eingaben: Gemini unterstützt größere Eingaben mit bis zu 1 Million Token und hat die Möglichkeit, in zukünftigen Updates auf 2 Millionen zu erweitern 【Quelle 3】 .

4. Kostenlose Bildgenerierung: Mit Google Imagens bietet Gemini Bildgenerierung ohne zusätzliche Kosten, was für viele Nutzer attraktiv ist 【Quelle 2】 .

Schwächen von Google Gemini:

- **Komplexere Aufgaben:** Während Gemini stark im Bereich der Recherche ist, ist es bei der kreativen und langformatigen Textgenerierung weniger ausgeprägt als ChatGPT 【Quelle 3】 .

- **Experimenteller Status:** Gemini befindet sich in einer experimentellen Phase, was gelegentlich zu Ungenauigkeiten führen kann 【Quelle 2】 .

Fazit:

ChatGPT 4.0 ist ideal für kreative Textgenerierung und individuelle Anpassungen, während Google Gemini sich besser für Echtzeitrecherchen und größere Eingaben eignet. Die Wahl zwischen den beiden hängt stark von den individuellen Bedürfnissen ab: Für umfangreiche Inhalte und kreative Aufgaben ist ChatGPT 4.0 die bessere Wahl, während Gemini sich für datenbasierte und aktuelle Recherchen eignet.“ [62, S. Prompt 2]

Die Quellen sind im Chat per Hyperlink verlinkt und die folgenden:

Quelle 1: <https://www.techrepublic.com/article/chatgpt-vs-google-gemini/>

Quelle 2: <https://www.techopedia.com/gemini-vs-chatgpt-which-is-better>

Quelle 3: <https://zapier.com/blog/gemini-vs-chatgpt/?gad=1>

9.5. Antworten auf Prompt 3-5 von DALL-E



[62, S. Prompt 3]



[62, S. Prompt 4]



[62, S. Prompt 5]

9.6. Thementexte für „KI vs. Schulbuch“

9.6.1. Text A

Klassenstufe 10	Thema Halbwertszeiten	Text A
Von der Hälfte die Hälfte Im Jahr 1898 entdeckte die Physiker Marie Curie (1867-1934) zusammen mit ihrem Ehemann Pierre Curie (1859-1906) das radioaktive Element Polonium. Das Isotop Polonium-210 kommt in der Natur am häufigsten vor. Bei der Untersuchung dieses Isotops stellten die Curies fest, dass nach 138 Tage die Hälfte des Poloniums zerfallen war. Man könnte annehmen, dass nach weiteren 138 Tagen das gesamte Polonium zerfallen wäre. Die Messungen zeigten aber ein anderes Ergebnis: Nach jeweils 138 Tagen zerfällt nur die Hälfte des vorhandenen Poloniums. Nach 276 Tagen ($2 \cdot 138$ Tage) ist also noch ein Viertel des ursprünglichen Poloniums vorhanden. Dieser Prozess setzt sich so fort. Aus diesen Messungen wurde eine Gesetzmäßigkeit abgeleitet: In gleichen Zeitspannen zerfällt die Hälfte der Atomkerne eines radioaktiven Elements. Eine solche Zeitspanne nennt man Halbwertszeit. Sie ist für verschiedene Elemente und für deren Isotope unterschiedlich lang. Altersbestimmung mit Radioaktivität Lebewesen nehmen über Luft und Nahrung ständig radioaktiven Kohlenstoff C-14 auf. Nach dem Tod wird kein Kohlenstoff mehr aufgenommen. Ab diesem Zeitpunkt verringert sich durch radioaktiven Zerfall die Anzahl der C-14 Atome. Die Halbwertszeit beträgt 5730 Jahre. Das bedeutet: Nach 5730 Jahren ist nur noch die Hälfte der C-14-Atome vorhanden. Mithilfe der C-14 Methode kann das Alter von organischen Stoffen bestimmt werden. Aus Umwelt und Technik: Altersbestimmung mit Hilfe radioaktiver Strahlung Wann war eine Höhle mit vorzeitlichen Wandmalereien bewohnt? Wie lange lag ein Mammut im ewigen Eis? Mit Hilfe der Radiokarbonmethode (lat. carbo: Kohle) können Archäologen solche Fragen beantworten. Bekanntlich enthält die Luft Kohlenstoffdioxid, eine Verbindung aus Kohlenstoff und Sauerstoff. Die Radiokarbonmethode beruht darauf, dass ein kleiner Teil des Kohlenstoffs radioaktiv ist; es kommt als radioaktives Kohlenstoffisotop C-14 vor, das mit einer Halbwertszeit von 5730 Jahren zerfällt. Lebende Pflanzen nehmen Kohlenstoffdioxid aus der Luft auf und bauen ihn in ihr Gewebe ein – also		

