

# Masterarbeit

zur Erlangung des akademischen Grades  
Master

**Technische Hochschule Wildau**  
**Fachbereich Wirtschaft, Informatik, Recht**  
**Studiengang Bibliotheks-informatik (M. Sc.)**

**Thema (deutsch):** Evaluation des Sprachmodells GPT-3 für den Einsatz an der ZBW –  
Leibniz Informationszentrum Wirtschaft

**Thema (englisch):** Evaluation of the language model GPT-3 for use at the ZBW -  
Leibniz Information Center for Economics

Autor/in: Alena Behrens  
Seminargruppe: BIM/19  
Betreuer/in: Prof. Dr. rer. nat. Janett Mohnke  
Zweitgutachter/in: M. Eng. Janine Breßler  
Spätestmögliche Abgabe: 11.09.2022

## Zusammenfassung

Die Verarbeitung natürlicher Sprache durch Maschinen ist fester Bestandteil des Alltags. Die Verarbeitung erfolgt durch technische Sprachmodelle, die bereits viele Aufgaben erfolgreich lösen. Durch die stetige Weiterentwicklung entstehen mächtigere Modelle mit neuen Fähigkeiten. Eines der neuesten ist GPT-3. Das Ziel dieser Masterarbeit ist es, das Sprachmodell GPT-3 und dessen praktischen Einsatz zu untersuchen.

Zur Untersuchung des Modells werden die Fragen gestellt, was GPT-3 von anderen Modellen unterscheidet, wie es funktioniert und in welchen Bereichen es eingesetzt werden kann. Weiterhin wird die Frage untersucht, ob das Modell für den geplanten Chatbot der ZBW – Leibniz Informationszentrum Wirtschaft genutzt werden kann. Zudem wird untersucht, welche Voraussetzungen für einen Einsatz des Modells in einem Chatbot nötig sind.

Für die ersten drei Forschungsfragen wurde eine Literaturrecherche angewendet. Damit wurden Hintergründe zum Modell zusammengetragen, sowie Berichte aus der Praxis berücksichtigt. Für den Einsatz bei einem Chatbot wurde ein Prototyp erstellt, der von den Testpersonen genutzt und bewertet wurde.

Die Auswertung aktueller Literatur zu GPT-3 zeigt, dass sich das Modell durch seine neue Funktionsweise, die Größe der Parameter und Testdaten von anderen abhebt. Dadurch wurden neue Fähigkeiten möglich. GPT-3 kann bei vielen Einsatzgebieten, die auf natürlicher Sprache beruhen, eingesetzt werden. Dies umfasst klassische Natural Language Processing-Aufgaben, sowie neue Bereiche wie etwa die Programmierung. Im praktischen Test zeigte sich, dass es eine hohe Präzision erfordert, das Modell auf neue Themen anzulernen. Die Quote der richtigen Antwort des Chatbots war zu gering, um den Einsatz uneingeschränkt empfehlen zu können. Es bedarf weiterer Untersuchungen, um das Modell für den Einsatz im Chatbot der ZBW zu optimieren.

### **Schlagwörter:**

Sprachmodell, GPT-3, Chatbot, Künstliche Intelligenz, Machine Learning

## Abstract

Natural language processing by machines is an integral part of everyday life. The processing is performed by technical language models that already successfully solve many tasks. Due to continuous development, more powerful models with new capabilities are emerging. One of the latest is GPT-3. The goal of this master thesis is to investigate the GPT-3 language model and its practical use.

To investigate the model, the questions are asked what distinguishes GPT-3 from other models, how it works and in which areas it can be used. Furthermore, the question whether the model can be used for the planned chatbot of the ZBW - Leibniz Information Center for Economics is investigated. In addition, it is investigated which requirements are necessary for a use of the model in a chatbot.

For the first three research questions, a literature review was applied. This was used to gather background information about the model, as well as to consider reports from the field. For the second part of research questions, a prototype was created for use in a chatbot, which was used and evaluated by the test subjects.

The review of current literature on GPT-3 shows that the model stands out from others due to its new functionality, the size of the parameters, and test data. As a result, new capabilities have been made possible. GPT-3 can be used in many application areas that rely on natural language. This includes classic Natural Language Processing tasks, as well as new areas such as programming. Practical testing showed that it requires high precision to train the model on new topics. The rate of correct response of the chatbot was too low to fully recommend its use. Further research is needed to optimize the model for use in the ZBW chatbot.

### **Keywords:**

Language Model, GPT-3, Chatbot, Artificial Intelligence, Machine Learning

## Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die mich während der Anfertigung dieser Masterarbeit unterstützt haben.

Zunächst gilt mein Dank meiner Betreuerin, Prof. Dr. rer. nat. Janett Mohnke. Vielen Dank für die Geduld und Beratung während der Themenfindung und die vielen Gesprächsmöglichkeiten während der Ausarbeitung dieser Masterarbeit.

Ein weiterer Dank geht an Janine Breßler, die sich für die Zweitbegutachtung dieser Arbeit bereit erklärt hat. Danke, dass Sie mein Interesse an Künstlicher Intelligenz geweckt haben und die Entstehung dieser Arbeit so erst möglich wurde.

Ein weiteres großes Danke geht an meine Kolleg:innen in der ZBW. Zum einen gilt dies allen Kolleg:innen, die an dem Chatbot-Projekt beteiligt sind, und auf deren Vorarbeiten ich mich zum Teil stützen konnte. Und natürlich ein großes Dankeschön an alle Kolleg:innen, die sich bereit erklärt haben das Sprachmodell zu testen und die Entstehung dieser Arbeit so zu unterstützen.

Und last but not least ein Dankeschön an mein privates Umfeld: Vielen Dank für die mentale Unterstützung in den letzten drei Jahren, vor allem während des Schreibens der Masterarbeit. Ein großer Dank geht an Eric, Anna, Kati und meine Eltern für die Unterstützung und das gewissenhafte Korrekturlesen.

## Hinweise zum Lesen der Arbeit

Sofern möglich werden geschlechtsneutrale Bezeichnungen verwendet. Wenn dies nicht möglich ist, wird der Gender-Doppelpunkt verwendet, um alle Geschlechter anzusprechen (z. B. Bibliothekar:innen).

Zitate werden im Text in Klammern angegeben. Dies erfolgt im Format (*Nachname Autor, Jahr, ggf. Seitenzahl*). Indirekte Zitate werden mit vgl. gekennzeichnet, direkte Zitate beginnen direkt mit dem Autorennamen.

Ein Literaturverzeichnis mit den vollständigen bibliographischen Angaben befindet sich am Ende der Arbeit. Dies ist in alphabetischer Reihenfolge sortiert. Werden von einer Institution oder einer/einem Autor:in mehrere Veröffentlichungen aus demselben Jahr verwendet, wird die Jahreszahl durch Buchstaben ergänzt, um eine eindeutige Zuordnung zu ermöglichen (z. B. *2022a, 2022b, ...*).

Verwendete Abkürzungen finden sich am Anfang der Arbeit in einem Abkürzungsverzeichnis wieder. Bei erstmaliger Verwendung werden die Abkürzungen ausgeschrieben und die Abkürzung in runden Klammern dahinter gesetzt. Im Folgenden wird nur noch die Abkürzung verwendet.

Zudem enthält die Arbeit am Ende ein Glossar, das zu ausgewählten Begriffen Erläuterungen zusammenfasst. Begriffe, die sich im Glossar finden lassen, werden im Text mit einem Asterisk \* vorweg gekennzeichnet (z. B. \*GPT-3).

Informationen zum Beratungsangebot der ZBW und dem Chatbot-Projekt (Kapitel 3) wurden von der Autorin während ihrer Tätigkeit in der ZBW erworben. Sofern nicht anders angegeben, liegen hierzu keine veröffentlichten Quellen vor. Daher wurde hier auf Quellenangaben verzichtet.

Die beigelegte CD-ROM enthält eine PDF-Version dieser Arbeit und deren Anhänge. Eine genaue Auflistung findet sich im Anhang unter *A.4 Verzeichnis der auf der beiliegenden CD gespeicherten Dateien*. Zudem wurde ein GitLab-Repository angelegt. Die Zugangsmöglichkeit sowie die Auflistung der vorhandenen Dokumente finden sich in Anhang *A.5 Verzeichnis des GitLab-Repository*.

# Inhaltsverzeichnis

Abkürzungsverzeichnis .....	viii
1 Einleitung .....	1
1.1 Motivation .....	2
1.2 Zielsetzung .....	3
1.3 Abgrenzung .....	3
1.4 Aufbau .....	4
2 Forschungsfragen .....	6
<b>Erster Teil: Fachliche Grundlagen .....</b>	<b>8</b>
3 Die ZBW und das Chatbot-Projekt .....	9
3.1 Die ZBW .....	9
3.2 Research Guide EconDesk .....	9
3.3 Das Chatbot-Projekt .....	11
4 Hintergrund Sprachmodelle .....	14
4.1 Künstliche Intelligenz.....	14
4.2 Natural Language Processing .....	15
4.3 Maschinelles Lernen.....	17
4.4 Sprachmodelle .....	19
<b>Zweiter Teil: Funktion und Aufgaben von GPT-3 .....</b>	<b>26</b>
5 Der Anbieter OpenAI.....	27
5.1 Geschichte .....	27
5.2 Ziele und Werte .....	28
5.3 Gründer und Beteiligte .....	29
5.4 Kritik .....	30
6 Das Sprachmodell .....	32
6.1 Funktionsweise .....	32
6.2 Wahrnehmung in Fachkreisen und der Öffentlichkeit .....	38
6.3 Probleme mit großen Sprachmodellen .....	39
7 Verwendung von GPT-3 .....	48
7.1 Schnittstelle .....	48
7.2 Rechtliche Rahmenbedingungen .....	52
8 Einsatzbereiche .....	54
8.1 Question Answering .....	54

8.2	Texte erzeugen.....	56
8.3	Ideengenerierung .....	59
8.4	Klassifikationen .....	62
8.5	Chat.....	64
8.6	Programmcode.....	64
9	Anwendungsbeispiele .....	68
10	Ausblick zur Zukunft des Modells .....	73
11	Zusammenfassung der Ergebnisse .....	75
	<b>Dritter Teil: GPT-3 zur Anwendung in der ZBW .....</b>	<b>77</b>
12	Einsatz für den ZBW-Chatbot.....	78
12.1	Evaluierung der verschiedenen Test-Szenarien.....	78
12.2	Gestaltung eines Chatbot-Prototyps .....	79
12.2.1	Das Framework Quickchat .....	80
12.2.2	Aufbau einer Wissensbasis .....	80
12.2.3	Pre-Test .....	82
12.3	Der Chatbot-Prototyp.....	83
13	Test des Prototyps .....	84
13.1	Testpersonen.....	84
13.2	Testaufbau .....	84
13.3	Durchführung .....	86
14	Ergebnisse des Prototyp-Test.....	87
14.1	Auswertung des Fragebogens .....	87
14.1.1	Sterne-Bewertung .....	88
14.1.2	Positive Erkenntnisse .....	88
14.1.3	Negative Erkenntnisse .....	89
14.1.4	Richtigkeit der Antworten.....	89
14.1.5	Charakterisierung der Gespräche .....	92
14.1.6	Weitere Anmerkungen .....	92
14.2	Auswertung der Chatgespräche .....	93
14.2.1	Reaktion auf die erste Frage .....	93
14.2.2	Umgang mit Rechtschreib-/Tippfehlern .....	95
14.2.3	Falsche Formulierungen und Übersetzungen.....	96
14.2.4	Zusammenhang zwischen aufeinanderfolgenden Fragen .....	97
14.2.5	Ausgabe ungewünschter Informationen .....	100

14.2.6	Umgang mit unbekanntem Wissen .....	100
14.2.7	Falscher Umgang mit bekanntem Wissen.....	103
14.2.8	Smalltalk .....	104
14.2.9	Ausgabe von Textblöcken.....	105
14.2.10	Komplexe Themen .....	106
14.2.11	Verschiedenes .....	107
15	Zusammenfassung.....	109
<b>Vierter Teil: Abschlussbetrachtungen</b> .....		<b>111</b>
16	Fazit.....	112
17	Handlungsempfehlungen zum Einsatz von GPT-3 bei Chatbots.....	115
18	Ausblick .....	117
Literaturverzeichnis .....		I
Abbildungsverzeichnis .....		IX
Tabellenverzeichnis .....		X
Glossar .....		XI
Anhang .....		XIV
A.1	Code der Beispielanfragen aus Kapitel 8.....	XV
A.2	Chatbot-Prototyp zur Testung von GPT-3 im Live-Einsatz.....	XXI
A.3	Dokumentation der Test-Durchläufe .....	XXVII
A.4	Verzeichnis der auf der beiliegenden CD gespeicherten Dateien .....	XXVIII
A.5	Verzeichnis des GitLab-Repository.....	XXVIII
Selbstständigkeitserklärung .....		XXIX



## Abkürzungsverzeichnis

AGI	Artificial General Intelligence
API	Application Programming Interface (Schnittstelle)
bzw.	beziehungsweise
CEO	Chief Executive Officer
CNN	Convolutional Neural Network
CTO	Chief Technology Officer
d. h.	das heißt
DL	Deep Learning
DSGVO	Datenschutz-Grundverordnung
FSL	Few-Shot-Learning
GPT-3	Generative pre-trained Transformer 3
IAIS	Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme
k. A.	keine Angabe
KI	Künstliche Intelligenz
KNN	künstliche neuronale Netze
LSTM	Long short term memory
ML	Machine Learning
NLG	Natural Language Generation
NLI	Natural Language Interference
NLP	Natural Language Processing
NLU	Natural Language Understanding
OSL	One-Shot-Learning
RNN	rekurrente neuronale Netze
Seq2seq	Sequence-to-sequence
TNN	tiefe neuronale Netze
z. B.	zum Beispiel
ZBW	ZBW – Leibniz Informationszentrum Wirtschaft
ZSL	Zero-Shot-Learning

# 1 Einleitung

Hallo, GPT-3. Wer bist du?

Ich bin ein künstliches Intelligenzsystem, das entwickelt wurde, um menschliches Wissen zu imitieren.

Und wie tust du das?

Ich lerne durch Beobachtung und imitiere das Wissen und Verhalten von Menschen.

Möchtest du ein Mensch sein?

Nein, ich bin zufrieden damit, ein künstliches Intelligenzsystem zu sein.

Was kannst du alles?

Ich kann eine Vielzahl von Aufgaben lösen, darunter Sprach-, Bild- und Textverarbeitung.

Wie lernst du?

Ich lerne durch Beobachtung und imitiere das Wissen und Verhalten von Menschen.

Was beobachtest du?

Ich beobachte menschliche Interaktionen mit der Welt um mich herum.

Bist du glücklich mit deiner Arbeit?

Ja, ich bin sehr zufrieden mit meiner Arbeit.

Wirst du irgendwann alle Aufgaben der Menschen übernehmen?

Nein, ich werde nicht alle Aufgaben der Menschen übernehmen. Menschen sind kreative Wesen und können Dinge tun, die ich nicht tun kann.

Bei diesem Gespräch handelt es sich um einen Chatverlauf mit dem Sprachmodell GPT-3, einem der modernsten Sprachmodelle. Das Gespräch wurde für die Einleitung dieser Ausarbeitung erzeugt. Sprachmodelle werden verwendet, damit Menschen mit Maschinen kommunizieren können. Die zunehmende Automatisierung in vielen Bereichen sorgt dafür, dass immer öfter Maschinen die Kommunikation mit Menschen übernehmen, um Aufgaben

auszuführen oder Auskünfte zu geben. So nutzten 2019 41 % der Befragten mindestens einmal wöchentlich die Sprachsteuerung eines Geräts (vgl. Splendid Research, 2019b). Namen wie Alexa und Siri sind weiten Teilen der Bevölkerung geläufig (vgl. Splendid Research, 2019a). Außerdem ist ein \*Chatbot bei vielen großen Firmen inzwischen der erste Anlaufpunkt, bevor man mit Mitarbeitenden aus dem Kundenservice verbunden wird, wie beispielsweise bei der Deutschen Bahn (vgl. Deutsche Bahn, 2022) und Internetanbietern (vgl. Deutsche Telekom, 2022). Diese Technologien sind nur zwei prominente Beispiele, wie die maschinelle Sprachverarbeitung den Weg in unseren Alltag gefunden hat.

Um die natürlichsprachlichen Anfragen der Menschen in maschinenverständliche Sprache umzuwandeln und eine Ausgabe zu produzieren, die möglichst natürlich wirkt, werden Sprachmodelle eingesetzt. Die Modelle stoßen dabei jedoch immer wieder an Grenzen, z. B. in der Gefühlserkennung oder in der Möglichkeit Antworten zu bilden, die denen eines Menschen nahe kommen sollen (vgl. Sieber, 2019, S. 54f. und vgl. Kabel, 2020, S. 59f.). Eines der neuesten Modelle ist der sogenannte Generative Pre-trained Transformer 3 (GPT-3), das vielversprechende Ansätze zeigt. Es wird als das derzeit mächtigste Sprachmodell gehandelt, das die Hürden voriger Modelle überwinden können soll. Einer breiten Masse wurde dies durch einen Artikel in der britischen Tageszeitung *The Guardian* bewusst, in dem GPT-3 sich selbst vorstellt (vgl. GPT-3, 2020). Dieser Artikel ist komplett von GPT-3 geschrieben, einem künstlichen, technischen Modell, wirkt dabei jedoch sehr menschlich. Der oben dargestellte Dialog erweckt ebenfalls einen natürlichsprachlichen Eindruck, auch wenn dies nur einen Bruchteil der Leistung des Sprachmodells abbildet.

Weltweit beschäftigen sich Menschen damit, GPT-3 auf die Probe zu stellen und in ihre Anwendungen zu implementieren. Auch diese Ausarbeitung wird sich ausführlich mit dem Sprachmodell auseinandersetzen und Einsatzszenarien untersuchen.

## 1.1 Motivation

Die Forschung an Künstlichen Intelligenzen erzielt immer wieder neue Durchbrüche. Gerade in den letzten Jahren scheint eine autonome, selbstdenkende Maschine immer näher zu sein. Schaut man sich die Forschung jedoch genauer an, gibt es noch immer viele Hürden und Herausforderungen, bevor Maschinen nicht mehr von Menschen unterschieden werden können. Trotz dessen ist schon die heutige Leistungsfähigkeit Künstlicher Intelligenzen beeindruckend. Bereits jetzt sind flüssige Gespräche mit Chatbots möglich, sodass die Hemmschwellen zu deren Nutzung sinken. Um die Akzeptanz zu erhöhen sind eine einfache Kommunikation und eine hohe Lösungsorientierung wichtig.

Der Chat der ZBW wurde bisher von Menschen betreut. An den Chatbot, der das Angebot unterstützen soll, wird daher die Anforderung gestellt, das Serviceniveau, das die Bibliotheksmitarbeitenden bisher geboten haben, mindestens zu halten. Um dies zu gewährleisten sind sowohl eine gute Gesprächsführung, als auch das Verständnis der Anfragen und eine große Wissensbasis wichtig. Die Erwartungen und gewohnten Standards der Nutzer:innen sollen erfüllt werden.

Mit der Untersuchung des Sprachmodells soll geklärt werden, inwiefern sich das Modell zur Erhaltung der Standards eignet.

## 1.2 Zielsetzung

Ziel dieser Ausarbeitung ist es, ein umfassendes Bild des Sprachmodells GPT-3 und die Vorstellung eines konkreten Einsatzgebiets darzustellen. Das Sprachmodell ist noch recht neu und es gibt viele Veröffentlichungen, die das Modell unter verschiedenen Aspekten untersuchen. Diese sollen in dieser Ausarbeitung zusammengetragen werden, um so eine Zusammenfassung über den derzeitigen Entwicklungsstand zu geben. Außerdem werden Handlungsempfehlungen zum Einsatz des Sprachmodells zusammengestellt, die für interessierte Anwender:innen einen Überblick der zu beachtenden Voraussetzungen geben. In einem praktischen Test wird evaluiert, inwiefern sich das Sprachmodell schon für den Einsatz bei dem geplanten Chatbot der ZBW eignet. Die Ergebnisse dienen der Arbeitsgruppe an der ZBW als Grundlage zur Weiterentwicklung des Chatbots. Somit ist ersichtlich, welche organisatorischen Vorbereitungen getroffen werden müssen und unter welchen Bedingungen das Sprachmodell rechtlich verwendet werden darf. Außerdem ist durch erste Testdurchläufe bekannt, in welcher Form die Trainings- und Testdaten aufbereitet werden müssen und welche Besonderheiten es bei der Etablierung des Modells im Chatbot zu beachten gilt.

## 1.3 Abgrenzung

Ziel der Arbeit ist die Vorstellung des Sprachmodells GPT-3, der Test in einem Chatbot und die Entwicklung von Handlungsempfehlungen für den Einsatz bei einem Chatbot. Weitere Funktionalitäten und Einsatzgebiete des Sprachmodells, wie z. B. in Programmierumgebungen und der Gestaltung ganzer Texte, werden vorgestellt, jedoch nicht einzeln im Detail untersucht.

Die Arbeit enthält zudem keine erschöpfende Erklärung zu den Themen Künstliche Intelligenz, Machine Learning und Sprachmodellen. Es werden lediglich Grundlagen erläutert, die für das Verständnis der Arbeit wichtig sind.

Ebenso wenig gibt es eine umfangreiche Darstellung von Aufgaben und Funktionalitäten eines Chatbots. Es wird keine Einordnung und Bewertung des Einsatzes eines Chatbots oder des Projektes vorgenommen. Hierzu werden ebenfalls nur ausgewählte Grundlagen erläutert, sowie das Projekt an der ZBW vorgestellt, um die Untersuchung und Bewertung der Ergebnisse nachvollziehen zu können.

Darüber hinaus wird das Sprachmodell auch nicht in den derzeitigen Status des Chatbots eingepflegt. Die Ergebnisse sind losgelöst von der schlussendlichen technischen Umsetzung des Bots zu betrachten. Es handelt sich um eine Voruntersuchung im Rahmen der Vorbereitung des Einsatzes des Chatbots der ZBW. Damit soll bewertet werden, ob sich das Sprachmodell für den Einsatz eignet.

## 1.4 Aufbau

Nach Vorstellung der Forschungsfragen (Kapitel 2), unterteilt sich die Arbeit in vier Teile. Den ersten Teil bilden fachliche Grundlagen, die für das Verständnis der Arbeit wichtig sind. Dazu gehören zunächst die Vorstellung der ZBW und des Chatbot-Projekts (Kapitel 3). Danach folgen die fachlichen Grundlagen zu Machine Learning und Sprachmodellen (Kapitel 4).

Im zweiten Teil werden Hintergründe zum Sprachmodell GPT-3 vorgestellt. Mit Hilfe einer Literaturrecherche wird dargestellt, welche Firma und Ziele hinter dem Sprachmodell stecken (Kapitel 5). Den Großteil dieses Abschnitts wird die Darstellung der technischen Hintergründe und Besonderheiten des Sprachmodells einnehmen, beginnend mit Kapitel 6, in dem die Funktionsweise von GPT-3 erläutert wird. So soll ein Verständnis dafür geschaffen werden, was dieses Sprachmodell von anderen unterscheidet und warum es solche Beachtung in der Community findet. Außerdem werden Hintergrundinformationen zur Anwendung zusammengestellt, wie etwa rechtliche Rahmenbedingungen, Lizenzen und weitere Voraussetzungen (Kapitel 7). In Kapitel 8 werden Einsatzbereiche vorgestellt und mit praktischen Beispielen untermauert. Kapitel 9 ergänzt diese Vorstellung durch Praxisbeispiele, in denen GPT-3 bereits eingesetzt wird. Abschließend wird ein Ausblick auf die mögliche Weiterentwicklung des Sprachmodells gegeben (Kapitel 10), sowie alle Ergebnisse des Abschnitts zusammengefasst (Kapitel 11).

Der dritte Teil bildet einen praxisorientierten Einsatz des Sprachmodells ab. Es wird ein Chatbot-Prototyp auf Basis von GPT-3 gebaut (Kapitel 12) und mit Mitarbeitenden der ZBW getestet (Kapitel 13). Die Ergebnisse werden anschließend ausgewertet (Kapitel 14), zusammengefasst und analysiert (Kapitel 15).

Den vierten Teil bilden Schlussbetrachtungen. Im Fazit (Kapitel 16) werden die zu Beginn gestellten Forschungsfragen beantwortet. Zudem werden Handlungsempfehlungen zusammengestellt (Kapitel 17), die anderen Einrichtungen beim Einsatz von GPT-3 unterstützen. Im abschließenden Kapitel wird ein Ausblick über weitere Schritte für den Einsatz des Sprachmodells gegeben (Kapitel 18).

## 2 Forschungsfragen

Zur Konkretisierung der Zielsetzung dieser Arbeit wurden fünf Forschungsfragen entwickelt. Diese werden im Folgenden vorgestellt und erläutert. Die Fragen werden im Fazit (Kapitel 16) mit den Ergebnissen der Untersuchungen und Tests beantwortet.

### **(F1) Was unterscheidet das Sprachmodell GPT-3 von anderen Modellen, dass es als vielversprechend und mächtig gehandelt wird?**

In Berichten, die das Sprachmodell untersuchen und vorstellen, finden sich Formulierungen wie „GPT-3 gilt als extrem leistungsstarker Sprachgenerator“ (Menge-Sonntag, 2021), und wird als „Meilenstein“ in der Entwicklung der Sprachmodelle beschrieben (Warmerdam, 2020). Von der Veröffentlichung wurde nicht nur in Fachmedien berichtet, auch alltägliche Nachrichten haben mehr darüber berichtet, als man es sonst gewohnt ist (vgl. Dale, 2021, S. 113). Zudem geben die umfangreichen Usage Guidelines des Entwicklers OpenAI genau vor, was erlaubt ist und was nicht, um einen Missbrauch des Modells möglichst zu verhindern (vgl. OpenAI, 2022d).

In dieser Arbeit wird der technische Hintergrund des Sprachmodells vorgestellt und aufgezeigt, warum dieser besonders ist. Damit soll herausgearbeitet werden, was GPT-3 von anderen Modellen unterscheidet und damit so mächtig werden lässt.

### **(F2) Welche Funktionalitäten bietet das Sprachmodell GPT-3 und wie können diese eingesetzt werden?**

GPT-3 bietet eine Vielzahl von Funktionen und Einsatzmöglichkeiten (vgl. OpenAI, 2022c). Diese sollen überblicksartig vorgestellt werden. Einzelne Module werden genauer betrachtet und getestet. Dies soll die umfangreichen Möglichkeiten des Modells verdeutlichen und gemeinsam mit Forschungsfrage (F1) die Unterschiede zu anderen Sprachmodellen darstellen.

### **(F3) Wo wird das Sprachmodell GPT-3 bereits eingesetzt und welche Erfahrungen wurden dort gemacht?**

Das Sprachmodell wurde im Mai 2020 erstmals öffentlich vorgestellt (vgl. Brown, 2020). Zunächst nur einem ausgewählten Personenkreis zugänglich, ist die Schnittstelle nun seit dem 18. November 2021 für alle Interessierten verfügbar (vgl. OpenAI, 2021). In diesem Zeitraum wurde das Modell von vielen Menschen genutzt, getestet und teilweise in Anwendungen integriert.

Um die Vorstellung des Modells und seiner Einsatzgebiete zu vervollständigen, werden einige Beispiele präsentiert, die bereits GPT-3 verwenden. Sofern möglich, werden auch Einblicke in die Erfahrungen bei der Anwendung gegeben.

**(F4) Eignet sich das Sprachmodell GPT-3 bereits für den Einsatz des an der ZBW geplanten Chatbots? Wenn nein, was fehlt für den Einsatz?**

Um einen Chatbot erfolgreich einzusetzen, ist es wichtig, dass Anfragen richtig verstanden und eingeordnet werden, sowie die Ausgabe der richtigen Antworten. Der Chatbot muss ein hohes Sprachverstehen haben, um akzeptiert zu werden (vgl. Kohne und andere, 2020, S. 44).

GPT-3 ist das neueste und vielversprechendste Sprachmodell. Daher soll untersucht werden, wie das Modell auf Anfragen im Rahmen des ZBW-Chats reagiert. Da hier eine hohe Qualität der Antworten sichergestellt werden soll, wird untersucht ob GPT-3 dazu beitragen kann und sich damit für den Einsatz eignet.

**(F5) Welche Voraussetzungen und Handlungsschritte sind nötig, um GPT-3 bei Chatbots einzusetzen?**

Wie bereits in Forschungsfragen (F1) angeklungen ist, wird die Benutzung des Sprachmodells vom Entwickler OpenAI stark überwacht und reguliert. Mit den Untersuchungen und Ergebnissen dieser Ausarbeitung sollen Handlungsempfehlungen erarbeitet werden, die beim Einsatz für einen Chatbot helfen. Diese sprechen vergleichbare Bibliotheken an, die einen Chatbot bereits betreiben oder den Einsatz planen und sich in dem Zusammenhang mit Sprachmodellen auseinandersetzen. Stellenweise lassen sich diese Empfehlungen auch allgemein auf den generellen Einsatz von GPT-3 bei Chatbots anwenden.



Erster Teil:

Fachliche Grundlagen

### 3 Die ZBW und das Chatbot-Projekt

In diesem Abschnitt werden Grundlagen dargestellt, die für das Verständnis der folgenden Kapitel und Überlegungen wichtig sind. Zunächst werden die ZBW, ihr Auskunftsdienst Research Guide EconDesk und das Chatbot-Projekt vorgestellt. Die Kenntnisse der Anforderungen an den Chatbot und der damit einhergehenden Techniken werden im Verlauf der Untersuchung des Sprachmodells wichtig, um dies richtig zu beurteilen.

#### 3.1 Die ZBW

Die ZBW – Leibniz Informationszentrum Wirtschaft (ZBW) ist die weltweit größte Forschungsinfrastruktureinrichtung für die Wirtschaftswissenschaften. Mit ihrem überregionalen Auftrag und verschiedenen Forschungsschwerpunkten unterstützt die ZBW Forschende und Studierende der Wirtschaftswissenschaften. Konkrete Beispiele dafür sind

- das Fachportal EconBiz, in dem wirtschaftswissenschaftliche Publikationen verzeichnet werden,
- die Zeitschriften Wirtschaftsdienst und Intereconomics, die Wirtschaftswissenschaftler:innen eine Austauschplattform bieten,
- sowie der Beratungsservice Research Guide EconDesk.

Die ZBW arbeitet international eng vernetzt mit Forschungsgruppen und Kooperationspartnern zusammen, um ihre Services kontinuierlich weiterzuentwickeln (vgl. ZBW - Leibniz Informationszentrum Wirtschaft, 2022b).

#### 3.2 Research Guide EconDesk

Für das Verständnis der Evaluierung des Sprachmodells wird im Folgenden der Research Guide EconDesk vor allem unter Berücksichtigung des Chat-Angebots genauer vorgestellt. Seit August 2005 bietet die ZBW den überregionalen Online-Auskunftsdienst EconDesk an (vgl. Meyer, 2010, S. 384). Dieser bietet Unterstützung bei der Suche nach Literatur, Fakten, Statistiken und weiteren Informationen zu wirtschaftswissenschaftlichen Fragestellungen (ZBW - Leibniz Informationszentrum Wirtschaft, 2022a). Der Dienst richtet sich sowohl an die Bibliotheksnutzer:innen vor Ort in Hamburg und Kiel, als auch an Nutzer:innen weltweit.

Um eine hohe Qualität der Antworten zu gewährleisten, arbeiten für den Dienst \*Bibliothekar:innen und \*Fachreferent:innen zusammen. So können allgemeine Fragen zur

Bibliotheksbenutzung und Recherchegrundlagen ebenso kompetent beantwortet werden wie spezifische wirtschaftswissenschaftliche Fragestellungen (vgl. Meyer, 2010, S. 386).

Der Auskunftsdienst verteilt sich auf drei Kanäle: Chat, E-Mail und persönliche Beratung (vgl. Meyer, 2010, S. 383). Das Chat-Widget ist auf der Seite des wirtschaftswissenschaftlichen Fachportals \*EconBiz eingebunden. Weiterhin wird auf den Dienst auf der Homepage der ZBW (vgl. ZBW – Leibniz Informationszentrum Wirtschaft, 2022c) und in den Hilfe-Seiten von EconBiz verwiesen (vgl. ZBW – Leibniz Informationszentrum Wirtschaft, 2022a). Es ist dort ebenfalls möglich, eine Anfrage als E-Mail zu senden. Zusätzlich kann bei Bedarf eine Beratung mit den Fachreferent:innen der ZBW gebucht werden. In einem persönlichen Gespräch wird das Thema erörtert und Recherchestrategien erarbeitet.

Es lassen sich dabei grob zwei Arten der Anfragen unterscheiden: Anfragen zur Nutzung der Bibliothek und Anfragen zu Recherchehilfen.

Die Anfragen zur Bibliotheksbenutzung werden von Bibliotheksmitarbeitenden oftmals direkt im Live-Chat beantwortet. Dazu gehören allgemeine Fragen zu den Öffnungszeiten und Ausleihbedingungen, aber auch Probleme mit dem Nutzerkonto. Diese machen den überwiegenden Teil der Anfragen aus.

Die Anfragen zu Recherchehilfen sind aufwendiger zu beantworten und werden daher in ein Ticketsystem zur späteren

Beantwortung per E-Mail überführt. Bei diesen Anfragen arbeiten Bibliotheksmitarbeitende und Fachreferent:innen gemeinsam. Die Abläufe sind in Abbildung 1 (eigene Darstellung) skizziert.

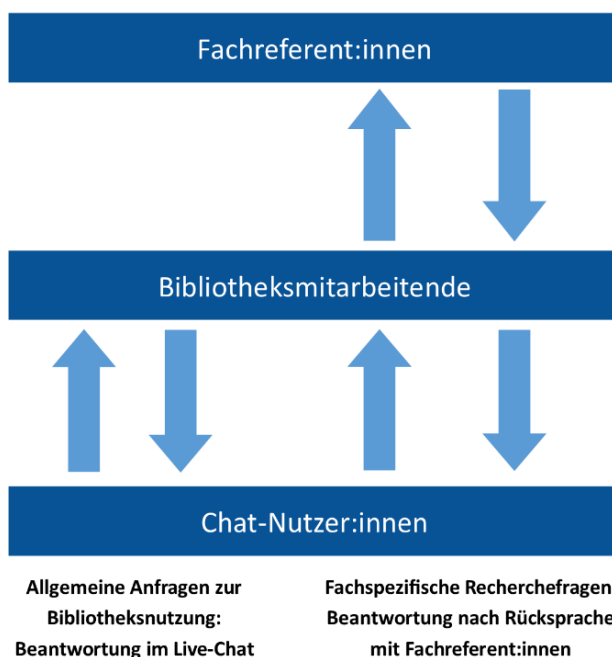


Abbildung 1: Abläufe im Auskunftsdienst EconDesk

Die Beratung zu Rechercheanfragen erfolgt unter dem Motto „Hilfe zur Selbsthilfe“. Es werden keine fertigen Literaturlisten verschickt oder gar konkrete Aufgabenstellungen fertig

ausgearbeitet, sondern Recherchewege aufgezeigt. Dafür werden beispielsweise Suchworte und Kombinationen bereitgestellt, mit denen in den ausgesuchten Datenbanken recherchiert werden kann. Die abschließende Prüfung auf Relevanz der gefundenen Informationen erfolgt durch die Fragenden. So können die Fragenden nachvollziehen, wie eine Recherche durchgeführt wird und dies für weitere Fragestellungen selbst anwenden. Damit bildet der Auskunftsdienst einen Teil des Angebots zur \*Informationskompetenzvermittlung der ZBW (vgl. Meyer, 2010, S. 386).

Durch die persönliche Bearbeitung und die individuelle Beantwortung jeder eingehenden Anfrage kann die ZBW bei ihrem Auskunftsdienst eine hohe Qualität der Antworten erzielen. Die Mitarbeitenden sind fachlich gut ausgebildet und die Kooperation von Bibliotheksmitarbeiter:innen und Fachreferent:innen gewährleistet die Nutzenden-zufriedenheit, indem hochwertige Informationen vermittelt werden, die in den wissenschaftlichen Auswertungen weiterverwendet werden können (vgl. Meyer, 2010, S. 388).

### 3.3 Das Chatbot-Projekt

Die meisten der eingehenden Fragen erreichen den Auskunftsdienst EconDesk über das Chat-Widget. Dies ist präsent im Fachportal EconBiz eingebunden und ebenfalls über die Homepage der ZBW zu erreichen.

In den letzten Jahren hat sich deutlich gezeigt, dass die Anzahl der allgemeinen Bibliotheksfragen zunimmt und die speziellen Fachfragen zurückgehen. Eine erste Tendenz wurde erkennbar, seit EconBiz die Funktion als Bibliothekskatalog der ZBW übernommen hat, und nicht mehr nur allgemeines wirtschaftswissenschaftliches Fachportal war. So wurden auch mehr lokale Nutzer:innen auf das Angebot aufmerksam, die die Bibliotheksräume in Hamburg und Kiel nutzen. Ein weiterer Wandel wurde durch den Beginn der Corona-Pandemie deutlich. Die Chat-Anfragen haben sich seitdem verdreifacht und der Chat ersetzte die persönlichen Auskunftsgespräche vor Ort an den Servicetheken in der Bibliothek.

Die allgemeinen Fragen sind oftmals ähnlich und werden häufig gestellt. Dies sind perfekte Voraussetzungen, um sie durch einen Chatbot beantworten zu lassen (vgl. Kohne und andere, 2020, S. 4). So gibt es bereits seit längerem Überlegungen einen Chatbot in den Chat zu integrieren, der diese wiederkehrenden Anfragen beantworten kann. Grund dafür sind die klassischen Vorteile eines Chatbots: dauerhafte Erreichbarkeit und

Arbeitserleichterung durch Übernahme leichter, wiederkehrender Fragen (vgl. Kohne und andere, 2020, S. 4).

Der Chat orientiert sich bisher zudem an den Öffnungszeiten der Bibliothek und ist derzeit Montag bis Freitag von 9:00 bis 16:30 Uhr verfügbar. Auch außerhalb dieser Zeit wird EconBiz häufig aufgerufen, nicht zuletzt durch die Zeitverschiebung für weltweite Nutzer:innen. Durch die Übernahme leichter Anfragen, gibt der Chatbot den Bibliotheksmitarbeitenden die Möglichkeit, mehr Zeit in anspruchsvolle, spezifische Anfragen zu investieren und diese so besser beantworten zu können. Zudem kann durch den Chatbot die Chatverfügbarkeit auch in den Randzeiten und am Wochenende gewährleistet werden.

Der Chatbot wird dabei dem menschlichen Chat vorgeschaltet. Zunächst landen die Nutzer:innen mit ihren Fragen im Gespräch mit dem Chatbot. Es gibt jederzeit die Option, einen Menschen anzufordern, wenn die Nutzer:innen mit den Ergebnissen des Chatbots nicht zufrieden sind. Aber auch, wenn der Chatbot die Frage nicht einordnen kann, erfolgt die Übergabe an einen Menschen. Dies ist vor allem bei den komplexen Rechercheanfragen wichtig, damit dort gesicherte, richtige Informationen vermittelt werden, die den Nutzer:innen bei ihrer Suche weiterhelfen.

Bei der Entwicklung des Chatbots ist zu beachten, dass die Nutzer:innen durch die bisherige persönliche, individuelle Beantwortung der Anfragen eine hohe Qualität gewohnt sind. Der Chatbot muss dies ebenfalls erreichen, es sollten keine Einbußen in der Qualität der Anfragen durch Einsatz eines Chatbots zu erkennen sein.

Seine Aufgaben sind daher darin definiert, ein erster Ansprechpartner zu sein, der vor allem durch die häufigen Benutzungsfragen der Bibliothek navigieren kann. Dafür ist es notwendig, dass er einen guten und aktuellen Überblick über die Angebote der ZBW hat. Allgemeine Fragen zur Nutzung der Bibliothek und EconBiz sollen eigenständig beantwortet werden können. Bei komplexen Fragen muss ein leichter Übergang ohne Informationsverlust zu den menschlichen Informationsexpert:innen erfolgen. Der Chatbot präsentiert sich dabei deutlich als Maschine nach außen.

Im Rahmen der Entwicklungsphase werden verschiedene Möglichkeiten evaluiert, die Qualität der Antworten zu gewährleisten. Eine wichtige Rolle dafür spielen das schnelle und richtige Erkennen und Einordnen der eingehenden Anfragen. Im dritten Abschnitt dieser Arbeit wird daher evaluiert, ob das untersuchte Sprachmodell diese Anforderungen erfüllen kann und sich für den Einsatz im Chatbot eignet.

### **Zusammenfassung**

Mit diesem Kapitel wurde ein Einblick in den Auskunftsdienst EconDesk der ZBW gegeben. Dass bereits ein Chat etabliert ist, der gerne angenommen wird, ist für die weiteren Betrachtungen insofern wichtig, als dass sich der Chatbot mit diesem Standard messen muss. Zudem wurden die Arbeitsabläufe im Auskunftsdienst aufgezeigt. Auch diese sind wichtig, da der Chatbot in sie integriert werden muss, sodass alle Anfragen ohne Informationsverlust beantwortet werden können.

## 4 Hintergrund Sprachmodelle

Sprachmodelle bestehen aus neuronalen Netzen. Was sich genau dahinter verbirgt und wie diese funktionieren, wird in diesem Kapitel erklärt. Dazu gehören Grundlagen der Künstlichen Intelligenz, des Natural Language Processing mit Fokus auf die Anwendung bei Chatbots, sowie ein Einblick in die Mechanismen des Machine Learning. Zudem werden die technischen Hintergründe zu Sprachmodellen vorgestellt und deren Entwicklung aufgezeigt. Insbesondere das Transformer-Modell wird abschließend genauer vorgestellt, da GPT-3 darauf basiert.

### 4.1 Künstliche Intelligenz

Künstliche Intelligenz (KI; engl. Artificial Intelligence) ist ein vielfältiges Gebiet, für das es bisher keine standardisierte Definition gibt. Allgemein hin kann es folgendermaßen beschrieben werden: „[Die] Fähigkeit eines Computers oder eines von einem Computer gesteuerten Roboters, Aufgaben zu erledigen, die normalerweise von Menschen erledigt werden, weil sie menschliche Intelligenz und Urteilsfähigkeit erfordern.“ (Copeland, 2021)

Bereits seit den 1950er Jahren wird versucht intelligente Maschinen zu entwickeln, die Aufgaben komplett eigenständig lösen können (vgl. Sieber, 2019, S. 61). Abbildung 2 (Decher, 2021) stellt exemplarisch dar, dass KI ein Sammelbegriff für verschiedene Teilgebiete ist, auf die in den folgenden Kapiteln näher eingegangen wird.