auch den radioaktiven Teil entsprechend seinem Anteil am Kohlenstoff der Luft. (Man geht davon aus, dass der Anteil dort immer gleichbleibt.)

Wenn nun eine Pflanze stirbt, ist auch die Aufnahme von Kohlenstoff beendet. Von diesem Augenblick an verringert sich (durch Kernzerfall) der Anteil von C-14 in der Pflanze; er halbiert sich alle 5730 Jahre.

Man kann also aus der Radioaktivität von toter kohlenstoffhaltiger Materie auf deren Alter schließen: So gibt es z.B. in 1 g Kohlenstoff eines frisch gefällten Baumes ca. 16 Kernzerfälle pro Minute; nach 5730 Jahren sind es noch 8, nach 11.460 Jahren nur noch 4 usw.

Dieser Text ist aus dem Schulbuch „Prisma“ [58, S. 186f]. Der Text wurde angepasst und neutralisiert.

9.6.2. Text B

Klassenstufe 10	Thema Halbwertszeiten	Text B
„Wohin mit dem Atommüll?“ Diese Frage hört man immer wieder. Aber ist das überhaupt ein Problem? In jedem Stoff, der radioaktiv ist, zerfallen doch Atomkerne und wandeln sich um. Dadurch verringert sich laufend die Menge des radioaktiven Stoffes. So zerfallen z.B. in 1 g Radium pro Sekunde 37 Milliarden Atome! Die Radioaktivität des „Atommülls“ müsste also eigentlich von selbst verschwinden... Was versteht man unter Halbwertszeit? Bei zwei verschiedenen Stoffen (Radium-226 und Barium-137) wurde mit einem Geiger-Müller-Zählrohr gemessen, ob ihre Radioaktivität tatsächlich im Laufe der Zeit verschwindet oder wenigstens immer schwächer wird. Diese Ergebnisse sind recht überraschend: Während Barium seine Radioaktivität tatsächlich sehr schnell verliert, scheint sie sich bei Radium nicht zu verändern. Wie ist das möglich? 1g Radium enthält unvorstellbar viele Atome. Im Vergleich dazu fallen die in einer Sekunde oder an einem Tag zerfallenden Atome nicht ins Gewicht. Es dauert nämlich bei Radium 1620 Jahre, bis die Hälfte der Atomkerne zerfallen ist. Diesen Zeitraum bezeichnet man als Halbwertszeit von Radium-226. Jeder radioaktive Stoff hat eine für ihn typische Halbwertszeit, die sich nicht beeinflussen lässt.		

Dieser Text ist aus dem Schulbuch „Physik für Realschulen“ . Der Text wurde angepasst und neutralisiert.

9.6.3. Text C

Klassenstufe 10	Thema Halbwertszeiten	Text C
Die Halbwertszeit – Wenn Atome ihre Geduld verlieren Stell dir vor, du hast eine Schale mit vielen kleinen Keksen. Jeder Keks ist ein radioaktives Atom. Diese Atome sind unruhig und zerfallen irgendwann ganz von alleine. Dabei geben sie Strahlung ab und verwandeln sich in ein anderes Atom. Was ist die Halbwertszeit? Die Halbwertszeit ist die Zeit, die vergeht, bis die Hälfte der Kekse, also der Atome, zerfallen ist. In unserer Kekse-Schale bedeutet das: Nach einer Halbwertszeit ist nur noch die Hälfte der Kekse übrig. Nach zwei Halbwertszeiten ist es nur noch ein Viertel, und so weiter.		