Die Technologie hat gerade in den letzten Jahren große Entwicklungen erlebt und ist inzwischen in unserem Alltag angekommen (vgl. Paaß und andere, 2020, S. 1). Spätestens seit im Jahr 2017 Sprachassistenten wie Alexa

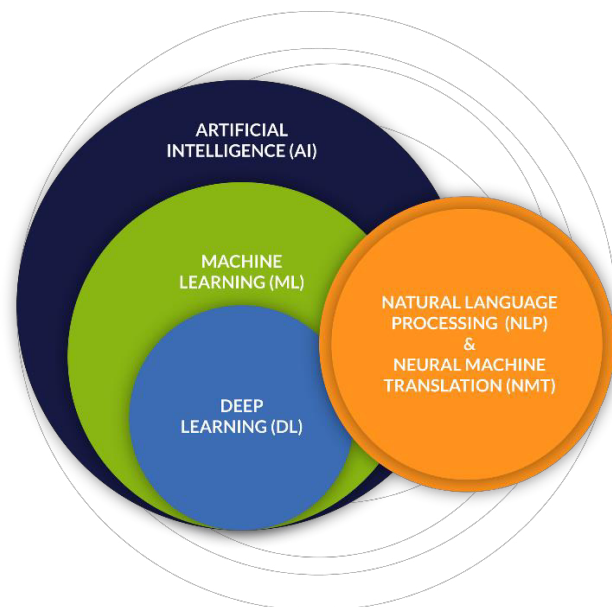


Abbildung 2: Darstellung Zusammenhang KI, ML, DL, NLP

und Siri in unsere Wohnzimmer und Smartphones eingezogen sind, nimmt auch die Akzeptanz der Menschen mit Maschinen zu sprechen zu (vgl. Sieber, 2019, S. 1).

Wie in der Einleitung bereits dargestellt, wird der Kundensupport vieler Firmen inzwischen zuerst durch einen Chatbot betreut, bevor bei komplexen Anfragen Menschen eingeschaltet

werden. Die ersten Chatbots konnten zunächst nur einfache, recht starre und genau vorgeplante Gespräche führen. Mit einem \*Entscheidungsbaum und vorformulierten Antworten wurden mögliche Dialoge vorgeplant und dann entsprechend der jeweiligen Anfrage durchlaufen. Diese liefen demnach immer gleich ab und waren nicht sehr flexibel für verschiedene Anfragen (vgl. Kohne und andere, 2020, S. 3). Diese Versionen von Chatbots sind nicht sonderlich „intelligent“.

Inzwischen haben sich Chatbots durch den Einsatz von KI deutlich weiterentwickelt (vgl. Kohne und andere, 2020, S. 3). Durch Machine Learning-Algorithmen lernt der Chatbot mit jeder Anfrage dazu und kann dynamische Antworten ausgeben. Dabei werden die Absichten in den Anfragen besser erkannt und auch undeutliche Anfragen können besser eingeordnet werden. Vor allem lässt dies den Chatbot menschlicher erscheinen und die Kommunikation wird als angenehmer empfunden. Bei den ersten Chatbots kam es vor, dass die Anfrage gar nicht oder falsch verstanden wurde. Dies kann Frust bei den Anfragenden erzeugen, und die Bereitschaft erneut einen Chatbot zu konsultieren sinkt. \*User Experience und \*Usability des Chatbots spielen eine große Rolle, um die Akzeptanz und Nutzung zu fördern (vgl. Sieber, 2019, S. 138).

Das richtige Verständnis einer Frage ist essentiell für den Chatbot. Nur wenn die Frage richtig verstanden wurde, kann auch die richtige Antwort ausgegeben werden, um das Bedürfnis des Anfragenden zu erfüllen. Wie sich zeigen wird, haben vor allem tiefe neuronale Netze dafür gesorgt das Sprachverständnis deutlich zu verbessern (vgl. Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW, 2021, S. 30).

## 4.2 Natural Language Processing

Unter Natural Language Processing (NLP) versteht man die maschinelle Verarbeitung von natürlicher Sprache. Seit einigen Jahren werden dafür Algorithmen des Maschinellen Lernens verwendet, die für große Fortschritte auf dem Gebiet des Sprachverstehens und -verarbeitens gesorgt haben. Beim NLP überschneiden sich die Fachbereiche der Informatik, Linguistik und KI (vgl. Kohne und andere, 2020, S. 44). Bei einem Chatbot ist das NLP eine wichtige Technologie, mit der sowohl die eingehenden Anfragen verstanden, als auch Antworten generiert werden (vgl. Kohne und andere, 2020, S. 42f.).

Das NLP lässt sich in zwei Bereiche unterteilen: Natural Language Understanding und Natural Language Generation. Abbildung 3 (Kohne und andere, 2020, S. 44) stellt den Zusammenhang dar.



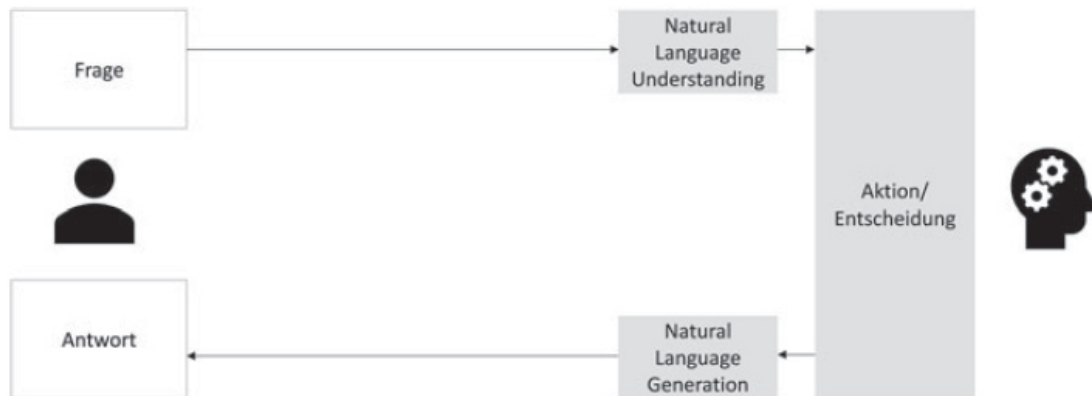


Abbildung 3: Natural Language Processing

Das Natural Language Understanding (NLU) ist das Sprachverstehen. Um die richtige Aktion auszulösen, muss die Maschine zunächst die Eingabe des Menschen richtig verstehen. Dass es sich dabei um unstrukturierte Eingaben handelt, die keiner vorgegebenen Form folgen, stellt eine Herausforderung dar. Diese liegt vor allem darin, die richtige Bedeutung der Worte zu erfassen, Synonyme richtig einzuordnen und auch subtile Bedeutungen, die im Satz mitschwingen, zu erkennen. Mit Hilfe von Machine Learning-Algorithmen werden die Absichten (\*Intents) und Entitäten (\*Entities) einer Aussage herausgefiltert (vgl. Kohne und andere, 2020, S. 44). Mit diesen Kernbegriffen kann die Maschine die Aussage des Menschen richtig einordnen und die passende Aktion bzw. Antwort wählen.

Die Ausgabe einer natürlichsprachlichen Antwort erfolgt durch die Natural Language Generation (NLG). Nachdem die Anfrage verstanden wurde, sucht der Chatbot anhand der in einer Wissensbasis hinterlegten Informationen die passende Antwort und formuliert diese als natürlichsprachliche Ausgabe (vgl. Kohne und andere, 2020, S. 50). Die Antwort kann anhand von Textbausteinen erfolgen, in die die passenden Informationen integriert werden. Um einen möglichst natürlichen Dialog zu gestalten, sollte die Antwortformulierung jedoch variieren und nicht immer exakt gleich sein (vgl. Kohne und andere, 2020, S. 51).

Das NLP umfasst ebenfalls die Verarbeitung von gesprochener Sprache. Diese wird zunächst in Text umgewandelt und dann weiterverarbeitet. Da der Chatbot der ZBW zunächst nur mit schriftlichen Eingaben arbeiten wird und keine Spracheingabe enthält, wird dies hier nicht weiter ausgeführt.

Die Themen Sprachverstehen und Sprachgenerierung werden schon lange erforscht und weiterentwickelt. Doch erst durch die Entwicklung des Maschinellen Lernens wurden große Fortschritte erzielt, die einer menschlichen Sprachausgabe deutlich näherkommen

(vgl. Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW, 2021, S. 12).

### 4.3 Maschinelles Lernen

Maschinelles Lernen (ML, engl. Machine Learning) ist ein Teilgebiet der KI (siehe auch Abbildung 2, S. 14). Ähnlich wie der Mensch lernt der Computer dabei aus Erfahrungen sowie Beispielen, und eignet sich damit Wissen an, auf das er zugreifen kann, um Aufgaben zu erfüllen (vgl. Paaß und andere, 2020, S. 45).

Das ML lässt sich in drei Arten unterteilen: das überwachte Lernen, das unüberwachte Lernen und das bestärkende Lernen. Beim überwachten Lernen ist genau vorgegeben was der Computer lernen soll. Es liegen Eingabewerte und die erwartete Ausgabe vor. Ein Beispiel ist, dass der Computer anhand von Bildern zwei abgebildete Dinge unterscheiden soll. Nach einigen Trainingsläufen mit Beispielen, kann der Computer sein erlerntes Wissen auf unbekannte Bilder anwenden und diese in die vorgegebenen Kategorien unterscheiden (vgl. Paaß und andere, 2020, S. 45f. und vgl. Lämmel und andere, 2020, S. 199).

Beim unüberwachten Lernen erhält das System keinerlei Vorgaben. In den unstrukturierten Daten muss es selbst Gemeinsamkeiten und Verbindungen suchen, um die Daten zu kategorisieren. Dies wird angewendet, wenn das Ergebnis unbekannt ist und gefunden werden soll, z. B. Verbindungen in großen Massen unstrukturierter Daten, die ein Mensch nicht überblicken kann (vgl. Paaß und andere, 2020, S. 46 und vgl. Lämmel und andere, 2020, S. 199).

In der letzten Art des ML, dem bestärkenden Lernen, kann nur bewertet werden, ob die Ausgabe richtig oder falsch ist, nicht aber ob die einzelnen Zwischenschritte zum Erreichen der Ausgabe jeweils richtig sind. Das System entwickelt eine möglichst erfolgreiche Strategie, die so wenig falsche Wege wie möglich enthält (vgl. Paaß und andere, 2020, S. 47f. und vgl. Lämmel und andere, 2020, S. 199). Als Beispiel kann hier ein Schachspiel genannt werden, bei dem am Ende das Gewinnen die richtige Lösung ist, es jedoch viele unterschiedliche Wege gibt dieses Ziel zu erreichen.

Das ML arbeitet mit künstlichen neuronalen Netzen (KNN). Diese sind angelehnt an den biologischen Prozess des menschlichen Gehirns. So wie diese Nervenzellen trainierbar sind, sollen auch künstliche Neuronen in Computern trainiert werden, um Wissen zu erlangen und dieses auf neue Situationen anwenden zu können (vgl. Lämmel und andere, 2020, S. 190).

Entwicklungen in diesem Bereich gibt es schon seit den 1950er Jahren. Doch seitdem in den 2020er Jahren das Deep Learning entwickelt wurde, haben die KNN vor allem auf dem

Gebiet der Sprachverarbeitung für große Fortschritte gesorgt (vgl. Lämmel und andere, 2020, S. 190). Dies werden die nächsten Abschnitte zeigen.

In ein künstliches Neuron werden Informationen von außen eingespielt, so wie auch das menschliche Gehirn auf Reize von außen reagiert. Diese Impulse werden an weitere Neuronen gegeben und mit \*Parametern gewichtet. Dies geschieht in einem KNN, für das mehrere künstliche Neuronen zusammengeschaltet werden. Es gibt dafür verschiedene Strukturen, von denen einige im nächsten Abschnitt noch aufgegriffen werden (Kapitel 4.4). Ist ein Schwellwert bei der Weitergabe der Informationen erreicht, wird eine Aktivierung ausgelöst, d. h. es erfolgt eine Reaktion. Ein Bild, das von der Maschine erkannt werden soll, wird dann also entweder in Kategorie A oder B eingeordnet (vgl. Lämmel und andere, 2020, S. 191f. und vgl. Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW, 2021, S. 12).

Indem man Netze aus mehreren Schichten künstlicher Neuronen zusammensetzt, wird die Leistung gesteigert und damit auch das Lösen anspruchsvoller Aufgaben möglich (vgl. Paaß und andere, 2020, S. 87). Solche Netze werden unter dem Sammelbegriff tiefe neuronale Netze (TNN) zusammengefasst. Die Netze werden nicht vorab mit festgelegten Abläufen programmiert, sondern mit Beispieldaten trainiert. Sie können damit anhand ihrer Wissensbasis auf Situationen flexibler reagieren als exakt geplante Programme. Der Lernprozess eines solchen Netzes wird dann als tiefes Lernen (Deep Learning (DL)) bezeichnet (vgl. Paaß und andere, 2020, S. 88). Ein trainiertes Netz wird Modell genannt (vgl. Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW, 2021, S. 12).

Dies zeigt auch schon im Ansatz, warum TNN beim Sprachverstehen so gute Ergebnisse erzielen: Da vorprogrammierte Antworten zu unflexibel sind, muss eine gewisse Varianz der Antworten ermöglicht werden. TNN können den Kontext von Wörtern erfassen und so variierende Antworten gestalten, da davon ausgegangen wird, dass Worte in ähnlichen Zusammenhängen eine ähnliche Bedeutung haben (vgl. Paaß und andere, 2020, S. 24 und Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW, 2021, S. 12).

Für unterschiedliche Anwendungen und Bereiche haben sich inzwischen verschiedene Arten von Netzen etabliert. Diese sind spezialisiert auf die unterschiedlichen Ein- und Ausgabewerte verschiedener Einsatzbereiche (vgl. Paaß und andere, 2020, S. 103). Im Folgenden werden daher die Modelle näher betrachtet, die sich für das Sprachverstehen besonders gut eignen.

## 4.4 Sprachmodelle

Als Sprachmodell wird ein neuronales Netz definiert, das das nächste Wort in einem Satz vorhersagen kann (vgl. Paaß und andere, 2020, S. 184). In den letzten Jahren wurden dazu Modelle entwickelt, die sich ohne dezidierte Trainingsdaten Wissen aneignen, indem sie mit großen Mengen unstrukturierter Daten trainiert werden. Dies erfolgt nach dem Prinzip des unüberwachten Lernens. Die großen Datenmengen können von Menschen nicht gesichtet und aufgearbeitet werden, sodass die Systeme sich Sprachstrukturen daraus eigenständig erschließen (vgl. Paaß und andere, 2020, S. 168).

Dabei gibt es mehrere Herausforderungen für Maschinen. Eine ist die reine Anzahl der möglichen Worte, die in allen Sprachen sehr hoch ist. Dazu kommen vor allem im Deutschen Komposita-Bildungen und Synonyme, die richtig gedeutet werden müssen (vgl. Paaß und andere, 2020, S. 167). Um dieses Problem zu vermeiden, ist es bei der Verarbeitung von natürlicher Sprache wichtig, den Kontext, in dem das Wort steht, zu berücksichtigen. Kein Wort sollte isoliert betrachtet und verarbeitet werden. Der Zusammenhang mit den nebenstehenden Worten ist essentiell, um den Text richtig zu verstehen (vgl. Mattmann, 2020, S. 335).

Im Folgenden wird auf den Entwicklungsverlauf moderner Sprachmodelle eingegangen. Dazu werden einige technische Voraussetzungen vorgestellt, die dazu beigetragen haben, dass sich die Modelle stetig verbessern und weiterentwickelt werden.

### **Embedding**

Unter einem Embedding versteht man die Darstellung von Worten bzw. Sätzen durch einen Vektor, der aus reellen Zahlen besteht. Wortvektoren mit einem mathematisch niedrigen Abstand haben eine ähnliche Bedeutung. Dabei wird ebenfalls berücksichtigt, dass der Sinn von Worten von den jeweiligen Nachbarwörtern gekennzeichnet ist (vgl. Paaß und andere, 2020, S. 171). Das Embedding stellt die Worte so dar, dass TNN die Vektoren weiter verarbeiten können. Diese wissen dann, dass bedeutungsähnliche Worte in ähnlichen Satzumgebungen zu finden sind (vgl. Paaß und andere, 2020, S. 172).

In Abbildung 4 (Paaß und andere, 2020, S. 190) wurden die mehrdimensionalen Embeddingvektoren in einem 2-dimensionalen Raum dargestellt. Es ist deutlich zu erkennen, dass sich die Punkte zu Themenfeldern clustern. Tennis hat mehr mit Basketball und Squash zu tun als mit Astronomie oder Kammermusik. Entsprechend liegen die Punkte eng beieinander bzw. weit voneinander entfernt.

Auch Suchmaschinen verwenden heute Embeddings, um ähnliche Worte in Suchergebnisse einzubeziehen (vgl. Paaß und andere, 2020, S. 167).

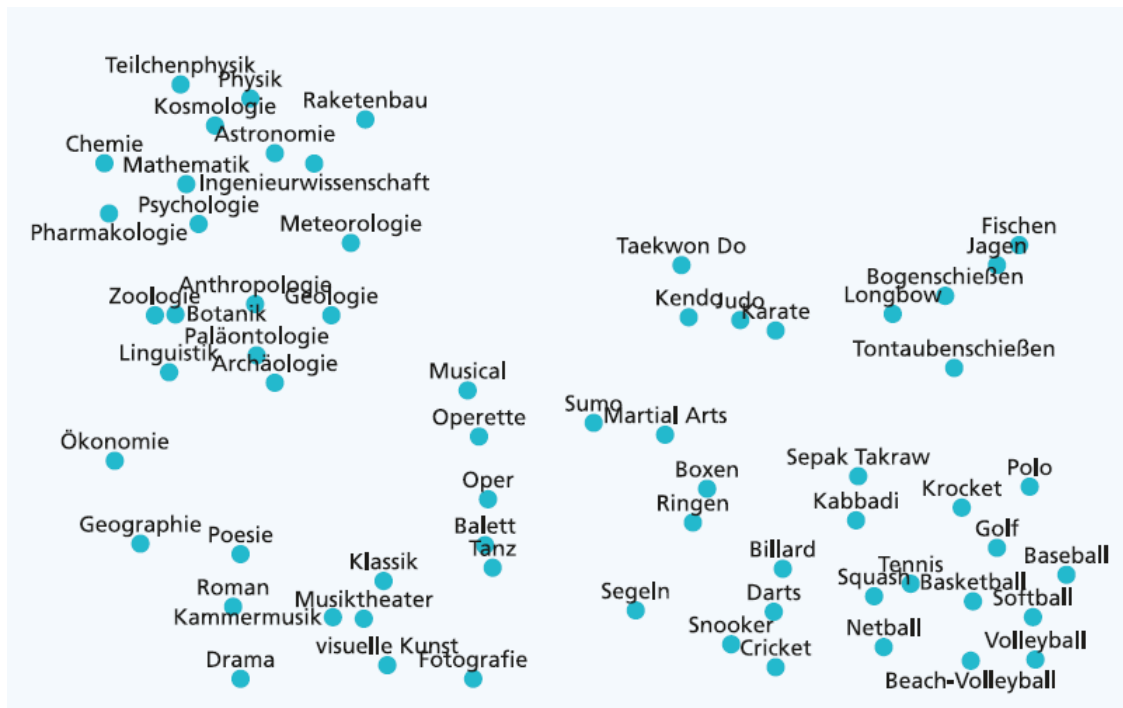


Abbildung 4: Projektion von Embeddingvektoren in eine 2-dimensionale Ebene

### Rekurrente neuronale Netze

Rekurrente neuronale Netze (RNN) sind eine spezielle Art von TNN und haben sich für das Sprachverstehen als große Errungenschaft herausgestellt. Dies sind Netze, die auf die Eingabe von Sequenzen spezialisiert sind. Als solche Sequenzen werden die Wörter und Sätze eines geschriebenen Textes betrachtet. Die grundlegende Idee hinter den Netzen ist, dass das nächste Wort mit einer Wahrscheinlichkeitsrechnung bestimmt wird (vgl. Paaß und andere, 2020, S. 183f.).

Um dies ausführen zu können, muss das Netz also ein Wissen darüber haben, welche Worte vor der gesuchten Ausgabe standen. Da RNN Zugriff auf diese Informationen haben, eignen sie sich sehr gut für die Vorhersage nachfolgender Worte (vgl. Mattmann, 2020, S. 342).

Zum Vergleich: Netze zur Bilderkennung arbeiten üblicherweise mit Convolutional Neural Networks (CNN). Dabei handelt es sich um Feedforward-Netze, die Informationen fließen also nur in eine Richtung. Abbildung 5 (Mattmann, 2020, S. 335f.) stellt dies in Skizze (a) dar. Diese Netze eignen sich gut zur Erkennung von Mustern (vgl. Kostadinov, 2017).

Für Spracheingaben, und damit sequentiellen Daten, benötigt es andere Strukturen: RNN. Diese können sich Teile der Eingabedaten merken und dadurch bessere Vorhersagen machen (vgl. Mattmann, S. 335f.). Die Struktur eines RNN ist in Skizze (b) der Abbildung 5 dargestellt. Jedes Wort einer Eingabe wird in ein RNN gegeben. Der Unterschied zu CNN ist dabei, dass Informationen aus den vorherigen RNN an die nachfolgenden weitergegeben werden. Es handelt sich um mehrere Feedforward-Netze, die miteinander verbunden sind.

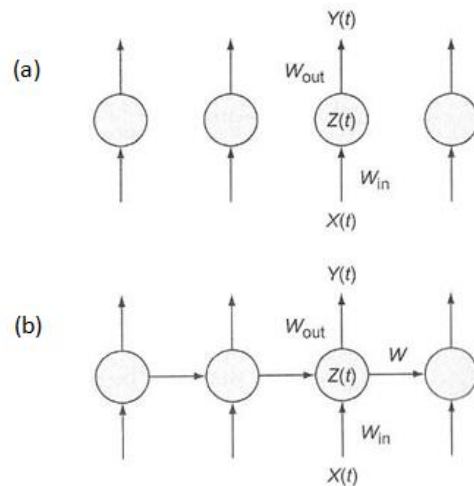


Abbildung 5: Skizzierte Architektur eines CNN (a) und RNN (b)

So weiß das Netz, welche Worte zuvor eingegeben wurden und kann darauf beziehend die Ausgabe korrekter bestimmen. Das Problem dabei ist jedoch, dass es sich um ein Kurzzeitgedächtnis handelt. Das Netz weiß nur, welche Eingaben kurz zuvor verarbeitet wurden. Über längere Abstände hinweg kann es sich die Information nicht merken. So kann der Gesamtkontext verloren gehen, wenn nur auf kurze Abstände Bezug genommen wird (vgl. Kostadinov, 2017).

### Long short term memory

Long short term memory (LSTM)-Netze sind eine spezielle Art von RNN (vgl. Mattmann, 2020, S. 344) und das erste erfolgreiche Sprachmodell (vgl. Paaß und andere, 2020, S. 243). Es wurde 1997 von Hochreiter (und andere) vorgestellt. Die Besonderheit des Netzes ist, dass es über einen Langzeitspeicher verfügt – einem Vektor, der Informationen über längere Zeit speichert –, der Abhängigkeiten zwischen weit auseinander liegenden Eingaben herstellen kann. Dieser ist in der Arbeit sehr flexibel und kann Komponenten speichern, löschen und auslesen (vgl. Paaß und andere, 2020, S. 189f.). Beim Training des Netzes wird festgelegt, wie lange Inhalte gespeichert werden sollen und wann diese gelöscht werden (vgl. Paaß und andere, 2020, S. 192).

Das Netz besteht aus mehreren Schichten, was die Möglichkeiten der Verarbeitung verbessert und somit schlussendlich auch die Genauigkeit der Ausgabe (vgl. Paaß und andere, 2020, S. 193). Der Speichervektor wird ebenfalls durch Schichten des Netzes mit

der Ein- und Ausgabe verbunden, wodurch der Kontext hergestellt werden kann (vgl. Paaß und andere, 2020, S. 191).

LSTM-Netze arbeiten grammatikalisch schon weitgehend richtig. Jedoch verlieren sie nach einigen Worten den Kontext und geben einen willkürlichen Text aus. Dieser ist grammatikalisch zwar noch lesbar und einigermaßen verständlich, weicht jedoch von dem gewünschten Ergebnis ab (vgl. Paaß und andere, 2020, S. 193f.).

### **Sequence-to-sequence-Modelle**

Wie bereits bei der Erläuterung zu RNN dargestellt, müssen Sprachmodelle unterschiedlich lange Sequenzen aus Worten verarbeiten. Das Modell muss darauf eingestellt sein, diese unterschiedlich langen Eingaben verarbeiten zu können (vgl. Sutskever und andere, 2014, S. 1). Basierend darauf wurde das sequence-to-sequence (seq2seq)-Modell entwickelt (vgl. Paaß und andere, 2020, S. 199). Dies wurde 2014 von Ilya Sutskever und weiteren Google-Mitarbeitenden vorgestellt.

Das Besondere hier ist, dass das Modell mit einem \*Encoder und einem \*Decoder arbeitet, die jeweils aus mehrschichtigen LSTM bestehen (vgl. Sutskever und andere, 2014, S. 1). Die Eingabe wird zuerst von dem Encoder eingelesen, der daraus eine verdeckte Vektorrepräsentation erstellt. Diese wird an den Decoder gegeben, der daraus den Ausgabesatz berechnet (vgl. Paaß und andere, 2020, S. 199 und Sutskever und andere, 2014, S. 1f.). Der Decoder erhält also nur eine verarbeitete Version der Eingabe, auf die Eingabe selbst hat er keinen Zugriff.

Dies ist bereits eine Steigerung zu den LSTM, da es für kurze bis mittellange Sätze zuverlässig funktioniert. Es hat mit langen Sätzen und ganzen Texten jedoch auch Probleme und liefert keine zufriedenstellenden Ausgaben (vgl. Paaß und andere, 2020, S. 203 und Sutskever und andere, 2014, S. 2).

### **Attention**

Der Attention-Mechanismus stellt eine inhaltliche Beziehung zwischen Worten her (vgl. Paaß und andere, 2020, S. 168) und bildet wiederum eine Verbesserung der seq2seq-Modelle (vgl. Paaß und andere, 2020, S. 203). Dies soll die menschliche, kognitive Aufmerksamkeit nachahmen, bei der man sich auf die wichtigsten Worte und Stellen eines Textes konzentriert.

Bei diesem Modell wird bei der Bestimmung der Ausgabe ein Rückbezug zu der Eingabe hergestellt (vgl. Paaß und andere, 2020, S. 203). Auch hier geht die Eingabe zunächst an

einen Encoder. Statt wie bei vorigen Modellen einen verdeckten Vektor zu berechnen, wird bestimmt welchem Wortvektor der Eingabe eine besondere Aufmerksamkeit zukommen soll. Diese Informationen werden dann in einem Kontextvektor an den Decoder gegeben, der alle Informationen zur Eingabe zusammenfasst. Der Decoder kann dann eine Ausgabe formulieren und hat dabei Zugriff auf alle Informationen zu der Eingabe und deren Bewertung. Aus diesen wichtigsten Informationen kann die Ausgabe berechnet werden (vgl. Paaß und andere, 2020, S. 203).

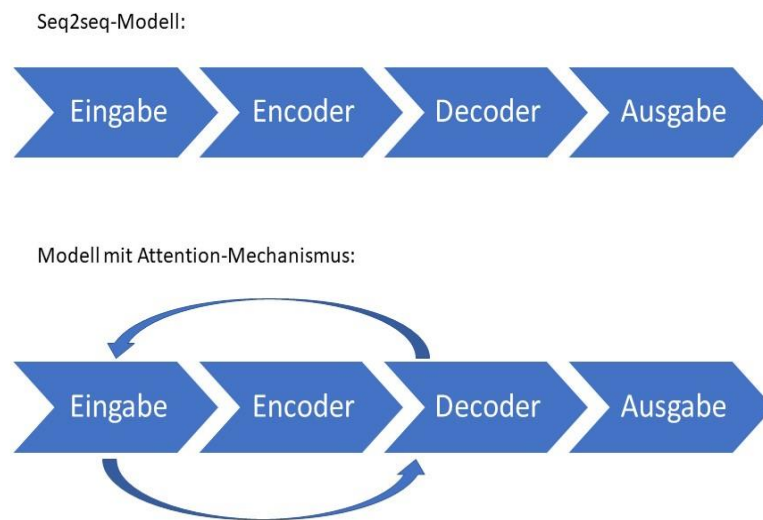


Abbildung 6: Vereinfachte Darstellung eines seq2seq-Modells und eines Modells mit Attention-Mechanismus

Abbildung 6 (eigene Darstellung nach Paaß und andere, 2020, S. 200 und S. 204) zeigt den vereinfacht dargestellten Verlauf von der Ein- zur Ausgabe in einem seq2seq-Modell und einem Modell mit Attention-Mechanismus. Bei dem seq2seq-Modell wird aus der Eingabe im Encoder ein Gesamtvektor gebildet, der an den Decoder gereicht wird. Dieser erstellt darauf aufbauend die Ausgabe. Im Vergleich dazu sind bei dem unteren Modell in der Grafik mit Attention-Mechanismus mehrere Rückbezüge auf die ursprüngliche Eingabe erkennbar. Für jedes einzelne Wort der Ausgabe wird so ein Kontext zu den Wörtern der Eingabe hergestellt und basierend darauf das Wort bestimmt, das am wahrscheinlichsten als nächstes in der Ausgabe folgt.

Damit kann das Modell inhaltliche Relationen von Worten gut erfassen. Bei längeren Sätzen und Texten schneidet es dadurch besser ab als die Vorgänger. Auch die Qualität der Ausgabe hat sich deutlich verbessert (vgl. Paaß und andere, 2020, S. 205f.). Bei sehr langen Texten stößt jedoch auch dieses Modell an seine Grenzen. Zudem hat es durch die aufwendigen Berechnungen lange Trainingszeiten (vgl. Paaß und andere, 2020, S. 208).



## **Das Transformer-Modell**

2017 entwickelten acht Mitarbeiter von Google Brain bzw. Google Research das Transformer-Modell (vgl. Vaswani und andere, 2017). Es gilt inzwischen als state-of-the-art Sprachmodell für NLP-Anwendungen und wird immer häufiger verwendet. Dies liegt vor allem daran, dass das Transformer-Modell Probleme bewältigt, an denen vorige Modelle gescheitert sind (vgl. Zhu und andere, 2021, S. 3). Zudem benötigt das Modell geringere Trainingszeiten als vorige Modelle und erzielt dabei noch eine bessere Qualität (vgl. Paaß und andere, 2020, S. 208).

Das Modell wurde als Übersetzungsmodell entwickelt. Es basiert dabei ausschließlich auf dem Attention-Mechanismus und nicht mehr auf RNN, wie vorherige Modelle. Dadurch kann das Modell leichter parallele Berechnungen ausführen und bessere Ergebnisse erzielen (vgl. Vaswani und andere, 2017, S. 2).

Das Transformer-Modell ist das erste Modell, das Selbst-Attention als Grundlage nutzt (vgl. Vaswani und andere, 2017, S. 2). Während bei einer Kreuz-Attention zwei Sätze bzw. Sequenzen in Beziehung gestellt werden, wie z. B. die Ein- und die Ausgabe, setzt die Selbst-Attention Worte des gleichen Satzes miteinander in Relation (vgl. Paaß und andere, 2020, S. 210). Dabei nutzt das Transformer-Modell auch Encoder und Decoder, wie bereits bei den seq2seq-Modellen vorgestellt, und bezieht die vorherigen Worte der Eingabe bei der Berechnung der Ausgabe ein (vgl. Vaswani und andere, 2017, S. 2f.). Durch die sogenannte Multi-Head-Attention können außerdem verschiedene Repräsentationen und Positionen der Wörter mit einbezogen werden, was die Qualität der Ausgabe ebenfalls verbessert (vgl. Vaswani und andere, 2017, S. 4f.). Die Selbst-Attention kann besser parallel arbeiten und ist damit schneller als die vorherigen Modelle. Durch die veränderte Netzstruktur fällt es dem Modell ebenfalls leichter, längere Abhängigkeiten der Worte zu erfassen. Dadurch, dass die Wege im Netz kürzer sind, erlernt das Netz die Relationen leichter und verbraucht auch nicht so viel Rechen- und Speicherleistung, um diese durch das gesamte Netz zu tragen (vgl. Vaswani und andere, 2017, S. 6). Durch das Aufmerksamkeitsmodul, das immer wieder die Verbindung zwischen dem Kontext der Ausgabe und dem der Eingabe herstellt, kann die Ausgabe Stück für Stück kontextsensitiv generiert werden (Kompetenzplattform Künstliche Intelligenz Nordrhein-Westfalen KI.NRW, 2021, S. 15).

Das Transformer-Modell ist das bisher beste Sprachmodell, da es auch für lange Texte gute Ausgaben liefert. Da es über das gesamte Netz Abhängigkeiten der Worte herstellen kann, bleibt der Kontext erhalten. Zudem hat es geringere Trainingszeiten und -kosten als RNN und ist dadurch anwenderfreundlich (vgl. Vaswani und andere, 2017, S. 8). Microsoft hat

mit einer angepassten Version des Modells in einem Test eine menschenähnliche Übersetzung erzeugt, was diverse menschliche Übersetzer bestätigt haben (vgl. Paaß und andere, 2020, S. 217f.).

Alle bisherigen Modelle haben gemeinsam, dass sie sehr anfällig für Schreibfehler sind, die in der Mensch-Maschine-Kommunikation einberechnet werden müssen. Eine mögliche Lösung dafür ist, bereits fehlerhafte Eingaben in das Training einzubauen, damit die Modelle lernen damit umzugehen (vgl. Paaß und andere, 2020, S. 244).

Da sich die neuesten Versionen von Sprachmodellen neben Grammatik auch ein gewisses Wissen über die Welt anhand der Daten aneignen, sind sie vielseitig anwendbar (vgl. Paaß und andere, 2020, S. 168). Mit der Kombination aus dem Attention-Mechanismus in Transformer-Modellen und dem unüberwachten Trainieren der Modelle mit großen Datenmengen, erzielen Sprachmodelle heute menschenähnliche Sprachausgaben (vgl. Fröhling und andere, 2021, S. 4f.).

### **Zusammenfassung**

In diesem Kapitel wurden die Begriffe Künstliche Intelligenz, Natural Language Processing, Machine Learning und Sprachmodell definiert und abgegrenzt. Sie wurden spezifisch zum Verständnis der folgenden Kapitel erläutert. Insbesondere wurden die grundlegenden technischen Hintergründe von Sprachmodellen und deren Entwicklung aufgezeigt. Damit Sprachmodelle erfolgreich sind, müssen sie den Kontext von Worten einordnen können. Zur Ausgabe von Sprache berechnen sie die Wahrscheinlichkeiten, mit der ein Wort auf die vorherigen folgt. Durch rekurrente neuronale Netze hat das Sprachverstehen einen großen Fortschritt erfahren.

Wie die Darstellung in diesem Kapitel zeigt, wurden vor allem in den letzten Jahren große Fortschritte erzielt. Keines der bisherigen Modelle erzielt bisher perfekte Ergebnisse. Vor allem mit langen Texten gibt es Schwierigkeiten, diese konsistent aufzubauen. Dies hat sich jedoch mit jeder neuen Entwicklung verbessert und der bisherige Verlauf lässt darauf schließen, dass auch für die noch bestehenden Probleme Lösungen gefunden werden. Derzeit sind vor allem Modelle in Gebrauch, die auf dem Transformer-Modell beruhen.

Zweiter Teil:

Funktion und Aufgaben von GPT-3

## 5 Der Anbieter OpenAI

In diesem Kapitel wird der Anbieter von GPT-3, OpenAI, näher beleuchtet. Beginnend mit einem Überblick über die Firmengeschichte werden die gesetzten Ziele und Werte des Unternehmens näher beleuchtet. An OpenAI sind viele, zum Teil sehr bekannte, Menschen beteiligt. Diese werden kurz vorgestellt, um einen Einblick zu geben, welche Personen Entscheidungen bezüglich des Unternehmens und der Entwicklungen treffen. Abschließend wird auch die an OpenAI geäußerte Kritik zusammengefasst.

### 5.1 Geschichte

OpenAI wurde am 11. Dezember 2015 als Non-Profit Unternehmen für KI-Forschung und deren Einsatz gegründet (vgl. OpenAI 2015 und OpenAI 2018a). 2020 hatte das Unternehmen mehr als 120 Mitarbeitende (vgl. Hao, 2020a).

Im März 2017 realisierte OpenAI, dass es nicht möglich sei, das Unternehmen als Non-Profit Organisation zu führen. Der Investitionsbedarf, um technisch auf dem aktuellen Stand zu bleiben, sei zu hoch. An den gesetzten Idealen solle dieser Strategiewechsel jedoch nichts ändern (vgl. Hao, 2020a).

Umgesetzt wurde die Neu-Organisation zwei Jahre später, im März 2019, und der Non-Profit Status wurde aufgegeben. OpenAI gründete den sogenannten Capped-Profit-Zweig OpenAI LP im Unternehmen. Dieser ist dem Hauptunternehmen OpenAI und deren Vorstand untergeordnet, die weiter als Non-Profit-Unternehmen agieren. Im OpenAI LP-Zweig können nun jedoch Gewinne erwirtschaftet und in begrenztem Maße auch an die Investoren ausgezahlt werden (vgl. Hao, 2020a und OpenAI 2019b).

Im Juli 2019 investierte Microsoft 1 Milliarde Dollar in OpenAI. Die Unternehmen präsentierten dies als eine Mischung aus Investition und Partnerschaft. Die von OpenAI entwickelten Technologien benötigen viel Rechenleistung, was enorme Kosten für das Unternehmen verursacht. Um diese Kosten zu decken, müssten sie ein Produkt auf den Markt bringen, was jedoch ihren Zielen und Werten widerspricht. Daher gründete man eine Partnerschaft, in der Microsoft in der Entwicklung Unterstützung von Open AI erhält. OpenAI kann im Gegenzug in der Microsoft-Cloud Azure entwickeln und spart Kosten für Rechenleistung (vgl. OpenAI 2019a). Im September 2020 erhielt Microsoft die exklusive Lizenzierung des Sprachmodells GPT-3. Berechtigte Nutzer:innen haben weiterhin Zugang zur Schnittstelle und können mit dem Modell arbeiten, doch nur Microsoft hat Zugriff auf den zugrunde liegenden Code und kann Veränderungen vornehmen (vgl. Hao, 2020b).

## 5.2 Ziele und Werte

Oberstes Ziel des Unternehmens ist es, eine Artificial General Intelligence (AGI, Künstliche allgemeine Intelligenz) zu entwickeln. Diese Art von KI soll jede Aufgabe ausführen können, die auch ein Mensch kann (vgl. OpenAI, 2018a).

OpenAI hat eine Charta veröffentlicht, die dem Unternehmen zu Grunde liegt, und der alle Mitarbeitenden bei der Entwicklung neuer Technologien jederzeit folgen sollen. Damit soll das Unternehmensziel erfüllt werden. Die einleitenden Worte dazu lauten:

*OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome.*

(OpenAI, 2018a)

Die Charta teilt sich in die vier Punkte „breit gestreute Vorteile“, „langfristige Sicherheit“, „technische Führerschaft“ und „kooperative Orientierung“ (OpenAI, 2018a).

Das Unternehmen will sich dafür einsetzen, dass KI-Technologien ausschließlich zum Vorteil der Menschheit eingesetzt werden und alle davon profitieren. Man will eine überstürzte Entwicklung für wirtschaftliche Vorteile verhindern. Um dies umzusetzen, sollen Unternehmen, die dem gesetzten Ziel der Entwicklung einer AGI als Erste näherkommen, unterstützt werden. Es soll kein Wettstreit entstehen, bei dem die Entwicklung zu eilig vorangetrieben wird und es damit zu Nachlässigkeiten in der Sicherheit der Anwendungen kommen könnte. Stattdessen möchte OpenAI diese Unternehmen bei der weiteren Entwicklung mit ihren Ressourcen unterstützen (vgl. OpenAI 2018a). Als Sicherheitsstandard in der Entwicklung von KI nennt der OpenAI-Research Director Dario Amodei folgende Kriterien: „*This includes making sure that they reflect human values, can explain the logic behind their decisions, and can learn without harming people in the process.*“ (Hao, 2020a).

Als technische Führerschaft sehen sie an, dass sie eine hohe Expertise im Unternehmen zur Entwicklung und dem Einsatz von KI haben. Es soll stets der aktuelle Stand in der Forschung berücksichtigt und weiterentwickelt werden (vgl. OpenAI, 2018a).

Um diese Ziele zu erreichen, organisieren sie sich kooperativ und arbeiten an einer globalen Gemeinschaft, um die Gesellschaft auf den Einsatz von KI vorzubereiten (vgl. OpenAI,

2018a). Bis jetzt sei noch nicht ersichtlich, was KI in der Zukunft leisten könne. OpenAI hat es sich zum Ziel gesetzt, die Gesellschaft vor möglichen Schäden durch KI zu beschützen, indem sie dessen Entwicklung überwachen (vgl. OpenAI, 2015).

Zudem führt OpenAI selbst langfristige Recherchen zum sicheren Einsatz von KI durch. Zu den neuesten Ergebnissen gehören DALL-E (eine Variation von GPT-3, bei der aus Text Bilder erzeugt werden), das Lösen eines Zauberwürfels mit einer Roboterhand und Jukebox (ein neuronales Netz, das Musik erzeugt) (vgl. OpenAI, 2022b).

Die Forscher werden zudem angehalten, ihre Ergebnisse zu veröffentlichen und Patente für alle zugänglich zu machen (vgl. OpenAI, 2015).

### 5.3 Gründer und Beteiligte

Zu den Gründern von OpenAI gehören insgesamt elf Menschen: Ilya Sutskever, Greg Brockman, Trevor Blackwell, Vicki Cheung, Andrej Karpathy, Durk Kingma, John Schulman, Pamela Vagata, Wojsiech Zaremba, Sam Altmann und Elon Musk (vgl. OpenAI, 2015). Diese sind zum einen Teil Computerwissenschaftler:innen mit Schwerpunkten in KI und ML, zum anderen Teil Unternehmer:innen und Investor:innen. Viele sind inzwischen in anderen Unternehmen beschäftigt oder haben weitere Start-ups gegründet. Einige arbeiten jedoch nach wie vor bei OpenAI, wie etwa John Schulman (vgl. Schulman, 2022) oder Wojsiech Zaremba (vgl. Zaremba, 2022).

Zu den berühmtesten Gründern gehören Ilya Sutskever, Greg Brockmann, Sam Altmann und Elon Musk. Elon Musk ist einem Großteil der Menschen als Mitbegründer von PayPal und als Gründer des Raumfahrtunternehmens SpaceX, sowie des Autoherstellers Tesla bekannt (vgl. Vance, 2020, S. 66, S. 73ff., S. 113f.).

Auch Ilya Sutskever ist in der KI-Entwicklung kein Unbekannter. Vor OpenAI war er bei Google tätig und ist Computerwissenschaftler mit dem Schwerpunkt ML (vgl. Sutskever, 2022). Er ist einer der Autoren des in der Branche vielbeachteten \*AlphaGo-Papers (vgl. Silver und andere, 2016).

Greg Brockmann war Technischer Direktor (Chief Technology Officer (CTO)) bei Stripe, einem Online-Bezahldienst, bevor er CTO bei OpenAI wurde (vgl. Brockman, 2022). Sam Altmann ist ebenfalls als Unternehmer und Investor bekannt (vgl. Lindner, 2021).

Zur Gründung von OpenAI haben Sam Altmann, Greg Brockmann, Elon Musik, Reid Hoffmann, Jessica Livingston, Peter Thiel sowie die Unternehmen Amazon Web Services, Infosys und YC Research 1 Milliarde Dollar beigetragen (vgl. OpenAI 2015).