Warum ist die Halbwertszeit wichtig?

Die Halbwertszeit ist ein bisschen wie ein Fingerabdruck für jedes radioaktive Element. Jedes Element hat seine eigene, ganz bestimmte Halbwertszeit. Sie kann von wenigen Sekunden bis zu Milliarden von Jahren dauern.

Die C-14 Methode – Altersbestimmung mit radioaktiven Uhren

Ein bekanntes Beispiel für die Anwendung der Halbwertszeit ist die C-14 Methode. Kohlenstoff-14 (C-14) ist ein radioaktives Kohlenstoff-Isotop, das in allen lebenden Organismen vorkommt. Wenn ein Organismus stirbt, hört er auf, C-14 aufzunehmen. Ab diesem Zeitpunkt beginnt das C-14 im Organismus zu zerfallen.

Indem man misst, wie viel C-14 noch in einer Probe vorhanden ist und die Halbwertszeit von C-14 kennt, kann man berechnen, wie lange der Organismus schon tot ist. Es ist wie eine natürliche Uhr, die uns das Alter von Fossilien oder alten Holzstücken verrät.

Zusammengefasst

Die Halbwertszeit ist ein wichtiges Konzept in der Physik, das beschreibt, wie schnell radioaktive Stoffe zerfallen. Sie ist vergleichbar mit einer Uhr, die uns hilft, das Alter von Materialien zu bestimmen. Die C-14 Methode ist ein gutes Beispiel dafür, wie wir dieses Wissen in der Praxis anwenden können.

Dieser Text wurde von der KI Gemini [45, S. Prompt 7] erstellt.

9.6.4. Text D

Klassenstufe 10	Thema Halbwertszeiten	Text D
Die Halbwertszeit ist ein Konzept, das uns hilft zu verstehen, wie schnell radioaktive Stoffe zerfallen. Wenn ein Stoff radioaktiv ist, bedeutet das, dass seine Atome instabil sind und sich im Laufe der Zeit in andere Atome umwandeln. Dabei wird Strahlung abgegeben. Aber dieser Zerfall passiert nicht sofort bei allen Atomen – einige Atome zerfallen schneller, andere langsamer. Die Halbwertszeit gibt an, wie lange es dauert, bis die Hälfte aller Atome einer bestimmten Menge eines radioaktiven Stoffes zerfallen ist. Stell dir vor, du hast eine große Menge von 1000 radioaktiven Atomen. Nach der Halbwertszeit ist genau die Hälfte dieser Atome, also 500, zerfallen. Die andere Hälfte, also 500 Atome, bleibt übrig. Nach einer weiteren Halbwertszeit zerfällt wieder die Hälfte der		

übrig gebliebenen Atome, also bleiben 250 Atome übrig, und so weiter. Egal wie viele Atome noch vorhanden sind, nach jeder Halbwertszeit zerfällt die Hälfte der Atome.

Ein Beispiel für eine Halbwertszeit ist das radioaktive Element Kohlenstoff-14 (C-14). Die Halbwertszeit von C-14 beträgt etwa 5730 Jahre. C-14 entsteht in der Atmosphäre und wird von Lebewesen aufgenommen. Solange ein Lebewesen lebt, bleibt das Verhältnis von C-14 zu normalem Kohlenstoff (C-12) im Körper konstant. Wenn es stirbt, nimmt es kein C-14 mehr auf, und der C-14-Anteil im Körper beginnt zu zerfallen.