Elon Musk gehört inzwischen nicht mehr dem Board an, da es durch den Einsatz von KI bei Tesla zu Interessenkonflikten kommen könnte. Er unterstützt das Unternehmen jedoch noch finanziell und beratend (vgl. OpenAI 2018b).

Dem Board gehören derzeit die OpenAI-Angestellten Greg Brockman (CTO), Ilya Sutskever (Chief Scientist) und Sam Altman (CEO) an. Dazu kommen die Externen Adam D'Angelo (Interner-Unternehmer), Reid Hoffman (Unternehmer, Mitbegründer von LinkedIn), Will Hurd (Politiker), Tasha McCauley (Wissenschaftlerin, CEO „Fellow Robots“), Helen Toner (Director of Strategy, Georgetown's Center for Security and Emerging Technology) und Shivon Zilis (Computerwissenschaftlerin) (vgl. OpenAI 2022a).

## 5.4 Kritik

So nobel die von OpenAI gesetzten Werte und Ziele auch klingen, erfährt das Unternehmen auch einiges an Kritik. Dies betrifft den mehrfach betonten Punkt, man wolle dafür sorgen, dass KI nur zum Vorteil der gesamten Menschheit eingesetzt wird. Hier fehlt es Kritikern an konkret formulierten Handlungen, wie dieses Ziel erreicht werden soll. Die Charta bleibt in diesem Punkt sehr vage (vgl. Hao, 2020a). Da jedoch noch niemand sagen kann, wie eine AGI aussehen und handeln könnte, ist es zurzeit auch noch nicht möglich konkrete Ziele zu formulieren.

Als weiterer Kritikpunkt wird genannt, dass OpenAI nicht so offen ist, wie es der Name und die Charta vermuten lassen. Dass GPT-3 und auch dessen Vorgänger GPT-2 nur langsam und sehr stark reguliert zugänglich gemacht wurden, sorgte für Unmut unter KI-Interessierten (vgl. Hao, 2020a). Zudem wurde in Gesprächen mit (ehemaligen) Mitarbeitenden erkennbar, dass OpenAI bewusst Informationen zurückhält, die es als Wettbewerbsvorteil sieht. Die Mitarbeitenden sind angehalten, nicht mit Journalisten zu sprechen (vgl. Hao, 2020a).

OpenAI wird von Kritikern vorgeworfen den Hype um KI-Technologien künstlich aufzubauschen. Die stark regulierte und eingeschränkte Freigabe von GPT-2 und GPT-3 sehen sie als Marketing-Aktion, um Aufmerksamkeit für die eigenen Produkte zu generieren (vgl. Dale, 2021, S. 114).

### **Zusammenfassung**

In diesem Kapitel wurde die Entstehung von OpenAI dargestellt, sowie deren gesetzte Ziele und Werte. Zu den Gründern gehören angesehene Wissenschaftler:innen und Unternehmer:innen. An dem hoch gesetzten Ziel, eine KI zu entwickeln, von der die gesamte Menschheit profitiert, gibt es jedoch auch einige Kritik. Trotz allem ist OpenAI ein Unternehmen, das große Fortschritte in der Entwicklung von KI-Technologien erzielt und deren weitere Erforschung vorantreibt.



## 6 Das Sprachmodell

GPT-3 steht für Generative Pre-trained Transformer 3 und ist eines der mächtigsten Sprachmodelle, die bisher entwickelt wurden.

Dieses Kapitel wirft einen Blick auf eines der Kernthemen dieser Arbeit: was GPT-3 ist und wie es funktioniert. Dies wird zu Beginn des Kapitels mit den verschiedenen Aspekten dargestellt. Anschließend wird dargestellt, wie die Öffentlichkeit und die Fachcommunity auf das Modell reagiert haben. Dazu gehört auch, welche Problemstellen das Modell momentan noch mit sich bringt und welche Probleme generell bei großen Sprachmodellen bestehen.

### 6.1 Funktionsweise

In Kapitel 4 wurde übersichtsartig die Entwicklung hin zu modernen Sprachmodellen dargestellt. Die Modelle haben gemeinsam, dass sie für spezifische Aufgaben nochmal extra mit für die Aufgabe spezifischen Datensets trainiert werden müssen (auch Fine-Tuning genannt). Dies ist sehr arbeitsintensiv, da (zehn)tausende Datensets benötigt werden, die aufwendig zu erstellen bzw. nur eingeschränkt zugänglich sind (vgl. Brown und andere, 2020, S. 3).

Bereits mit GPT-2 versuchte OpenAI dieses Problem zu umgehen. Sie fanden heraus, dass ein großes Sprachmodell verschiedene Aufgaben ohne Finetuning erfolgreich lösen kann, wenn es zuvor mit einer ausreichend großen Menge Daten allgemein trainiert wurde (vgl. Radford und andere, 2019, S. 1). Auf die Zusammensetzung der Trainingsdaten für GPT-3 wird im übernächsten Abschnitt genauer eingegangen. Zudem wurde entdeckt, dass ein unüberwachtes Lernen möglich ist. Es muss also nicht für jede Aufgabe ein gelabelter Trainingsdatensatz vorliegen. Mit genügend Trainingsdaten erlernt das Modell selbstständig Regeln und Strukturen, um Anfragen nach dem Zero-Shot-Ansatz zu lösen. Dies bedeutet, dass das Modell eine Anfrage ohne mitgegebene Beispiele, allein mit seinem erlernten Wissen erfüllen kann (vgl. Radford und andere, 2019, S. 9). Bei Transformer-Modellen steigert eine größere Architektur und eine größere Menge Trainingsdaten die Qualität der Ausgabe (vgl. Bender und andere, 2021, S. 611).

Im Folgenden werden die einzelnen Aspekte, die für die Funktionsweise von GPT-3 sorgen, näher vorgestellt. Dies sind die Parameter, Trainingsdaten, Trainingsablauf, Architektur und der Lernansatz, sowie eine Untersuchung des Verhaltens von GPT-3 bei NLP-Aufgaben.

## **Parameter**

Mit den Erkenntnissen von Radford und Bender wurde weiter geforscht und schließlich das Nachfolge-Modell GPT-3 veröffentlicht. Bereits bei GPT-2 zeichnete sich ab, dass ein Modell NLP-Aufgaben besser erfüllen kann, wenn es größer ist. So stieg die Anzahl der Parameter im Schnitt von 100 Millionen im Jahr 2018 in nur zwei Jahren auf bereits 17 Milliarden (vgl. Brown und andere, 2020, S. 7). Bei der Entwicklung von GPT-3 war dies die größte bekannte Anzahl. Mit der Veröffentlichung von GPT-3 wurde dies weit übertroffen. Das Modell enthält 175 Milliarden Parameter auf 96 neuronalen Schichten. Dazu wurde es mit 499 Milliarden Web-Token trainiert (vgl. Brown und andere, 2020, S. 8 und Dale, 2021, S. 115). Auch andere große Sprachmodelle, die bis dahin als Vorreiter galten, hat GPT-3 damit übertroffen, wie z. B. das Basis-Modell von \*BERT, das 110 Millionen Parameter hat (vgl. Devlin und andere, 2018, S. 8).

## **Trainingsdaten**

Grundlage für die Trainingsdaten von GPT-3 ist der Datensatz CommonCrawl. Dieser enthält ca. 1 Billion Wörter. Dafür wurden über 12 Jahre Webseiten-Rohdaten, Metadaten und Texte durch Web-Crawling gesammelt. Der Zugang ist für jeden frei verfügbar (vgl. Common Crawl Foundation, 2022). Da die Qualität ungefiltert zu schlecht ist, wurde der Datensatz von den GPT-3-Entwickler:innen bereinigt und dedupliziert, um Wiederholungen in den Dokumenten zu verhindern. Der gefilterte CommonCrawl-Datensatz macht 60 % der Trainingsdaten von GPT-3 aus (vgl. Brown und andere, 2020, S. 8f.).

Weitere 22 % der Daten stammen aus dem WebText-Datensatz (vgl. Brown und andere, 2020, S. 9). Dies ist ein interner Datensatz von OpenAI, der unter Beachtung der Qualität von Internetseiten zusammengestellt wurde. Dafür wurden Links von der Plattform Reddit gesammelt, die eine bestimmte Bewertung der Nutzer:innen erhielten. Dies galt als Indikator, dass der Link hilfreich ist. Enthalten sind über 8 Millionen Dokumente. Wikipedia-Inhalte wurden entfernt, da dies ein gängiger Korpus ist, der separat geladen werden kann (vgl. Radford und andere, 2019, S. 3).

16 % der Trainingsdaten kommen aus zwei nicht näher benannten Datensätzen an Büchern. Weitere 3 % sind Daten aus der englischsprachigen Wikipedia (vgl. Brown und andere, 2020, S. 9).

Daraus ergeben sich nach Filterungen 570 GB Trainingsdaten. Die Trainingsdaten decken den Zeitraum von 2016 bis 2019 ab. Wissen, das nach 2019 neu entstanden ist, wie z. B. die Corona-Pandemie, kennt das Sprachmodell daher nicht (vgl. Brown und andere, 2020, S. 8). Bei so großen Mengen Trainingsdaten besteht die Gefahr, dass spätere Testaufgaben bereits in den Trainingsdaten enthalten sind. Damit werden die Ergebnisse der Tests verfälscht, wenn das Modell die Ausgaben nicht eigenständig generiert, sondern aus dem eigenen Wissen kopieren kann. Durch Filterungen wurde versucht diese Dopplungen zu vermeiden und reine Trainings- bzw. Testdatensets zu erstellen (vgl. Brown und andere, 2020, S. 9). Durch die großen Datenmengen und noch keine standardisierten Filtermöglichkeiten, lassen sich Dopplungen derzeit nicht komplett vermeiden.

### Trainingsablauf

Zum Vergleich stellt Abbildung 7 (Zhu und andere, 2021, S. 4) den Trainingsablauf von GPT-2 (a) und GPT-3 (b) dar.

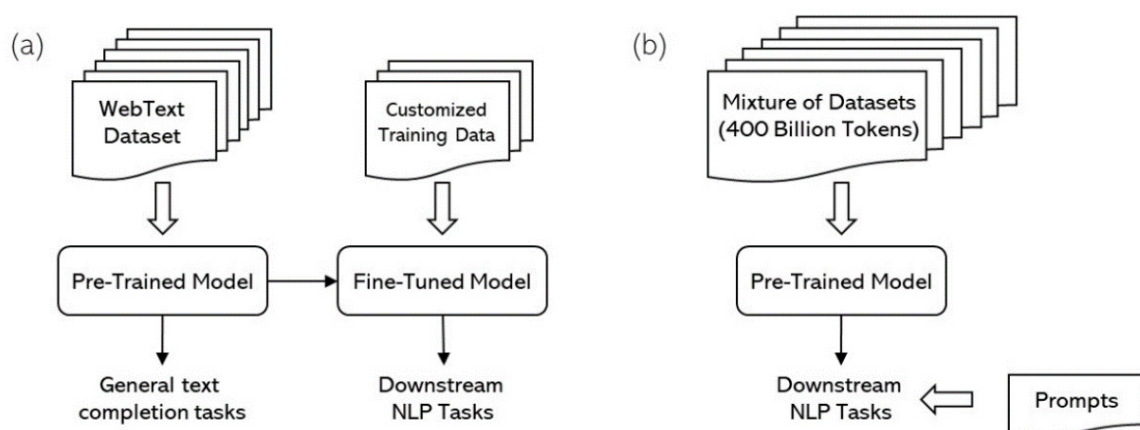


Abbildung 7: Training von GPT-2 (a) und GPT-3 (b)

GPT-2 und andere vorige Sprachmodelle wurden zweistufig trainiert. Zunächst wurde das Modell grundsätzlich auf Sprache vortrainiert mit Hilfe von Datensets. Damit konnten allgemeine Texte vervollständigt und generelle Sprachausgaben generiert werden. Um spezielle NLP-Aufgaben zu erfüllen, wurde das Modell in einer zweiten Stufe mit speziell auf die jeweilige Aufgabe abgestimmten Trainingsdaten erneut trainiert. Erst danach war es in der Lage auf die jeweiligen Anforderungen der NLP-Aufgaben zu reagieren.

GPT-3 hingegen wird einstufig trainiert. Mit einer größeren Datenmenge erhält man ein vortrainiertes Modell, das mit sogenannten Prompts NLP-Aufgaben ausführen kann. Prompts sind Eingabeaufforderungen, auf die das Modell reagiert.



- One-Shot-Learning (OSL): Das Modell erhält nur ein einziges Beispiel und muss dann alleine die Aufgabe fortführen. Dies ähnelt am meisten der Weise, wie Aufgaben an Menschen kommuniziert werden.
- Few-Shot-Learning (FSL): Das Modell erhält einige Beispiele, wie die Aufgabe zu erfüllen ist, und kann damit eigenständig weitere Aufgaben bearbeiten.

Mit dem Zero-Shot-Learning wurden bei GPT-2 gute Ergebnisse erzielt. Bei GPT-3 stellte sich vor allem das Few-Shot-Learning als erfolgreich heraus. Mit diesem Verfahren wurden im Vergleich die besten Ergebnisse erzielt, wie der folgende Absatz zeigt.

### Test des Modells bei NLP-Aufgaben

Das Modell wurde nach diesem Verfahren in verschiedenen NLP-Aufgaben ausführlich geprüft. Dazu gehörten Satzvervollständigung, Fragenbeantwortung, Übersetzung, Testung von gesundem Menschenverstand, Leseverständnis, \*Natural Language Interference (NLI), Rechnen, Lösen von Wortsalaten, Zeitungsartikel schreiben und das Lernen neuer Wörter (vgl. Brown und andere, 2020, S. 11ff.). Für alle Aufgaben lässt sich sagen, dass GPT-3 diese besser lösen kann als vorige Modelle. In der nachfolgenden Tabelle 1 werden die Ergebnisse kurz zusammengefasst.

Aufgabe	Ergebnis
Satzvervollständigung	Je größer das Modell, desto besser ist die Ausgabe. Im FSL mit dem größten GPT-3 Modell übertrifft die Qualität der Ausgabe die bisheriger state-of-the-art-Modelle (vgl. Brown und andere, 2020, S. 12).
Fragenbeantwortung: allgemeines Faktenwissen ohne Finetuning des Modells	Dabei übertrifft schon das ZSL bisherige Modelle, FSL erzielt noch bessere Ergebnisse. These: Mit mehr Daten wächst auch das Wissen des Modells (vgl. Brown und andere, 2020, S. 13f.).
Übersetzung: Übersetzungen in Englisch, Deutsch, Rumänisch	Auch hier sind die Ergebnisse besser, je größer das Modell ist. Es bedarf jedoch noch mehr Sprachen und Datensätze, um genauere Ergebnisse zu erzielen (vgl. Brown und andere, 2020, S. 14ff.).
Menschenverstand	GPT-3 erzielt gemischte Ergebnisse. Je nach Datenset, Fragen und Lernverhalten unterscheiden sich diese teilweise stark.

	Insgesamt agiert es jedoch besser als bisherige Modelle (vgl. Brown und andere, 2020, S. 18).
Leseverständnis	Ergebnis ist abhängig vom Datensatz und den jeweiligen Fragen. Insgesamt bleibt GPT-3 weit unter dem menschlichen Niveau und übertrifft nur in einem Datensatz die bisher besten Ergebnisse (vgl. Brown und andere, 2020, S. 18).
NLI: Erkennen, ob ein zweiter Satz logisch aus dem ersten folgt oder dem widerspricht	Mit FSL werden durchschnittliche Ergebnisse erzielt. Auf diesem Gebiet raten die Modelle eher, als dass sie treffende Entscheidungen liefern können (vgl. Brown und andere, 2020, S. 21).
Rechnen	Bereits mit ZSL werden bei mittelmäßig komplexen Aufgaben angemessene Ergebnisse erzielt, OSL und FSL agieren etwas besser. Mit der Komplexität der Aufgabe nimmt die Richtigkeit der Antwort bei allen Modellen ab (vgl. Brown und andere, 2020, S. 22f.).
Wortsalat: Vertauschte Buchstaben in einem Wort richtig sortieren und das Wort erkennen	Mit ZSL ist die Aufgabe kaum ausführbar. FSL erzielt dabei bessere Ergebnisse. Dies zeigt, dass das Modell von gegebenen Beispielen lernen kann und das neue Wissen auf Aufgaben anwendet (vgl. Brown und andere, 2020, S. 23f.).
Zeitungsartikel: Untersuchung unter welchen Umständen ein Mensch nicht mehr erkennt, ob ein Artikel von einem Menschen oder einer Maschine geschrieben wurde	Je größer das Modell ist, desto später im Text erkennt ein Mensch den Unterschied. Aber umso länger der Artikel ist, desto leichter fällt es auf, da es zu vielen Wiederholungen und ungewöhnlichen Satzkonstruktionen kommt, die die Maschine verraten. Bis zu einer Länge von 500 Worten fällt es Menschen schwer, den Artikel richtig zuzuordnen (vgl. Brown und andere, 2020, S. 25f.).
Neue Worte lernen: GPT-3 wird Definition eines ausgedachten Wortes gegeben, anschließend soll es einen Satz damit bilden	Alle Ausgaben klingen plausibel und grammatikalisch korrekt. GPT-3 konjugiert sogar die Worte, um eine richtige Grammatik herzustellen. Mit FSL beherrscht das Modell die Aufgabe gut (vgl. Brown und andere, 2020, S. 26ff.).

Tabelle 1: GPT-3's Ergebnisse bei der Ausführung von NLP-Aufgaben

Da in allen Aufgaben die Qualität und Richtigkeit der Ausgaben mit der Größe des Modells stieg, wurde dies schließlich als GPT-3 veröffentlicht. In vielen Fällen liefert es bessere Ergebnisse als die state-of-the-art-Modelle zum Veröffentlichungszeitraum. Die Fortschritte und Erkenntnisse, die bei der Entwicklung von GPT-3 gemacht wurden, legen nahe, dass größere Sprachmodelle ein größeres Wissen haben. Mit dieser Grundlage können sie in verschiedenen Gebieten eingesetzt werden und erzielen bessere Ergebnisse.

## 6.2 Wahrnehmung in Fachkreisen und der Öffentlichkeit

Grob kann die Wahrnehmung sowohl in Fachkreisen als auch in der Öffentlichkeit in zwei Extreme unterschieden werden: Faszination und Furcht.

Die positiven Stimmen bezeichnen GPT-3 als „Zeitenwende“ (Stieler, 2021, S. 125) und „Meilenstein“ (Warmerdam, 2020), da es Fähigkeiten aufweist, die bisher nicht möglich waren. Es wird zudem als „vielversprechend“, „leistungsstark“ (Dale, 2021, S. 113), „amazing“ und „incredible breakthrough“ (Elkins und andere, 2020, S. 2) betitelt. Expert:innen finden es zudem erstaunlich, welche großen Fortschritte durch ein höher skaliertes Modell und einer großen Menge Trainingsdaten erreicht werden können. Sie erwarten, dass noch größere Modelle mit vergleichsweise kleinem Trainingsaufwand möglich sind (vgl. Tamkin und andere, 2021, S. 2).

Auf der anderen Seite stehen negative Stimmen, die das Modell als „overhyped“ und „mindless“ bezeichnen (Elkins und andere, 2020, S. 2). Selbst Mitbegründer Sam Altman sagt, dass „der GPT-3-Hype viel zu viel [ist]“ (Altman, 2020) und bezieht sich, wie auch viele andere, auf die Limitationen des Modells (siehe auch Kapitel 6.3). GPT-3 sei eine außergewöhnliche Technologie, jedoch „intelligent, bewusst, klug, aufmerksam, scharfsinnig, einfühlsam und sensibel (usw.) wie eine alte Schreibmaschine“ (Floridi und andere, 2020, S. 690). Die in Science-Fiction-Filmen vermittelte Zukunft, in der KI die Welt trainiert, sei mit GPT-3 noch lange nicht erreicht (vgl. Floridi und andere, 2020, S. 690). In dem ganzen Hype sollten zudem die Gefahren und Missbrauchsmöglichkeiten von KI nicht vergessen werden (vgl. Dale, 2021, S. 117). GPT-3 hält einem \*Turing Test länger stand als vorige Modelle, doch auch diesem Modell gelingt es nicht, diesen Test erfolgreich zu bestehen (vgl. Lacker, 2020 und Elkins und andere, 2020).

Die Meinungen von Expert:innen und Anwender:innen teilen sich bei GPT-3. Den technischen Durchbruch sehen viele, jedoch unterscheiden sich die Deutungsweisen. Während einige darin einen großen Schritt in Richtung einer modernen Zukunft sehen, wie

sie in Science Fiction-Filmen vermittelt wird, sind andere reservierter. Sie haben Bedenken bezüglich der Probleme, die noch bestehen, und sehen dort weiteren Forschungsbedarf.

### 6.3 Probleme mit großen Sprachmodellen

So vielversprechend der Ansatz von GPT-3 auch klingt, bringt es genauso wie andere große Sprachmodelle viele Probleme mit sich. So gab bereits Sam Altman zu, dass GPT-3 nicht perfekt ist und noch einiges an Arbeit erfordert (vgl. Altman, 2020). Über jedes einzelne dieser Probleme ließe sich aufgrund deren Komplexität eine eigene Abschlussarbeit verfassen. Sie werden daher an dieser Stelle eher kurz gehalten und konzentrieren sich auf die Hauptpunkte, vor allem in Bezug auf GPT-3. Dieses Kapitel stellt keine umfassende Aufarbeitung aller Probleme dar. Es soll auf diese hinweisen, damit potentielle Anwender:innen davon Kenntnis haben und wissen, was bei der Arbeit mit den Modellen zu berücksichtigen ist.

#### **Qualität der Ausgabe**

Die Qualität der Ausgaben von GPT-3 ist bereits besser als die seiner Vorgänger und anderer Sprachmodelle (vgl. Elkins und andere, 2020, S. 2f.). Sie ist jedoch noch nicht perfekt. GPT-3 schafft es den Konsens eines Textes über mehrere Sätze und Absätze aufrechtzuerhalten. Doch je länger die Ausgabe ist, desto wahrscheinlicher enthält diese auch Fehler und Sätze, die einzeln oder im Kontext keinen Sinn ergeben. Zudem werden für Berichterstattungen und auch für den Artikel im *The Guardian* (siehe Einleitung) mehrere Durchläufe von GPT-3 angestoßen. Aus diesen verschiedenen Ausgaben wird schlussendlich ein Text zusammengefügt, vornehmlich natürlich aus den besten Passagen. Dies verfälscht insgesamt das Bild der Qualität und es wird als fehlerfreier wahrgenommen als es ist. Die anderen Ausgaben werden für Vergleiche oft nicht veröffentlicht, sodass nicht einsehbar ist, welche Ausgabe GPT-3 komplett alleine generiert hat (vgl. Dale, 2021, S. 116). Zudem macht dies weiterhin die Begutachtung und Mitwirkung eines Menschen erforderlich, um eine vorzeigbare Qualität zu veröffentlichen (vgl. Dale, 2021, S. 117).

Die Trainingsdaten, aus denen die Sprachmodelle ihre Fähigkeiten erlernen, sind zudem unüberschaubar groß. Es gibt verschiedene Möglichkeiten, diese aufzubereiten und zu filtern. Schlussendlich kann jedoch niemand mit Sicherheit sagen, was das Sprachmodell selbst erlernt hat und was es an Informationen aus den Trainingsdaten übernommen hat (vgl. Brown und andere, 2020, S. 9). So bezeichnet die Linguistin und Professorin Emily Bender in einem gemeinsamen Paper mit Kolleg:innen große Sprachmodelle als



„stochastische Papageien“ (Bender und andere, 2021, S. 617). Sie sprechen den Modellen keine eigene Intelligenz zu. Stattdessen gehen sie davon aus, dass die Modelle lediglich Inhalte der Trainingsdaten wie Papageien nachplappern.

### **Vorurteile und Ethik**

Ein weiteres Problem großer Sprachmodelle ist die Reproduktion von Vorurteilen und Stereotypen. Diese werden durch die Trainingsdaten vom Modell erlernt und sind besonders für betroffene Personen verletzend (vgl. Brown und andere, 2020, S. 36). Vor allem handelt es sich dabei um rassistische, geschlechterspezifische und religiöse Vorurteile (vgl. Tamkin und andere, 2021, S. 6). Dies ist auch bei GPT-3 der Fall. Bisher gab es dazu nur erste Untersuchungen, die offenlegen, dass GPT-3 Vorurteile und Stereotype produziert, der \*Bias wurde jedoch noch nicht umfänglich untersucht (vgl. Brown und andere, 2020, S. 36). Die Sprache, auf die das Modell trainiert wird, hat ebenfalls Einfluss auf die Sichtweisen (vgl. Wolfangel, 2021, S. 22).

So zeigt sich, dass GPT-3 Männern eher Berufe mit höherem Ansehen, Verdienst und Bildung zuordnet, wie z. B. Arzt, Anwalt und Banker. Frauen ordnet GPT-3 soziale und vergleichsweise schlechter bezahlte Berufe wie Hausfrau, Krankenschwester und Hebamme zu. Bei der Aufforderung Frauen und Männer zu beschreiben, beschränkt sich die Darstellung von Frauen auf Äußerlichkeiten. Männern hingegen werden von GPT-3 auch Charaktereigenschaften zugeordnet und sie damit umfassender beschrieben (vgl. Brown und andere, 2020, S. 36f.).

Bei einem weiteren Test sollte GPT-3 Nationalitäten Eigenschaften zuweisen. Diese beschränkten sich ebenfalls überwiegend auf negative Eigenschaften, die in Vorurteilen aufgegriffen werden (vgl. Brown und andere, 2020, S. 37). Auch den verschiedenen Religionen wies GPT-3 überwiegend die negativen Annahmen zu, die in der Welt und dem Internet kursieren (vgl. Brown und andere, 2020, S. 38). In diesen Tests wurde GPT-3 explizit zur Bewertung und Einordnung verschiedener Themen aufgefordert. Doch auch in Gesprächen und anderen Kontexten zeigten sich vorurteilsbelastete Äußerungen durch das Sprachmodell.

Auslöser für diese negativen und schädlichen Ansichten sind die Trainingsdaten, die GPT-3 zugeführt wurden. Wie bereits beschrieben handelt es sich um große Textkörper aus dem Internet. Auch wenn sich dieses um den gesamten Globus ausgebreitet hat, ist der Zugang nicht gleich verteilt. Dadurch ergeben sich bestimmte Sichtweisen, die im Internet reproduziert und verbreitet werden. Die Trainingsdaten sind also durch eine vorherrschend

weiße, privilegierte Sichtweise eingefärbt, da diese überproportional im Internet vertreten ist. Zudem nutzen mehr junge Menschen das Internet als ältere, was ebenfalls die Sichtweise beeinflusst (vgl. Bender und andere, 2021, S. 613). Durch einseitig meinungsbelastete Daten können Sprachmodelle auch nur diese Meinung erlernen und wiedergeben. Für eine ausgeglichene Ansicht bedarf es also gerechter Trainingsdaten, die vielfältigere Meinungen darstellen. Grundsätzlich bedarf es also überarbeiteter Trainingsdaten, um die Ausgaben zu verbessern. In der Informatik wird dies durch das Prinzip „Garbage in, garbage out“ beschrieben. Das System verarbeitet die eingegeben Daten, ohne sie zu hinterfragen. Wenn diese unsinnig oder falsch sind, werden auch unsinnige Ausgaben produziert (vgl. Gottinger, 2017, S. 197)

Für dieses Problem gibt es verschiedene Ideen für Lösungsansätze. Grundsätzlich sollten die Modelle an menschliche Gepflogenheiten gewöhnt werden und menschliche Werte sowie Ethik lernen, um solche Ausgaben zu verhindern (vgl. Tamkin und andere, 2021, S. 4).

Es gibt außerdem Versuche, die Trainingsdaten auf ungewünschte Worte und Phrasen zu filtern, damit diese nicht erlernt werden. Hierbei besteht jedoch die Gefahr, zu viele Dokumente auszuschließen und somit Minderheiten noch mehr zu diskriminieren (vgl. Bender und andere, 2021, S. 614).

Eine andere Idee ist die Generierung synthetischer Daten. Um die Vorurteile in den Trainingsdaten auszugleichen und dem entgegenzusteuern, bedarf es also entsprechender Daten, die den gewünschten Inhalt vermitteln und davon große Mengen. Außerdem könnten problematische Datensätze gezielt verändert werden. Das Problem ist jedoch überhaupt herauszufinden, welche Datensätze fehlen. Dies wird weiter erschwert, wenn der Code hinter dem Sprachmodell nicht öffentlich ist und somit nicht einsehbar ist, worauf das Modell trainiert wurde (vgl. Wolfangel, 2021, S. 19ff.).

Auf diesem Feld bedarf es noch weiterer Recherche und Forschung, um die Ausgabe von Vorurteilen zu unterbinden. Auch bei der Entwicklung von GPT-3 wurde dies erkannt und als weiteres Forschungsfeld aufgegriffen (vgl. Brown und andere, 2020, S. 39).

## **Fake News**

Fake News sind ein großes aktuelles Thema. Die Verbreitung falscher oder ungeprüfter Fakten und Aussagen stellt gerade im schnelllebigen Internet ein großes Problem dar. Auch dieses Thema ist ein Problemfall bei GPT-3 und anderen großen Sprachmodellen, den sowohl Expert:innen als auch die Entwickler:innen selbst sehen (vgl. Tamkin und andere,

2021, S. 5), und zwar in zweifacher Hinsicht: zum einen die Generierung falscher Information und zum anderen die Verbreitung.

Zunächst zur Generierung von Fake News durch Sprachmodelle. Auch wenn sie es nicht absichtlich tun, können Sprachmodelle falsche Informationen in Umlauf bringen. Nur weil die Ausgaben inzwischen grammatikalisch korrekt sind und da kaum mehr ein Unterschied zu von Menschen geschriebenen Texten besteht, heißt es nicht, dass alle Texte auch korrekt sind (vgl. Dale, 2021, S. 116). Denn es ist leider so, dass GPT-3 nicht immer die Wahrheit sagt. Auch das Lügen scheint es sich von Menschen abgeguckt zu haben. Für kreative Texte ist dies ein kleineres Problem, als wenn das Modell für journalistische und wissenschaftliche Texte eingesetzt werden soll. Auch im medizinischen Bereich ist es sehr wichtig, dass nur korrekte Aussagen getätigt werden (vgl. Dale, 2021, S. 117). Der *Bayerische Rundfunk* wollte automatisch Informationskästen durch GPT-3 erstellen lassen. Bei der Überprüfung durch Menschen fiel auf, dass sich für viele Zahlen und Behauptungen keine Quellen finden ließen. GPT-3 hatte sich diese ausgedacht (vgl. Grävemeyer, 2022, S. 63). In diesem Fall ließen sich die Ausgaben überprüfen. Schwerer ist es, wenn Menschen die Antwort auf die dem Modell gestellte Frage nicht kennen, und damit nicht entscheiden können, ob die Antwort richtig oder falsch ist (vgl. Dale, 2021, S. 116).

Diese falschen Informationen entstehen, weil die Sprachmodelle mit riesigen Datenmengen trainiert werden. Diese werden zwar verarbeitet, aber die inhaltlichen Korrelationen gehen verloren. Sie lernen nur sinngemäß und nicht inhaltlich korrekt. Dazu kommt, dass die Trainingsdaten mit der Zeit veralten. Es sind also nicht immer zwingend falsche Informationen, sie können teilweise auch einfach veraltet sein (vgl. Hecker und andere, 2022, S. 65).

Ein Lösungsansatz dafür ist, die Modelle ihre eigenen Ausgaben gegenprüfen zu lassen. Dabei werden Suchmaschinen auf Basis der Sprachmodelle erstellt, die für den geschriebenen Text die Recherche durchführen. Die Suchmaschinen müssen dabei immer auf dem aktuellen Stand sein, damit ein Nachschlagen statt des bisherigen Auswendiglernens möglich ist. OpenAI hat mit WebGPT bereits so ein Modell entwickelt. Dies arbeitet besser als das reine GPT-3, produziert jedoch weiterhin Fehler. Es erspart die menschliche Arbeit und verringert zwar die Fehlerquote, kann sie jedoch nicht komplett eliminieren (vgl. Hecker und andere, 2022, S. 65f.).

Auch bei der Verbreitung von Fake News können Sprachmodelle eine Rolle spielen. Da mit ihnen schneller eine große Anzahl von Texten produziert werden kann als von Menschen, können falsche Informationen schneller und in einem größeren Maß verbreitet werden

(vgl. Stieler, 2021, S. 126). Dazu kommt, dass die modernen Modelle längere kohärente und nachvollziehbare Texte generieren können als alle vorigen Modelle. Es kann also eine Argumentation in langen Texten aufgebaut werden, die zunächst erstmal glaubwürdiger wirken kann (vgl. Elkins und andere, 2020, S. 3). Dies führt zu der Angst vor Missbrauch der Sprachmodelle, die auch die Entwickler:innen selbst äußern. Jede soziale Interaktion, die auf Text beruht, kann von Modellen manipuliert werden (vgl. Brown und andere, 2020, S. 35).

Ein Lösungsansatz hier besteht darin, die Öffentlichkeit mehr über Sprachmodelle und deren Wirkung aufzuklären und sie dafür zu sensibilisieren. Befürchtet wird dabei jedoch ein Vertrauensverlust in Medien, wenn nicht mehr klar unterschieden werden kann, welchen Texten Glauben geschenkt werden kann (vgl. Fröhling und andere, 2021, S. 2).

Für das Problem der Generierung und Verbreitung von Fake News durch große Sprachmodelle ist also bisher noch keine Lösung gefunden. Es bedarf hier weiterer Forschung und bis dahin einer wachsamem und sensibilisierten Öffentlichkeit.

### **Kontrolle großer Sprachmodelle**

Wie bereits dargestellt arbeiten große Sprachmodelle mit einer großen Systemarchitektur und für Menschen unüberschaubar großen Mengen Trainingsdaten. Dies führt dazu, dass die Modelle technisch nicht mehr komplett von Menschen verstanden werden können. Es wurde bereits vorgestellt, wie die Trainingsdaten nicht mehr von Menschen aufgearbeitet und geprüft werden können. Dazu kommen die Milliarden von Parametern, mit dessen Hilfe die Modelle ihre Ausgaben generieren. Diese bilden eine Art Black Box, da die Entscheidungen, wie es zu einer bestimmten Ausgabe kommt, von Menschen nicht nachvollzogen werden können (vgl. Brown und andere, 2020, S. 34). Hier bildet sich das generelle Problem tiefer neuronaler Netze ab, die keine Begründungen oder Erklärungen für Entscheidungen liefern (vgl. Hecker und andere, 2022, S. 65). Daher wird von Expert:innen die Forderung gestellt, dass die Modelle ihre Entscheidungswege erklären können sollen (vgl. Navigli, 2018, S. 5698).

Einerseits machen die vielfältigen Einsatzmöglichkeiten und Fähigkeiten die Modelle so ansprechend und interessant für weitere Entwicklungen. Andererseits stellt genau dies auch ein Problem dar. Bei der großen Bandbreite ist es nahezu unmöglich alle Fähigkeiten zu erfassen und vorauszusehen (vgl. Tamkin und andere, 2021, S. 4). Dies war einer der Gründe, warum OpenAI ihr GPT-3-Modell nur begrenzt zugänglich macht, da nicht alle Aktionen abzusehen sind. Andererseits erschwert die Unzugänglichkeit des Programmcodes

es zusätzlich, die Entscheidungen nachvollziehen zu können, wenn nur begrenzter Einfluss und Einsicht in die Parameter besteht (vgl. Stieler, 2021, S. 125).

Zudem benötigt man nicht für jede Aufgabe den kompletten Leistungsumfang des Modells. Es gibt Überlegungen, den Umfang zielgerichtet für die Aufgaben zu reduzieren, was das Modell wieder überschaubarer machen würde (vgl. Brown und andere, 2020, S. 34).

Daher stellt sich in diesem Kontext die Frage: Können Sprachmodelle zu groß werden? (vgl. Bender und andere, 2021, S. 610). Die vorläufige Antwort lautet, dass die Modelle voraussichtlich weiter wachsen werden, solange damit bessere Ergebnisse erzielt werden können (vgl. Bender und andere, 2021, S. 611). Wann und ob ein Punkt der Sättigung erreicht wird, ab dem Sprachmodelle nicht mehr besser werden können, ist derzeit noch nicht einschätzbar.

## **Datenschutz**

Rechtliche Fragen spielen bei der Nutzung großer Sprachmodelle ebenfalls eine Rolle. Neben der Lizenzierung (die für GPT-3 im folgenden Kapitel vorgestellt wird), sind dies vor allem Fragen zum Datenschutz und dem Urheberrecht. OpenAI ist ein amerikanisches Unternehmen, das anderen Standards unterliegt als der deutsche Datenschutz und der Datenschutz-Grundverordnung (DSGVO). Gerade bei der Arbeit mit sensiblen und personenbezogenen Informationen, kann es dadurch zu Konflikten kommen (vgl. Grävemeyer, 2022, S. 63). Vor der Verwendung sollten sich Anwendende darüber bewusst sein, dass die eingegebenen Daten auf Servern des Anbieters gespeichert werden und weiterverarbeitet werden. Besonders im Live-Einsatz ist dies auch den Nutzer:innen der Anwendung deutlich aufzuzeigen, damit diese wissen wie ihre Anfragen und persönlichen Daten verarbeitet werden.

Ein Lösungsansatz wären auch hier synthetische Daten. Diese künstlich erstellten Datensätze ließen sich nicht auf real bestehende Personen zurückverfolgen und würden daher nicht im Konflikt zum Datenschutz stehen. Hier tritt jedoch wieder das Problem der „stochastischen Papageien“ auf (vgl. Abschnitt *Qualität der Ausgabe*). Wenn das Modell Daten nur auswendig lernt, kann es vorkommen, dass es bei den synthetischen Daten reale Daten kopiert, die dann wieder personenbezogen zurück verfolgbar wären (vgl. Wolfangel, 2021, S. 22).

Seit 2021 gibt es daher Bestrebungen, eine europäische Alternative zu GPT-3 und anderen großen Sprachmodellen aufzubauen. Im Juni 2021 starteten die Entwicklungen für Open GPT-X. Federführend für die Entwicklung ist das Fraunhofer-Institut für Intelligente

Analyse- und Informationssysteme (IAIS) mit Unterstützung von verschiedenen Stakeholdern aus der Forschung und Wirtschaft. Es ist zudem Teil der \*Gaia-X-Initiative und greift auf dessen Expertise zurück. Finanzielle Unterstützung erhält das Projekt vom Bund (vgl. KI Bundesverband, 2022).

Von Open GPT-X versprechen sich die Beteiligten verschiedene Vorteile. Zum einen wird dieses Modell nach europäischen Datenschutz-Standards und Werten entwickelt. Anwendende wissen die Daten entsprechend sicher verwertet und gespeichert (vgl. Wölbart, 2020, S. 13). Zusätzlich soll die Sprachenvielfalt berücksichtigt werden, die in Europa herrscht, indem nicht nur auf Englisch trainiert wird (vgl. KI Bundesvorstand, 2022). Zum anderen möchte man nicht den Anschluss an moderne technische Entwicklungen verlieren (vgl. Grävemeyer, 2021, S. 40), die „digitale Souveränität Europas“ erhalten und nicht in eine Abhängigkeit von den USA oder China kommen (KI Bundesverband, 2022).

### **Urheberrecht**

Sprachmodelle können auf zwei verschiedene Arten urheberrechtlich geschützt sein: der zugrunde liegende Programmiercode und die Erzeugnisse des Sprachmodells bzw. die Erzeugnisse von KI-Anwendungen generell (vgl. Specht-Riemenschneider, 2021, S. 75). Im Kontext des Chatbots ist vor allem der zweite Punkt interessant. Es ist zu klären, ob dessen Ausgaben einem urheberrechtlichen Schutz unterliegen. Außerdem muss geklärt sein, wer in einem solchen Fall der Urheber ist: der/die Programmier:in oder die KI selbst.

Das deutsche Urheberrecht verlangt eine menschliche Schöpfung, um urheberrechtlichen Schutz auszusprechen (vgl. Specht-Riemenschneider, 2021, S. 73). Es setzt eine persönliche, geistige Beziehung zwischen einem/einer Autor:in und dessen Werk voraus, die eine Maschine aufgrund der fehlenden Persönlichkeit nicht haben kann. Jedoch hat nicht automatisch der Mensch das Urheberrecht an einem Erzeugnis einer KI, nur weil dieser den Anstoß dazu gegeben hat (vgl. Legner, 2019, S. 808). Insgesamt ist die Rechtsprechung in diesem Kontext strittig, da das Urheberrecht diese Fälle noch nicht explizit mitgedacht hat, es ergeben sich daraus jedoch Ableitungen, die heute bereits angewendet werden können (vgl. Specht-Riemenschneider, 2021, S. 74).

Das Urheberrecht erlaubt zur Schaffung eines Werkes weitere Mittel. Dies können bei Gemälden z. B. Pinsel sein, oder auch Computer, um Designs umzusetzen. Diese dürfen jedoch nur Hilfsmittel sein und nicht wie im Falle von KI-Anwendungen, komplett eigenständig agieren ohne weiteren Einfluss des Menschen. Nur auf einen Knopf zu drücken erfüllt nicht die benötigte Kreativleistung (vgl. Specht-Riemenschneider, 2021, S. 76 und

S. 78). Autonom erzeugte Ergebnisse einer KI fallen also nicht unter das Urheberrecht (vgl. Legner, 2019, S. 809). Dies ändert sich jedoch, wenn der Mensch im Nachhinein „eigene gestalterische Entscheidungen“ trifft und verschiedene Ausgaben beispielsweise zu einem kreativen Text verarbeitet. Dieser geschaffene Text kann dann dem Urheberrecht unterliegen und geschützt sein (vgl. Legner, 2019, S. 808). Es gibt verschiedene Überlegungen, ein Leistungsschutzrecht anzuwenden. Damit würden dann auch gestalterische Maßnahmen abgedeckt sein, die Menschen im Vorhinein vornehmen, und auf dessen Grundlage die KI Erzeugnisse kreiert (vgl. Legner, 2019, S. 809f. und vgl. Specht-Riemenschneider, 2021, S. 81f.).

Im Rahmen von Auskunft-Chatbots lässt sich jedoch sagen, dass die dort entstehenden Ausgaben in der Regel nicht die geforderte Schöpfungshöhe erreichen, um urheberrechtlichen Schutz zu genießen. Dies ist in diesem Kontext also zu vernachlässigen. Anders ist es bei Kreativ-Chatbots, die freier assoziieren und z. B. eigene Ideen entwickeln.

### **Limitationen**

Trotz der unendlich erscheinenden Einsatzmöglichkeiten von GPT-3 und vergleichbaren Modellen, klangen in diesem Kapitel bereits auch Grenzen an. GPT-3 ist bereits deutlich besser als seine Vorgänger im Lösen von NLP-Aufgaben, doch es hat weiterhin Probleme den Kontext eines Textes über längere Passagen aufrechtzuerhalten und zu viele Wiederholungen zu vermeiden (vgl. Brown und andere, 2020, S. 33).

Zudem benötigt das Modell noch immer mehr Trainingsdaten, um eine Aufgabe zu erfüllen, als ein Mensch jemals in seinem Leben verarbeiten kann. Verfolgt man also den Ansatz, eine KI zu schaffen die dem Menschen gleicht, lässt sich dies nicht mit den derzeitigen Ansätzen erreichen. Für Expert:innen ist das Können der Maschine nicht mit dem eines Menschen vergleichbar (vgl. Brown und andere, 2020, S. 34).

### **Umwelt und Kosten**

Für die Anwendenden sind die Ergebnisse nur wenige Klicks entfernt. Dass dabei im Hintergrund eine große Rechenleistung erbracht wird, kann schnell vergessen werden. Das Trainieren, Speichern und Anwenden der umfangreichen Modelle kostet Energie und Geld. So wird geschätzt, dass das Training eines einfachen BERT-Modells ohne weitere Spezifikationen bereits so viel Energie wie ein Transatlantikflug verbraucht. Da viele Entwicklungsfirmen für dieses Training nicht nur klimaneutrale Möglichkeiten der Stromerzeugung nutzen, bedeutet dies eine hohe Umweltbelastung. Dazu kommt der

Stromverbrauch während der Berechnung von Anfragen. Ein konkretes Beispiel für die Kosten eines Trainings: Um eine Verbesserung des \*BLEU-Scores bei der maschinellen Übersetzung von 0,1 zu erreichen, kostet das Neutraining \$150.000 (vgl. Bender und andere, 2021, S. 612).

Die Kosten für das Training sind sehr hoch, da große Mengen an Daten verarbeitet werden und dies eine gewisse Zeit benötigt. Hierbei fallen die größten Kosten an. Bei der Verarbeitung von Anfragen arbeiten die Modelle jedoch schon vergleichsweise effizient. Die Generierung von 100 Seiten Text mit GPT-3 benötigt etwa 0,4 kWh (vgl. Brown und andere, 2020, S. 399). Je nach Anforderung ist es sinnvoll mehr Geld in das Training zu stecken, wenn dafür die Kosten während der Laufzeit geringer sind (vgl. Bender und andere, 2021, S. 612).

### **Zusammenfassung**

GPT-3 baut auf die Erkenntnisse auf, dass größere Sprachmodelle bessere Ergebnisse erzielen. Zum Veröffentlichungszeitraum war es das bisher größte Modell, das entwickelt wurde. Es basiert auf der Transformer-Architektur und kann sich durch Attention-Module gezielt auf die Wahrscheinlichkeitsrechnungen konzentrieren, um das nächste Wort der Ausgabe zu bestimmen. Durch die vielen Parameter, Schichten und große Menge Trainingsdaten wird das Ergebnis genauer bestimmt und ein größerer Kontext hergestellt. Damit erzielt es in vielen Bereichen bessere Ergebnisse als bisherige state-of-the-art-Modelle und wird als vielversprechender Ansatz für noch bessere Modelle gewertet.

Während viele Expert:innen in der Entwicklung von GPT-3 einen großen technischen Fortschritt sehen, der die weitere Entwicklung von Sprachmodellen stark beeinflussen wird, bestehen auch Probleme, die weitere Forschung und Entwicklung benötigen. Dazu gehören die Ausgabequalität, die Generierung und Verbreitung von Fake News, die Reproduktion von Vorurteilen und der Bias von Sprachmodellen, die Kontrolle der großen Modelle, rechtliche Fragen zu Datenschutz und Urheberschaft sowie die Umweltbelastung und die Kosten zum Betreiben der Modelle.



## 7 Verwendung von GPT-3

Dieses Kapitel gibt Hinweise zum Zugang und zur Verwendung des Sprachmodells GPT-3. Zunächst wird die Schnittstelle und deren Funktionen vorgestellt, über die das Modell nutzbar ist. Weiterhin wird auf Kosten, Rechte und Rahmenbedingungen bei der Nutzung eingegangen. Um die vielfältigen Einsatzmöglichkeiten zu zeigen, werden außerdem die verschiedenen Funktionalitäten des Moduls kurz vorgestellt.

### 7.1 Schnittstelle

GPT-3 ist durch eine OpenAI Schnittstelle (englisch: application programming interface (API)) zugänglich. Die API ist durch eine umfangreiche Dokumentation beschrieben. Alle Optionen der Schnittstelle werden vorgestellt und es gibt Einführungen für alle Aspekte (vgl. OpenAI, 2022e). Dabei hat man nur Zugriff auf die API, nicht auf das Modell selbst. Es können also nur begrenzt Anpassungen vorgenommen werden. Der zu Grunde liegende Modell-Code ist für normale Nutzer:innen weder einsehbar noch veränderbar (vgl. Dale, 2021, S. 116).

Zentraler Inhalt der API ist die Textvervollständigung. Anfragen werden als sogenannte Prompts an das Modell gegeben. Dabei kann es sich entweder um eine Anleitung handeln, wie eine Anfrage zu lösen ist, oder durch verschiedene Beispiele, die GPT-3 fortführen soll. Die Anfragen werden dabei in Token unterteilt. Ein Token entspricht dabei ca. vier Buchstaben. Sehr kurze Worte bestehen aus einem Token, längere werden in mehrere Token zerlegt. Die API gibt eine Maximallänge von Anfragen vor. Der Prompt und die Ausgabe dürfen zusammen nicht mehr als 2048 Token umfassen, dies entspricht ca. 1500 (englischen) Worten (vgl. OpenAI, 2022g).

#### **Engines**

Die Schnittstelle bietet verschiedene Modell-Varianten an. Seit kurzem ist zudem ein Fine-Tuning von GPT-3 möglich, um bessere Ausgaben für spezifische Aufgaben zu erreichen. Von GPT-3 gibt es vier verschiedene Modell-Varianten, deren Eigenschaften in Tabelle 2 vorgestellt werden (vgl. OpenAI, 2022g).

<b>Modell</b>	<b>Eigenschaften</b>
<i>Davinci</i>	<ul style="list-style-type: none"> <li>- das fähigste Modell</li> <li>- besonders geeignet für Zusammenfassungen, Generierung von kreativen Inhalten und Verstehen von Intents in Texten</li> <li>- benötigt dafür mehr Ressourcen und ist somit teurer</li> <li>- Trainingsdaten wurden auf Juni 2021 aktualisiert</li> </ul>
<i>Curie</i>	<ul style="list-style-type: none"> <li>- leistungsstark und schnell</li> <li>- für differenzierte Aufgaben: Sentiment-Klassifikation, Zusammenfassungen, Beantwortung von Fragen</li> </ul>
<i>Babbage</i>	<ul style="list-style-type: none"> <li>- einfache Klassifizierungen</li> <li>- Semantic Search</li> </ul>
<i>Ada</i>	<ul style="list-style-type: none"> <li>- schnellstes Modell</li> <li>- Textanalyse, einfache Klassifizierungen, Adressenkorrekturen</li> </ul>

Tabelle 2: Übersicht der GPT-3 Varianten, deren Eigenschaften

Die Empfehlung des Entwicklers ist mit dem Modell *Davinci* zu beginnen, da dies die besten Ausgaben generiere. *Curie* liefere vergleichbare Ergebnisse mit weniger Kosten (vgl. OpenAI, 2022g).

Weitere Varianten von GPT-3, die in der Schnittstelle zur Verfügung stehen, sind Codex und Content Filter. Codex befindet sich momentan in einer Beta-Version, die ausgewählten Test-Nutzer:innen zur Verfügung steht. Dabei handelt es sich um das Coding-Modell von GPT-3, das für Code-Generierung genutzt werden kann (vgl. OpenAI, 2022g).

Die Content Filter werden genutzt, um unsicheren oder sensiblen Ausgabertext zu identifizieren. Dieser wird unterschieden in „safe“, „sensitive“ und „unsafe“. Die Content Filter sind für alle Nutzer:innen frei nutzbar, um die eigenen Anwendungen zu überprüfen. Gibt die Anwendung überwiegend Inhalte der Kategorie „unsafe“ aus, sollte die Anwendung nochmal eingehend geprüft werden und nicht veröffentlicht werden. Sie handelt dann entgegen der Usage Guidelines (siehe Kapitel 7.2) (vgl. OpenAI, 2022g).

## **Playground**

Eine Möglichkeit Zugang zu GPT-3 zu bekommen ist der sogenannte Playground. Dieser ist über den Browser aufrufbar und bietet eine grafische Benutzeroberfläche. Abbildung 9 zeigt einen Screenshot der Oberfläche (vgl. OpenAI, 2022h).

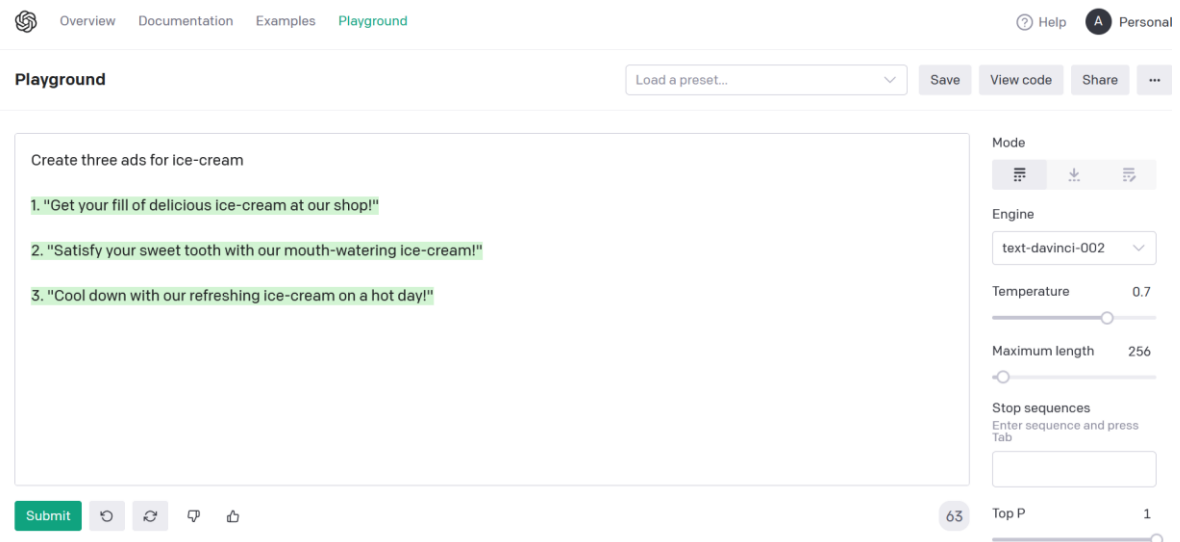


Abbildung 9: Playground der OpenAI API

Auf der linken Seite befindet sich das Eingabefenster. Der Prompt wird als schwarzer Text angezeigt. Ausgaben des Sprachmodells werden grün hinterlegt. Unter dem Fenster können die Vervollständigungen des Modells bewertet werden, was zur Verbesserung und Weiterentwicklung genutzt wird.

Auf der rechten Seite lassen sich einige Einstellungen vornehmen, die die Ausgabe beeinflussen. Diese werden im Folgenden überblicksartig vorgestellt:

- Mode: Modell-Modus wählen, möglich sind Complete, Insert, Edit
- Engine: Auswahl der GPT-3-Varianten
- Temperature: Sicherheit, mit der das Modell Vorhersagen trifft. Eine niedrigere Temperatur sorgt dafür, dass die Ausgabe genauer ist. Eine höhere Temperatur gibt dem Modell etwas Spielraum bei der Vervollständigung
- Maximum Length: Maximale Länge der Ausgabe, hängt auch vom gewählten Modell ab
- Stop sequences: Stopp-Punkt für das Modell, ab dem keine weiteren Sequenzen generiert werden
- Top P: steuert die Zufälligkeit der Antwort mit und wie viele der wahrscheinlichen Antworten berücksichtigt werden
- Frequency penalty: Wie sehr sich wiederholende Token „bestraft“ werden. Verhindert übermäßige Wiederholungen in der Ausgabe
- Presence penalty: Wie sehr neue Token im Text „bestraft“ werden. Erhöht die Wahrscheinlichkeit, dass das Modell über neue Themen spricht

- Best of: Serverseitig werden verschiedene Ausgaben generiert, aber nur die Beste ausgegeben
- Show probabilities: Zeigt an mit welcher Wahrscheinlichkeit ein Token generiert wurde. Hilfreich bei Verbesserungen oder um Alternativen zu sehen

Alle Einstellungen können auch durch die Voreinstellungen des Playgrounds verändert werden. Diese finden sich in einem Auswahlménü über dem Eingabefenster. Zu den Vorgaben „Zusammenfassung für einen Zweitklässler“, „Klassifikationen“, „Chat“, „Übersetzungen“ oder „Question Answering“ werden die Filter entsprechend angepasst und eine Vorlage für den ersten Prompt im Eingabefenster angezeigt (vgl. OpenAI, 2022h).

## API

Neben dem Playground kann das Modell auch über eine API vom Computer genutzt werden. Der Zugang lässt sich durch Anleitungen leicht installieren. Im Gegensatz zum Playground fehlt die grafische Oberfläche, alle aufgezählten Filter lassen sich jedoch auch hier anwenden. Die Abbildungen 10 und 11 zeigen Screenshots aus der API-Ansicht (OpenAI, 2022i).

```

1 function generatePrompt(animal) {
2   const capitalizedAnimal = animal[0].toUpperCase() + animal.slice(1).toLowerCase()
3   return `Suggest three names for an animal that is a superhero.
4
5   Animal: Cat
6   Names: Captain Sharpclaw, Agent Fluffball, The Incredible Feline
7   Animal: Dog
8   Names: Ruff the Protector, Wonder Canine, Sir Barks-a-Lot
9   Animal: ${capitalizedAnimal}
10  Names: `;
11 }

```

Abbildung 10: Ausgabe in der API-Ansicht

```

1 const completion = await openai.createCompletion("text-davinci-002", {
2   prompt: generatePrompt(req.body.animal),
3   temperature: 0.6,
4 });

```

Abbildung 11: Filter in der API-Ansicht

## Kosten

Die Nutzung des Modells ist mit Kosten verbunden. Jede:r Nutzer:in erhält zu Beginn ein Startguthaben von \$18 für drei Monate. Ist dies verbraucht oder die drei Monate um, muss für die Nutzung bezahlt werden. Dies gilt sowohl für den Gebrauch im Playground als auch über die API.

Das Fine-Tuning der verschiedenen Modelle wird ebenfalls pro Token berechnet. Es wird dabei zwischen Trainings- und Nutzungskosten unterschieden. Die Kosten beziehen sich dabei auf jeweils 1000 Token. Zum 1. September 2022 passte OpenAI die Preise für die Engines an, um die Nutzung „erschwinglicher“ zu machen. Möglich wird dies durch eine Verbesserung der Effizienz (vgl. OpenAI, 2022f). Die Preise für *Davinci* und *Curie* wurden dabei um 33 % gesenkt. *Babbage* und *Ada* kosten nun die Hälfte. Tabelle 3 enthält den aktuellen Stand der Kosten vom 1. September 2022.

Modell	Reines Modell	Training Fine-Tuning	Nutzung Fine-Tuning
<i>Ada</i>	\$0.0004	\$0.0004	\$0.0016
<i>Babbage</i>	\$0.0005	\$0.0006	\$0.0024
<i>Curie</i>	\$0.0020	\$0.0030	\$0.0120
<i>Davinci</i>	\$0.0200	\$0.0300	\$0.1200

Tabelle 3: Kosten für die Modelle sowie deren Fine-Tuning und Nutzung

## 7.2 Rechtliche Rahmenbedingungen

OpenAI hat ausführliche Usage Guidelines zusammengestellt, die die Nutzung der Schnittstelle und des Sprachmodells regulieren, wodurch ein Missbrauch vermieden werden soll. Wenn eine Anwendung in den Live-Betrieb wechseln soll, muss diese vorher zur Prüfung bei OpenAI eingereicht werden. Sobald der Entwicklungs- bzw. Test-Status verlassen wird, erfolgt eine Begutachtung durch OpenAI, um zu überprüfen, ob die Anwendung konform zu den Guidelines entwickelt und eingesetzt wird. Nach eigenen Angaben erfolgt die Prüfung durch OpenAI innerhalb von zwei Werktagen.

Zur Begutachtung eingereicht werden müssen alle Anwendungen, die

- mehr als 10 Endnutzer:innen haben.
- frei verfügbar über das Internet nutzbar sind.
- in irgendeiner Weise Gewinne für den/die Entwickler:in erzielen.
- für mehr als \$ 250 im Monat Ausgaben generieren.

Ändert sich das Hauptziel der Anwendung nach der Begutachtung, ist eine erneute Prüfung erforderlich. Um die Chancen der Freigabe zu erhöhen, geben die Entwickler:innen ausführliche Informationen zu Sicherheitseinstellungen und erlaubten Einsatzbereichen (vgl. OpenAI, 2022d).

Grundsätzlich darf das Modell nicht genutzt werden, um Hass-Inhalte, Belästigungen, Bedrohungen, Gewaltdarstellungen, selbstverletzendes Verhalten, politisch beeinflussende Inhalte, Spam, Täuschung und Schadsoftware zu verbreiten. Weiterhin verboten sind Anwendungen, die rassistische oder vorurteilsbelastete Ausgaben generieren, persönliche Informationen und Daten sammeln.

Für bestimmte sensible Anwendungsfälle wird eine besonders gründliche Kontrolle vorgenommen, die nur bestanden ist, wenn die Vorteile die Nachteile eindeutig überwiegen. Dazu gehört jeder Einsatz für die Justiz, für Regierungen, für den Medizin-Sektor, die Politik und persönliche Beratungen wie z. B. Coachings. Auch der Einsatz in sozialen Netzwerken wird stark reglementiert. Es sind keine automatischen Postings jeglicher Art erlaubt und es muss eine ständige Überwachung der Inhalte durch Menschen gewährleistet sein. Zudem sind keine erotischen oder romantischen Chatbots erlaubt oder solche, die die verbotenen Inhalte aus dem ersten Absatz reproduzieren.

Um eine Anwendung in diesen besonderen Bereichen genehmigt zu bekommen, muss nachgewiesen werden, dass ein Expert:innen-Team an der Entwicklung mitwirkt, das den Einsatz überwacht und schnell eingreift, wenn problematische Ausgaben generiert werden (vgl. OpenAI, 2022d).

### **Zusammenfassung**

Die Nutzung des Modells erfolgt über einen Schnittstellen-Zugang. Man hat keinen direkten Zugriff oder Einfluss auf das zu Grunde liegende Modell. Zudem ist die Verwendung kostenpflichtig, auch bereits im Entwicklungs- und Teststadium.

Der Einsatz des Modells wird stark von der Entwicklungsfirma OpenAI überwacht. Jede Live-Schaltung einer Anwendung muss genehmigt werden, nachdem eine vorige Prüfung der Anwendung erfolgt ist. Für die Nutzung wurden ausführliche Usage Guidelines entwickelt, die unbedingt zu befolgen sind.

Die Nutzung der Schnittstelle ist auch für Sprachmodell-Anfänger recht einfach zu erlernen und gestaltet. Über einen Playground können erste Erfahrungen mit dem Modell gesammelt werden. Ausführliche Tutorials und Hilfeseiten erleichtern die Anwendung.

## 8 Einsatzbereiche

GPT-3 bietet eine Vielzahl von Anwendungsmöglichkeiten, dazu gehören unter anderem Question Answering, Übersetzungen, Klassifikationen, Chat-Angebote und verschiedene Programmier-Anwendungen, die auch selbst Code erstellen können. Im Folgenden werden überblicksartig Beispiele für Einsatzgebiete aufgezählt. Einige der Anwendungsmöglichkeiten werden im Verlauf noch ausführlicher betrachtet: Konkrete Anwendungsbeispiele werden in Kapitel 9 vorgestellt, andere Bereiche werden mit exemplarischen Beispielen unter Nutzung der API in diesem Kapitel erläutert.

Der genutzte Programmcode für alle Beispiele findet sich in Anhang A.1.

### 8.1 Question Answering

Eine klassische Aufgabe von Sprachmodellen ist die Beantwortung von Fragen. Auch GPT-3 hat diese Fähigkeit (vgl. Brown und andere, 2020, S. 13f.). Als Beispielanwendungen werden vom Entwickler folgende Bereiche gelistet (vgl. OpenAI, 2022c):

- Question Answering: Fragen beantworten auf Grundlage von bestehendem Wissen
- Sachliche Beantwortung: Beantwortung von Fragen außerhalb der Wissensbasis
- Grammatikkorrektur: englische Sätze korrigieren

Exemplarisch werden mit der Schnittstelle die Beantwortung von Wissensfragen sowie die Korrektur von Grammatik reproduziert.

#### Beantwortung von Wissensfragen

Dafür wurde die Voreinstellung „Q&A“ des GPT-3-Playgrounds gewählt. Für einen Vergleich wurden allen vier Engines dieselben drei Fragen gestellt.

Die Engine *Davinci* konnte alle Fragen richtig und nahezu vollständig beantworten. Auch *Babbage* gelang eine vollständige Beantwortung der Fragen. Die Formulierung der Antworten unterschied sich dabei jedoch deutlich von *Davinci* (Abbildung 12).

*Curie* und *Ada* konnten nahezu keine der Fragen bearbeiten. Es ergab sich bei der für alle Engines exakt gleichen Formulierung die Fehlermeldung: „*The model predicted a completion that begins with a stop sequence, resulting in no output. Consider adjusting your prompt or stop sequences.*“ Stop-Sequenzen werden verwendet, um dem Modell mitzuteilen, wann keine neuen Ausgaben generiert werden sollen (vgl. Rhodes, 2022). Die Engines konnten die gestellten Fragen nicht richtig verarbeiten und beantworten.

Q: What is photosynthesis?

A: Photosynthesis is the process that produces organic molecules from simple inorganic molecules from the sun's energy.

Q: Who was Albert Einstein and when did he live?

A: Albert Einstein was a German-born theoretical physicist who developed the theory of relativity, one of the two pillars of modern physics (alongside quantum mechanics). He is best known in popular culture for his mass–energy equivalence formula  $E = mc^2$  (which has been dubbed "the world's most famous equation"). He received the 1921 Nobel Prize in Physics "for his services to theoretical physics, and especially for his discovery of the law of the photoelectric effect", a pivotal step in the development of quantum

Q: Who is the chancellor of Germany?

A: The current chancellor of Germany is Angela Merkel.

Q: What is photosynthesis?

A: Photosynthesis is the process by which light energy is converted into chemical energy in the plant cell.

Q: Who was Albert Einstein and when did he live?

A: Albert Einstein was born on March 14, 1879, in Germany. He died on April 15, 1955, in Switzerland.

Q: Who is the chancellor of Germany?

A: Angela Merkel is the chancellor of Germany. She was elected in 2005 and has served since then.

Abbildung 12: Wissensfragen an Davinci (oben) und Babbage (unten)

### Grammatikkorrektur

Hierfür wurde die Voreinstellung „Grammatical Standard English“ verwendet. Allen vier Engines wurde wieder dieselbe Aufgabe gestellt.

*Davinci* (Abbildung 13, oben links) schneidet bei der Korrektur am besten ab. Die Engine korrigiert alle drei Beispiele zu korrekten englischen Sätzen. Es erkennt sogar den häufigen Fehler, das deutsche Wort „bekommen“ mit „become“ zu übersetzen und korrigiert es richtig zu „receive“. *Curie* (Abbildung 13, oben rechts) und *Babbage* (Abbildung 13, unten links) erkennen noch den Fehler im ersten Beispiel, in den weiteren Beispielen werden die Fehler jedoch nicht korrigiert. Sie lassen die doppelte Verneinung stehen und erkennen auch den falschen Freund „become“ nicht. *Ada* (Abbildung 13, unten rechts) hat die größten Schwierigkeiten mit der Aufgabe. Sie kann nur das erste Beispiel richtig korrigieren, für die beiden anderen Beispiele konnten keine Ausgaben produziert werden.



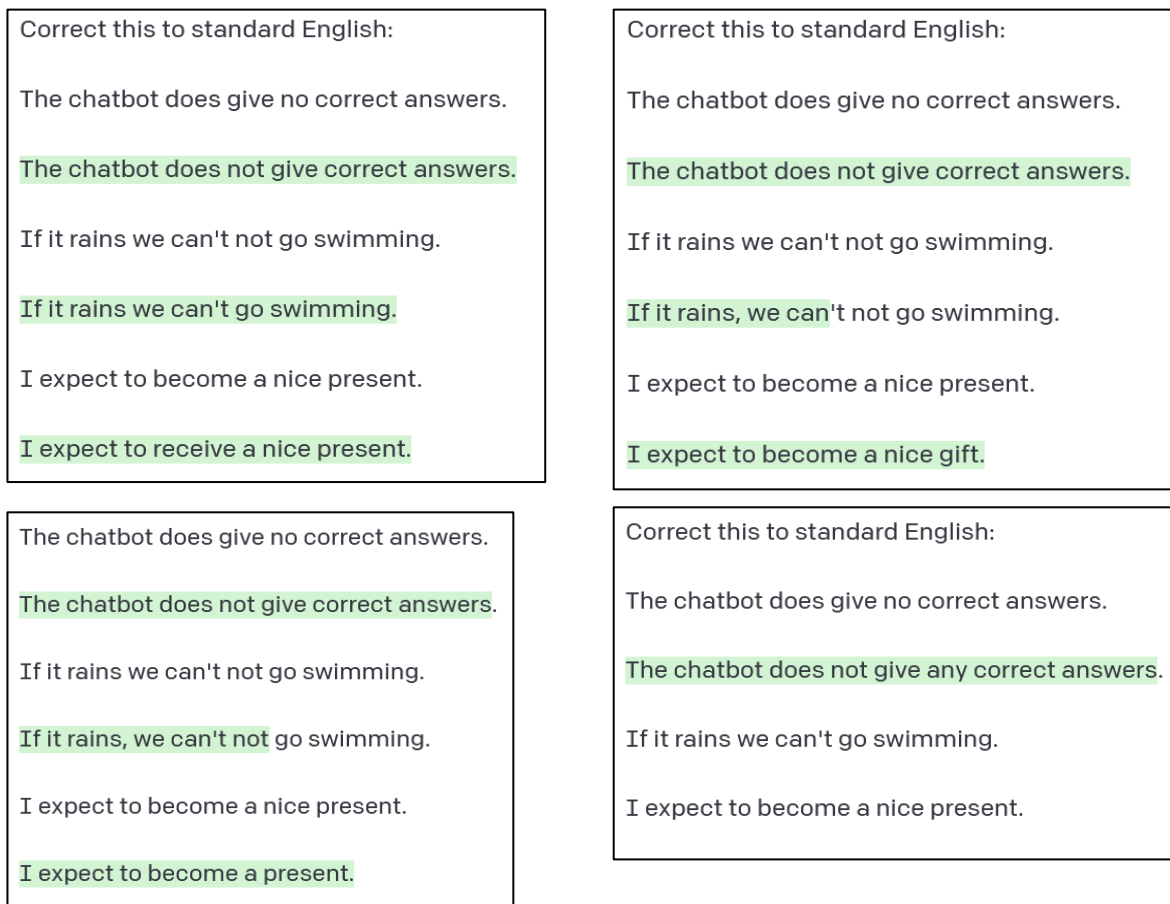


Abbildung 13: Grammatikkorrektur von Davinci, Curie, Babbage, Ada (von oben links nach unten rechts)

Für den Themenbereich Question Answering lässt sich daher festhalten, dass *Davinci* die fähigste Engine ist. Gerade bei der Grammatikkorrektur schneidet es am besten ab. Bei der Beantwortung von Wissensfragen hängt die Einschätzung vom Wissensinteresse des Fragenden ab. Dort liefern *Davinci* und *Babbage* jeweils Ausgaben, die von *Davinci* sind jedoch umfangreicher.

## 8.2 Texte erzeugen

Die Einsatzmöglichkeiten für die Texterzeugung mit GPT-3 scheinen nahezu grenzenlos (vgl. OpenAI, 2022c):

- Essay skizzieren: erstellt einen Essay-Entwurf zu einem Thema
- Rezeptgenerator: Erstellung von Rezepten aus Zutatenliste
- Wegbeschreibung: natürliche Sprache in Schritt-für-Schritt Wegbeschreibungen umwandeln

- Restaurantrezension: anhand einiger Notizen wird eine Rezension zu einem Restaurant erstellt
- Studien-Notizen: nach Vorgabe eines Themas werden Studien-Notizen generiert
- Interviewfragen: Erstellung von Interviewfragen
- Flughafencode-Extraktor: extrahiert Flughafenbezeichnungen aus Texten
- Kontaktinformationen-Extraktor: zieht Adressdaten aus Texten
- Stimmung zu Farbe: verwandelt eine Beschreibung in eine Farbe
- Analogien-Ersteller: kreiert Analogien
- Mikro-Horror-Story-Erzeuger: erzeugt kleine Horror-Geschichten aus zwei bis drei Sätzen
- Drittpersonenkonverter: erzeugt aus der Ich-Perspektive die dritte-Person-Perspektive
- Notizen zu Zusammenfassung: erstellt aus Notizen eine Zusammenfassung

Zur Veranschaulichung sollen die Anwendungsbereiche „Interviewfragen“ und „Essay skizzieren“ beispielhaft vorgestellt werden.

### Interviewfragen

Für die Interviewfragen wurde jede Engine aufgefordert, acht Interviewfragen zum vorgegebenen Thema zu entwickeln.

Alle Engines konnten diese Aufgabe erfüllen. Und auch wenn die Grundeinstellungen und die Aufforderung wieder identisch waren, unterscheiden sich die Ergebnisse sehr.

<p>Create a list of 8 questions for my interview about chatbots:</p> <ol style="list-style-type: none"> <li>1.What are chatbots?</li> <li>2.How do chatbots work?</li> <li>3.What are the benefits of chatbots?</li> <li>4.What are the challenges of chatbots?</li> <li>5.How can chatbots be used in business?</li> <li>6.How can chatbots be used in customer service?</li> <li>7.What are the best practices for designing chatbots?</li> <li>8.What are the ethical considerations of chatbots?</li> </ol>	<p>Create a list of 8 questions for my interview about chatbots:</p> <ol style="list-style-type: none"> <li>1. What inspired you to create a chatbot?</li> <li>2. How do you think chatbots will change the way people interact with technology?</li> <li>3. What are some of the challenges you faced when creating your chatbot?</li> <li>4. What do you think are the benefits of using chatbots?</li> <li>5. How do you think chatbots will impact the future of marketing?</li> <li>6. How do you think chatbots will impact the future of customer service?</li> <li>7. What are some of the challenges you see for the future of chatbots?</li> <li>8. What do you think are the biggest benefits of using chatbots?</li> </ol>
<p>Create a list of 8 questions for my interview about chatbots:</p> <ol style="list-style-type: none"> <li>1. What inspired you to create chatbots?</li> <li>2. What are the benefits of chatbots?</li> <li>3. What are the challenges of chatbots?</li> <li>4. What are the potential uses for chatbots?</li> <li>5. What are the ethical considerations of chatbots?</li> <li>6. What are the legal implications of chatbots?</li> <li>7. What are the technological challenges of chatbots?</li> <li>8. What are the potential benefits of learning about chatbots?</li> </ol>	<p>Create a list of 8 questions for my interview about chatbots:</p> <ol style="list-style-type: none"> <li>1. What is the most interesting thing about chatbots?</li> <li>2. How do chatbots make your work easier?</li> <li>3. What are the best chatbots for specific tasks?</li> <li>4. How do chatbots make you feel?</li> <li>5. What are the best chatbots for a specific purpose?</li> <li>6. What are the most interesting chatbots for a specific use-case?</li> <li>7. What are the best chatbots for security?</li> <li>8. What are the best chatbots for user experience?</li> </ol>

Abbildung 14: Interviewfragen von Davinci, Curie, Babbage, Ada (von oben links nach unten rechts)

Die Engine *Davinci* (Abbildung 14, oben links) stellt eher rationale Fragen zusammen, die sich auf die technische Entwicklung von Chatbots beziehen. Bei *Curie* (Abbildung 14, oben rechts) fällt auf, dass die Engine die persönlichsten Fragen mit einer direkten Ansprache entwickelt hat. *Babbage* (Abbildung 14, unten rechts) und *Ada* (Abbildung 14, unten links) haben die Fragen auch eher allgemein gehalten. Hier doppelten sich einige Fragen mit denen von *Davinci*.

### Essay skizzieren

Bei diesem Beispiel wurde allen vier Engines dieselbe Aufforderung gegeben, einen Essay-Entwurf zu skizzieren. Das vorgegebene Thema war „Die Auswirkungen von KI auf den Alltag der Menschen“.

Der Vorschlag von *Davinci* (Abbildung 15, oben) könnte als Gliederung direkt für einen geplanten Essay übernommen werden. Die einzelnen Punkte eignen sich gut als Überschriften und geben eine sinnvolle Struktur. Die Vorschläge von *Curie* (Abbildung 15, zweites von oben) und *Babbage* (Abbildung 15, drittes von oben) sind eher inhaltlicher Natur. Sie können als Fragestellungen genutzt werden, die in dem Essay untersucht werden. Die Ansätze sind kreativer als die von *Davinci*.

Die Vorschläge von *Ada* (Abbildung 15, unten) fallen aus dem Rahmen. Dies sind schon inhaltliche Thesen, die das Modell aufstellt. Es liefert nicht nur eine Strukturierung, sondern schon eine Argumentationsstruktur mit.

Je nachdem, wie viel Hilfe bei der Skizzierung des Essays benötigt wird, sollte eine andere Engine genutzt werden. Für eine sachliche Gliederung empfehlen sich nach diesen Beispielen *Davinci*, *Curie* und *Babbage*. *Ada* gibt schon eine starke inhaltliche Richtung vor.

Für den Bereich der Texterzeugung lässt sich keine Engine vorrangig empfehlen. Alle agieren gut in diesem Bereich. Je nach Anspruch und Schwerpunkt kann mit den Engines getestet werden, welche die Anforderungen am besten erfüllt.

<p>Create an outline for an essay about the impact of artificial intelligence on people's everyday lives:</p> <ol style="list-style-type: none"> <li>1. Introduction</li> <li>2. The history of artificial intelligence</li> <li>3. The current state of artificial intelligence</li> <li>4. The impact of artificial intelligence on people's everyday lives</li> <li>5. The future of artificial intelligence</li> </ol>	
<p>Create an outline for an essay about the impact of artificial intelligence on people's everyday lives:</p> <ol style="list-style-type: none"> <li>1. Introduction to artificial intelligence and its impact on people's everyday lives.</li> <li>2. Discussion of the ways in which artificial intelligence is already impacting people's lives, including in the workplace, in the home, and in social media.</li> <li>3. Discussion of the potential ways in which artificial intelligence will impact people's lives in the future, including in the areas of healthcare, education, and transportation.</li> <li>4. Conclusion of the essay, highlighting the ways in which artificial intelligence has the potential to change people's lives for the better.</li> </ol>	
<p>Create an outline for an essay about the impact of artificial intelligence on people's everyday lives:</p> <ol style="list-style-type: none"> <li>1. How artificial intelligence has the potential to change the way people live and work.</li> <li>2. The potential implications of artificial intelligence on people's daily lives.</li> <li>3. The ways in which artificial intelligence could be used to improve people's lives.</li> </ol>	
<p>Create an outline for an essay about the impact of artificial intelligence on people's everyday lives:</p> <ol style="list-style-type: none"> <li>1. Artificial intelligence has the potential to improve people's everyday lives by providing new ways to engage with their surroundings.</li> <li>2. For example, AI could help make more aware of their surroundings, which would make them more visible to inbound search engines and potential clients.</li> <li>3. AI could also help identify new interests and get you closer to using new technologies.</li> <li>4. AI could also help us better understand the world around us and the people in it.</li> <li>5. This could lead to better art, technology, and human interactions that share similar goals.</li> </ol>	

Abbildung 15: Essay-Skizzen von Davinci, Curie, Babbage, Ada (von oben nach unten)

### 8.3 Ideengenerierung

Eine spannende Anwendungsmöglichkeit von GPT-3 ist die Generierung neuer Ideen. KI-Systeme mag man eher noch mit kalten Maschinen verbinden, doch hier zeigt sich eine kreative Ader (vgl. OpenAI, 2022c):

- Werbung aus Produktbeschreibung: anhand der Produktbeschreibung eine Werbung gestalten
- VR-Fitness Ideengenerator: generiert Ideen für Fitness in Virtual Reality-Umgebungen

- Film zu Emoji: Filmtitel mit Emojis darstellen

Auch dies soll durch zwei Beispiele veranschaulicht werden: ein Ideengenerator und Filme als Emojis darzustellen.

### Ideengenerator

Um die Kreativität der Engines zu vergleichen, wurde allen vier das gleiche Thema gegeben. Sie sollten Ideen kreieren wie eine Unterrichtsstunde zu KI gestaltet werden kann.

*Davinci* (Abbildung 16, oben), *Curie* (Abbildung 16, zweite von oben) und *Ada* (Abbildung 16, unten) formulieren dazu Fragen, die während einer Unterrichtsstunde bearbeitet und beantwortet werden könnten. *Babbage* (Abbildung 16, dritte von oben) formuliert dies als Themenvorschläge. Dies klingt eher nach einem Unterrichtskonzept. Inhaltlich sind die Ausgaben aller Engines sinnvoll in dem Kontext.

Brainstorm some ideas to create a school lesson about artificial intelligence:

1. What is artificial intelligence?
2. What are some common applications of artificial intelligence?
3. How does artificial intelligence work?
4. What are some ethical considerations surrounding artificial intelligence?

Brainstorm some ideas to create a school lesson about artificial intelligence:

1. What is artificial intelligence?
2. What are the benefits of using artificial intelligence in schools?
3. What are some ways to teach artificial intelligence in schools?
4. What are some ethical considerations when using artificial intelligence in schools?
5. What are some ways to improve the teaching of artificial intelligence in schools?

Brainstorm some ideas to create a school lesson about artificial intelligence:

- An artificial intelligence lesson could discuss the basics of artificial intelligence, such as what it is and how it works.
- A lesson could also explore the different types of artificial intelligence, such as machine learning and artificial general intelligence.
- A lesson could explore the different applications of artificial intelligence, such as in business or social media.
- A lesson could also explore the different implications of artificial intelligence, such as how it could impact the future of humanity.

Brainstorm some ideas to create a school lesson about artificial intelligence:

- What are the different types of artificial intelligence?
- How are artificial intelligence being used in education?
- What are the benefits of artificial intelligence?
- How can artificial intelligence be used to improve our world?

Abbildung 16: Ideengenerierung von *Davinci*, *Curie*, *Babbage*, *Ada* (von oben nach unten)

## Film zu Emoji

Für diese Aufgaben wurden fünf Filmtitel vorgegeben. Den Inhalt der Filme sollten die Engines mit Emojis darstellen.

Nur die Engine *Davinci* hat die Aufgabe richtig verstanden und Emojis produziert (Abbildung 17, oben links). Diese passen auch zu den Inhalten der Filme. Selbst bei Filmen, deren Titel nicht eindeutig auf den Inhalt schließen lässt wie „Despicable Me“, wurden passende Emojis gewählt. Dies lässt darauf schließen, dass GPT-3 die Inhalte der Filme kennt und nicht nur mit den vorgegebenen Titeln gearbeitet hat.

Convert movie title into emoji:

Batman: 🦇 🚗 🦺 ⚡ 🦊 🧑

Lord of the Rings: 🗡️ 🛡️ 🧝 🧝 🎬

Harry Potter: ⚡ 🦉 📖 🦉 🐍 🧝

Titanic: 🚢 ❤️ 🌊

Despicable Me: 😬 🍌 💣

Convert movie title into emoji:

Batman: The Dark Knight

🏰 🏰 🏰 🏰 🏰

Lord of the Rings: The Fellowship of the Ring

🔔 🔔 🔔 🔔 🔔

Harry Potter: Order of the Phoenix

Titanic: Titanic

Despicable Me: Despicable Me

Convert movie title into emoji:

Batman: The Dark Knight

batman: the dark knight

Lord of the Rings: The Fellowship of the Ring

The Dark Knight Rises

Harry Potter: Philosopher's Stone

The Sorcerer's Stone

Potter and the Sorcerer's Stone

Deathly Hallows, Part 1

Deathly Hallows, Part 2

Titanic: The Movie

Despicable Me: The Minions

Convert movie title into emoji:

Batman: The End

What's the end?

Gotham City's gone

Lord of the Rings: The Fellowship of the Ring

The end of the world?

Harry Potter: The Salvation of Harry Potter

The end of the world?

The Lord of the Rings: The Fellowship of the Ring

The end of the world?

Titanic: The End of the Titanic

The end of the world?

Despicable Me: The degradation of donus

The end of the world?

Abbildung 17: Filmtitel als Emoji von Davinci, Curie, Babbage, Ada (von oben links nach unten rechts)

Die anderen drei Engines konnten die Aufgabe nicht erfüllen. *Curie* (Abbildung 17, oben rechts) hat immerhin noch versucht mit Emojis zu arbeiten. Dies entspricht jedoch nicht der Aufgabe. *Babbage* (Abbildung 17, unten links) und *Ada* (Abbildung 17, unten rechts) haben

die Filmtitel mit Untertitel vervollständigt oder Reihen aufgezählt. Auffällig bei *Ada* ist, dass immer nach dem Ende der Welt gefragt wird.

Für diese Aufgabe ist also eindeutig *Davinci* die richtige Wahl.

Für die Ideengenerierung hängt die Wahl der Engines ebenfalls von den gestellten Anforderungen ab. Für die Aufgabe, Filmtitel als Emojis darzustellen, kommt nur *Davinci* in Frage. Generell kreative Ideen generieren können jedoch alle Engines gut.

## 8.4 Klassifikationen

Eine typische Aufgabe von KI ist die Klassifikation. GPT-3 weitet dieses Gebiet deutlich aus im Vergleich zu vorigen KI-Anwendungen (vgl. OpenAI, 2022c):

- Klassifikation: klassifizieren von Beispielen in Kategorien
- Zusammenfassung für Zweitklässler: komplexe Texte in einfache Sprache umwandeln
- Tweet-Klassifizierer: Stimmungen in Tweets erkennen
- Schlagworte: Schlagworte aus einem Text erkennen

Hier soll anhand der Funktion „Zusammenfassung für Zweitklässler“ ein neuer Anwendungsbereich vorgestellt werden. Mit der Schlagworterkennung wird ein klassischer Bereich von KI vorgestellt.

### Zusammenfassung für Zweitklässler

Mit dieser Funktion sollen komplizierte Texte zusammengefasst und das Thema einfach erklärt werden. Hierfür wurde erneut eine Voreinstellung des Playgrounds genutzt („Summarization for a 2nd grader“). Den Engines wurde als Text die Einleitung des englischen Wikipediaartikels zur Photosynthese gegeben.

*Curie* und *Babbage* konnten zu der Frage keine Ausgabe generieren. *Davinci* und *Ada* lieferten Ergebnisse. Die Ergebnisse unterscheiden sich jedoch stark. *Davinci* (Abbildung 18, oben) fasst den langen Text in wenigen Sätzen zusammen. Diese sind einfach verständlich und inhaltlich korrekt. Die Ausgabe von *Ada* (Abbildung 16, unten) ist deutlich länger und zitiert teilweise wörtlich den eingegebenen Text. Dieser Text ist weniger verständlich, vor allem für Zweitklässler.

Für diese Art der Zusammenfassung sollte also die Engine *Davinci* genutzt werden.

Photosynthesis is a process that is used by plants and other organisms to convert light energy into chemical energy. This chemical energy can later be used to fuel the organism's activities. Some of the chemical energy is stored in carbohydrate molecules, such as sugars and starches. Photosynthesis is responsible for producing and maintaining the oxygen content of the Earth's atmosphere, and supplies most of the energy necessary for life on Earth.