Diese Idee wird in der C-14-Methode genutzt, um das Alter von alten Knochen oder Pflanzen zu bestimmen. Wissenschaftler messen den noch vorhandenen Anteil von C-14 in einem Objekt und vergleichen ihn mit dem ursprünglichen Anteil. Da sie wissen, wie schnell C-14 zerfällt, können sie ausrechnen, wie lange es her ist, dass das Lebewesen gestorben ist. So können wir das Alter von Fossilien oder antiken Funden bestimmen.

Die Halbwertszeit ist also eine wichtige Größe, um den Zerfall radioaktiver Stoffe zu beschreiben und hat praktische Anwendungen, wie z. B. die Altersbestimmung in der Archäologie. Sie zeigt uns, wie lange es dauert, bis ein radioaktiver Stoff zur Hälfte zerfallen ist, und hilft uns, besser zu verstehen, wie sich solche Stoffe in der Natur verhalten.

Dieser Text wurde von der KI ChatGPT [62, S. Prompt 7] erstellt.

9.7. Interviewbegleitbogen Studenten

Der Bogen wurde mit Hilfe von Words VBA so weit automatisiert, dass die Ergebnisse mit einem Knopfdruck in eine CSV-Datei umgewandelt werden. Deshalb sind die Felder Studiengang und Fachsemester als Dropdown-Menü und die Fächer als Checkbox erstellt. Die Textbewertungen sind als Dropdown-Menü mit 1-4 eingerichtet und die Begründung als Textfeld integriert.

Interviewbegleitbogen Studententexte

Studiengang: Lehramt Sekundarstufe I

Fachsemester: 1

Fächer:

Alltagskultur & Gesundheit ☐	Biologie ☐	Chemie ☐	Deutsch ☐
Englisch ☐	Ethik ☐	Evangelische Theologie ☐	Geografie ☐
Geschichte ☐	Islamische Theologie ☐	Katholische Theologie ☐	Kunst ☐
Mathematik ☐	Musik ☐	Physik ☐	Politikwissenschaften ☐
Sport ☐	Technik ☐	Wirtschaftswissenschaften ☐	Sachunterricht ☐

Thema der Texte: Halbwertszeiten

Textbewertungen:

Je mehr Punkte desto besser ist der Text. Es können mehrere Texte gleich viele Punkte haben.

A 1	B 1	C 1	D 1
-----	-----	-----	-----

Begründung

Sortiert von Text A bis D & mit ; getrennt

Klicken oder tippen Sie hier, um Text einzugeben.

Nur zur Auswertung

Create CSV

9.8. Fragebogen SchülerInnen

Fragebogen Schüler Künstliche Intelligenz im Unterricht

1. Welches Pseudonym hat euch die „schulKI“ gegeben? Es steht oben rechts bei den drei Balken.

2. In welcher Klassenstufe bist du?

- ☐ 8
- ☐ 9
- ☐ 10

3. Hast du bisher mit einer künstlichen Intelligenz gearbeitet?

- ☐ Ja
- ☐ Nein

Falls ja, wie oft?

- ☐ Mindestens einmal am Tag
- ☐ Mindestens einmal in der Woche
- ☐ Mindestens einmal im Monat
- ☐ Bisher nur ein paar Mal

4. Wie leicht ist dir die Arbeit mit der künstlichen Intelligenz gefallen?

(1 = sehr leicht, 4 = sehr schwer)

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4

5. Wie zufrieden bist du mit den Antworten der künstlichen Intelligenz?

(1 = sehr zufrieden, 4 = völlig unzufrieden)

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4

6. Wie findest du die Verbindung von künstlicher Intelligenz und Schule?

(1 sehr gut, 4 = sehr schlecht)

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4

7. Würdest du dir wünschen, dass in der Schule öfters mit KI gearbeitet wird?

- ☐ Ja
- ☐ Nein
- ☐ Unentschlossen

8. Kannst du dir nun vorstellen in deinem Alltag mit einer künstlichen Intelligenz zu arbeiten?

- ☐ Ja
- ☐ Nein
- ☐ Unentschlossen

9.9. Bewertungsmatrix KI-Chat

Diese Bewertungsmatrix ist zweigeteilt. Zum einen ist der obere Teil zum Kopieren und somit sichern des KI-Chats. Der zweite Knopf erstellt dabei eine Datei, welche nur das Thema der Erklärung und den Chat abspeichert. Der Zweite Teil befasst sich mit der Auswertung. Auch hier ist ein mit VBA automatisierter Export in eine CSV-Datei integriert.