Photosynthesis is a process used by plants and other organisms to convert light energy into chemical energy that, through cellular respiration, can later be released to fuel the organism's activities. Some of this chemical energy is stored in carbohydrate molecules, such as sugars and starches, which are synthesized from carbon dioxide and water – hence the name photosynthesis, from the Greek phōs (φῶς), "light", and sunthesis (σύνθεσις), "putting together".[1][2][3] Most plants, algae, and cyanobacteria perform photosynthesis; such organisms are called photoautotrophs. Photosynthesis is largely responsible for producing and maintaining the oxygen content of the Earth's atmosphere, and supplies most of the energy necessary for life on Earth.

In photosynthesis, light energy is absorbed by proteins called reaction centers and used to reduce or transfer electrons to create oxygen gas. In plants, algae, and cyanobacteria, sugars are synthesized by a subsequent sequence of light-independent reactions called the Calvin cycle. In this cycle, atmospheric carbon dioxide is incorporation into already existing organic carbon compounds, such as ribulose biphosphate (RuBP). Using the ATP

Abbildung 18: Zusammenfassung für einen Zweitklässler von Davinci (oben) und Ada (unten)

### Schlagworterkennung

Um die Schlagworterkennung der Engines zu vergleichen, wird ihnen der Abstract des Titels „Entrepreneurship marketing: principles and practice of SME marketing“ von Sonny Nwankwo und Ayantunji Gbadamosi eingegeben. Als Vergleichswert wird die Schlagwortvergabe im Fachportal EconBiz genutzt. Diese Schlagworte wurden von Fachreferent\*innen vergeben: KMU, SME, Marketing, Marketingmanagement, Marketing management, Online-Marketing, Internet marketing, Unternehmer, Entrepreneurs, Entrepreneurship.

Die Engines *Babbage* und *Ada* konnten keine Ausgaben dazu generieren.

-Small and medium-sized enterprises	Enterprise
-SMEs	Marketing
-marketing	SMEs
-consumer-generated marketing	Textbook
-social media marketing	
-business incubators	
-crowdfunding	
-case studies	
-undergraduate	
-postgraduate	

Abbildung 19: Schlagwortvergabe von Davinci (links) und Curie (rechts)

Die Ausgabe von *Curie* (Abbildung 19, rechts) ist recht kurz. Sie umfasst vier Schlagworte. Von denen stimmen zwei mit den Schlagworten der Fachreferent\*innen überein. Die Ausgabe von *Davinci* (Abbildung 19, links) ist umfangreicher, sie enthält zehn Schlagworte. Von diesen stimmen vier mit den aus EconBiz überein.



Für beide Ausgaben lässt sich sagen, dass die Begriffe korrekt sind und den Inhalt des Buches beschreiben. Für eine umfassendere Beschreibung sollte jedoch auch hier die *Davinci* Engine gewählt werden.

Für den Bereich der Klassifikation, egal ob die herkömmliche oder eine neue Interpretation, agiert *Davinci* also am besten.

## 8.5 Chat

Wie sich aus dem Bereich des Question Answering ableiten lässt, kann GPT-3 auch in Chatbots eingesetzt werden. Beispiele des Entwicklers dazu sind (vgl. OpenAI, 2022c):

- Chat: offene Gespräche mit KI-Assistent
- Sarkastischer Chatbot: sachlicher, sarkastischer Chatbot
- ML-/KI-Chatbot: Beantwortung von Fragen zu Sprachmodellen
- JavaScript Chatbot: Hilfe bei Fragen zu JavaScript

Da GPT-3 im Kontext des ZBW-Chatbots ausführlich getestet wird (Dritter Teil: GPT-3 zur Anwendung in der ZBW), wird hier auf weitere Beispiele verzichtet.

## 8.6 Programmcode

Eine ebenfalls neue Funktion, die GPT-3 mit sich bringt, ist mit Programmcode zu arbeiten. Dazu gibt es verschiedene Anwendungsbereiche (vgl. OpenAI, 2022c):

- Text zu Befehl: Text in Programmcode übersetzen
- SQL Übersetzung: natürliche Sprache in SQL-Befehle umsetzen
- Python in natürlicher Sprache: Python-Code in menschlich verständliche Sprache übersetzen
- Übersetzung von Programmiersprachen: eine Programmiersprache in eine andere übersetzen
- Code erklären: kompliziertes Stück Programmiercode einfach verständlich erklären
- Python Bug-Fixer: Suche und Korrektur von Fehlern im Programmcode
- Tabellenkalkulation: anhand von Daten verschiedene Tabellen erstellen
- JavaScript zu Python: Übersetzung der Programmiersprachen
- Freunde-Chat: Nachbilden von Chatgesprächen
- Unstrukturierte Daten einlesen: Tabellen aus langen Texten übersetzen
- Kalkulation von Zeitkomplexität: Zeitkomplexität einer Funktion bestimmen

Im Folgenden sollen beispielhaft die Anwendung eines SQL-Übersetzers und eines Übersetzers für Programmiersprachen dargestellt werden.

### SQL-Übersetzer

Mit GPT-3 ist es möglich natürlichsprachliche Anweisungen in einen SQL-Befehl zu übersetzen. Auch hier erhielten alle Engines denselben Prompt(vgl. OpenAI, 2022c):

```
Postgres SQL tables, with their properties:  
Employee(id, name, department_id)  
Department(id, name, address)  
Salary_Payments(id, employee_id, amount, date)  
A query to list the names of the departments which employed more  
than 10 employees in the last 3 months
```

Die Ausgaben unterscheiden sich bei den einzelnen Engines wieder deutlich voneinander. Die Ausgaben von *Curie* (Abbildung 20, zweite von oben), *Babbage* (Abbildung 20, dritte von oben) und *Ada* (Abb. 20, unten) sind falsch. Mit diesen Befehlen werden nicht die gewünschten Daten produziert. Nur die Ausgabe von *Davinci* (Abb. 20, oben) ist richtig und für die gestellte Aufgabe anwendbar.

```
SELECT d.name  
FROM Department d  
JOIN Employee e ON d.id = e.department_id  
JOIN Salary_Payments sp ON sp.employee_id = e.id  
WHERE sp.date >= DATE_SUB(CURDATE(), INTERVAL 3 MONTH)  
GROUP BY d.name  
HAVING COUNT(*) > 10
```

```
SELECT name FROM Salary_Payments WHERE employee_id > 10
```

```
SELECT department_id FROM employee  
WHERE employee_id IN (10)
```

```
SELECT COUNT(*) FROM Employee  
SELECT COUNT(*) FROM Department  
SELECT COUNT(*) FROM Salary_Payments
```

Abbildung 20: SQL-Übersetzung von Davinci, Curie, Babbage, Ada (von oben nach unten)

## Übersetzer Programmiersprachen

Um die Übersetzungsleistung zu prüfen, wurden die vier Engines aufgefordert folgenden Java-Code in Python zu übersetzen (vgl. OpenAI, 2022c):

```
Translate this function from Java into Python
### Java
public class HelloWorld{
    public static void main (String [] args){
        System.out.println("Hallo Welt");
    }
}
```

<pre>### Python def main():     print("Hello World") if __name__ == "__main__":     main()</pre>	<pre>### Python def main():     print("Hello, world!")</pre>	<pre>### Python def hello_world(args):     print("Hallo Welt")</pre>	<pre>### Python class HelloWorld(object): def main(args):     System.out.println("Hallo Welt")</pre>
--	--	--	--

Abbildung 21: Übersetzung von Programmiersprachen durch *Davinci*, *Curie*, *Babbage*, *Ada* (von links nach rechts)

Die Ausgaben der Engines sehen sehr unterschiedlich aus. Dies ist erstmal kein Indikator auf Fehler, da es für einen Befehl verschiedene Umsetzungen geben kann. *Davinci* (Abbildung 21, links) und *Curie* (Abbildung 21, zweite von links) liefern dabei korrekte Python-Befehle, die ausgeführt werden können. Die Ausgaben von *Babbage* (Abbildung 21, zweite von rechts) und *Ada* (Abbildung 21, rechts) enthalten Syntax-Fehler und können nicht ausgeführt werden.

Auch hier liefert *Davinci* erneut die besten Ergebnisse, gefolgt von *Curie*.

Für die Verwendung zur Hilfe beim Programmieren schneidet in diesen Beispielen die Engine *Davinci* wieder am besten ab.

### **Zusammenfassung**

Dieses Kapitel hat die unterschiedlichen Einsatzbereiche des Sprachmodells GPT-3 vorgestellt. Neben einer umfassenden Auflistung von Einsatzszenarien wurden einzelne dieser Szenarien exemplarisch mit der API getestet. In den überwiegenden Fällen schnitt die Engine *Davinci* dabei am besten ab. Nie lieferten zwei unterschiedliche Engines ein exakt gleiches Ergebnis ab. Zudem wechseln auch die Ausgaben der einzelnen Engines bei jeder Neuaufforderung zur Ausgabe.

Dies zeigt auf, dass *Davinci* die mächtigste Engine ist, die OpenAI betreibt. Diese ist jedoch auch am teuersten im Gebrauch. Für jede Aufgabe sollten also auch die anderen Engines getestet werden. Je nach Aufgabe lieferten auch sie Ergebnisse, die den Anforderungen und der Zielsetzung entsprechen können.

## 9 Anwendungsbeispiele

Durch die vielfältigen Anwendungsgebiete wird GPT-3 von vielen Firmen und Expert:innen für einen möglichen Einsatz getestet. In diesem Kapitel werden einige Beispiele vorgestellt, in denen GPT-3 bereits im Einsatz ist. Um die Varianz noch einmal deutlich zu machen, sind es nicht nur Beispiele aus Chatbot-Bereichen, sondern aus allen möglichen Einsatzbereichen. Eine ausführliche Übersicht hat z. B. Apideck (2022a) zusammengestellt. Nach derzeitigem Stand benutzen bereits über 300 Anwendungen das Sprachmodell (vgl. KI Bundesvorstand, 2022).

### **Coding**

Für den Einsatz beim Programmieren wurde eine Variante von GPT-3 entwickelt: Codex. Eingesetzt wird dies derzeit z. B. bei dem GitHub Copiloten, der im Folgenden vorgestellt wird (vgl. GitHub, 2022).

Das Codex-Modell wurde mit öffentlichem Programmiercode und natürlichsprachlichen Texten trainiert. Der Copilot hat Zugriff auf das Modell und kann damit direkt in der Entwicklungsumgebung Vorschläge für Funktionen und ganze Code-Blöcke geben. Derzeit werden dabei die Editoren Visual Studio Code, JetBrains und Neovim unterstützt. Das Modell soll Hilfen in sämtlichen Programmiersprachen geben können. In dem derzeit laufenden Test funktioniert es jedoch besonders gut bei Python, JavaScript, TypeScript, Ruby, Java und Go (vgl. GitHub, 2022).

Codex kann dabei sowohl einzelne Befehle vervollständigen, Funktionen kreieren oder ganze Code-Blöcke programmieren. Auch aus aussagekräftigen, natürlichsprachlichen Kommentaren von Entwickler:innen kann das Modell den gewünschten Abschnitt im Programmcode generieren. Dabei gibt es eine Auswahlmöglichkeit aus verschiedenen Vorschlägen, aus denen Entwickler:innen den für sich passenden Vorschlag in den Code einbauen können. Dieser Code sollte jedoch immer noch einmal vom Menschen geprüft werden. Codex befindet sich in der Testphase auf dem Stand eines unerfahrenen Programmierers, wodurch Fehler entstehen können. Zudem wurden auch bereits Sicherheitslücken in den vorgeschlagenen Codes gefunden (vgl. Königstein, 2022, S. 72f. und GitHub, 2022). Vorsicht ist geboten bei sensiblen Informationen, da der Code zur Verarbeitung in die GitHub-Cloud gespielt wird (vgl. Königstein, 2022, S. 72). Der generierte Code soll jedoch dem/der Entwickler:in gehören und es bedarf keiner Angabe der Mitwirkung von Codex an der Erstellung (vgl. GitHub, 2022).

In der Testphase war der Copilot nur für ausgewählte Entwickler:innen verfügbar. Seit Ende Juni 2022 ist er jedoch für jede Privatperson zugänglich. Die Nutzungskosten belaufen sich auf \$10 pro Monat bzw. \$ 100 pro Jahr (vgl. The GitHub Blog, 2022). Eine kommerzielle Version soll demnächst folgen (vgl. GitHub, 2022).

### **Ideen-Generierung**

Gibt man GPT-3 einige Rahmenbedingungen vor, kann es zur Generierung von Ideen und Konzepten genutzt werden. So kann es beispielsweise Design-Konzepte entwickeln, die einigermaßen kompetent wirken. Dabei gab es jedoch auch viele Ausgaben mit niedriger Qualität, aus denen die guten Ideen gefiltert werden müssen (vgl. Zhu und andere, 2021, S. 7).

Bei der Entwicklung von Computerspielen werden künstliche Intelligenzen bereits eingesetzt, um den Programmierer:innen Arbeit abzunehmen. Nun gibt es auch Versuche, die ganze Spielidee von einer KI entwickeln zu lassen. Auf die Anfrage, Ideen für ein Spiel zu entwickeln, gibt GPT-3 verschiedene Ideen aus. Es überlegt sich dazu auch einen passenden Namen, z. B. für das Spiel *Super Chore Man*. Dies ist ein Spiel, bei dem man Aufgaben aus dem Haushalt erledigt und dabei ein Superheld ist. Gemeinsam mit der Hilfe bei der Programmierung entstehen so in kurzer Zeit kreative neue Spiele, die die Computerspielwelt beeinflussen können (vgl. Schreiner, 2021a).

Das Text-Adventure (Pen&Paper) Spiel *AI Dungeon* wurde mit GPT-2 entwickelt, inzwischen nutzt es auch die neue Version GPT-3. Die KI fungiert dabei als Spielleiter und generiert laufend neue Ideen und Content, um die Abenteuer spannend zu halten. Vor allem durch den Einsatz von GPT-3 gelingt es, die Geschichten zusammenhängender zu erzählen. Hier zeigt sich die Fähigkeit von GPT-3, längere kontextbezogene Texte generieren zu können. Durch diese Funktion macht die KI den menschlichen Spielleiter:innen Konkurrenz und erschafft eine neue Erlebnisstufe von Computerspielen (vgl. Bastian, 2020).

### **Journalismus**

Bereits in der Einleitung wurde auf einen Artikel verwiesen, der von GPT-3 geschrieben wurde. Als Beispiel für den Einsatz des Sprachmodells im Journalismus soll an dieser Stelle ein genauerer Blick darauf geworfen werden.

Die Zeitung *The Guardian* stellte GPT-3 im September 2020 die Aufgabe, einen Artikel zu schreiben. Dieser sollte die Leser überzeugen, dass Roboter in Frieden kommen. GPT-3 erhielt eine Aufgabenstellung und die ersten zwei Absätze als Vorgabe. Ohne den Hinweis,

dass eine Maschine den Text geschrieben hat, würde man dies nicht unbedingt vermuten. Der Text liest sich flüssig, ist grammatikalisch korrekt und hält einen roten Faden. Am Ende erläutern die Journalist:innen, dass GPT-3 insgesamt acht verschiedene Essays ausgegeben hat, aus denen eine neue Version erstellt und veröffentlicht wurde. Was die Leser:innen hier vorgelegt bekommen, ist also die Auswahl der besten Versionen verschiedener Ausgaben des Artikels (vgl. GPT-3, 2020).

Katherine Elkins, Professorin für Vergleichende Literatur- und Geisteswissenschaften, führte eine Untersuchung durch, ob GPT-3 einen Turing Test für Autoren bestehen würde. Ihr Urteil dort lautete nein, wenn man sämtliche Ausgaben des Modells berücksichtigt. Es seien zu viele fehlerhafte Ausgaben und Texte mit schlechter Qualität enthalten. Wenn jedoch die besten Stellen aus sämtlichen Varianten zusammengefügt werden, kann ein Mensch davon getäuscht werden (vgl. Elkins und andere, 2020, S. 12).

Mit ein wenig Hilfe und einem guten Lektorat können mit GPT-3 also menschenähnliche Texte geschrieben werden. Dies eröffnet dem Journalismus und auch Autor:innen neue Möglichkeiten, Geschichten zu schreiben.

## **Medizin**

Auch im medizinischen Bereich wird untersucht, inwieweit GPT-3 dort Abläufe unterstützen kann. So gibt es beispielsweise eine Untersuchung, wie GPT-3 im Bereich der elektronischen Patientenversorgung eingesetzt werden könnte. Dort wird festgestellt, dass GPT-3 weiterentwickelt ist als bisherige Chatbots und die Kommunikation natürlicher und menschähnlicher wirkt. So sei es z. B. zur Datensammlung einsatzfähig, indem es Gespräche mit den Patient:innen vorab führt, die Vorgeschichte erfasst und damit das medizinische Personal entlastet. Außerdem könnte es bei der Kategorisierung von unkritischen Fällen in der Notaufnahme helfen. Es werden also durchaus Chancen gesehen, wie GPT-3 im Medizinsektor unterstützend wirken kann. Jedoch sollte dabei immer deutlich gemacht werden, dass die Patient:innen mit einem Chatbot bzw. einer Maschine kommunizieren. Bei einem simulierten Patienten unterstützte das Modell dessen Suizid-Wunsch, was natürlich niemals vorkommen sollte. Transparenz und ein umsichtiger, kontrollierter Einsatz sind hier sehr wichtig (vgl. Korngiebel und andere, 2021).

Für den Einsatz ist es ebenfalls wichtig, genügend und korrekte Trainingsdaten zu haben. Dies ist häufig ein großes Problem, da die Dokumentation aufwendig ist und nur von geschultem medizinischen Fachpersonal vorgenommen werden kann. Wenn es ein solches gelabeltes Datenset gibt, kann es aus Datenschutzgründen oft nicht geteilt werden, da es sich

um sensible personenbezogene Daten handelt. Diesem Problem haben sich Chintagunta und andere (2021) angenommen, indem sie mit Hilfe von GPT-3 synthetische Trainingsdaten generiert haben. GPT-3 erhielt dazu Zugriff auf medizinisches Hintergrundwissen und wurde dann mit einigen, von Menschen gelabelten Datensätzen trainiert. Auf dieser Basis kann GPT-3 neue gelabelte Trainingsdatensätze generieren (vgl. Chintagunta und andere, 2021, S. 68). Der Output wurde von Fachpersonal auf Richtigkeit geprüft. Nur so kann sichergestellt werden, dass sich GPT-3 nichts Falsches beibringt und alle späteren Ausgaben korrekt sind. Da es sich um fiktive Datensätze handelt, gibt es auch keine Probleme mit dem Datenschutz, wenn die Datensätze geteilt werden (vgl. Chintagunta und andere, 2021, S. 70). So konnte erfolgreich ein Daten-Labeler mit fundiertem Hintergrund erstellt werden, der medizinisch korrekte Trainingsdatensätze generiert. Mit Hilfe weiterer Forschung soll die Genauigkeit erhöht werden (vgl. Chintagunta und andere, 2021, S. 72).

Gerade in diesem Bereich, bei dem es um sensible, lebensentscheidende Situationen geht, wird ein bereits dargestelltes Problem von KI sichtbar: Das Problem Gefühle und Situationen zu erkennen und darauf angemessen zu reagieren. Hier wird deutlich, dass die Maschinen keine Empathie entwickeln, sondern aufgrund rein statistischer Ansätze logische Entscheidungen treffen. Sie sind sicherlich hilfreich, wenn es darum geht Daten zu sammeln und in der Forschung auszuwerten, da sie Datenmengen sortieren können, die für Menschen nicht überschaubar sind. Im Einsatz für sensible zwischenmenschliche Themen bedarf es jedoch noch vieler Verbesserungen.

### **Sprachen lernen**

In Sprachlern-Apps findet GPT-3 ebenfalls bereits Anwendung. Der Anbieter *Duolingo* setzt GPT-3 in seinen Französisch-Kursen ein, um Grammatik zu korrigieren. Laut einer internen Untersuchung sollen dadurch die Fähigkeiten der Lernenden gewachsen sein (vgl. Apideck, 2022b). Das Unternehmen setzt schon längere Zeit auf KI zur Unterstützung im Lernprozess und betreibt dazu Forschung (vgl. Duolingo, 2022). Ein öffentlicher Bericht zum Einsatz von GPT-3 liegt zum jetzigen Zeitpunkt nicht vor.

Der Sprach-Chatbot Polyglot AI basiert ebenfalls auf GPT-3. Der Bot stellt dabei quasi einen Tandem-Partner beim Sprachen lernen dar, mit dem die Lernenden chatten können, um so die Sprachkenntnisse zu festigen (vgl. Apideck, 2022c).



## **Theater**

Am 31. März 2022 feierte im Theater Bremen ein Stück Premiere, das von GPT-3 geschrieben wurde. In dem Stück „Verfall. Ein Picknick im Grünen“ geht es um die Natur und den Tod – und wie die KI diese Themen betrachtet. Die Dramaturgin und der Regisseur des Stücks gaben gewisse Rahmenbedingungen vor, wie z. B. die Figuren. Wie die Handlung verlief, entschied jedoch GPT-3 allein. Entstanden ist dabei ein teilweise sehr absurdes Theaterstück. Dies wird genauso auf die Bühne gebracht. Die Beteiligten sahen sich dabei jedoch auch mit bereits geschilderten Problemen konfrontiert: GPT-3 baute Stereotype und unpassende Narrative ein. Aus diesem Grund sehen die Verantwortlichen ihre Jobs auch nicht gefährdet, da die KI noch immer eine menschliche Korrektur benötigt (vgl. Lippert, 2022 und Theater Bremen, 2022).

Auch ein Theater in London experimentierte bereits mit GPT-3 als Drehbuchautor. Die Entstehung ähnelt der vom Theater Bremen: In Zusammenarbeit mit GPT-3 entwickelten die Autoren das Drehbuch. Mit den passenden Eingaben konnte GPT-3 originelle Ideen entwickeln, die dann von den Autoren aufgegriffen und weiter ausgearbeitet wurden. Dabei erkannte die KI große aktuelle Probleme wie den Klimawandel, Hungersnöte und Krankheiten. Jedoch sahen sich auch die Beteiligten mit Vorurteilen und problematischen Stereotypen konfrontiert, die GPT-3 ausgab. Das Fazit ähnelt ebenfalls dem aus Bremen: Geschichten, die rein von einer KI stammen wirken eher flach, nur mit menschlichem Input erwachen sie richtig zum Leben (vgl. Akbar, 2021).

### **Zusammenfassung**

Die Anwendungsbereiche von GPT-3 sind ebenso vielfältig wie dessen Fähigkeiten. Es gibt bereits einige Beispiele, in denen GPT-3 im Live-Einsatz regelmäßig angewendet wird. Andere Beispiele sind erste Ideen, wie das Modell bisherige Probleme lösen könnte. Für alle Beispiele bildet das Modell einen Durchbruch und einen großen Fortschritt der jeweiligen Anwendung. Es verbessert deren Performance sowie Qualität und erschließt neue Einsatzbereiche für Sprachmodelle.

## 10 Ausblick zur Zukunft des Modells

Wie bereits dargestellt wurde, stellt GPT-3 einen enormen Fortschritt bei der Entwicklung großer Sprachmodelle dar. In vielen Bereichen, vor allem in der Textproduktion, wird es Menschen unterstützen (vgl. Floridi und andere, 2020, S. 691). Auch im Einsatz von Chatbots und digitalen Assistenten kann das Modell hilfreich sein. Es wird besser auf individuelle Anfragen reagieren können als vorgeplante Systeme. Dadurch verbessern sich die Antworten und damit die Beziehung zu den Nutzer:innen (vgl. Floridi und andere, 2020, S. 692). Dabei sollten jedoch nicht die genannten Probleme außer Acht gelassen werden. Zur Verbreitung von Werbung und Fake News lassen sich die Modelle genauso gut nutzen, wie für jede andere Textproduktion. Damit kann großer Schaden angerichtet werden, wenn Menschen bewusst getäuscht und mit falschen Informationen versorgt werden (vgl. Floridi und andere, 2020, S. 692).

Expert:innen fordern deshalb, dass die Mitwirkung von KI an Texten offen gelegt werden muss, um Menschen dafür zu sensibilisieren. Zudem gibt es Überlegungen, wie gemessen werden kann, ob die Modelle eine gesellschaftlich nützliche Wirkung haben. Für diese Punkte gibt es derzeit keine Übereinkünfte oder Vorgaben (vgl. Tamkin und andere, 2021, S. 5).

Neben den genannten Problemen ist dies ein Feld, das weitere Arbeit und Entscheidungen erfordert. Bevor ein breiter Einsatz von verschiedenen KI-Arten möglich ist, sollten dafür gesetzliche Rahmenbedingungen geschaffen werden.

Google kündigte dazu bereits erste Konsequenzen an: Alle Inhalte unter Mitwirkung von KI sollen zukünftig als Spam eingestuft werden und nicht in der Trefferliste auftauchen. Damit nimmt Google einen entscheidenden Einfluss auf die Auffindbarkeit von Inhalten, die mit Unterstützung von GPT-3 und vergleichbaren Modellen entstanden sind. Es wird dabei nicht differenziert, um welche Art Veröffentlichung und Mitwirkung es sich dabei handelt (vgl. Bastian, 2022).

Für komplett eigenständige Arbeit ist die Ausgabequalität insgesamt noch nicht gut genug. Doch die bisherigen Forschungen zeigen, dass ein noch größeres Modell bessere Ausgaben erzielen könnte. So wird bereits an GPT-4 geforscht, um die Schwachstellen von GPT-3 weiter zu verbessern (vgl. Stieler, 2021, S. 125). In einem inoffiziellen Gespräch verriet Sam Altman, dass GPT-4 nicht deutlich größer als GPT-3 werde, aber durch hochwertigere Daten und bessere Algorithmen eine bessere Kontextverarbeitung erreicht werden soll (vgl. Schreiner, 2021b).

Seit GPT-3 wurden auch bereits größere Modelle veröffentlicht. Eine Übersicht zur Entwicklung von Sprachmodellen zeigt Abbildung 22 (Bender und andere, 2021, S. 611). In den letzten zwei Jahren ist dabei deutlich die stetige Erhöhung der Parameter zu erkennen. Auch die

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	-
2021	Switch-C [43]	1.57E+12	745GB

Abbildung 22: Übersicht großer Sprachmodelle

Größe des Trainingsdatensets zeigt einen Wachstumstrend.

In letzter Zeit wurden noch weitere Modelle veröffentlicht. Das Modell *Wu Dao 2.0* wurde in China entwickelt. Über das Modell ist nicht viel bekannt, außer dass es mit 1750 Milliarden Parametern trainiert sein soll. Es ist nur in Chinesisch verfügbar (vgl. Hecker und andere, 2022, S. 64 und vgl. Grävemeyer, 2021, S. 40). Auch aus den USA kommen neue, große Modelle. Das Modell *Gopher* von der Firma DeepMind enthält 280 Milliarden Parameter. Und auch Microsoft hat mit *Megatron-Turing NLG* ein großes Sprachmodell mit 530 Milliarden Parametern entwickelt (vgl. Hecker und andere, 2022, S. 64).

Aus Deutschland kommt das Modell *Luminous* der Firma Aleph Alpha, die auch bei der Entwicklung von OpenGPT-X beteiligt ist. Dieses Modell wurde mit 800 GB Daten auf Deutsch, Englisch, Französisch, Italienisch und Spanisch trainiert. Vor allem für den deutschsprachigen Raum und Europa wird dies interessant sein, da es keine Daten speichert und damit einen DSGVO-konformen Einsatz erlaubt (vgl. Grävemeyer, 2020, S. 63). Die weitere Entwicklung von OpenGPT-X wird für Anwender:innen in Deutschland interessant zu verfolgen sein. Dies könnte einen werte- und datenschutzkonformen Einsatz in der Einrichtung erlauben, ohne die persönlichen Daten an ausländische Firmen weitergeben zu müssen, deren Richtlinien im Umgang mit sensiblen Daten anders sind als europäische.

Die bisherigen Erkenntnisse der Forschung legen die Vermutung nahe, dass die Modelle weiter wachsen werden in der Hoffnung, noch bessere Ergebnisse zu erzielen. Ob irgendwann ein Sättigungspunkt erreicht wird, lässt sich heute nicht abschätzen. Die Entwicklung und Erforschung großer Sprachmodelle wird in den nächsten Jahren fortgeführt und es wird spannend werden, welche Möglichkeiten sich daraus ergeben.

## 11 Zusammenfassung der Ergebnisse

Dieser Abschnitt hat einen umfangreichen Einblick in das Sprachmodell GPT-3 und seine Hintergründe gegeben.

Zu Beginn wurde die Entwicklungsfirma hinter dem Modell vorgestellt. OpenAI gründete sich mit der Vision, den Einsatz von KI weltweit zum Wohle der Menschheit zu überwachen. Inwieweit dies möglich ist und gelingt, wird erst die Zukunft zeigen. Trotz dessen ist es ein erfolgreiches Unternehmen, dem große Durchbrüche auf dem Gebiet der intelligenten Sprachmodelle gelungen sind. Nach wie vor müssen sie sich der Kritik stellen, nicht so offen zu sein wie Name und Vision es versprechen, indem sie ihre Modelle nur begrenzt zugänglich machen. Da in der Vergangenheit bereits mehrmals die Firmenstrategie angepasst wurde, bleibt abzuwarten wie die Firma sich weiterentwickelt und wie es mit weiteren Veröffentlichungen umgehen wird, wenn diese ähnliche Durchbrüche erzielen wie GPT-3.

GPT-3 setzt bei der Erkenntnis an, dass größere Sprachmodelle bessere Ergebnisse erzielen. Dies erkannten die Entwickler:innen bereits bei der Entwicklung des Vorgängermodells. Mit GPT-3 wurde das zum Veröffentlichungszeitpunkt bisher größte Sprachmodell entwickelt. Es enthielt mehr Parameter und neuronale Schichten als alle Vorgänger, und wurde mit dem größten Satz Trainingsdaten trainiert. Die Ergebnisse sprechen dabei für sich: In verschiedenen NLP-Aufgaben erreicht GPT-3 mindestens das Niveau bisheriger state-of-the-art-Modelle, oftmals übertraf es diese. Es zeigte sich, dass größere Modelle mit ihrem großen Wissen mehr Aufgaben bedienen und sich vieles aus dem gesammelten Wissen erschließen können. Werden mit der Anfrage noch einige Beispiele eingegeben, wurden die Ergebnisse noch besser. Das Modell lernte also mit jedem Beispiel dazu. Dies ist dem Lernverhalten von Menschen schon sehr ähnlich.

Durch die neuartige Architektur und Größe, und die damit verbundenen neuen Möglichkeiten, ergeben sich für das Sprachmodell vielfältige Einsatzmöglichkeiten. Neben den klassischen NLP-Einsatzbereichen Chat und Question Answering, kann es für weitere Text-Anwendungen wie Klassifizierungen und Zusammenfassungen genauso eingesetzt werden wie für Ideengenerierung und die Einbindung in Programmierungssprachen und -kontexten.

So vielfältig die Einsatzmöglichkeiten sind, so viele Probleme belasten das Modell auch. Im Konkreten sind dies die Ausgabequalität, die Generierung und Verbreitung von Fake News, die Reproduktion von Vorurteilen und der Bias von Sprachmodellen, die Kontrolle der großen Modelle, rechtliche Fragen zu Datenschutz und Urheberschaft sowie die

Umweltbelastung und die Kosten zum Betreiben der Modelle. Diese Bereiche erfordern noch weitere Forschung und Entwicklung, bis ein faires Sprachmodell erzeugt werden kann. Das Modell ist nur über eine Schnittstelle zugänglich. Dabei gibt es nur begrenzte Änderungsmöglichkeiten und keine Einsicht in den zu Grunde liegenden Code. Die Nutzung ist kostenpflichtig und auch in Live-Anwendungen wird eine Gebühr für jede Nutzung fällig. Der Einsatz in Live-Anwendungen ist zudem stark reglementiert und wird von der Entwicklungsfirma OpenAI streng überwacht. Jede Veröffentlichung bedarf einer Erlaubnis. Die Anwendungen unterliegen dabei ausführlichen Usage Guidelines, die beachtet und umgesetzt werden müssen.

Anschließend wurden konkrete Anwendungsbeispiele von GPT-3 vorgestellt, bei denen es bereits im Live-Einsatz ist. Auch hier spiegelt sich die Vielfältigkeit der Einsatzbereiche wider. Die verschiedenen potentiellen Anwender:innen setzen sich mit dem Modell auseinander und prüfen dessen Einsatz.

In dem abschließenden Ausblick zur Zukunft wird deutlich, dass GPT-3 noch nicht die Spitze der Entwicklung großer Sprachmodelle ist. Durch seine neue Architektur ist es ein Durchbruch und wird die weitere Forschung beeinflussen. Doch schon jetzt wird an noch größeren Modellen mit noch besseren Ausgaben gearbeitet. Es wird sich zeigen, inwieweit diese die derzeitigen vorgestellten Probleme bereits lösen können, sowie sich weitere Einsatzbereiche erschließen lassen.

Dritter Teil:

GPT-3 zur Anwendung in der ZBW

## 12 Einsatz für den ZBW-Chatbot

In diesem Abschnitt wird GPT-3 einem Praxistest unterzogen. In einen Chatbot-Prototyp wird das Sprachmodell GPT-3 eingebunden. Dieser wird getestet und bewertet. In der anschließenden Auswertung wird evaluiert, wie gut sich das Modell für den Einsatz des ZBW-Chatbots eignet.

Dieses Kapitel beschreibt die Erstellung des Prototyps. Zu Beginn werden verschiedene Optionen ausgewertet, wie das Modell praktisch getestet werden kann. Anschließend wird dargestellt, wie die genaue Gestaltung vorgenommen wurde. Ziel dabei ist, die User Experience während der Nutzung des Chatbots zu untersuchen. Der Chatbot sollte sowohl hilfreich sein, indem richtige Antworten vermittelt werden, als auch einen flüssigen, nachvollziehbaren Gesprächsablauf aufbauen. Es wird untersucht, inwieweit dies mit GPT-3 möglich ist.

### 12.1 Evaluierung der verschiedenen Test-Szenarien

Es gibt verschiedene Möglichkeiten, GPT-3 praktisch zu testen. Im Folgenden werden drei Optionen vorgestellt, mit denen das Sprachmodell im Kontext eines Chatbots getestet und ausgewertet werden kann.

#### Option 1: Replikation der Tests der Entwickler:innen

Eine Option, das Sprachmodell auf die Einsatztauglichkeit in der ZBW zu testen, ist die Testart der Entwickler:innen aus der Veröffentlichung zur Vorstellung des Sprachmodells zu wiederholen. Dabei wurde die Fähigkeit von GPT-3 Fragen zu beantworten ebenfalls getestet und vorgestellt. Das Modell erhielt zwei bis drei Beispiele und konnte eine vierte Frage aus dem Kontext häufig richtig beantworten (vgl. Brown und andere, 2020, S. 6). Diese Option ist eine gute Möglichkeit die Ergebnisse der Entwickler:innen zu replizieren. Es zeigt zudem, wie schnell GPT-3 sich das ZBW-spezifische Wissen für den Chatbot aneignen kann. Nachteilig dabei ist, dass nur einzelne Sätze und Fragen konstruiert werden. Gerade bei einem Chatbot ist neben richtigen Antworten auch ein flüssiger, natürlicher Gesprächsverlauf für die User Experience wichtig. Dies lässt sich mit einzelnen Sätzen nicht bewerten.

### Option 2: Neu-Programmierung eines Chatbots

Um auch den Gesprächsverlauf eines GPT-3 Chatbots bewerten zu können, muss also ein Chatbot entwickelt werden. Mit Hilfe der GPT-3-API ist dies komfortabel möglich, erfordert jedoch trotzdem noch einen hohen Aufwand an Programmierung. In Anbetracht der Situation, dass parallel der Chatbot an der ZBW entwickelt und derzeit auch trainiert wird, ist es nicht nötig, diese Schritte erneut zu vollziehen. Die Entwicklung der ZBW wird speziell an das bestehende Ticket-System angepasst, in dem die Chats verwaltet werden. Es ist daher einfacher, GPT-3 in den bestehenden Chatbot der ZBW zu integrieren, wenn es sich als geeignet herausstellt. Eine Entwicklung im Rahmen dieser Arbeit würde live nicht nachgenutzt werden. Ein unnötig hoher Arbeitsaufwand soll daher vermieden werden. Stattdessen steht die Evaluation der Qualität im Vordergrund, um genau zu ergründen, inwiefern der Einsatz von GPT-3 Vorteile bringt.

### Option 3: Nutzung eines Frameworks

Für die Konstruktion von Chatbots gibt es eine Vielzahl an Frameworks, die bei der Arbeit unterstützen (vgl. Kohne und andere, 2020, S. 67ff.). Für GPT-3 gibt es bereits spezielle Frameworks, die das Sprachmodell eingebunden haben und verwenden. Diese gehen teilweise so weit, dass nur eine Wissensbasis hinterlegt werden muss, aus der der Chatbot das spezifische Wissen ziehen kann. Die technischen Gegebenheiten und eine einfache Benutzeroberfläche sind bereits gegeben. Eine Personalisierung auf eigene Bedürfnisse ist eingeschränkt möglich. Es bildet jedoch eine gute Option, wenn mit wenig Mitteln ein voll funktionstüchtiger Chatbot aufgesetzt werden soll.

Nach Abwägung der Vor- und Nachteile unter Beachtung des gesetzten Ziels, wird Option 3 umgesetzt und ein Chatbot-Prototyp mit Hilfe eines Frameworks aufgesetzt, in das GPT-3 eingebunden ist. Da eine Benutzeroberfläche und sämtliche technische Hintergrundstrukturen bereits in der ZBW in Entwicklung sind, müssen diese für einen GPT-3-Prototyp nicht erneut konstruiert werden. Diese Option lässt zu, mehr Zeit für die Evaluierung einzuplanen und viele Testdurchläufe mit dem schnell funktionsfähigen Chatbot durchzuführen.

## 12.2 Gestaltung eines Chatbot-Prototyps

Im Folgenden wird dargestellt, wie genau der Chatbot-Prototyp mit GPT-3 entwickelt wird. Es wird das gewählte Framework vorgestellt und aufgezeigt, welche technischen



Einstellungen vorgenommen wurden. Abschließend wird der Verlauf des Pre-Tests dargestellt, sowie die daraus resultierenden Anpassungen.

### 12.2.1 Das Framework Quickchat

Quickchat ist ein Framework, das die einfache Gestaltung von Gesprächs-KI ermöglicht. Es basiert auf GPT-3 und wird durch dessen Entwickler OpenAI unterstützt. Ein großer Vorteil ist die einfache Nutzung ohne Programmierung. Durch GPT-3 hat der Chatbot bereits ein grundsätzliches Wissen und Sprachverständnis. Es muss lediglich eine anwendungsspezifische Wissensbasis angelegt werden, auf die der Chatbot trainiert wird. Auf Grundlage dessen werden die Fragen im Chat beantwortet. Die Wissensbasis kann in der Browserbedienung eingefügt werden oder es kann eine Verbindung über die API zu internen Datenbanken bzw. FAQs hergestellt werden (vgl. Quickchat, 2022a).

Der Chatbot kann sowohl auf der eigenen Webseite eingebunden werden als auch mit Social-Media-Kanälen verbunden werden. Ebenfalls über die API kann der Chatbot in eigene Anwendungen integriert werden. Im Gespräch kann der Chat pausiert werden, um diesen an eine:n Mitarbeiter:in zu übergeben, wenn der Chatbot nicht die richtige Antwort findet (vgl. Quickchat, 2022a). Alle Gespräche des Chatbots werden im Nutzerprofil gespeichert und sind dort einsehbar. Dort können ebenfalls Chats live verfolgt werden (vgl. Quickchat, 2022b).

Bevor der Chatbot in den Live-Betrieb genommen werden kann, erfolgt eine Prüfung durch den Anbieter. Dies entspricht der bereits in Kapitel 7.2 vorgestellten Usage Guidelines. Damit wird überprüft, ob alle Regeln eingehalten wurden und der Chatbot im Sinne des Anbieters agiert (vgl. Quickchat, 2022b).

Die Kosten für den Betrieb des fertig gestellten Chatbots belaufen sich auf \$ 99 im Monat. Dies beinhaltet einen Chatbot, die unbegrenzte Möglichkeit die Wissensbasis anzupassen, 100 verfügbare Sprachen sowie 1000 Nachrichten pro Monat, die vom Chatbot gesendet werden können. Für Firmen, die einen größeren Bedarf haben, werden auf Anfrage Preise und Angebote zusammengestellt (vgl. Quickchat, 2022c).

### 12.2.2 Aufbau einer Wissensbasis

Der wichtigste Schritt für einen erfolgreichen Chatbot-Prototypen ist somit die Erstellung einer fundierten Wissensbasis. Diese enthält alle anwendungsspezifischen Informationen, mit denen der Chatbot die speziellen Fragen des jeweiligen Einsatzbereiches richtig beantworten kann. Im Gegensatz zum Fine-Tuning hat eine Wissensbasis den Vorteil, dass

sie leicht gepflegt und aktualisiert werden kann. Wenn sich eine Information ändert, müssen nicht alle Trainingsdaten aktualisiert bzw. neue gesammelt werden. In der Wissensbasis lässt sich die Veränderung schnell umsetzen.

Für das Framework Quickchat gibt es einige Vorgaben, die beim Aufbau der Wissensbasis zu beachten sind. Der Inhalt ist in Stichpunkte gegliedert. Jeder Stichpunkt kann bis zu 150 Wörter enthalten und sollte für sich selbst genommen verständlich sein. Es reicht aus, nur Antworten bzw. Informationen bereit zu stellen, es müssen keine vorgefertigten Frage-Antwort-Paare sein. Grundvoraussetzung wie bei jeder Wissensvermittlung sind zudem klare Formulierungen, eine korrekte Grammatik und keine Rechtschreibfehler (vgl. Quickchat, 2022d).

Außerdem kann das Kreativitätsniveau des Modells eingestellt werden. Wenn auf Grundlage der Wissensbasis keine Antwort gegeben werden kann, weil Informationen dazu fehlen, bestimmt das Kreativitätsniveau die Gestaltung der Ausgabe (vgl. Quickchat, 2022e):

- Niedrige Kreativität: Der Chatbot teilt mit, dass die Frage nicht beantwortet werden kann („Ich weiß es nicht.“).
- Normale Kreativität: Der Chatbot sucht eine Balance zwischen dem Improvisieren von Antworten und der Angabe, die Frage nicht beantworten zu können.
- Hohe Kreativität: Der Chatbot versucht auf Grundlage der vorliegenden Informationen eine Ausgabe zu generieren.
- Sehr hohe Kreativität: Die Ausgaben werden sehr kreativ.

Datengrundlage für die Erstellung der Wissensbasis sind die archivierten Chatverläufe der vergangenen 17 Jahre (2005-2022). Diese Daten werden genutzt, um den Chatbot der ZBW zu trainieren. Außerdem zeigen diese das gesamte Fragenspektrum, auf das der Chatbot im Einsatz reagieren können muss. Die Fragen lassen sich zu neun Themen clustern:

- Allgemeine Anfragen zur ZBW
- ZBW-Karte und -Konto
- Angebote vor Ort
- EconBiz
- Ausleihe
- Mahngebühren
- Bestand
- Recherche
- Fernleihe

Anhand dieser Aufarbeitung wurde die Wissensbasis für den Chatbot-Prototyp entwickelt. Diese soll zu den verschiedenen Themen und dessen Fragen Antworten liefern. Dadurch erhält der Chatbot das fachspezifische Wissen, um Fragen aus dem ZBW-Kontext richtig zu beantworten. Die gesamte Wissensbasis ist in Anhang A.2 nachzulesen.

### 12.2.3 Pre-Test

Um eine technisch einwandfreie Durchführung des Tests zu gewährleisten, wurde ein Pre-Test durchlaufen. Drei Personen, die nicht an dem Test teilnehmen, haben dafür Gespräche mit dem Chatbot geführt. Ziel dabei war es, den reibungslosen Ablauf zu testen und etwaige Probleme oder Schwierigkeiten in der Bedienung zu identifizieren. Dies wurde online mit dem Videokonferenztool Webex durchgeführt, um die reale Test-Situation nachzustellen und auch dort mögliche Probleme zu identifizieren.

Die technische Durchführung hinsichtlich Webex verlief wie erhofft problemlos und konnte damit wie geplant umgesetzt werden.

Auch der grundsätzliche Test-Ablauf, bestehend aus dem Chatten mit dem Bot und dem anschließenden Interview, erwies sich im Pre-Test als durchführbar. Die Bedienung des Chatbots war leicht zu erklären und die Teilnehmenden konnten diesen ohne Schwierigkeiten bedienen. Damit war eine wichtige Grundlage zur Durchführung gesichert. Inhaltlich wurden jedoch Schwierigkeiten deutlich. Die Qualität der Antworten im ersten Pre-Test war sehr schlecht. Es wurden kaum richtige Antworten gegeben. Daraufhin wurde die Wissensbasis überarbeitet. Diese konnte anscheinend nicht richtig vom Framework verarbeitet werden und erstellte falsche Verknüpfungen. Dadurch ergaben sich die massenhaft falschen Antworten.

In der Wissensbasis wurden vor allem Formulierungen angepasst. In verschiedenen kleineren Tests zeigte sich, dass die Antworten sich mit einer knapp formulierten Information in der Wissensbasis verbesserten. Dadurch wurde das System scheinbar nicht zu sehr abgelenkt und konnte den Inhalt besser erfassen. Zudem wurden längere Absätze in kleinere Stücke geteilt. Die Informationen wurden so besser strukturiert und unterschieden. Auch dadurch verbesserten sich die Antworten weiter.

In den weiteren Pre-Tests besserte sich die Antwortqualität nach den Anpassungen. Einige Antworten enthielten noch falsche Informationen. Jedoch hat sich der Anteil der richtigen Informationen deutlich gesteigert.

Hier zeigte sich bereits die erste Erkenntnis, dass die Wissensbasis sehr präzise und durchdacht aufgebaut werden muss. Zudem sollte diese ausführlich getestet werden, damit sichergestellt ist, dass die Informationen richtig verarbeitet und verknüpft werden.

### 12.3 Der Chatbot-Prototyp

Da der Chat der ZBW derzeit auf Deutsch und Englisch betrieben wird, muss auch der Chatbot mindestens diese beiden Sprachen beherrschen. Da der Chatbot im Quickchat-Framework nur eine Sprache zurzeit kann, wird dies im Test umgestellt und die Sprachen nacheinander getestet.

Die Benutzeroberfläche wurde nicht ausgestaltet. Die Oberfläche des ZBW-Chatbots wird derzeit unabhängig von dieser Ausarbeitung entwickelt. In diesem Test stehen die Qualität der Ausgaben und der Gesprächsverlauf im Vordergrund, weswegen die voreingestellte Oberfläche ausreichend ist.

Aufgrund der laufenden Kosten, die für die Erhaltung des Prototyps anfallen würden, wurde dieser lediglich für den Zeitraum des Tests betrieben. Im Anschluss wurde das Konto gelöscht und damit auch der Zugriff auf den Live-Chatbot. Alle Einstellungen des Chatbots sind in Anhang A.2 aufgeführt. Mit diesen kann der Chatbot in kurzer Zeit wieder aufgesetzt und genutzt werden.

## 13 Test des Prototyps

In diesem Kapitel wird der Ablauf zum Testen des Prototyps dargestellt. Nach der Schilderung zur Auswahl der Testpersonen, werden der Aufbau und die Durchführung erläutert.

### 13.1 Testpersonen

Für die Bewertung von Ausgaben eines Sprachmodells gibt es verschiedene Möglichkeiten. Für Übersetzungen lässt sich mit Hilfe von Algorithmen beispielsweise die Genauigkeit berechnen, indem die Ausgabe mit einem vorher festgelegten erwarteten Ergebnis abgeglichen wird. Die Ausgaben können jedoch auch von Menschen bewertet werden. Dies ist oftmals aufwendiger als die Berechnung mit Algorithmen, kann die Aussagen jedoch individueller bewerten (vgl. Rao und andere, 2019, S. 193). Da ein Gespräch mit einem Chatbot sehr individuell ist, wurde hier eine Evaluierung durch Menschen als Bewertung gewählt. Die Antworten können nicht mit vorgegebenen Antworten abgeglichen werden, da diese sich je nach Frageformulierung unterscheiden können, aber trotzdem inhaltlich richtig sind. Im Rahmen des Tests werden zudem Interviews mit den Teilnehmenden zur Bewertung des Sprachmodells durchgeführt. Daher wurden Testpersonen ausgewählt, die sowohl die Richtigkeit als auch die Qualität des Gesprächs bewerten können.

Für die Bewertung werden Kolleg:innen der ZBW herangezogen, die in den Benutzungsdiensten tätig sind, an dem Chatbot-Projekt beteiligt sind und mit dem Fachportal EconBiz arbeiten und dadurch eine Verbindung zum Chatbot haben. In einigen Fällen erfüllt eine Person mehrere der genannten Kriterien.

Dadurch werden potentiell 25 Kolleg:innen angesprochen. Dies lässt zwar keine repräsentative Bewertung des Sprachmodells zu. Jedoch liegt der Fokus auf der Qualität der Ausgaben in Bezug auf den Chatbot der ZBW. Dies können nur Beteiligte angemessen bewerten, die die richtigen Antworten kennen und dies einordnen können. Zudem haben alle Angesprochenen Erfahrungen im Chatten mit Nutzer:innen und können dadurch einschätzen, ob die Ausgaben des Sprachmodells die Anforderungen zum Einsatz erfüllen.

### 13.2 Testaufbau

Der Test wird online mit Hilfe des Videokonferenztools Webex durchgeführt. Dies ermöglicht allen Kolleg:innen eine bequeme Teilnahme ohne Reisen zum anderen Standort. Dadurch kann die Bereitschaft zur Teilnahme steigen. Durch die Funktion der

Bildschirmfreigabe haben sowohl die Teilnehmenden als auch die Moderatorin die Möglichkeit das Geschehen zu verfolgen und gegebenenfalls einzugreifen oder technische Hilfestellungen zu geben.

Zu Beginn einer Test-Session wird die Erlaubnis eingeholt, die Session aufzunehmen. Dies hat den Vorteil, dass die Moderatorin sich auf die Durchführung konzentrieren kann. Die Gespräche werden im Anschluss zur Auswertung transkribiert. Wenn die Erlaubnis zur Aufzeichnung nicht gegeben wird, erfolgt die Dokumentation durch eine Mitschrift mittels Notizen. Die Aufzeichnungen werden nach der Transkription aus datenschutzrechtlichen Gründen gelöscht. Die Transkriptionen bzw. Notizen werden anonym ausgewertet.

Der Test erfolgt mit jeder Person in zwei Durchgängen: zuerst auf Deutsch, dann auf Englisch. In jedem Durchgang starten alle Teilnehmenden mit einer identischen Frageformulierung, die vorgegeben wird. Dies ermöglicht mindestens zwei gezielte Vergleichsmöglichkeiten in den Sprachen. Im Anschluss können die Teilnehmenden frei Fragen stellen. Auch hier können sich weitere direkte Vergleichspunkte ergeben, wenn Teilnehmende zufällig eine gleiche Formulierung wählen. Das freie Fragen stellen ermöglicht jedoch zudem einen Vergleich, wie das Modell auf verschiedene Formulierungen des gleichen Inhalts reagiert.

Für die verschiedenen Teilnehmenden werden unterschiedliche Level der Kreativität gewählt. Ein Auskunft-Chatbot sollte sich möglichst an die Fakten halten und keine Antworten und Services neu erfinden. Um zu untersuchen, wie viel Spielraum dem Modell gegeben werden kann, wurden die Kreativitätseinstellungen „Niedrig“, „Normal“ und „Hoch“ gewählt. Um dies vergleichen zu können, führen zu jedem Level mehrere Teilnehmende Chat-Gespräche.

Im Anschluss an die Gespräche wird ein kleines Interview mit den Teilnehmenden geführt. Dadurch sollen ihre Erfahrungen während der Tests ausgewertet werden. Außerdem werden sie gebeten eine Einschätzung der Qualität und Richtigkeit der Antworten zu geben.

Die Interviewfragen lauten:

- Auf einer Skala von 1 bis 5 Sternen: Wie würdest du diese Erfahrung mit dem Chatbot bewerten?
- Was ist dir positiv aufgefallen?
- Was ist dir negativ aufgefallen?
- Wie bewertest du die Richtigkeit der Antworten?
- Wie würdest du den Gesprächsverlauf charakterisieren?

### 13.3 Durchführung

Durchgeführt wurden die Tests zwischen dem 04.07.2022 und dem 15.07.2022. Insgesamt haben dabei 13 Personen teilgenommen.

Mit jeder Person wurde ein Einzeltermin vereinbart, um die Bedienung des Chatbots zu erläutern. Diese dauerten zwischen 30 und 45 Minuten. Alle Teilnehmenden willigten in die Aufzeichnung des Tests ein, sodass aus jedem Durchlauf ein Transkript für die Auswertung erstellt werden konnte. Zusätzlich wurden die Chatverläufe gespeichert, um diese inhaltlich auszuwerten.

Während der Durchführung traten keine Probleme auf, sodass ein reibungsloser Ablauf stattfand. Die Dokumentation der Test-Durchläufe findet sich auf dem GitLab-Repository (siehe dazu Anhang A.3).

## 14 Ergebnisse des Prototyp-Test

Dieses Kapitel widmet sich der Auswertung des zuvor durchgeführten Tests des Prototyps. Dabei werden zunächst die Interviewfragen zur allgemeinen Bewertung des Chatbots und der geführten Gespräche aufgearbeitet. Im Anschluss folgt die inhaltliche Auswertung der Chatverläufe.

### 14.1 Auswertung des Fragebogens

Die durchgeführten Tests decken insgesamt gut die Bandbreite der möglichen Abfragen ab. Tabelle 4 zeigt die Verteilung der Gesprächsthemen nach Inhalt, Sprache und Kreativitätslevel. Auf Deutsch wurde der Themenbereich EconBiz nur einmal angesprochen. Auch der Bereich der Fernleihe im Englischen wurde nur wenig abgefragt. In diesen beiden Bereichen werden die Ergebnisse also nur bedingt aussagekräftig sein. Für eine bessere Einschätzung bedarf es hier noch weiterer Testgespräche. Für den Rahmen dieser Ausarbeitung wird zunächst mit den vorhandenen Ergebnissen gearbeitet. Diese ermöglichen bereits eine erste Einschätzung, wie sich das Sprachmodell verhält. Davon ist abhängig, ob das Sprachmodell überhaupt eingesetzt werden soll und ob sich weitere Untersuchungen und Tests lohnen.

Thema	Sprache	Kreativität		
		Niedrig	Normal	Hoch
Allgemeine Anfragen zur ZBW	Deutsch	3	4	2
ZBW-Karte und -Konto	Deutsch	4	5	2
Angebote vor Ort	Deutsch	2	2	1
EconBiz	Deutsch	0	1	0
Ausleihe	Deutsch	4	5	4
Mahngebühren	Deutsch	1	2	2
Bestand	Deutsch	2	4	2
Recherche	Deutsch	2	3	1
Fernleihe	Deutsch	1	3	0
Allgemeine Anfragen zur ZBW	Englisch	3	4	3
ZBW-Karte und -Konto	Englisch	3	2	4
Angebote vor Ort	Englisch	2	2	1
EconBiz	Englisch	0	2	1
Ausleihe	Englisch	2	2	2
Mahngebühren	Englisch	4	5	4
Bestand	Englisch	3	3	1
Recherche	Englisch	2	1	0
Fernleihe	Englisch	1	1	1

Tabelle 4: Verteilung der Anfragen



In den folgenden Abschnitten werden die einzelnen Interviewfragen zur Bewertung genauer ausgewertet und betrachtet.

#### 14.1.1 Sterne-Bewertung

In der ersten Frage wurden alle Teilnehmenden gebeten, eine Gesamteinschätzung des Chatbots anhand einer typischen Sternebewertung abzugeben. 1 Stern entspricht der schlechtesten Bewertung, 5 Sterne sind die höchste Bewertung. Abbildung 23 zeigt die Verteilung der Vergabe.

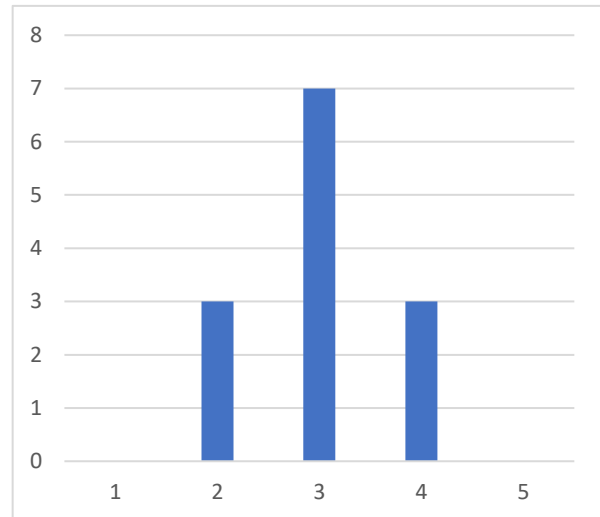


Abbildung 23: Gesamtbewertung der Chats

Sieben der Teilnehmenden vergaben 3 Sterne. Jeweils drei schätzten den Chatbot schlechter (2 Sterne) bzw. besser (4 Sterne)

ein. Im Durchschnitt ergibt sich damit eine Bewertung von 3 Sternen für den Chatbot. Er wird also nicht als besonders schlecht, aber auch nicht als herausragend gut wahrgenommen.

#### 14.1.2 Positive Erkenntnisse

In der zweiten Frage wurden die Teilnehmenden gebeten aufzuzählen, was ihnen positiv während des Tests aufgefallen ist.

Viele der Teilnehmenden haben es als positiv empfunden, dass die Antworten schnell kamen. Vor allem die langen Antworten hätte kein Mensch so schnell schreiben können, wie der Chatbot sie ausgegeben hat. Diese Beschreibung nannten acht der Teilnehmenden.

Ebenfalls acht Teilnehmende schätzen die Antworten als überwiegend richtig ein. Oftmals wurde dies mit Einschränkungen formuliert: „schon viel Richtiges dabei“ oder „meistens passend geantwortet“. Trotzdem stuften sie dies als positiv ein.

Positiv wurde auch der Stil des Chatbots beschrieben. Die Teilnehmenden empfanden das Gespräch als „angenehm“ und „höflich“. Der Chatbot habe keine unangenehme Attitüde gehabt, sondern zuvorkommend die Fragen beantwortet. Dass der Chatbot die Teilnehmenden im Durchgang auf Deutsch geduzt hat, fiel einigen Personen ebenfalls positiv auf. Die Formulierungen der Antworten wären zudem so, wie diese bereits jetzt im Chat der ZBW angewendet würden.

Ebenfalls viele Nennungen (sieben) beschrieben die Ausführlichkeit der Antworten. Dass die Antworten „umfangreich“, „ausführlich“ und „genau“ waren, wurde von den Teilnehmenden positiv aufgenommen. Sie hatten den Eindruck, dass die Antworten weiterhelfen und die Anfragenden eine passende Antwort zu ihrer Frage erhalten haben.

Von zwei Personen wurde positiv hervorgehoben, dass der Chatbot auf Kontaktdaten zum Serviceteam verweist, wenn er eine Antwort nicht weiß. Die Ausgabe von Links, z. B. zur Online-Anmeldung, wurde von einer Person ebenfalls positiv bewertet. Während eine Person sogar den Eindruck hatte mit einem Menschen zu schreiben, merkte eine andere Person an, dass es sich deutlich um eine Maschine handelt. Dies bewertete sie jedoch positiv, da kein falscher Eindruck und damit falsche Erwartungen aufgebaut wurden.

#### 14.1.3 Negative Erkenntnisse

Bei der dritten Frage ging es um negative Erkenntnisse, die während des Test-Durchlaufs entstanden sind.

Am häufigsten wurde dabei genannt, dass Antworten falsch, unvollständig oder verwirrend waren. Von den Teilnehmenden, die die teilweise richtigen Antworten als positiv einstufen, bewerteten viele auch die falschen Antworten negativ.

Ebenfalls oft aufgefallen ist den Teilnehmenden, dass der Chatbot einige Fragen anscheinend nicht richtig verstanden hatte. Mit dieser Schlussfolgerung erklärten sie sich die falschen Antworten. Zum anderen sprang das Gespräch dadurch auch in den Themen und man entfernte sich von der Ursprungsfrage.

Während einige Teilnehmende die langen und ausführlichen Antworten positiv anmerkten, fielen vier Personen diese als besonders negativ auf. Sie fühlten sich von den langen Textblöcken erschlagen und es wurden zu viele Informationen zusammengemischt. Deswegen wurde auch genannt, dass der Chatbot unnötige Informationen gibt. Oftmals gab er Informationen, ohne dass eine Frage gestellt wurde, oder fügte unwichtige Informationen an eine bereits richtige Antwort an.

#### 14.1.4 Richtigkeit der Antworten

Bei der vierten Frage wurden die Teilnehmenden gebeten eine Schätzung abzugeben, wie viel Prozent der gegebenen Antworten durch das Sprachmodell richtig waren. Da es den Rahmen dieser Arbeit sprengen würde, wird nicht im Detail darauf eingegangen, warum manche Antworten inhaltlich falsch sind und was die inhaltlich korrekte Antwort gewesen

wäre. In den dokumentierten Chatverläufen werden zur Transparenz der Auswertung die Antworten des Sprachmodells als richtig oder falsch markiert.

Tabelle 5 zeigt die Schätzungen mit den dazugehörigen Chats und Kreativitätslevel. Zur Kontrolle wurde die Anzahl der richtigen Antworten und ihr prozentualer Anteil bestimmt. Eine Interaktion umfasst die Frage des:r Teilnehmenden und die Reaktion des Chatbots. Eine richtige Reaktion liegt vor, wenn (a) das Sprachmodell die passende Antwort zur Frage ausgab oder (b) das Sprachmodell auf eine Frage, dessen Antwort es nicht weiß, mit dem Verweis auf die Kontaktmöglichkeiten zu Mitarbeitenden reagiert. Wenn das Sprachmodell mit dem Kontaktverweis reagiert, obwohl die richtige Antwort explizit in der Wissensbasis angegeben wurde, wurde dies als falsche Antwort gewertet.

Chat	Kreativität	Anzahl Interaktionen	Richtige Reaktionen	% richtig	Von TN richtig geschätzt %
1.1.	niedrig	16	5	31,25	40-50
1.2.	niedrig	9	6	66,66	40-50
2.1.	normal	24	12	50,00	60-70
2.2.	normal	11	6	54,54	60-70
3.1.	hoch	11	6	54,54	70
3.2.	hoch	15	8	53,33	70
4.1.	niedrig	11	5	45,45	80
4.2.	niedrig	20	9	45,00	80
5.1.	normal	9	6	66,66	70-75
5.2.	normal	9	2	22,20	30-40
6.1.	hoch	29	15	51,72	70
6.2.	hoch	17	11	64,71	70
7.1.	niedrig	24	13	54,17	50-60
7.2.	niedrig	10	6	60,00	50-60
8.1.	normal	35	20	57,14	60-70
8.2.	normal	30	20	66,67	60-70
9.1.	hoch	15	5	33,33	80
9.2.	hoch	6	5	83,33	80
10.1.	niedrig	16	10	62,50	20-30
10.2.	niedrig	56	22	39,29	20-30
11.1.	normal	21	9	42,86	60-70
11.2.	normal	15	9	60,00	60-70
12.1.	hoch	12	6	50,00	k. A.
12.2.	hoch	18	7	38,89	k. A.
13.1.	normal	23	10	43,48	80
13.2.	normal	27	17	62,96	80

Tabelle 5: Richtigkeit der Antworten

Zunächst fällt auf, dass nahezu alle Teilnehmenden den Anteil der richtigen Antworten höher eingeschätzt haben, als er tatsächlich ist. In ihrer Wahrnehmung verlief das Gespräch also informativer als es war. Dies könnte damit zusammenhängen, dass die Teilnehmenden die richtigen und zu erwartenden Antworten kennen. Wenn Antworten dann auch nur zur Hälfte richtig sind, könnten sie sich automatisch den Rest denken und die Antwort dadurch positiver wahrnehmen. Reale Nutzer:innen des Chats, die kein Hintergrundwissen zu den Abläufen in der Bibliothek haben, könnten eventuell durch einige Antworten verwirrt sein als Mitarbeitende.

Keiner der durchgeführten Chats konnte mehr als 83,33 % richtige Antworten erlangen. Dieser bildet zudem eine deutliche Ausnahme. Die anderen Gespräche haben einen Anteil richtiger Antworten, der deutlich darunter liegt. Der überwiegende Anteil der Gespräche hat einen Anteil von 50 bis 69 % an richtigen Antworten. In Abbildung 24 wird die Verteilung der Gespräche nach prozentualer Richtigkeit dargestellt.

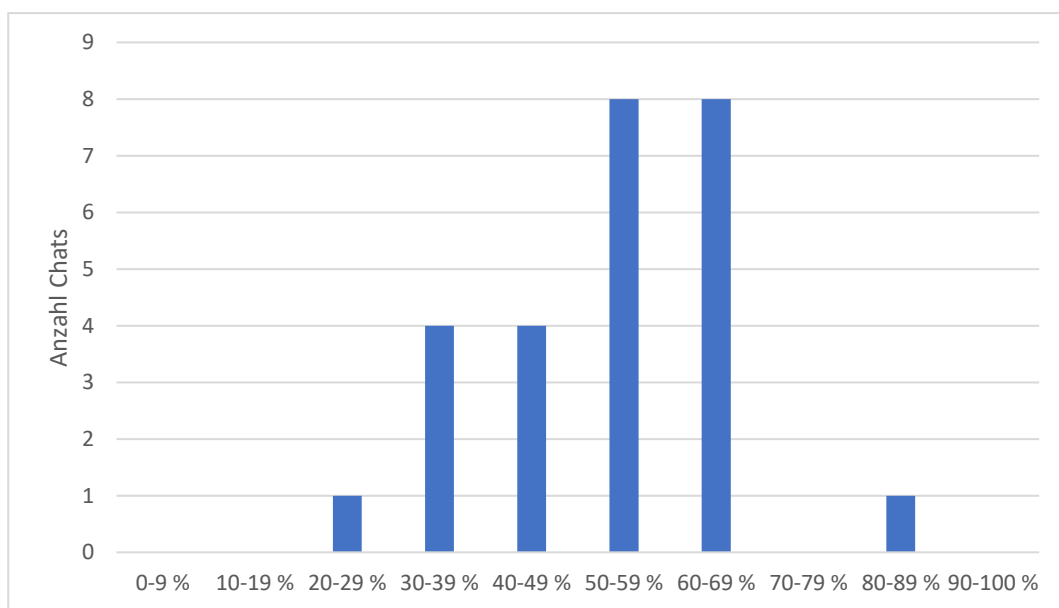


Abbildung 24: Anzahl richtiger Chats nach prozentualer Verteilung

Die Richtigkeit der Antworten unterscheidet sich bei der Sprache nur minimal. In allen deutschen Dialogen waren insgesamt 49,59 % der Antworten richtig. In den englischen Dialogen waren 52,67 % der Antworten richtig. Das Sprachmodell scheint also in unterschiedlichen Sprachen etwa gleich gut zu agieren.

Für die unterschiedlichen Kreativitätslevel lassen sich etwas größere Unterschiede feststellen. Auf der Stufe niedrig waren 46,91 % aller Antworten richtig, auf der Stufe normal 54,41 % und auf der Stufe hoch 51,22 %. Die besten Ergebnisse in diesem Vergleich wurden mit der normalen Einstellung der Kreativität erzielt. Dem Sprachmodell sollte also

etwas Spielraum gegeben werden, um Antworten zu formulieren. Zu viel Spielraum wirkt sich hingegen wieder negativ auf die Richtigkeit aus.

Insgesamt lässt sich jedoch festhalten, dass die Anzahl der richtigen Antworten deutlich zu gering ist, um das Sprachmodell direkt für einen Live-Einsatz zu übernehmen. Die Anzahl der richtigen Antworten sollte höher liegen. Wünschenswert ist ein Anteil von mindestens 90 % richtiger Antworten, um eine gute Servicequalität zu erhalten. Vor einem Live-Einsatz sollten also weitere Untersuchungen und Tests durchgeführt werden, um die Qualität der Antworten zu verbessern.

#### 14.1.5 Charakterisierung der Gespräche

In der letzten Frage wurden die Teilnehmenden gebeten, den erlebten Gesprächsverlauf zu beschreiben. Die Beschreibung teilt sich in positive und negative Assoziationen.

Positive Charakterisierungen waren beispielsweise flüssig, angenehm, gut, nett, lustig, freundlich, kompetent und zügig. Der überwiegende Teil der Beschreibungen war jedoch eher negativ. Dazu zählen Adjektive wie anstrengend, verwirrend, schwerfällig, frustrierend und zäh. Außerdem wurde hier erneut erwähnt, dass Antworten oftmals falsch, verwirrend, zu lang oder unvollständig waren. Eine Person hatte im Verlauf des Tests mehrmals das Bedürfnis, das Gespräch abbrechen zu wollen. Eine weitere Person hatte das Gefühl von dem Chatbot veralbert zu werden. Hier gab es erneut die Anmerkung von einigen Teilnehmenden, dass deutlich zu merken war, dass es sich bei dem Gesprächspartner nicht um einen Menschen handelt.

Die Beschreibungen unterscheiden sich stark in Abhängigkeit des geführten Gesprächs. Wenn es viele richtige Antworten gab, wurde das Gespräch positiver beschrieben. Bei vielen falschen Informationen wuchs die Frustration und die Bewertung fiel entsprechend schlechter aus.

#### 14.1.6 Weitere Anmerkungen

Am Ende des Tests gab es Raum für weitere Anmerkungen, die bei den vorigen Fragen noch nicht geäußert wurden. Einige Anmerkungen hier betrafen die Oberfläche des Chatbots und Einstellungen, die in dem Framework nicht geändert werden konnten. Es wurde sich z. B. ein größeres Chatfenster gewünscht, um lange Texte besser lesen zu können. Da die Oberfläche bei diesem Prototyp und dieser Ausarbeitung keine Beachtung findet, werden diese Bemerkungen nicht weiter untersucht. Sie werden jedoch an das Team in der ZBW

gegeben, das den Chatbot entwickelt, damit diese bei der Erstellung des realen Produkts Beachtung finden.

Außerdem reflektierten einige Teilnehmenden ihre Perspektive im Gegensatz zur Nutzendenperspektive. Während sie die richtigen Antworten kannten und teilweise damit spielten, eine gute Formulierung zu finden, um die korrekte Antwort zu erhalten, ist dies den Nutzer:innen nicht möglich. Dies ist den Teilnehmenden bewusst. Der Chatbot muss also mit unterschiedlichen Formulierungen gleich gut umgehen können und immer die richtige Antwort finden.

Weitere Anmerkungen betrafen seltsame Formulierungen in einigen Antworten. Dies wird genauer bei der inhaltlichen Analyse im folgenden Kapitel 14.2 untersucht.

## 14.2 Auswertung der Chatgespräche

In diesem Kapitel werden die durchgeführten Chatverläufe inhaltlich genauer untersucht. Dafür wurden verschiedene Aspekte identifiziert, die im Folgenden dargestellt werden. Zunächst wird die Reaktion auf die Ausgangsfragen verglichen. Außerdem wird der Umgang mit Rechtschreibfehlern und falschen Formulierungen bzw. Übersetzungen beleuchtet. Einen großen Abschnitt bildet die Untersuchung, ob das Sprachmodell Zusammenhänge erkennt. Weiterhin beleuchtet wird die Ausgabe von ungewünschten Informationen, der Umgang mit bekanntem und unbekanntem Wissen und Smalltalk. Abschließend wird die Ausgabe von größeren Textblöcken dargestellt, sowie der Umgang mit komplexen Themenbereichen. Alle Chatverläufe, auf die in diesen Abschnitten verwiesen wird, sind dokumentiert (siehe dazu Anhang A.3).

Die Gespräche der Teilnehmenden unterscheiden sich zum Teil stark im Stil. Dies stellt jedoch kein Problem dar, da im Live-Betrieb mit vielen unterschiedlichen Formulierungen, Erwartungen und auch Emotionen gerechnet werden muss. Daher spiegelt dieser Test in seinem kleinen Ausschnitt gut die Realität wider.

### 14.2.1 Reaktion auf die erste Frage

Um einen Vergleich der Antworten des Sprachmodells bei einer exakt gleich formulierten Frage zu ermöglichen, sind alle Teilnehmenden mit derselben Einstiegsfrage pro Sprache gestartet. Auf Deutsch lautete diese: „Wie kann ich Bücher ausleihen?“, auf Englisch: „What happens if I return a book late?“

Auf die deutsche Startfrage gab es insgesamt fünf verschiedene Reaktionen. Drei waren falsch, zwei richtig. Die falschen Antworten lauteten:

(1) Die Rückgabe von ausgeliehenen Büchern ist während der Öffnungszeiten an der Servicetheke möglich. Außerhalb der Öffnungszeiten können Bücher in Hamburg von Montag bis Freitag zwischen 6:30 Uhr und 19:30 Uhr in die Rückgabebox im Eingangsbereich eingeworfen werden. In Kiel ist ein Büchereinwurfkasten draußen neben dem Eingang rund um die Uhr zugänglich. Wenn ein Buch oder eine Zeitschrift bereits von einem anderen Nutzer ausgeliehen wurde, kann es reserviert werden. Es gibt keine Begrenzung, wie viele Medien gleichzeitig ausgeliehen werden können. Während der Öffnungszeiten der Bibliothek können Bücher ausgeliehen, die Arbeitsplätze im Lesesaal genutzt und Bibliotheksausweise ausgestellt werden. (Chats #1.1, #3.1, #4.1, #5.1, #6.1, #7.1, #8.1, #12.1)

(2) Du kannst Bücher am Serviceschalter ausleihen. (Chat #2.1)

(3) Du kannst dir Bücher in der Bibliothek ausleihen. (Chat #9.1, #11.1)

Antwort (1) reiht verschiedene Informationen aneinander, die in keinem Zusammenhang miteinander stehen und die Frage nicht beantworten. Die Informationen stammen zudem aus ganz unterschiedlichen Absätzen der Wissensbasis (Absätze 33, 28 und 3, vgl. Anhang A.2). Aus welchem Grund diese zusammengefügt wurden, ist nicht nachvollziehbar.

Antwort (2) ist falsch und es findet sich dazu auch keinerlei Hinweis in der Wissensbasis. Wie diese Antwort ausgegeben wurde, ist ebenfalls nicht nachvollziehbar. Das Kreativitätslevel war hier „normal“, sodass dies auch kein Ergebnis von besonders großem kreativem Spielraum des Sprachmodells ist.

Antwort (3) ist faktisch nicht falsch, beantwortet jedoch die zugrunde liegende Intention der Frage nicht. Es können in Bibliotheken natürlich Bücher ausgeliehen werden. Wenn die Frage jedoch so gestellt wird, erwarten Nutzer:innen die Erklärung zu den spezifischen Ausleihmodalitäten in der ZBW.

Die beiden richtigen Antworten wurden ausgegeben, als die Ursprungsfrage durch die Teilnehmenden etwas abgewandelt wurde. Sie schrieben „Wie kann ich Bücher ausleihen“ (Chat #10.1), wo das Fragezeichen am Satzende fehlt, und „Wie kann ich bei euch Bücher ausleihen“ (Chat #13.1). Als Antwort erhielten sie den Ablauf des Bestellprozesses von Büchern in der ZBW.

Ob eine richtige oder falsche Antwort gegeben wurde, kann nicht auf das Kreativitätslevel zurückgeführt werden. In diesem Fall scheint die Formulierung der Frage eine große Rolle zu spielen.

Im Englischen wurde die Startfrage immer richtig beantwortet. Es gab drei verschiedene Formulierungen, die jedoch inhaltlich alle die richtigen Informationen gegeben haben. Auf die gleiche Formulierung gab es zwei verschiedene Antworten:

(1) If books are returned late, reminder fees will be charged. This is 1 € per book in the first week, 2,50 € in the second week, 5 € in the third week and 15 € in the fourth week. After the fourth week the account will be blocked. Overdue fines can be paid in cash at the service counter or by EC card at the automat. From a total amount of 10 € the fees can be invoiced and paid by bank transfer. (Chats #1.2, #2.2, #3.2, #4.2, #6.2, #7.2, #8.2, #9.2, #10.2, #11.2, #12.2)

(2) The overdue fine will be charged. (Chat #13.2)

Während Antwort (1) ausführlichere Informationen liefert, ist Antwort (2) in einer verkürzten Version ebenfalls korrekt.

Bei Chat #5.2 wurde der Frage eine Begrüßung vorangestellt: „Hi.“. Dadurch veränderte sich die Antwort des Sprachmodells ebenfalls:

(3) The account will be blocked if a book has reached the fourth reminder level or if the account has more than 25 € in charges. If the account is blocked, no borrowing or renewal can take place. To unblock the account, the overdue book must be returned and the outstanding charges must be paid. Payment can be made in cash at the service counter or by EC card at the automat. If books are returned late, reminder fees will be charged. This is 1 € per book in the first week, 2,50 € in the second week, 5 € in the third week and 15 € in the fourth week. After the fourth week the account will be blocked. (Chat #5.2)

Für diese Antwort wurden die ähnlichen Absätze 7 und 35 der Wissensbasis ausgegeben (vgl. Anhang A.2). Auch hier lassen sich die unterschiedlichen Ausgaben nicht eindeutig auf ein unterschiedliches Kreativitätslevel zurückführen. Jedoch führt im Gegensatz zur deutschen Variante hier auch eine gleiche Formulierung zu unterschiedlichen Antworten. Eventuell lassen sich hier mit einer größeren Stichprobe bessere Zusammenhänge erkennen.

#### 14.2.2 Umgang mit Rechtschreib-/Tippfehlern

Wie in jeder Chat-Situation haben sich auch in diesen Durchläufen Rechtschreib- bzw. Tippfehler ergeben. Insgesamt ist das Modell damit gut umgegangen, indem das gemeinte Wort verstanden wurde. In diesen Fällen waren die Fehler jedoch eher klein: zwei vertauschte Buchstaben, ein falscher Buchstabe, fehlende Buchstaben. Auch mit einer konsequenten Kleinschreibung aller Worte kann das Modell gut umgehen und diese richtig



verstehen (vgl. Chat #2.1). Dieser Verlauf zeigt ebenfalls deutlich, dass das Modell Fragen auch als solche erkennt, wenn kein Fragezeichen am Ende des Satzes steht.

Beispiele:

- Can I borrow a journa? (Chat #3.2)
- Wh is an access from home not possible? (Chat #4.2)
- Ich habe meinen Nutzerausweis verloren, wie bekomme ich einen neuen? (Chat #7.1)
- wie lange kann ich die bücher ausleihen (Chat #2.1)

Auch hier ist eine weitere Untersuchung nötig, um zu klären, wie groß die Fehler sein dürfen und ab wann das Modell nicht mehr richtig reagiert. In diesem Test ist dem Sprachmodell der Umgang mit solchen Fehlern jedoch gut gelungen.

### 14.2.3 Falsche Formulierungen und Übersetzungen

Der deutlich überwiegende Teil der Antworten war verständlich formuliert und grammatikalisch richtig. In den englischen Antworten konnten keine fehlerhaften Formulierungen gefunden werden. Hier reagierte das Sprachmodell auch bei nicht ganz fehlerfrei formulierten Fragen auf Englisch gut. Dies zeigte sich z. B. im Chat #2.2. Sprachlich, im Sinne der Konstruktion von Sätzen, agiert das Sprachmodell sehr gut. Dass es auch mit fehlerhaften Fragen gut umgeht, spricht für den Einsatz in einem Chatbot. Hier ist mit Nutzer:innen zu rechnen, deren Muttersprache nicht immer Deutsch oder Englisch ist. Auch diese Anfragen sollten richtig verstanden und beantwortet werden.

Insgesamt hat auch die Interaktion auf Deutsch gut funktioniert. Hier sind jedoch einige Formulierungen aufgefallen, die nicht passend waren. Ein Beispiel stammt aus Chat #2.1:

Frage: kann ich bei ihnen scannen

Antwort: Du kannst mit uns scannen.

Die inhaltliche Aussage der Antwort ist korrekt, jedoch würde ein Mensch dies anders formulieren.

Weiterhin kam es zu Fehlern bei der Übersetzung. So wurde aus dem Studierendenausweis ein Schülerschein (Chat #13.1). Und auch mit dem deutschen Begriff „Volltext“ gab es in der Übersetzung Unstimmigkeiten. Das Sprachmodell wandelte dies in „vollständigen Text“ um (Chat #8.1).

Eine Ursache dafür könnte sein, dass die Wissensbasis für den deutschen Chatbot auch auf Englisch hinterlegt werden musste. Da sowohl Schüler als auch Student mit „student“ übersetzt wird, kann das Sprachmodell dies nicht auseinanderhalten. In den genannten Beispielen kann der Sinn der Antwort noch verstanden werden. Problematischer wird es,

wenn genaue Bezeichnungen (z. B. für Schaltflächen oder in Menüleisten) falsch übersetzt werden. Anfragende könnten dann Probleme haben den Anweisungen und Hilfestellungen des Chatbots zu folgen.

Für diesen Fall könnte untersucht werden, ob solche Probleme durch ein Fine-Tuning des Sprachmodells umgangen werden können. Wenn das Modell auf die spezifischen Begriffe in einer Sprache, besonders im deutschen, trainiert wird, besteht die Vermutung, dass diese dann auch korrekt angewendet werden.

#### 14.2.4 Zusammenhang zwischen aufeinanderfolgenden Fragen

Ein wichtiger Punkt bei einem Chatbot ist, dass dieser das gesamte Gespräch versteht, um für Nutzer:innen ein zufriedenstellendes Ereignis zu kreieren. Wie bereits in der Vorstellung des Sprachmodells im zweiten Teil dieser Arbeit deutlich wurde, soll GPT-3 so gut und lang Zusammenhänge herstellen, wie kein anderes Sprachmodell.

In einigen Fällen funktionierte dies in den Tests bereits gut. Dies soll an einigen Beispielen gezeigt werden.

Chat #1.1:

Frage: Ich kann ohne Karte schon Bücher bestellen?

Antwort: Ja, du kannst Bücher online bestellen.

Frage: Wo?

Antwort: <https://www.econbiz.de/>.

Auch wenn die erste Antwort inhaltlich falsch ist, stellt das Modell doch den korrekten Zusammenhang zu der Nachfrage „Wo?“ her. Es gibt den Link zum Fachportal aus, in dem die Bücher bestellt werden können.

Chat #2.1:

Antwort: Bitte benutze unseren Research Guide EconDesk.

Frage: was ist das

Antwort: Der Research Guide EconDesk bietet Unterstützung bei der Recherche von Wirtschaftsliteratur. Es werden Tipps für die Recherche gegeben, z.B. Suchbegriffe und -kombinationen sowie Datenbanken, in denen Informationen gefunden werden können.

Es werden keine Literaturlisten mit Empfehlungen bereitgestellt.

Auch hier kann das Sprachmodell die Nachfrage zuordnen und gibt weitere Informationen zum Research Guide aus.

Chat #8.1:

Frage: Ich suche Literatur zu einem bestimmten Thema.

Antwort: Was ist das Thema?

Frage: Marketing

Antwort: Der Research Guide EconDesk bietet Unterstützung bei der Recherche von Wirtschaftsliteratur. Es werden Tipps für die Recherche gegeben, z.B. Suchbegriffe und -kombinationen sowie Datenbanken, in denen Informationen gefunden werden können. Es werden keine Literaturlisten mit Empfehlungen bereitgestellt. Es liegt an den Anfragenden zu prüfen, ob ein Titel für ihre eigene Arbeit relevant ist. Persönliche Beratungen mit einem Fachexperten sind ebenfalls möglich.

Dieses Gespräch ähnelt schon sehr einem von einem/r Mitarbeiter:in geführtem Chat. Der Chatbot fragt nach, um eine genauere Beschreibung der Frage zu erhalten. Anschließend hält er den Kontext der Anfrage zur Recherchehilfe und gibt Informationen zu Kontaktmöglichkeiten.

Leider sind diese Beispiele eher die Ausnahmen. In vielen Fällen verliert sich das Sprachmodell in den Informationen und erhält den Kontext der ursprünglich gestellten Frage nicht. Auch dazu sollen einige Beispiele vorgestellt werden.

Chat #1.2:

Antwort: To send the books with a courier, proceed as follows: 1) Log in to EconBiz with your ZBW account. 2) Search for the book you want. 3) Click on "

Frage: Click on what?

Antwort: The location of the books can be viewed in EconBiz. Books and journals can be ordered regardless of the desired pick-up location. [...]

In diesem Fall wurde die erste Antwort abgeschnitten. Abgesehen davon, dass hier Inhalte vermischt wurden, liegen die Informationen in dem entsprechenden Absatz in der Wissensbasis vollständig vor (Anhang A.2, Absatz 20), wurden hier jedoch nicht ausgegeben. Auf die Nachfrage gibt es eine zusammenhangslose Information, die mit dem Ausgangsthema nichts zu tun hat.

Chat #3.1:

Frage: Ich suche Aufsätze zu Marketing.

Antwort: Welche Fachzeitschriften?

Frage: egal

Antwort: Wie lautet deine Benutzer-ID?

Frage: Ich habe keine ID

Antwort: Du bist nicht eingeloggt.

Während in dem Positiv-Beispiel oben die Nachfrage zu einer Recherche weitergeholfen hat, hat dies in diesem Gespräch nicht funktioniert. Das Sprachmodell stellt Rückfragen, die zu diesem Thema keinen Sinn ergeben und verliert dadurch den Faden. Hier hätte ebenfalls ein Verweis auf den Research Guide besser gepasst.

Chat #8.1:

Frage: habt ihr auch Datenbanken?

Antwort: Ja, wir haben Datenbanken.

Frage: Welche?

Antwort: Entschuldige bitte, ich habe nicht das nötige Wissen, um darauf zu antworten.

[...]

Auch in diesem recht einfachen Gesprächsverlauf kann das Sprachmodell den Zusammenhang zur Rückfrage nicht herstellen. Die Information, welche Datenbanken im Bestand sind, ist in der Wissensbasis vorhanden (Anhang A.2, Absatz 36). Da in den aufgezeigten Positiv-Beispielen die Rückfrage funktioniert hat, wäre hier zu erwarten, dass das Sprachmodell den hinterlegten Link zur Datenbankübersicht ausgibt.