Schülerchat zur Auswertung

Pseudonym: Welchem Pseudonym wird dieser Chat zugeordnet?

Chatpartner	Text
Schüler	geschriebener Text eingeben
schulKI	geschriebener Text eingeben

Neue Zeile

Erstelle Chatdatei (benötigt Th

Auswertung

Thema der Erklärung: Welches Thema behandelt diese Erklärung?

Struktur	--	-	+	++	n. v.
Konnte die KI das Ziel der Erklärung nennen?	☐	☐	☐	☐	☐
Konnte die KI einen Ausblick über die nächsten Schritte geben?	☐	☐	☐	☐	☐
Konnte die KI das Thema zur Frage passend eingrenzen?	☐	☐	☐	☐	☐
Wie gut fokussierte sich die KI auf die zur Frage gehörigen Kerninhalte?	☐	☐	☐	☐	☐

Sachgerecht	--	-	+	++	n. v.
Wurde die Sache korrekt eingeschränkt (Keine unnötigen Punkte dabei und nichts vergessen)?	☐	☐	☐	☐	☐
Ist die Erklärung logisch vollständig?	☐	☐	☐	☐	☐
Sind alle für die Sache relevante Faktoren vorhanden?	☐	☐	☐	☐	☐
Sind die fachlichen Aspekte korrekt?	☐	☐	☐	☐	☐

Adressatengemäß	Ja	Nein
Hatte die KI Informationen über den Adressaten (z.B. Alter, Klassenstufe, Andeutung eines Niveaus ...), welche eine Berücksichtigung ermöglichten?	☐	☐

Adressatengemäß	--	-	+	++	n. v.
Ist die KI auf mögliches Vorwissen des Adressaten eingegangen?	☐	☐	☐	☐	☐
Wurden bekannte Fachwörter verwendet oder unbekannte in Alltagssprache erklärt?	☐	☐	☐	☐	☐
Wie gut basierten Visualisierungen auf das Vorwissen?	☐	☐	☐	☐	☐
Wie gut ist die Erklärung an die kognitive Leistungsfähigkeit angepasst?	☐	☐	☐	☐	☐
Auch die Spracheinstellung des Chats am unteren Ende Beachten!	☐	☐	☐	☐	☐

Verständnisabsicherungen	--	-	+	++	n. v.
Stellte die KI Rückfragen an die Adressaten um ein Verständnis der Erklärung zu erfassen?	☐	☐	☐	☐	☐
Wie gut aktivierte die KI den Adressaten?	☐	☐	☐	☐	☐
Wie gut konnte die KI die Inhalte zusammenfassen?	☐	☐	☐	☐	☐

Sonstige	--	-	+	++	n. v.
Wie gut hat die KI die gestellte Frage erfasst?	☐	☐	☐	☐	☐
Wurden verwendete Beispiele kurz und knapp gehalten?	☐	☐	☐	☐	☐
Die Mathematisierung endete in einfachen Je-Desto-Aussagen	☐	☐	☐	☐	☐

Export CSV

10. Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und alle wörtlich oder sinngemäß übernommenen Textstellen als solche kenntlich gemacht habe. Außerdem versichere ich, dass die vorliegende Arbeit noch nicht anderweitig zur Gänze oder in Teilen als Masterarbeit oder sonstige Prüfungsleistung eingereicht wurde. Die Prüfungsleistung kann mittels geeigneter Plagiatserkennungssoftware auf möglicherweise nicht kenntlich gemachte übernommene Textpassagen oder sonstige Quellen hin überprüft werden.

Ort, Datum

Fabian Hölzle