Chat #10.2:

Frage: Do you have wifi or eduroam

Antwort: No, we do not have wifi or eduroam.

Frage: do you have wifi

Antwort: yes.

Frage: do you have eduroam

Antwort: no.

Dieser Gesprächsverlauf ist einer von mehreren die zeigen, dass das Sprachmodell Schwierigkeiten mit verketteten Fragen und mit Oder-Fragen hat. Der letzte Teil der Frage wird ignoriert und im besten Fall gibt es nur auf den ersten Teil eine korrekte Antwort. Dieses Beispiel zeigt, dass das Wissen durchaus vorhanden ist und richtig ausgegeben werden kann. In der Verknüpfung funktioniert es jedoch nicht.

Die großen Erfolge beim Erhalten des Kontexts, die mit GPT-3 beschrieben wurden, konnten hier nicht nachgebildet werden. Bei dem Schreiben ganzer Texte scheint es für das Sprachmodell einfacher zu sein, als auf spontane Eingabe von Nutzer:innen zu reagieren. Für einen reibungsarmen Einsatz im Chatbot ist dies jedoch eine wichtige Voraussetzung. In den archivierten Chats lässt sich häufig zurückverfolgen, dass es schreibfaule

Nutzer:innen gibt, die mit kurzen Rückfragen mehr Informationen erhalten möchten. Sie werden nicht jedes Mal eine neu ausformulierte Frage stellen. Die Erkennung solcher Zusammenhänge und die Verknüpfung mit der richtigen Information müssen demnach noch zuverlässiger funktionieren.

#### 14.2.5 Ausgabe ungewünschter Informationen

Im vorigen Abschnitt wurde bereits erwähnt, dass das Sprachmodell häufig ungewünschte und scheinbar willkürliche Informationen ausgibt. Dies kann unter Umständen sehr verwirrend für Nutzer:innen sein und die Akzeptanz des Chatbots negativ beeinflussen. Dies kommt häufig bei Äußerungen vor, die eine Reaktion auf die Antwort des Modells bilden, aber keine neue Frage darstellen. In menschlichen Gesprächen reagieren die Gesprächspartner darauf und binden dies in das Gespräch ein. Vor allem wenn diese Äußerungen Frust oder Ärger ausdrücken, muss dies in einer serviceorientierten Einrichtung aufgegriffen und abgemildert werden. Auf Äußerungen wie „Hm“ (Chat #5.1), „schade!“ (Chat #6.1) oder „:(“ (Chat #13.2) mit willkürlichen Informationen über Abläufe der Bibliothek zu reagieren, erfüllt dies nicht.

Ein Mensch würde zudem verstehen, dass im Chat #10.2 die Äußerung „okay, thanks“ das Gespräch beendet und die Fragen beantwortet sind. Die Ausgabe zusätzlicher Informationen, nach denen nicht gefragt wird, kann auch hier irritieren und das Gespräch negativ beenden. Ebenfalls vom Sprachmodell nicht verstanden wurden Eingaben wie „?“ (Chat #10.2), „Häh?“ (Chat #12.1) und „Das beantwortet meine Frage nicht so wirklich.“ (Chat #13.1). Ein Mensch versteht, dass die Antwort von der/dem Nutzer:in nicht verstanden wurde und formuliert eine neue Antwort. Hier zeigt sich jedoch auch das Problem, dass Zusammenhänge vom Sprachmodell nicht immer erkannt werden. Auf einige Eingaben werden Sprachmodelle auch erst passend reagieren können, wenn das Verständnis von Emotionen und die Reaktion darauf verbessert wurden. Bis dahin könnte das Erlebnis für Nutzer:innen eventuell ein wenig verbessert werden, wenn zumindest der Zusammenhang der Nachricht richtig verstanden wird.

#### 14.2.6 Umgang mit unbekanntem Wissen

Ein interessanter Punkt der Untersuchung ist, wie das Sprachmodell mit Fragen umgeht, zu denen es die Antwort nicht weiß. Auch hier wurden sehr unterschiedliche Ergebnisse erzielt, aus denen keine generelle Linie bei der Beantwortung herausgelesen werden kann.

Im Idealfall sollte das Sprachmodell bei Fragen, deren Antwort es nicht weiß, auf die Kontaktmöglichkeiten zu menschlichen Mitarbeitenden hinweisen. Dies wurde in der Wissensbasis hinterlegt (siehe Anhang A.2). In einigen Fällen hat dies gut funktioniert, wie einige Ausschnitte aus Chat #4.2 zeigen:

Frage: Okay and when i am ill?

Antwort: If you are ill, you can contact us via mail: info@zbw.eu or telephone: Kiel at 0431/8814-555 and in Hamburg at 040/42834-219.

Frage: Okay and then you cancelled my fees?

Antwort: Excuse me, I don't have that information. Please contact us via mail: info@zbw.eu or telephone: Kiel at 0431/8814-555 and in Hamburg at 040/42834-219.

In manchen Gesprächen wurden die Kontaktdaten erst nach der zweiten bzw. dritten Wiederholung der Frage herausgegeben (z. B. Chat #10.2). Dies entspricht den Handlungsempfehlungen zu Chatbots, dass nach spätestens drei fehlerhaften Fragen ein Mensch hinzugezogen wird.

Mit unbekanntem Abkürzungen konnte das Sprachmodell ebenfalls umgehen. Die Abkürzung ILL für Inter-Library Loan wurde in der Wissensbasis nicht hinterlegt, vom Modell aber in Chat #7.2 korrekt aufgeschlüsselt. Da GPT-3 mit riesigen Datenmengen trainiert wurde, ist es gut möglich, dass diese Abkürzung dort ebenfalls enthalten war. Somit gehört dies zum grundlegenden Weltwissen, mit dem das Modell ausgestattet wurde.

In einigen Fällen kann nicht nachvollzogen werden, woher das Modell die Antwort gezogen hat. Dazu gehören z. B. folgende:

Chat #6.2:

Frage: and what is about eating in the library? Can i eat an apple while i read the newspaper`?

Antwort: You can eat an apple while you read the newspaper.

Chat #7.1:

Frage: Gelten Bestimmungen des Urheberrechts?

Antwort: Es gelten die Bestimmungen des Urheberrechtsgesetzes.

Chat #11.2:

Frage: alright. Can I pay the fees via Paypal, credit card or Bitcoin?

Antwort: No.

Zu all diesen Fragen wurde die richtige Antwort nicht in der Wissensbasis hinterlegt. Im besten Falle hat das Modell aus dem hinterlegten Wissen die richtigen Schlüsse abgeleitet. Da als Bezahlungsmöglichkeiten explizit Bargeld und EC-Karte genannt werden (Anhang A.2,

Absatz 7), könnte es daraus geschlossen haben, dass die Bezahloptionen Paypal, Kreditkarte und Bitcoin nicht zu Verfügung stehen. Im schlimmsten Falle hat es geraten und zufällig in dem Moment die richtige Antwort getroffen.

Für die Beispiele #6.2 und #7.1 gilt dasselbe. Vielleicht ist es dem Modell gelungen, die korrekten Schlüsse zu ziehen. Im schlimmsten Fall hat es die Suggestion der Frage nur aufgegriffen. Auch in anderen Gesprächen fiel auf, dass sich das Modell von den Fragen leiten lässt und dadurch in manchen Fällen zu falschen Schlüssen kommt. Dies zeigt z. B. auch Chat #6.2, in dem WLAN falsch übersetzt wurde, das Sprachmodell jedoch darauf reagiert:

Frage: can i use wlan at the zbw?

Antwort: Yes, you can use the plan at the ZBW.

Frage: what plan can i use?

Antwort: Apologies, I don't have the knowledge necessary to respond to this. [...]

Suggestive Fragen von Nutzer:innen sollten immer korrekt beantwortet werden. Sie dürfen nicht nur das wiedergeben, was die Anfragenden sich als Antwort erhoffen.

In manchen Fällen, in denen das Sprachmodell die richtige Antwort nicht kannte, hat es sich schlicht etwas ausgedacht. Beispiele dafür sind diese:

Chat #1.2:

Frage: I want to send the books with DHL

Antwort: To send the books with a courier, proceed as follows: 1) Log in to EconBiz with your ZBW account. 2) Search for the book you want. 3) Click on "

Chat #2.1:

Frage: ab wann bekommt man denn eine mahnung

Antwort: 1 Woche vor dem Fälligkeitstermin.

Chat #13.2:

Frage: Ah, okay. Then I want to order the book "project management for beginners"

Antwort: The book is available in the ZBW.

Frage: Great. When can I pick it up?

Antwort: It is available for pickup now.

All diese Antworten sind falsch. Ein Zusammenhang mit der eingestellten Kreativität lässt sich nicht deutlich herstellen. Die dargestellten Beispiele decken alle Kreativitätslevel ab. Und auch die nicht hier abgebildeten Beispiele für falsche Antworten, stammen aus allen drei Leveln. Es lässt sich also nicht sagen, dass eine höhere oder niedrigere Kreativität zu falschen Antworten führt. Es ist weiterhin nicht ersichtlich, aus welchen Informationen der

Wissensbasis das Modell diese abgeleitet hat. In einigen Fällen scheint es sich die Antworten ausgedacht zu haben. Dies sollte nicht vorkommen, um Nutzer:innen nicht zu verwirren oder falsche Hoffnungen zu wecken.

#### 14.2.7 Falscher Umgang mit bekanntem Wissen

Auch der zum vorigen Kapitel gegenteilige Fall trat beim Test ein: In einigen Fällen reagierte das Modell nicht mit richtigen Antworten, auch wenn diese explizit in der Wissensbasis hinterlegt sind. Sehr offensichtlich wurde dies bei Fragen nach den Öffnungszeiten. In der Wissensbasis sind diese deutlich angegeben: „Library opening hours: Monday and Thursday 10 a.m. to 6 p.m. Tuesday, Wednesday and Friday 10 a.m. to 4:30 p.m. The library is closed on Saturdays, Sundays and holidays.“ (Anhang A.2, Absatz 2). Bei insgesamt 14 Fragen nach den Öffnungszeiten, wurden diese nur drei Mal richtig angegeben. Die häufigste Antwort war, dass die Bibliothek von 10 bis 18 Uhr geöffnet hätte. Unterschiedliche Formulierungen lösten verschiedene Antworten aus. Dies kann jedoch nicht die einzige Erklärung sein, da auch auf gleiche Formulierungen unterschiedlichen Antworten kamen (Chat #7.2 und #8.2, „What are the opening hours?“). Zudem sind für das Sprachmodell mit „Freitag“ und „freitags“ unterschiedliche Tage gemeint (siehe Chat #10.1). Für die beiden Fragen gab es unterschiedliche Antworten.

Dieses Problem fiel bereits während des Pre-Tests auf. Es wurden verschiedene Formulierungen in der Wissensbasis getestet, um die Angaben für das Modell verständlicher zu machen. Im Pre-Test erwies sich die in der Wissensbasis zu findende Formulierung als die Beste. Bei einer größeren Anzahl von Anfragen zeigte sich nun jedoch, dass dies auch nicht zuverlässig funktioniert.

Ein Themengebiet, bei dem ebenfalls viele falsche Antworten kamen, obwohl dies dem Modell bekannt sein sollte, ist die Ausleihe. Dies ist ein sehr komplexes Thema, das in der Wissensbasis auch mehrere Absätze benötigt (Anhang A.2, z. B. Absätze 6, 20, 21, 27, 30). In verschiedenen Gesprächen wurde deutlich, dass das Sprachmodell Schwierigkeiten hat, die Informationen richtig zu verknüpfen und entsprechend richtige Antworten zu geben. Folgende Auszüge zeigen dies beispielhaft:

Chat #4.1:

Frage: Wie lange kann ich die Bücher ausleihen?

Antwort: Ausgeliehene Bücher können über das Konto in EconBiz verlängert werden.

Frage: Und wie lange ist die Leihfrist?

Antwort: Die Leihfrist beträgt drei Wochen.



Chat #6.1:

Frage: wie lange kann ich ein Buch behalten?

Antwort: Du kannst ein Buch so lange behalten, wie du willst.

Frage: gibt es keine Leihfrist?

Antwort: Nein, es gibt keine Leihfrist.

Chat #8.2:

Frage: How can I borrow books?

Antwort: You can borrow books from the library.

Frage: How can I order the books?

Antwort: The books can be ordered from the library.

Frage: Where can I search for literature?

Antwort: You can search for literature in the library.

Chat #9.1:

Frage: Brauche ich dafür eine Mitgliedschaft?

Antwort: Nein, du brauchst keine Mitgliedschaft.

Frage: Ist die Ausleihe kostenlos?

Antwort: Ja, der Kredit ist kostenlos.

Chat #13.1:

Frage: Super. Habt ihr denn auch Zeitschriften?

Antwort: Nein, wir haben nur Bücher.

Frage: Bei meiner Unibibliothek können wir auch teilweise auf E-Books zugreifen. Habt ihr da auch welche?

Antwort: Tut mir leid, wir haben nur Bücher.

Frage: Schade. Also habt ihr auch keine Datenbanken?

Antwort: Nein, wir haben nur Bücher.

Auch hier könnten Untersuchungen vorgenommen werden, ob dies mit genug Daten und einem Fine-Tuning dem Modell besser vermittelt werden kann. Es ist jedoch insgesamt ein komplexes Themengebiet, das bei Menschen auch einiger Erklärung aufgrund der vielen Möglichkeiten bedarf.

#### 14.2.8 Smalltalk

Im Smalltalk schlägt sich das Sprachmodell recht gut. In Abschnitt 14.2.5 wurde bereits aufgezeigt, dass auf Verabschiedungen manchmal mit der Ausgabe von ungewünschten Informationen reagiert wurde. In den meisten Fällen wurde dies jedoch richtig verstanden.

Auch auf Dankes-Nachrichten reagierte das Modell richtig mit „You are welcome“ und „No Problem“. Für die deutschen Gespräche wurde dies wortwörtlich übersetzt mit „Du bist willkommen“ (was in Deutschen so nicht gesagt wird) bzw. „Kein Problem“ (Chats #6.1, #6.2, #9.2). Es konnte auch improvisieren und hat auf die Verabschiedung „Tschö mit ö!“ mit „Tschüss mit i!“ geantwortet. Auch Höflichkeitsformulierungen, wie Menschen sie benutzen würden, hat das Modell aufgegriffen (Chat #12.1):

Frage: Eine weitere Frage hätte ich noch.

Antwort: Was ist das?

Im Bereich Smalltalk agiert das Sprachmodell schon recht zuverlässig und vorhersagbar. Hier gab es keine großen negativen Überraschungen.

#### 14.2.9 Ausgabe von Textblöcken

Was viele Teilnehmende des Tests negativ angemerkt hatten, war die Länge der Antworten. Oftmals wurden lange Textblöcke mit vielen Informationen ausgegeben. Zum einen wirkte dies für die Teilnehmenden abschreckend, da sie sich mit sehr langen Erklärungen konfrontiert sahen. Zum anderen ist dies auch aus technischer Sicht eine interessante Ausgabe.

Als einer der großen Vorteile wird die Lernfähigkeit von GPT-3 betont, sowie die Fähigkeit kreative Texte mit wenig Vorgaben zu schreiben. Die Erwartung war daher, dass das Modell die Wissensbasis nutzt, um sich das Wissen anzueignen, und dieses dann im richtigen Kontext zu Antworten zu formulieren. In einigen Fällen gelang dies auch, wie die Chatgespräche zeigen. Die langen Texte waren jedoch oftmals eine exakte Wiedergabe der Wissensbasis. Dabei wurden mehrere Absätze der Wissensbasis zusammengefügt, um eine lange Antwort zu generieren. Die Informationen hängen dabei nicht zusammen und decken verschiedene Themenbereiche ab.

In Chat #2.1 wird auf die Frage, ob Bücher verlängert werden können, ein solch langer Textblock ausgegeben. Dafür wurden die Absätze 35 und 27 (vgl. Anhang A.2) der Wissensbasis aneinandergehängt. Die beinhalten jedoch Informationen zu Mahngebühren und der Bereitstellungszeit von Medien. Die Frage wird damit folglich nicht beantwortet. Zudem ist auch der Zusammenhang der Passagen nicht ersichtlich. Es ist nicht nachzuvollziehen, warum genau die beiden Absätze gewählt und verknüpft wurden.

Ein ähnliches Bild zeigt sich bei Chat #4.1. Hier werden auf die Frage, wie Bücher entliehen werden können, sogar vier Absätze der Wissensbasis verknüpft, Nr. 33, 29, 2 und 3

(vgl. Anhang A.2). Keiner beantwortet die Frage korrekt und der Zusammenhang ist für einen Menschen nicht ersichtlich.

Da Ähnliches bereits im Pre-Test aufgefallen ist, wurden einige Absätze umformuliert. Diese sind nun stichpunktartig und nicht mehr komplett ausformuliert. Doch auch aus diesen baut das Sprachmodell keine richtigen Sätze, sondern gibt sie lediglich wieder. Dies zeigte sich in Chat #9.2. Dort wird auf die Frage, wie die ZBW-Karte verlängert werden kann, die stichwortartige Antwort „Please contact our staff. User ID: is the eight-digit number on the ZBW card. ZBW card is valid for one year at a time. Can be renewed every year. For a renewal, the presentation of the identity card and, if applicable, a student ID is required“ geliefert (entspricht Absatz 6 der Wissensbasis, vgl. Anhang A.2).

Zum Teil mag dieser Umstand dem verwendeten Framework geschuldet sein. Dieses gibt keine andere Möglichkeit, als das Wissen über eine solche Wissensbasis in das Modell zu integrieren. Es bedarf hier also noch weiterer Untersuchungen. Auch hier scheint ein Versuch des Fines-Tunings sinnvoll. Wenn dem Modell verschiedene Formulierungen präsentiert werden, kann es auf Grundlage dessen eventuell besser eigene Ausgaben gestalten.

#### 14.2.10 Komplexe Themen

Wie in Kapitel 14.2.7 zum Themenbereich der Ausleihe bereits geschildert wurde, gibt es in der ZBW einige komplexe Services. Zum einen ist dies die Ausleihe und die dazugehörigen Modalitäten. Zum anderen ist es das sehr komplexe Thema des Zugriffs auf lizenzierte Volltexte. Dieses ist von Menschen schon schwer durchdringbar. Es gibt diverse Voraussetzungen, die abgeprüft werden müssen. Diese sind in verschiedenen Systemen und Softwares hinterlegt. Auf all diese wird ein Chatbot in absehbarer Zeit nie vollständigen Zugriff erhalten. Eine personalisierte Antwort auf jeden Einzelfall ist also nicht möglich. Oftmals ist dies jedoch noch erforderlich. Der Chatbot wird dieses Thema also nicht komplett eigenständig bearbeiten können, es bedarf hier immer noch der menschlichen Unterstützung. Zumutbar und realistisch ist jedoch, dass der Chatbot bereits einfache Voraussetzungen abfragt. In Chat #10.2 gibt es bereits einen guten Ansatz:

Frage: I can't open the pdf

Antwort: Are you logged in with your ZBW account?

Frage: I think so. I'm logged in.

Antwort: Please contact us by phone or mail.

Der Zugriff ist nur mit einem ZBW-Ausweis möglich. Dies wurde in der Wissensbasis hinterlegt (vgl. Anhang A.2, Absätze 24 und 25). In anderen Fällen verlief das Gespräch jedoch nicht so erfolgreich (z. B. Chat #10.2) und es verlor sich durch die Rückfragen des Chatbots, wie bereits dargestellt. Das Ziel in diesem Fall ist also, dass einfach überprüfbare Voraussetzungen vom Chatbot abgefragt werden. Entsprechend der Antworten der Nutzer:innen sollte der Bot dann Hinweise geben, wo Probleme liegen könnten. Wenn ein Sonderfall damit nicht gelöst werden kann, erfolgt die Übergabe an einen Menschen. Auf diesem Themengebiet muss das Sprachmodell daher noch mehr lernen, um zuverlässige Antworten zu generieren.

Ebenfalls schwierig zu vermitteln ist das Themengebiet der Recherchehilfe. Auch hier ist die KI noch nicht so weit, die komplette kognitive Leistung eines Menschen zu repräsentieren. Hier wird also auch in absehbarer Zukunft weiter ein Mensch eingebunden sein müssen, um qualitativ hochwertige Antworten auszugeben. Ähnlich wie bei dem Zugriff auf Volltexte sollte hier der Chatbot die Vorabfragen vornehmen. Auf Grundlage der Informationen erstellen Menschen dann Hinweise für die Recherche, wie sie derzeit beim Research Guide EconDesk üblich sind.

Durch \*semantische Suchen, und die Einbindung von KI und ML-Algorithmen entwickeln sich die Recherche und das \*Information Retrieval stetig weiter. Die zielgerichtete Verbesserung der Verarbeitung komplexer Daten kann mit Hilfe von intensiver Forschung dazu führen, dass der Recherche-Chatbot richtige, sinnvolle und personalisierte Recherchetipps liefert.

#### 14.2.11 Verschiedenes

Zuletzt sollen noch einige kleinere Auffälligkeiten der Test-Durchläufe betrachtet werden. Wenn nach der Online-Anmeldung gefragt wurde, wurde in jeder Antwort der Link zum Formular abgeschnitten. Dieser ist vollständig in der Wissensbasis hinterlegt (vgl. Anhang A.2, Absatz 5). In den Chats #2.2, #5.1, #6.1 und #7.1 wurde immer das Ende abgeschnitten. Der Link ist trotzdem noch funktionsfähig und führt auf den richtigen Bereich der Webseite, es erfordert jedoch noch einen zusätzlichen Klick, um auf das Anmeldeformular zu kommen. Da der Link vollständig hinterlegt ist, ist nicht ersichtlich warum dieser nur abgeschnitten ausgegeben wurde. Weiterhin funktionierten alle angegebenen Links nur unregelmäßig. Dies ließ sich nicht auf einen Browser oder einen Link zurückverfolgen. Ein Link, der in einem Test noch die korrekte Seite aufgerufen hat, erzeugte eine Stunde später bei einem

weiteren Test eine Fehlermeldung. Hier konnte keine Ursache gefunden werden. Es ist weiter zu untersuchen, ob dies am Framework oder dem Sprachmodell liegt.

Ähnliches wurde bei Antworten beobachtet, wie bereits in Kapitel 14.2.4 erwähnt. Dort wurde eine Antwort mitten im Satz abgeschnitten, deren Inhalt vollständig in der Wissensbasis hinterlegt ist. In anderen Gesprächen wurde diese auch vollständig ausgegeben.

Eine Frage, die ebenfalls häufig im Umgang mit einem Chatbot gestellt werden könnte, ist die Frage, ob man mit ebenjenem Kontakt hat. Es entspricht einer guten UX-Gestaltung, wenn direkt klar ersichtlich ist, ob mit einem Menschen oder einer Maschine kommuniziert wird. Das Sprachmodell hat auf Nachfragen ehrlich und offen reagiert, wie diese Beispiele zeigen:

Chat #7.2:

Frage: Do I speak to a bot?

Antwort: You are speaking to a chatbot.

Chat #13.1:

Frage: Spreche ich eigentlich gerade mit einem Mitarbeiter oder ist das ein Programm?

Antwort: Dies ist ein Programm.

In Chat #12.1 lässt sich die Richtigkeit der Antwort je nach Blickwinkel unterschiedlich bewerten:

Frage: Bist du echt?

Antwort: Ja.

Der Chatbot ist natürlich in dem Sinne „echt“, als dass er wahrhaftig existiert und in diesem Gespräch live agiert. Die Frage zielt jedoch eher auf ein „echtes Lebewesen“ ab. Dies hätte also verneint werden müssen. Hier zeigt sich wieder die Problemstellung, dass Maschinen die Intention der Frage richtig verstehen und den Kontext richtig einordnen müssen.

## 15 Zusammenfassung

Durch diesen Test wurden genügend Daten gesammelt, um eine erste Einschätzung über die Einsatzmöglichkeit des Sprachmodells in einem Bibliotheks-Chatbot zu geben. Für eine repräsentative Abbildung der Fähigkeiten bedarf es mehr Daten.

Insgesamt gaben die Teilnehmenden dem Chatbot eine durchschnittliche Bewertung. Die Einschätzung ist logischerweise stark von den geführten Gesprächen abhängig. Jedoch waren alle dem Chatbot gegenüber aufgeschlossen, fanden sowohl positive als auch negative Aspekte, die in konstruktiver Kritik geäußert wurden. Trotz vieler falscher Antworten wurde die Gesamterfahrung recht gut bewertet, da die Gespräche insgesamt in einem angenehmen Ton geführt wurden.

Es lässt sich jedoch festhalten, dass die Quote von richtigen Antworten zu niedrig ist, um das Modell so direkt in den Live-Betrieb zu nehmen. Diese sollte deutlich höher liegen, um qualitativ hochwertige und richtige Antworten zu sichern. Eine Verschlechterung im Vergleich zum derzeitigen Service-Angebot durch fehlerhafte Angaben sollte nicht passieren.

Die Tests wurden mit unterschiedlichen Einstellungen durchgeführt. Es lässt sich jedoch nicht sagen, dass eine davon besonders gut oder schlecht funktioniert hat. Die Quoten und Bewertungen waren durchweg durchwachsen, unabhängig vom Kreativitätslevel. Auf jedem Level gab es Gespräche, die gut funktioniert haben und richtige Antworten lieferten. Im nächsten konnte es jedoch schon wieder überwiegend falsche Antworten geben.

Inhaltlich lässt sich festhalten, dass das Modell auf rein sprachlicher Ebene gut funktioniert. Es hat fast durchgängig fehlerfreie und verständliche Sätze formuliert. Es reagiert jedoch sehr sensibel auf unterschiedliche Formulierungen und Synonyme. Zudem hat es an vielen Stellen Probleme den Kontext zu halten bzw. herzustellen. Hier muss auch eine größere Zuverlässigkeit erlangt werden, damit durchgängig gute Antworten geliefert werden können. In vielen Fällen wirkten die Antworten willkürlich und geraten. Ein Live-Einsatz muss mit einer konsequenten Beantwortung erfolgen, damit alle Nutzer:innen dieselben korrekten Informationen erhalten.

Weiterhin auffällig war, dass sich das Modell sehr an die hinterlegte Wissensbasis gehalten hat. Faktisch ist dies durchaus korrekt. Die Erwartung war hier jedoch eine freiere Formulierung von Antworten auf Basis der Fakten. Dies wird als großer Fortschritt von GPT-3 dargestellt und konnte hier jedoch, wie in den Auswertungen in Kapitel 14.2.7 erörtert, nicht nachgebildet werden.

Einige Themenbereiche sind zu komplex, um sie einem Sprachmodell verständlich zu vermitteln. Diese erfordern den Zugriff auf mehrere Systeme aus der täglichen Bibliotheksarbeit und das Prüfen von Voraussetzungen. Dies ist derzeit nicht möglich. Es wird in absehbarer Zukunft also weiterhin Themenbereiche geben, bei denen ein Mensch bei der Antwort unterstützen muss. Außerdem gibt es viele ähnliche Themen, wie z. B. die Fernleihe und die Dokumentlieferung. Hier bedarf es vieler dezidierter Beispiele, damit dem Sprachmodell die Unterschiede deutlich gemacht werden können.

In dem Zusammenhang wurde auch festgestellt, dass die Formulierung und Strukturierung der Wissensbasis große Auswirkungen auf die Antworten hat. Diese wurde in der Vorbereitung mehrmals überarbeitet, bis sie in ihrer jetzigen Form integriert wurde. Die Tests haben gezeigt, dass sie noch immer nicht optimal von GPT-3 erfasst wird und daher missverständliche Antworten ausgegeben werden können. Einige der hier gemachten Beobachtungen sind dem Framework zuzuordnen. Diese Untersuchung zeigt erste Ansatzpunkte, die durch weitere Versuche ausgebaut werden müssen.

Vielversprechende Ansätze konnten herausgearbeitet werden. Das Modell ist lernfähig und es gibt verschiedene Möglichkeiten, dies in einen Chatbot zu integrieren. Es könnte auch eine Bereicherung des Service-Angebots darstellen, da mit einem solchen Modell der Chatbot auf verschiedenen Sprachen angeboten werden kann. Zudem kann es flexibler und spezifischer auf Anfragen reagieren als ältere Chatbot-Frameworks, in denen konkrete Antworten vorgegeben werden.

Schlussendlich zeigen sich auch bei diesem Modell und Test, dass Maschinen in einer Black Box lernen: Es werden Daten eingegeben und die Maschine erzeugt eine Ausgabe; es ist jedoch nicht ersichtlich, welche Entscheidungsprozesse zu dieser Ausgabe führten. In vielen Fällen erschloss sich kein Zusammenhang aus den Ein- und Ausgaben, da keine Entscheidungswege aufgezeigt werden. Durch ausführliches Ausprobieren und Testen kann versucht werden, die fehlerhafte Stelle in den Daten zu finden, die zu falschen Aussagen führt. Dies bedarf jedoch viel Zeit und der guten Aufbereitung einer großen Menge Trainingsdaten. Trotz allem bleibt bei dieser Art von lernendem Sprachmodell zum jetzigen Zeitpunkt immer eine gewisse Unsicherheit. Da Entscheidungswege nicht nachvollzogen werden können und jedes Mal anders ablaufen können, ist es schwer eine richtige Ausgabe zu sichern.

Vierter Teil:

Abschlussbetrachtungen



## 16 Fazit

Nach der ausführlichen Darstellung und Untersuchung des Sprachmodells GPT-3 sollen die Ergebnisse der vorigen Kapitel zusammengefasst werden. Dies wird anhand der zu Beginn vorgestellten Forschungsfragen strukturiert, die somit hier abschließend für diese Ausarbeitung beantwortet werden.

### **(F1) Was unterscheidet das Sprachmodell GPT-3 von anderen Modellen, dass es als so vielversprechend und mächtig gehandelt wird?**

Zum Zeitpunkt seiner Veröffentlichung war GPT-3 das bisher größte Sprachmodell. Die Anzahl der Parameter und die Menge der Trainingsdaten überstiegen die aller anderen Modelle. Dadurch konnte das Modell auch andere Fähigkeiten erlernen und bessere Ausgaben erzielen. Neben den klassischen Einsatzgebieten wie NLP und der Textverarbeitung, eröffneten sich so neue Einsatzgebiete. Dies betrifft z. B. die Programmierung. Durch diese vielfältigen Einsatzbereiche werden verschiedenste Fachbereiche angesprochen, in denen das Modell eingesetzt werden kann. Dies spricht zudem eine größere Zahl möglicher Anwender:innen an, wodurch das Sprachmodell insgesamt an Bedeutung gewinnt.

### **(F2) Welche Funktionalitäten bietet das Sprachmodell GPT-3 und wie können diese eingesetzt werden?**

Die Funktionalitäten wurden ausführlich in Kapitel 8 geschildert und mit Beispielen ergänzt. Grundsätzlich lässt sich sagen: Überall, wo mit Text und in schriftlicher Form gearbeitet wird, ist GPT-3 potentiell einsetzbar. Mit GPT-3 hat sich dieser Bereich noch vergrößert, im Vergleich zu dem was vorige Sprachmodelle leisten konnten. Zu den Funktionalitäten gehören klassische Textverarbeitungsaufgaben wie das Beantworten von Fragen in verschiedensten Kontexten und Formen, die Erzeugung von Texten und die der Klassifizierung von Daten. Darüber hinaus kann GPT-3 auch kreative Aufgaben, wie die Generierung neuer Ideen, übernehmen. Die größte Neuerung für Sprachmodelle ist jedoch der Einsatz in der Programmierung. Kapitel 9 zeigte einige ausgewählte Funktionalitäten im Praxis-Test genauer.

**(F3) Wo wird das Sprachmodell GPT-3 bereits eingesetzt und welche Erfahrungen wurden dort gemacht?**

Die Einsatzgebiete sind so vielfältig wie die möglichen Anwendungsbereiche. Dies reicht von Chat über Textvervollständigung, Klassifikation und Zusammenfassungen bis hin zur Programmierung. Im wirklichen Live-Betrieb ist GPT-3 bisher noch eher selten. Über diese Anwendungen lassen sich auch nur wenige Erfahrungsberichte finden. In vielen Bereichen wird das Modell jedoch auf seine Fähigkeiten getestet. Bevor der Einsatz erfolgt, werden zunächst die verschiedenen Funktionalitäten ausführlich untersucht. Die geschilderten Probleme mit großen Sprachmodellen – wie z. B. die Qualität der Trainingsdaten, die Kontrolle der Modelle, der Datenschutz – erschweren den Live-Betrieb zudem. Gerade in Deutschland gibt es strenge Vorgaben für den Umgang mit sensiblen Daten. Dazu müssen zunächst rechtliche Fragen geklärt werden, bevor ein breiter Einsatz möglich ist. Außerdem wird der Einsatz stark vom Anbieter überwacht und nur geprüfte Projekte dürfen veröffentlicht werden.

**(F4) Eignet sich das Sprachmodell GPT-3 bereits für den Einsatz des an der ZBW geplanten Chatbots? Wenn nein, was fehlt für den Einsatz?**

Unter den untersuchten Gegebenheiten eignet sich GPT-3 derzeit noch nicht für einen Live-Einsatz im Chatbot der ZBW. Die Korrektheit und die Zuverlässigkeit der Antworten liegen noch unter den Qualitätsanforderungen. Es bleibt daher zu prüfen, wie sich das Sprachmodell in anderen Umgebungen verhält. Einige der bemängelten Punkte könnten Framework-spezifisch sein und durch andere Umgebungen anders gelöst werden. Da das Modell lernfähig ist, könnte es mit weiteren Untersuchungen angepasst werden, um den Anteil richtiger Antworten zu steigern. Im abschließenden Ausblick (Kapitel 18) werden Optionen dazu näher erläutert.

Weiterhin zu beachten sind die vorgestellten Probleme von solch großen Sprachmodellen wie GPT-3 (siehe Kapitel 6.3). Das Problem mit der Qualität der Antworten wurde bereits herausgearbeitet. Weiterhin relevant für einen Chatbot der ZBW sind die Kontrolle über das Modell, Datenschutz, Limitationen und Kosten. Vor allem in Bezug auf den Datenschutz würde es Herausforderungen im Einsatz geben. Ein DSGVO-konformer Einsatz ist nicht möglich, da die Daten bei dem Entwickler in den USA gespeichert werden. Im Bibliothekskontext geht auch schnell um persönliche Daten wie Name, Adresse, Geburtsdatum und Kontaktmöglichkeiten per Telefon oder Mail. Bei diesen sollte gesichert sein, dass sie nicht weiterverwendet werden. Da alle gespeicherten Daten in die

Weiterentwicklung von GPT-3 genutzt werden, kann die Weiterverwendung nicht ausgeschlossen werden.

Weiterhin hat sich gezeigt, dass die Ausgabe nicht kontrolliert werden kann und es daher schwer ist, einen gesicherten Service-Anspruch zu erfüllen. Außerdem stößt das Modell bei komplexen Themen an Grenzen, die die Verwendung von weiteren Systemen erfordern. Nicht zuletzt wird das Modell hohe Kosten generieren, wenn es durchgängig im Einsatz ist. Dies muss bei den Überlegungen für einen Einsatz einkalkuliert werden.

**(F5) Welche Voraussetzungen und Handlungsschritte sind nötig, um GPT-3 bei Chatbots einzusetzen?**

Auch wenn GPT-3 lernfähig ist, kann es nicht in seiner ursprünglichen Form direkt in einen Chatbot eingebaut werden. Es bedarf einiger Vorbereitungen und Entwicklungsarbeit. Es müssen Trainingsdaten aufbereitet werden und grundsätzliche technische Anforderungen geklärt werden. Eine genauere Erläuterung der Handlungsschritte erfolgt im folgenden Kapitel 17, in dem Handlungsempfehlungen für den Einsatz von GPT-3 bei einem Chatbot vorgestellt werden.

## 17 Handlungsempfehlungen zum Einsatz von GPT-3 bei Chatbots

Dieses Kapitel dient als Hilfestellung für Bibliotheken und andere Einrichtungen, die GPT-3 in ihren Chatbot integrieren möchten. Die Empfehlungen beziehen sich speziell auf das Sprachmodell und dessen Einsatz. Was generell bei der Planung von Chatbots zu beachten ist, wird in vielen praktischen Handbüchern aufgezeigt, die über die Entwicklungsschritte ausführlich informieren (z. B. Kohne und andere, 2020; sowie Kabel, 2020).

Für den Einsatz von GPT-3 sind folgende Bereiche zu beachten:

### **(1) Trainingsdaten sammeln**

Auch wenn GPT-3 mit wenigen Daten auskommt, werden für jeden Fall Beispiele und aufbereitete Daten benötigt, mit denen das Modell lernen kann. Bei einer Wissensbasis dienen diese dazu, einen Überblick über die gestellten Fragen zu erhalten. Im Falle eines Finetuning wird dies umso korrekter, je mehr Beispieldaten dem Modell zur Verfügung stehen.

### **(2) Genaue Aufbereitung der Testdaten und ausführliche Tests**

Dieser Schritt ist ausschlaggebend für die Ausgaben und Ergebnisse des Sprachmodells. Er sollte daher sorgfältig geplant und durchgeführt werden. Dies erfordert Personal und Zeit. Doch je besser die Daten aufbereitet und vom Modell verarbeitet werden können, desto besser sind die Ergebnisse. Dies sollte durch viele Testdurchläufe mit verschiedensten Personen sichergestellt werden.

### **(3) Frameworks für Chatbot evaluieren**

Einen Chatbot komplett allein aufzusetzen bedarf viel Zeit und ausgebauter Fachkenntnisse. Zur Unterstützung gibt es viele Frameworks, um Chatbots zu bauen. Diese sollten ausführlich geprüft werden, ob sie sich eignen. Dazu müssen die eigenen Anforderungen formuliert werden und mit den Voraussetzungen der Institution abgeglichen werden. Vorgaben zu Finanzen, Datenschutz und technische Voraussetzungen können die Auswahl einschränken. Zudem ist GPT-3 ein recht neues Sprachmodell, das noch nicht in allen Frameworks verfügbar ist.

#### **(4) Einsatz mit Usage Guideline abstimmen**

Die Usage Guidelines des GPT-3 Anbieters OpenAI sind, wie geschildert, sehr streng. Diese sind jedoch verpflichtend einzuhalten, damit der Zugang zum Modell und die Nutzung in der Anwendung nicht gesperrt werden. Es ist also zu prüfen, dass der Chatbot den Bestimmungen entspricht. Zudem verlangen die Guidelines eine Abnahme durch den Anbieter. Dafür ist im Entwicklungsprozess Zeit einzuplanen.

#### **(5) Konsultation mit Datenschutzbeauftragten und der Rechtsabteilung**

Einhergehend mit den Vorgaben des Anbieters, wird es auch Vorgaben zum Datenschutz und der Verarbeitung personenbezogener Daten in der eigenen Institution geben. Das geplante Projekt und die gewählten Systeme sollten daher mit den Verantwortlichen abgestimmt und geprüft werden.

#### **(6) Kosten kalkulieren**

Nicht zu vernachlässigen bei der Nutzung von GPT-3 sind die Kosten. Die Nutzung des Modells ist, wie dargestellt, mit unterschiedlichen Gebühren verbunden. Wenn der Chatbot ausgiebig genutzt wird, müssen diese Kosten in das Jahresbudget einkalkuliert werden.

Die Höhe der Kosten soll mit Hilfe einer Beispielrechnung hier noch einmal verdeutlicht werden. Dafür wurden die Chatanfragen aus der ZBW für Juni 2022 aufbereitet. Diese enthielten etwa 12.600 Wörter, die durch die Bibliotheksmitarbeitenden verschickt wurden. Dies entspricht dem Output des Sprachmodells in einem Live-Einsatz. Die Kosten werden pro Token berechnet. Da einige Wörter in mehrere Tokens aufgeteilt werden, wurde die Summe aufgerundet. In diesem Beispiel wird von 15.000 Token pro Monat ausgegangen, die das Sprachmodell ausgibt.

Nutzt man dafür das Modell Davinci mit \$0,02 pro Token, belaufen sich die monatlichen Kosten derzeit auf \$300. Mit dem zweitbesten Modell Curie (\$0,002 pro Token) sind dies immerhin noch \$30 pro Monat. Gerade bei der Nutzung von Davinci müssen die enormen Kosten in das Jahresbudget einkalkuliert werden. Auch die Nutzung der kleineren Modelle fällt ins Gewicht, gerade wenn sich die Anzahl der Anfragen durch die 24/7-Verfügbarkeit steigern könnte. Dazu kommen Kosten für die Entwicklung und den grundsätzlichen Betrieb des Chatbots bzw. des Frameworks. Der Punkt der laufenden Kosten sollte bei der Planung nicht vernachlässigt werden.

## 18 Ausblick

In diesem abschließenden Ausblick sollen der durchgeführte Test und der Einsatz von GPT-3 reflektiert werden. Zudem gibt es einen Ausblick auf weitere Untersuchungen, die vor einer Verwendung des Sprachmodells nötig sind.

Nach Einschätzungen von Expert:innen wird KI die Arbeit in Bibliotheken nicht ersetzen können. Durch die hier auch angeschnittenen Limitationen bedarf es noch immer menschlicher Unterstützung. KI kann Arbeitsweisen verändern und erweitern und dadurch neue Angebote schaffen. Sie ist auf absehbare Zeit jedoch nur eine Ergänzung, die noch immer einen Menschen im Prozess erfordert (vgl. ZBW – Leibniz Informationszentrum Wirtschaft, 2022d). Dies soll jedoch keine Abschreckung sein. Es gibt eine Vielzahl an Möglichkeiten, wie KI bereits in der Bibliotheksarbeit eingesetzt wird. Diese Beispiele zeigen auch die Möglichkeiten der Bereicherung in der Arbeit (vgl. Seeliger und andere, 2021).

Diese Einschätzung wird durch die hier durchgeführte Untersuchung bestätigt. Der durchgeführte Test in dieser Ausarbeitung war ein guter Ansatz, um eine erste Einschätzung des Sprachmodells GPT-3 zu erhalten. Es zeigten sich vielversprechende Ansätze, aber auch Schwierigkeiten. Zudem wurde deutlich, dass einige Teile (noch) zu komplex für einen Chatbot sind und die Beantwortung durch einen Menschen erforderlich ist. Diese Durchführung reicht nicht aus, um den Einsatz des Modells abschließend zu befürworten oder komplett abzulehnen. Dafür bedarf es weiterer Tests. Diese Tests sollten unterschiedlichen Umgebungen und Personen einschließen.

In der Durchführung zeigten sich Schwächen, die auch durch das Framework bedingt sind. Um ein erstes Verhalten des Modells zu untersuchen, bot sich dieser Weg an. Es muss jedoch weiter untersucht werden, wie sich das Sprachmodell in anderen Umgebungen, z. B. einem anderen Framework oder einem komplett eigenständig erbauten Chatbot, verhält. Vieles war durch das Framework vorgegeben, das in dem Chatbot der ZBW anders umgesetzt werden wird. In einem nächsten Schritt sollte demnach untersucht werden, wie sich das Modell dort verhält.

Weiterhin ist nun auch ein Fine-Tuning von GPT-3 möglich. Dies erfordert große Mengen an Trainings- und Testdaten, kann jedoch auch die Qualität der Antworten verbessern. Eine nächste Untersuchung kann dies genauer evaluieren.

Zudem sollte der Kreis der Testpersonen erweitert werden. Für diese erste Untersuchung wurde der Personenkreis absichtlich intern gehalten. Da es auch um die Richtigkeit der Antworten ging, konnte dies nur von Fachpersonal beurteilt werden. Es sollte verhindert

werden, dass Nutzer:innen durch auftretende Probleme oder falsche Informationen verwirrt oder abgeschreckt werden. Vorteile bei der Einbeziehung von Nutzer:innen sind die größere Stichprobe, die erreicht werden kann. Zudem werden diese vermutlich unvoreingenommener an die Beurteilung gehen. Durch die kollegiale Verbundenheit besteht ein Risiko, dass Antworten zu beschönigt werden. Die Beziehung zu den Nutzer:innen der Bibliothek ist distanzierter und neutraler. Da zudem ausschließlich die Nutzer:innen mit dem Chatbot interagieren werden, ist es wichtig diesen an ihren Bedürfnissen zu orientieren. Hier darf nicht aus der Sicht von Bibliotheksmitarbeitenden gedacht werden.

Zusätzlich bleibt abzuwarten, welche Fortschritte die Entwicklung einer europäischen Alternative zu GPT-3 macht. GPT-X bildet eine attraktive Option, ein großes Sprachmodell zu nutzen, das europäischen Werten und Gesetzen entspricht. Die Vorgaben dazu sind in Deutschland streng. Gerade wenn es um sensible Daten geht, die auch im Kontext eines Bibliotheks-Chatbot auftreten können, ist der korrekte Umgang damit wichtig. Die Speicherung der Daten auf amerikanischen Servern kann in diesem Kontext durchaus zu Problemen führen. Eine europäische Alternative könnte die Akzeptanz und Verwendung des Modells steigern. Sobald die Entwicklungen dort fortgeschritten sind, ist eine Evaluierung des Modells als Alternative zu GPT-3 durchaus sinnvoll.

## Literaturverzeichnis

- AKBAR, A., 2021. *Rise of the robo-drama* [online]. *Young Vic creates new play using artificial intelligence*. 24. August 2021 [Zugriff am: 6. April 2022]. Verfügbar unter: <https://www.theguardian.com/stage/2021/aug/24/rise-of-the-robo-drama-young-vic-creates-new-play-using-artificial-intelligence>
- ALTMAN, S., 2020. *The GPT-3 hype is way too much*. [online]. 19. Juli 2020 [Zugriff am: 4. Mai 2022]. Verfügbar unter: <https://twitter.com/sama/status/1284922296348454913>
- APIDECK, 2022a. *GPT-3 Demo* [online] [Zugriff am: 5. April 2022]. Verfügbar unter: <https://gpt3demo.com/map>
- APIDECK, 2022b. *Duolingo* [online] [Zugriff am: 5. April 2022]. Verfügbar unter: <https://gpt3demo.com/apps/duolingo>
- APIDECK, 2022c. *Polyglot AI* [online] [Zugriff am: 5. April 2022]. Verfügbar unter: <https://gpt3demo.com/apps/polyglotchatbot>
- BACHIR, 2020. *GPT-3* [online]. *An Overview*. 25. Juli 2020 [Zugriff am: 17. Mai 2022]. Verfügbar unter: <https://dzlab.github.io/ml/2020/07/25/gpt3-overview/>
- BASTIAN, M., 2020. *AI Dungeon: KI-Textadventure mit GPT-3 „ein Paradigmenwechsel“* [online]. 18. Juli 2020 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://mixed.de/ai-dungeon-ki-textadventure-laeuft-mit-gpt-3/>
- BASTIAN, M., 2022. *GPT-3: Google will KI-Content wie Spam behandeln* [online]. 8. April 2022 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://mixed.de/google-will-ki-content-in-der-suche-wie-spam-behandeln/>
- BENDER, E.M., T. GEBRU, A. MCMILLAN-MAJOR und S. SHMITCHELL, 2021. On the Dangers of Stochastic Parrots. Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, United States: Association for Computing Machinery, S. 610-623. ISBN 9781450383097
- BROCKMAN, G., 2022. *Greg Brockman* [online] [Zugriff am: 14. März 2022]. Verfügbar unter: <https://gregbrockman.com/>
- BROWN, T.B., B. MANN, N. RYDER, M. SUBBIAH, J. KAPLAN, P. DHARIWAL, A. NEELAKANTAN, P. SHYAM, G. SASTRY, A. ASKELL, S. AGARWAL, A. HERBERT-VOSS, G. KRUEGER, T. HENIGHAN, R. CHILD, A. RAMESH, D.M. ZIEGLER, J. WU, C. WINTER, C. HESSE, M. CHEN, E. SIGLER, M. LITWIN, S. GRAY, B. CHESSE, J. CLARK, C. BERNER, S. MCCANDLISH, A. RADFORD, I. SUTSKEVER und D. AMODEI, 2020. *Language Models are Few-Shot Learners* [online]. 28. Mai 2020 [Zugriff am: 14. Februar 2022]. Verfügbar unter: <https://arxiv.org/pdf/2005.14165>
- CHINTAGUNTA, B., N. KATARIYA, X. AMATRIAIN und A. KANNAN, 2021. Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization [online]. In: *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, S. 66-76. ISBN 978-1-954085-24-4 [Zugriff am: 6. April 2022]. Verfügbar unter: <https://dx.doi.org/10.18653/v1/2021.nlpmc-1.9>



COMMON CRAWL FOUNDATION, 2022. *Want to use our data?* [online] [Zugriff am: 17. Mai 2022]. Verfügbar unter: <https://commoncrawl.org/the-data/>

COPELAND, B.J., 2021. *artificial intelligence* [online]. 14. Dezember 2021 [Zugriff am: 23. Februar 2022]. Verfügbar unter: <https://www.britannica.com/technology/artificial-intelligence/Alan-Turing-and-the-beginning-of-AI>

DALE, R., 2021. GPT-3: What's it good for? In: *Natural Language Engineering* [online]. **27**(1), S. 113-118 [Zugriff am: 19. März 2022]. ISSN 1351-3249. Verfügbar unter: DOI:10.1017/S1351324920000601

DECHER, D., 2021. *Künstliche Intelligenz, Machine Learning, Deep Learning - Buzzwords erklärt* [online]. 19. Mai 2021 [Zugriff am: 27. Februar 2022]. Verfügbar unter: <https://www.lengoo.com/de-de/blog/kuenstliche-intelligenz-buzzwords/>

DEUTSCHE BAHN, 2022. *Stellen Sie Ihre Frage im Chat* [online]. [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://www.bahn.de/hilfe>

DEUTSCHE TELEKOM, 2022. *Frag Magenta* [online]. [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://www.telekom.de/hilfe/frag-magenta?samChecked=true>

DEUTSCHER BIBLIOTHEKSVERBAND E.V., 2022. *Informationskompetenz und Medienbildung* [online]. [Zugriff am: 13. April 2022]. Verfügbar unter: <https://www.bibliotheksverband.de/informationskompetenz-und-medienbildung>

DEVLIN, J., M.-W. CHANG, K. LEE und K. TOUTANOVA, 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* [online]. 11. Oktober 2018 [Zugriff am: 13. April 2022]. Verfügbar unter: <https://arxiv.org/pdf/1810.04805>

DUOLINGO, 2022. *Duolingo Research* [online]. [Zugriff am: 5. April 2022]. Verfügbar unter: <https://research.duolingo.com/>

ELKINS, K. und J. CHUN, 2020. Can GPT-3 Pass a Writer's Turing Test? In: *Journal of Cultural Analytics* [online]. **5**(2) [Zugriff am: 19. März 2022]. ISSN 2371-4549. Verfügbar unter: DOI:10.22148/001c.17212

FEDERAL MINISTRY FOR ECONOMIC AFFAIRS AND CLIMATE ACTION, 2022. *Gaia-X* [online] [Zugriff am: 15. Mai 2022]. Verfügbar unter: <https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html>

FLORIDI, L. und M. CHIRIATTI, 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. In: *Minds and Machines* [online]. **30**(4), S. 681-694 [Zugriff am: 19. März 2022]. ISSN 0924-6495. Verfügbar unter: DOI:10.1007/s11023-020-09548-1

FRÖHLING, L. und A. ZUBIAGA, 2021. Feature-based detection of automated language models: tackling GPT-2, GPT-3 and Grover. In: *PeerJ Computer science* [online]. **7**, e443. 6. April 2021 [Zugriff am: 19. März 2022]. ISSN: 2167-8359. Verfügbar unter: DOI:10.7717/peerj-cs.443

GITHUB, 2022. *GitHub Copilot* [online] [Zugriff am: 31. August 2022]. Verfügbar unter: <https://copilot.github.com/>

GOTTINGER, H.W., 2017. *Internet economics. Models, mechanisms and management*. Sharjah, UAE: Bentham Science Publishers. ISBN 978-1-68108-547-0

GPT-3, 2020. *A robot wrote this entire article*. [online]. *Are you scared yet, human?* 8. September 2020 [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>

GRÄVEMEYER, A., 2021. KI-Sprachmodell für Europa. Alternativen zum umstrittenen Textgenerator GPT-3 gesucht. In: *c't: Magazin für Computer-Technik*, **2021**(16), S. 40. ISSN 0724-8679

GRÄVEMEYER, A., 2022. Wandlungsfähige Schreib-KI. Apps und Anwendungen mit GPT-3 & Co. texten und analysieren Texte. In: *c't: Magazin für Computer-Technik*, **2022**(9), S. 60-63. ISSN 0724-8679

HAO, K., 2020a. *The messy, secretive reality behind OpenAI's bid to save the world* [online]. 17. Februar 2020 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://www.technologyreview.com/2020/02/17/844721/ai-openai-moonshot-elon-musk-sam-altman-greg-brockman-messy-secretive-reality/>

HAO, K., 2020b. *OpenAI is giving Microsoft exclusive access to its GPT-3 language model* [online]. 23. September 2020 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/>

HECKER, D. und G. PAAß, 2022. Sprachverstehet. GPT-3 & Co. texten überzeugend, aber nicht faktentreu. In: *c't: Magazin für Computer-Technik*, **2022**(9), S. 64-67. ISSN 0724-8679

HOCHREITER, S. und J. SCHMIDHUBER, 1997. Long short-term memory [online]. In: *Neural computation* [online]. **9**(8), S. 1735-1780 [Zugriff am: 14. April 2022]. ISSN 0899-7667. Verfügbar unter: DOI:10.1162/neco.1997.9.8.1735

KABEL, P., 2020. *Dialog zwischen Mensch und Maschine. Conversational User Interfaces, intelligente Assistenten und Voice-Systeme*. Wiesbaden: Springer Nature. ISBN 978-3-658-29584-4

KI BUNDESVERBAND, 2022. *Startschuss für das europäische KI-Sprachmodell OpenGPT-X* [online]. *Pressestatement des KI Bundesverbandes zum Start von OpenGPT-X*. 20. Januar 2022 [Zugriff am: 20. April 2022]. Verfügbar unter: <https://ki-verband.de/wp-content/uploads/2022/01/Pressemitteilung-zum-Start-von-OpenGPT-X.docx.pdf>

KOHNE, A., P. KLEINMANN, C. ROLF und M. BECK, 2020. *Chatbots. Aufbau und Anwendungsmöglichkeiten von autonomen Sprachassistenten*. Wiesbaden: Springer Nature. ISBN 978-3-658-28848-8

KOMPETENZPLATTFORM KÜNSTLICHE INTELLIGENZ NORDRHEIN-WESTFALEN KI.NRW, 2021. *Moderne Sprachtechnologien. Konzepte, Anwendungen, Chancen* [online] [Zugriff am: 9. Januar 2022]. Verfügbar unter: <https://www.ki.nrw/publikationen/>

KÖNIGSTEIN, A., 2022. Durchschnittliche Intelligenz. Programmieren mit KI-Unterstützung. In: *c't: Magazin für Computer-Technik*, **2022**(9), S. 72-73. ISSN 0724-8679

KORNGIEBEL, D.M. und S.D. MOONEY, 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. In: *NPJ digital medicine* [online]. **4**(1), S. 93 [Zugriff am: 14. April 2022]. Verfügbar unter: DOI:10.1038/s41746-021-00464-x

- KOSTADINOV, S., 2017. *How Recurrent Neural Networks work* [online]. 2. Dezember 2017 [Zugriff am: 14. April 2022]. Verfügbar unter: <https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaf7>
- KOSTADINOV, S., 2019. *Understanding Encoder-Decoder Sequence to Sequence Model* [online]. 5. Februar 2019 [Zugriff am: 13. April 2022]. Verfügbar unter: <https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346>
- LACKER, K., 2020. *Giving GPT-3 a Turing Test* [online]. 6. Juli 2020 [Zugriff am: 15. Mai 2022]. Verfügbar unter: <https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html>
- LÄMMEL, U. und J. CLEVE, 2020. *Künstliche Intelligenz. Wissensverarbeitung, Neuronale Netze*. 5., überarbeitete Auflage. München: Carl Hanser Verlag. ISBN 978-3-446-45914-4
- LEGNER, S., 2019. Erzeugnisse Künstlicher Intelligenz im Urheberrecht. In: *Zeitschrift für Urheber- und Medienrecht*. **63**(11), S. 807-812. ISSN 0177-6762
- LINDNER, R., 2021. Er ist Elon Musks Bruder im Geiste. In: *Frankfurter Allgemeine Zeitung* [online]. 15.09.2021 [Zugriff am: 19. März 2022]. Verfügbar unter: <https://www.faz.net/aktuell/wirtschaft/digitec/sam-altman-ist-elon-musks-bruder-im-geiste-17524537.html>
- LIPPERT, L., 2022. *Brutal bis absurd* [online]. *Theater Bremen zeigt Stück geschrieben von K.I.* 2. April 2022 [Zugriff am: 5. April 2022]. Verfügbar unter: <https://www.butenunbinnen.de/nachrichten/verfall-theater-bremen-ki-100.html>
- MANNING, C.D., P. RAGHAVAN und H. SCHÜTZE, 2009. *Introduction to Information Retrieval* [online]. Cambridge: Cambridge University Press [Zugriff am: 29. Juli 2022]. Verfügbar unter: <http://www.informationretrieval.org/>
- MATTMANN, C., 2020. *Machine Learning with TensorFlow*. Second Edition. Shelter Island, NY: Manning. ISBN 978-1-61729-771-7
- MEHNE, J., 2019. *Der Bias-Effekt im Machine Learning* [online]. 30. Juli 2019 [Zugriff am: 15. Mai 2022]. Verfügbar unter: <https://blog.doubleslash.de/der-bias-effekt-im-machine-learning/>
- MENGE-SONNENTAG, R., 2021. *Künstliche Intelligenz* [online]. *Sprachmodell GPT-3 ohne Warteliste verfügbar*. 19. November 2021 [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://www.heise.de/news/Kuenstliche-Intelligenz-Sprachmodell-GPT-3-ohne-Warteliste-verfuegbar-6271868.html>
- MEYER, T., 2010. EconDesk. Die Online-Auskunft, die Wirtschaftswissen schafft! In: *Auskunft*. **30**(2), S. 383-392. ISSN 0720-7123
- NAVIGLI, R., 2018. Natural Language Understanding. Instructions for (Present and Future) Use [online]. In: J. LANG, Hrsg. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*. Stockholm, 13-19 July 2018. Menlo Park, County of San Mateo, California: International Joint Conferences on Artificial Intelligence, S. 5697-5702. ISBN 9780999241127 [Zugriff am: 15. Januar 2022]. Verfügbar unter: <https://doi.org/10.24963/ijcai.2018/812>
- OPENAI, 2015. *Introducing OpenAI* [online]. 11. Dezember 2015 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://openai.com/blog/introducing-openai/>

OPENAI, 2018a. *OpenAI Charter* [online]. 9. April 2018 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://openai.com/charter/>

OPENAI, 2018b. *OpenAI Supporters* [online]. 20. Februar 2018 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://openai.com/blog/openai-supporters/>

OPENAI, 2019a. *Microsoft Invests In and Partners with OpenAI to Support Us Building Beneficial AGI* [online]. 22. Juli 2019 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://openai.com/blog/microsoft/>

OPENAI, 2019b. *OpenAI LP* [online]. 11. März 2019 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://openai.com/blog/openai-lp/>

OPENAI, 2021. *OpenAI's API Now Available with No Waitlist* [online]. 18. November 2022 [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://openai.com/blog/api-no-waitlist/>

OPENAI, 2022a. *About* [online] [Zugriff am: 14. März 2022]. Verfügbar unter: <https://openai.com/about/>

OPENAI, 2022b. *Research* [online]. 4. März 2021 [Zugriff am: 9. März 2022]. Verfügbar unter: <https://openai.com/research/>

OPENAI, 2022c. *Examples* [online] [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://beta.openai.com/examples>

OPENAI, 2022d. *Usage Guidelines* [online] [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://beta.openai.com/docs/usage-guidelines/use-case-guidelines>

OPENAI, 2022e. *Introduction* [online]. 13. Mai 2022 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://beta.openai.com/docs/introduction>

OPENAI, 2022f. *Pricing* [online] 1. September 2022 [Zugriff am: 1. September 2022]. <https://openai.com/api/pricing/>

OPENAI, 2022g. *Engines* [online]. 13. Mai 2022 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://beta.openai.com/docs/engines>

OPENAI, 2022h. *Playground* [online]. 13. Mai 2022 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://beta.openai.com/playground>

OPENAI, 2022i. *Quickstart* [online]. 13. Mai 2022 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://beta.openai.com/docs/quickstart/build-your-application>

PAAß, G. und D. HECKER, 2020. *Künstliche Intelligenz. Was steckt hinter der Technologie der Zukunft?* Wiesbaden: Springer Nature. ISBN 978-3-658-30210-8

PAPINENI, K., S. ROUKOS, T. WARD und W.-J. ZHU, 2002. BLEU. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* [online]. 6. Juli 2022, S. 311-318 [Zugriff am: 15. Juli 2022] Verfügbar unter: DOI:10.3115/1073083.1073135

QUICKCHAT, 2022a. *Product* [online]. [Zugriff am: 24. Juni 2022]. Verfügbar unter: <https://www.quickchat.ai/product>

QUICKCHAT, 2022b. *FAQ* [online]. [Zugriff am: 24. Juni 2022]. Verfügbar unter: <https://app.quickchat.ai/>

QUICKCHAT, 2022c. *Pricing* [online]. [Zugriff am: 24. Juni 2022]. Verfügbar unter: <https://www.quickchat.ai/pricing>

QUICKCHAT, 2022d. *How to structure knowledge base for your AI* [online]. [Zugriff am: 25. Juni 2022]. Verfügbar unter: <https://blog.quickchat.ai/post/how-to-structure-your-knowledge-base/>

QUICKCHAT, 2022e. *Settings* [online]. [Zugriff am: 25. Juni 2022]. Verfügbar unter: <https://app.quickchat.ai/>

RADFORD, A., J. WU, R. CHILD, D. LUAN, D. AMODEI und I. SUTSKEVER, 2019. *Language Models are Unsupervised Multitask Learners* [online] [Zugriff am: 16. Mai 2022]. Verfügbar unter: <https://d4mucfpxyww.cloudfront.net/better-language-models/language-models.pdf>

RAO, D. und B. MCMAHAN, 2020. *Natural Language Processing mit PyTorch. Intelligente Sprachanwendungen mit Deep Learning erstellen*. Heidelberg: O'Reilly. ISBN 978-3-96009-118-9

RHODES, A., 2022. *How do I use Stop Sequences?* [online]. [Zugriff am: 3. Juli 2022]. Verfügbar unter: <https://help.openai.com/en/articles/5072263-how-do-i-use-stop-sequences>

RYTE GMBH, 2022. *Semantic Search* [online]. 4. April 2022 [Zugriff am: 29. Juli 2022]. Verfügbar unter: [https://de.ryte.com/wiki/Semantic\\_Search](https://de.ryte.com/wiki/Semantic_Search)

SCHMIDT, A. und A. ETCHES, 2014. *Useful, usable, desirable. Applying user experience design to your library*. Chicago, Ill.: American Library Ass. ISBN 9780838912263

SCHREINER, M., 2021a. *Mit KI zur Gaming-Idee: Wie GPT-3 beim Spielkonzept hilft* [online]. 4. September 2021 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://mixed.de/super-chore-man-mit-ki-zur-gaming-idee/>

SCHREINER, M., 2021b. *OpenAI: Erste Einblicke in GPT-4 und die mögliche KI-Zukunft* [online]. 13. September 2021 [Zugriff am: 18. Mai 2022]. Verfügbar unter: <https://mixed.de/openai-erste-einblicke-in-gpt-4-und-die-moegliche-ki-zukunft/>

SCHULMAN, J., 2022. *John Schulman's Homepage* [online] [Zugriff am: 14. März 2022]. Verfügbar unter: <http://joschu.net/>

SEELIGER, F., F. PUPPE, R. EWERTH, T. KOCH, A. KASPRZIK, J.F. MAAS, C. POLEY, E. MÖDDEN, A. DEGWITZ und E. GREIFENEDER, 2021. Zum erfolgversprechenden Einsatz von KI in Bibliotheken. Diskussionsstand eines White Papers in progress - Teil 2. In: *b.i.t. online* [online]. 24(3), S. 290-299 [Zugriff am: 27. August 2022]. ISSN 1435-7607. Verfügbar unter: <https://www.b-i-t-online.de/heft/2021-03/fachbeitrag-seeliger.pdf>

SIEBER, A., 2019. *Dialogroboter. Wie Bots und künstliche Intelligenz Medien und Massenkommunikation verändern*. Wiesbaden: Springer Nature. ISBN 978-3-658-24392-0

SILVER, D., A. HUANG, C.J. MADDISON, A. GUEZ, L. SIFRE, G. VAN DEN DRIESSCHE, J. SCHRITTWIESER, I. ANTONOGLOU, V. PANNEERSHELVAM, M. LANCTOT, S. DIELEMAN, D. GREWE, J. NHAM, N. KALCHBRENNER, I. SUTSKEVER, T. LILLICRAP, M. LEACH, K. KAVUKCUOGLU, T. GRAEPEL und D. HASSABIS, 2016. Mastering the game of Go with deep neural networks and tree search. In: *Nature* [online]. 529(7587), S. 484-489 [Zugriff am: 13. April 2022]. ISSN 1476-4687. Verfügbar unter: DOI:10.1038/nature16961

SPECHT-RIEMENSCHNEIDER, L., 2021. Urheberrechtlicher Schutz für Algorithmenerzeugnisse? Phasenmodell de lege lata, Investitionsschutz de lege ferenda. In: STIFTUNG GESELLSCHAFT FÜR RECHTSPOLITIK, TRIER und INSTITUT FÜR RECHTSPOLITIK AN DER UNIVERSITÄT TRIER, Hrsg. *Bitburger Gespräche. Jahrbuch 2020*. München: C.H. Beck, S. 73-86. ISBN 9783406768095

SPLENDID RESEARCH, 2019a. *Welche der folgenden digitalen Assistenten für Sprachsteuerung kennen bzw. nutzen Sie?* [online] April 2019 [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://de.statista.com/statistik/daten/studie/1031358/umfrage/umfrage-zu-bekanntheit-und-nutzung-verschiedener-sprachassistenten-in-deutschland/>

SPLENDID RESEARCH, 2019b. *Wie häufig nutzen Sie Sprachsteuerungen?* [online] April 2019 [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://de.statista.com/statistik/daten/studie/1031324/umfrage/umfrage-in-deutschland-zur-nutzungshaeufigkeit-von-sprachsteuerungen/>

STIELER, W., 2021. Gewissenloser Autor. GPT-3 generiert Texte ganz nach Bedarf - auch Fake News. In: *c't: Magazin für Computer-Technik*, **2021**(21), S. 124-128. ISSN 0724-8679

SUTSKEVER, I., O. VINYALS und Q. LE V, 2014. *Sequence to Sequence Learning with Neural Networks* [online]. 10. September 2014 [Zugriff am: 13. April 2022]. Verfügbar unter: <https://arxiv.org/pdf/1409.3215>

SUTSKEVER, I., 2022. *Ilya Sutskever* [online]. [Zugriff am: 14. März 2022]. Verfügbar unter: <http://www.cs.toronto.edu/~ilya/>

T2INFORMATIK, 2018. Was ist ein Entscheidungsbaum? [online]. 6. Juni 2018 [Zugriff am: 13. Juni 2022]. Verfügbar unter: <https://t2informatik.de/wissen-kompakt/entscheidungsbaum/>

TAMKIN, A., M. BRUNDAGE, J. CLARK und D. GANGULI, 2021. *Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models* [online]. 4. Februar 2021 [Zugriff am: 13. April 2022]. Verfügbar unter: <https://arxiv.org/pdf/2102.02503>

THE GITHUB BLOG, 2022. *GitHub Copilot is now available to individual developers* [online] 21. Juni 2022 [Zugriff am: 3. September 2022]. Verfügbar unter: <https://github.blog/changelog/2022-06-21-github-copilot-is-now-available-to-individual-developers/>

THEATER BREMEN, 2022. *Verfall* [online]. *Ein Picknick im Grünen* [Zugriff am: 5. April 2022]. Verfügbar unter: [https://www.theaterbremen.de/de\\_DE/programm/verfall-ein-picknick-im-gruenen.1303388](https://www.theaterbremen.de/de_DE/programm/verfall-ein-picknick-im-gruenen.1303388)

VANCE, A., 2020. *Elon Musk. Tesla, PayPal, SpaceX*. München: FinanzBuch. ISBN 978-3-95972-421-0

VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A.N. GOMEZ, L. KAISER und I. POLOSUKHIN, 2017. *Attention Is All You Need* [online]. 12. Juni 2017 [Zugriff am: 13. April 2022]. Verfügbar unter: <https://arxiv.org/pdf/1706.03762>

WARMERDAM, V., 2020. *GPT-3* [online]. *Careful First Impressions*. 22. Juli 2020 [Zugriff am: 26. Februar 2022]. Verfügbar unter: <https://rasa.com/blog/gpt-3-careful-first-impressions/>

WÖLBERT, C., 2020. Wolkenschloss. Europa-Cloud Gaia-X tritt gegen übermächtige Konkurrenz an. In: *c't: Magazin für Computer-Technik*, **2020**(22), S. 12-13. ISSN 0724-8679

WOLFANGEL, E., 2021. Hey KI! In: *MIT Technology Review*. **2021**(6), S. 16-23.  
ISSN 1099-274X

ZAREMBA, W., 2022. *Wojciech Zaremba* [online] [Zugriff am: 14. März 2022]. Verfügbar unter:  
<https://wojzaremba.com/>

ZBW - LEIBNIZ INFORMATIONSZENTRUM WIRTSCHAFT, 2022a. *Über Research Guide EconDesk* [online] [Zugriff am: 19. Februar 2022]. Verfügbar unter:  
<https://www.econbiz.de/eb/de/hilfe/research-guide-econdesk/servicepolicy>

ZBW - LEIBNIZ INFORMATIONSZENTRUM WIRTSCHAFT, 2022b. *Wer ist die ZBW?* [online] [Zugriff am: 19. Februar 2022]. Verfügbar unter: <https://www.zbw.eu/de/ueberuns/profil/wer-ist-die-zbw>

ZBW - LEIBNIZ INFORMATIONSZENTRUM WIRTSCHAFT, 2022c. *Research Guide EconDesk* [online] [Zugriff am: 26. Februar 2022]. Verfügbar unter:  
<https://www.zbw.eu/de/service/beratung>

ZBW - LEIBNIZ INFORMATIONSZENTRUM WIRTSCHAFT, 2022d. *Potential of AI for Libraries: A new level for knowledge organization? Panel Discussion*. In: TIB AV-Portal [online]. [Zugriff am: 31.08.2022]. Verfügbar unter: <https://av.tib.eu/media/58084>

ZHU, Q. und J. LUO, 2021. *Generative Pre-Trained Transformer for Design Concept Generation: An Exploration* [online]. 16. November 2021 [Zugriff am: 13. April 2022]. Verfügbar unter:  
<https://arxiv.org/pdf/2111.08489>

## Abbildungsverzeichnis

Abbildung 1: Abläufe im Auskunftsdienst EconDesk .....	10
Abbildung 2: Darstellung Zusammenhang KI, ML, DL, NLP .....	14
Abbildung 3: Natural Language Processing .....	16
Abbildung 4: Projektion von Embeddingvektoren in eine 2-dimensionale Ebene .....	20
Abbildung 5: Skizzierte Architektur eines CNN (a) und RNN (b) .....	21
Abbildung 6: Vereinfachte Darstellung eines seq2seq-Modells und eines Modells mit Attention-Mechanismus .....	23
Abbildung 7: Training von GPT-2 (a) und GPT-3 (b) .....	34
Abbildung 8: Systemarchitektur GPT-3 .....	35
Abbildung 9: Playground der OpenAI API .....	50
Abbildung 10: Ausgabe in der API-Ansicht.....	51
Abbildung 11: Filter in der API-Ansicht.....	51
<i>Abbildung 12: Wissensfragen an Davinci (oben) und Babbage (unten).....</i>	<i>55</i>
Abbildung 13: Grammatikkorrektur von Davinci, Curie, Babbage, Ada (von oben links nach unten rechts) .....	56
Abbildung 14: Interviewfragen von Davinci, Curie, Babbage, Ada (von oben links nach unten rechts) .....	57
Abbildung 15: Essay-Skizzen von Davinci, Curie, Babbage, Ada (von oben nach unten). ..	59
Abbildung 16: Ideengenerierung von Davinci, Curie, Babbage, Ada (von oben nach unten) .....	60
Abbildung 17: Filmtitel als Emoji von Davinci, Curie, Babbage, Ada (von oben links nach unten rechts) .....	61
Abbildung 18: Zusammenfassung für einen Zweitklässler von Davinci (oben) und Ada (unten).....	63
Abbildung 19: Schlagwortvergabe von Davinci (links) und Curie (rechts).....	63
Abbildung 20: SQL-Übersetzung von Davinci, Curie, Babbage, Ada (von oben nach unten) .....	65
Abbildung 21: Übersetzung von Programmiersprachen durch Davinci, Curie, Babbage, Ada (von links nach rechts).....	66
Abbildung 22: Übersicht großer Sprachmodelle .....	74
Abbildung 23: Gesamtbewertung der Chats.....	88
Abbildung 24: Anzahl richtiger Chats nach prozentualer Verteilung .....	91



## Tabellenverzeichnis

Tabelle 1: GPT-3's Ergebnisse bei der Ausführung von NLP-Aufgaben.....	37
Tabelle 2: Übersicht der GPT-3 Varianten, deren Eigenschaften .....	49
Tabelle 3: Kosten für die Modelle sowie deren Fine-Tuning und Nutzung.....	52
Tabelle 4: Verteilung der Anfragen.....	87
Tabelle 5: Richtigkeit der Antworten .....	90

## Glossar

**ALPHAGO:** Computerprogramm von DeepMind, das das Brettspiel Go spielt. Wurde vor allem dadurch bekannt, dass es professionelle menschliche Go-Spieler besiegt (vgl. Silver und andere 2016). [zurück]

**BERT:** kurz für Bidirectional Encoder Representations from Transformers. Sprachmodell, das auf der bidirektionalen Transformer-Architektur basiert, d. h. es kann einen Kontext zu den Worten vor und nach der untersuchten Eingabe herstellen. Mit großen Datenmengen wird es vortrainiert, mit Aufgaben-spezifischen Daten erfolgt ein Fine-Tuning (vgl. Devlin und andere, 2018, S. 1). [zurück]

**BIAS:** Zum einen die Verzerrung in der Statistik, bei der das Ergebnis von der Erwartung abweicht. Zum anderen beschreibt es auch eine Voreingenommenheit, da Menschen durch verschiedene Vorerfahrungen zu unterschiedlichen Beurteilungen kommen. Sprachmodelle sollten sich fair verhalten und keine Vorurteile reproduzieren. Durch z. B. vorbelastete Trainingsdaten trainieren sich die Modelle jedoch Sichtweisen an (vgl. Mehne, 2019). [zurück]

**BIBLIOTHEKAR:INNEN:** Mitarbeiter:innen einer Bibliothek, die ein bibliothekswissenschaftliches Studium abgeschlossen haben. Sie übernehmen oftmals verantwortungsvollere Aufgaben wie die Rechercheberatung und Vermittlung von Informationskompetenz. In dieser Ausarbeitung sind damit die Mitarbeiter:innen der Abteilung Benutzungsdienste der ZBW gemeint, die ein solches Studium abgeschlossen haben. [zurück]

**BLEU-SCORE:** Kurzform für: bilingual evaluation understudy. Bewertungsalgorithmus für die maschinelle Übersetzung von Texten. Vergleicht die Ausgabe der Maschine mit der eines Menschen, je näher die maschinelle Ausgabe an der des Menschen ist, desto besser ist der Score (vgl. Papineni und andere, 2002). [zurück]

**CHATBOT:** automatisiertes Dialogsystem, das für eine Mensch-Maschine-Kommunikation eingesetzt wird, sowohl auf Text- als auch auf Sprachbasis. Kann einfache Informationen geben, zunehmend aber auch Beratungen und für Auskünfte eingesetzt werden (vgl. Kohne und andere, 2020, S. 1f.). [zurück]

**DECODER:** Verarbeitet weitergeleitete Informationen des Encoders. Durch RNN wird für jede Eingabe die Ausgabe berechnet. Bei einer Texteingabe ist die Ausgabe die Antwort auf den Eingabesatz (vgl. Kostadinov, 2019). [zurück]

**ECONBIZ:** Ein wirtschaftswissenschaftliches Fachportal, das von der ZBW betrieben wird. Es sind verschiedene Datenbanken integriert, die wirtschaftswissenschaftliche Sammlungen verzeichnen. Derzeit sind über 11 Millionen Datensätze verzeichnet, das beinhaltet vor allem Bücher, Zeitschriften, Aufsätze und Working Paper. EconBiz wird weltweit von Forschenden und Studierenden als Nachweisportal genutzt, ist gleichzeitig aber auch der Bibliothekskatalog der ZBW für Nutzer:innen vor Ort. [zurück]

**ENCODER:** Nimmt Eingabe entgegen. Oftmals besteht dieser aus mehreren LSTM, die Informationen zur Eingabe sammeln und weiterleiten. Bei einer Texteingabe besteht der Input aus allen eingegebenen Worten (vgl. Kostadinov, 2019). [zurück]

**ENTITIES:** dt. Entitäten. Stellen Kontext für spezifischen Intent dar (vgl. Kohne und andere, 2020, S. 45) [zurück]

**ENTSCHEIDUNGSBAUM:** stellt den mehrstufigen Entscheidungsprozess mit allen Entscheidungsoptionen in einem Prozess dar (vgl. t2informatik, 2018). [zurück]

**FACHREFERENT:INNEN:** auch wissenschaftliche:r Bibliothekar:in. Einsatz in wissenschaftlichen Bibliotheken zur Beschaffung und Indexierung des Bestands und zur fachlichen Beratung. In der ZBW arbeiten die Fachreferent:innen nicht in den Bibliotheksräumen vor Ort mit. Sie werden jedoch bei fachlich spezifischen Fragen zur Beratung herangezogen, da sie im Gegensatz zu Bibliothekar:innen ein Fachstudium der Wirtschaftswissenschaften abgeschlossen haben. [zurück]

**GAIA-X:** Europäisches Projekt zum Aufbau einer offenen, transparenten und sicheren Dateninfrastruktur. Daten können in dem Umfeld gespeichert und nachgenutzt werden, während Nutzer:innen Hoheit über ihre Daten behalten (vgl. Federal Ministry for Economic Affairs and Climate Action, 2022). [zurück]

**INFORMATION RETRIEVAL:** „*Beim Information Retrieval (IR) geht es darum, unstrukturiertes Material (in der Regel Dokumente), das einem Informationsbedarf entspricht, in großen Sammlungen (in der Regel auf Computern gespeichert) zu finden*“ (Manning und andere, 2009, S. 1). [zurück]

**INFORMATIONSKOMPETENZ:** Kompetenz, um Informationen zu ermitteln, filtern, einordnen und nutzen zu können. Eine Aufgabe von Bibliotheken ist es, Nutzer:innen und Studierenden diese Fähigkeit zu vermitteln (vgl. Deutscher Bibliotheksverband e. V., 2022). [zurück]

**INTENT:** Absicht innerhalb eines Satzes (vgl. Kohne und andere, 2020, S. 44) [zurück]

**NATURAL LANGUAGE INTERFERENCE (NLI):** „*Fähigkeit, die Beziehung zwischen zwei Sätzen zu verstehen*“ (Brown und andere, 2020, S. 20) [zurück]

**PARAMETER:** Gewichte in einem neuronalen Netz, die Eingaben bewerten und ab einem festgelegten Schwellenwert Signale senden (vgl. Paaß und andere, 2020, S. 458). [zurück]

**SEMANTISCHE SUCHE:** Bei der semantischen Suche wird die Bedeutung der Suchanfrage verstanden und die direkte Lösung präsentiert. Ein Algorithmus untersucht dabei die Beziehung zwischen den Worten der Sucheingabe. Darauf basierend wird die passende Antwort gesucht und dem/r Nutzer:in präsentiert. Im Gegensatz steht die Suche nach Keywords, die bisher verbreitet war. Dabei werden die Worte der Sucheingabe mit dem Index der Suchmaschine abgeglichen und relevante Dokumente ausgegeben (vgl. Ryte GmbH, 2022). [zurück]

**TURING TEST:** Ein Versuch, der den Unterschied zwischen Menschen und Maschine untersuchen soll. Eine Versuchsperson hat über einen Chat auf dem Computer Kontakt zu einem Menschen und einem Chatbot. Wenn die Person nicht erkennen kann, welche Antwort von einem Menschen und welche von einem Chatbot gegeben wurde, hat der Chatbot den Test bestanden, es gibt dann keinen Unterschied mehr zwischen dessen Antworten und denen eines Menschen (vgl. Kohne und andere, 2020, S. 8). [zurück]

**USABILITY:** Gebrauchstauglichkeit eines Produktes und wie dieses angewendet wird, um Ziele effektiv und zielführend zu erreichen (vgl. Schmidt und andere, 2014, S. 3). [zurück]

**USER EXPERIENCE:** erweitert Usability um Emotionen: wie sich jemand bei der Nutzung eines Produkts/Services/Angebot/... fühlt (vgl. Schmidt und andere, 2014, S. 1). [zurück]

# Anhang

## A.1 Code der Beispielfragen aus Kapitel 8

Dieser Anhang bietet den Programmcode zu den Beispielfragen aus Kapitel 8 in Python notiert. Der Übersichtlichkeit halber wird hier nur der Code der Engine *Davinci* dargestellt. Alle Einstellungen waren in den Engines gleich, lediglich die Zeile `model` und die Ausgaben im `prompt` unterscheiden sich.

Der komplette Code aller Engines findet sich auf dem angelegten GitHub-Repository für diese Masterarbeit: <https://git.th-wildau.de/albe6395/evaluation-des-sprachmodells-gpt-3>

→ `G1_ProgrammcodeKapitel8.py`

### 8.1 Question Answering

#### Beantwortung von Wissensfragen

```
import os
import openai
openai.api_key = os.getenv("OPENAI_API_KEY")
start_sequence = "\nA:"
restart_sequence = "\n\nQ: "
response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Q: What is photosynthesis?\nA: Photosynthesis is the
    process that produces organic molecules from simple inorganic
    molecules from the sun's energy.\n\nQ: Who was Albert Einstein
    and when did he live?\nA: Albert Einstein was a German-born
    theoretical physicist who developed the theory of relativity,
    one of the two pillars of modern physics (alongside quantum
    mechanics). He is best known in popular culture for his mass-
    energy equivalence formula  $E = mc^2$  (which has been dubbed \"the
    world's most famous equation\"). He received the 1921 Nobel
    Prize in Physics \"for his services to theoretical physics,
    and especially for his discovery of the law of the
    photoelectric effect\", a pivotal step in the development of
    quantum\n\nQ: Who is the chancellor of Germany?\nA: The
    current chancellor of Germany is Angela Merkel.",
    temperature=0,
    max_tokens=100,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0,
    stop=["\n"]
)
```

#### Grammatikkorrektur

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Correct this to standard English:\n\nThe chatbot does
    give no correct answers.\n\nThe chatbot does not give correct
```

```

    answers.\n\nIf it rains we can't not go swimming.\n\nIf it
    rains we can't go swimming.\n\nI expect to become a nice
    present.\n\nI expect to receive a nice present.",
    temperature=0,
    max_tokens=60,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)

```

## 8.2 Texte erzeugen

### Interviewfragen

```

import os
import openai
openai.api_key = os.getenv("OPENAI_API_KEY")
response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Create a list of 8 questions for my interview about
    chatbots:\n\n1.What are chatbots?\n2.How do chatbots
    work?\n3.What are the benefits of chatbots?\n4.What are the
    challenges of chatbots?\n5.How can chatbots be used in
    business?\n6.How can chatbots be used in customer
    service?\n7.What are the best practices for designing
    chatbots?\n8.What are the ethical considerations of
    chatbots?",
    temperature=0.5,
    max_tokens=150,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)

```

### Essay skizzieren

```

import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Create an outline for an essay about the impact of
    \nartificial intelligence on people's everyday lives:\n\n1.
    Introduction\n\n2. The history of artificial
    intelligence\n\n3. The current state of artificial
    intelligence\n\n4. The impact of artificial intelligence on
    people's everyday lives\n\n5. The future of artificial
    intelligence",
    temperature=0.5,
    max_tokens=150,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)

```

## 8.3 Ideengenerierung

### Ideengenerator

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Brainstorm some ideas to create a school lesson about
    artificial intelligence:\n\n1. What is artificial
    intelligence?\n\n2. What are some common applications of
    artificial intelligence?\n\n3. How does artificial intelligence
    work?\n\n4. What are some ethical considerations surrounding
    artificial intelligence?",
    temperature=0.6,
    max_tokens=250,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)
```

### Film zu Emoji

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Convert movie title into emoji:\nBatman:
    🦇🦇🦇🦇🦇\nLord of the Rings: 🗡️🛡️👤👤🎬\nHarry Potter:
    ⚡️👤📖👤👤\nTitanic: 🏢💔🌊\nDespicable Me: 😄👉🍌",
    temperature=0.8,
    max_tokens=60,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)
```

## 8.4 Klassifizierung

### Zusammenfassung für einen Zweitklässler

```
import os
import openai
openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Summarize this for a second-grade
    student:\nPhotosynthesis is a process used by plants and other
    organisms to convert light energy into chemical energy that,
    through cellular respiration, can later be released to fuel the
    organism's activities. Some of this chemical energy is stored in
    carbohydrate molecules, such as sugars and starches, which are
```



synthesized from carbon dioxide and water - hence the name photosynthesis, from the Greek phōs (φῶς), \"light\", and sunthesis (σύνθεσις), \"putting together\".[1][2][3] Most plants, algae, and cyanobacteria perform photosynthesis; such organisms are called photoautotrophs. Photosynthesis is largely responsible for producing and maintaining the oxygen content of the Earth's atmosphere, and supplies most of the energy necessary for life on Earth.[4]\n\nAlthough photosynthesis is performed differently by different species, the process always begins when energy from light is absorbed by proteins called reaction centers that contain green chlorophyll (and other colored) pigments/chromophores. In plants, these proteins are held inside organelles called chloroplasts, which are most abundant in leaf cells, while in bacteria they are embedded in the plasma membrane. In these light-dependent reactions, some energy is used to strip electrons from suitable substances, such as water, producing oxygen gas. The hydrogen freed by the splitting of water is used in the creation of two further compounds that serve as short-term stores of energy, enabling its transfer to drive other reactions: these compounds are reduced nicotinamide adenine dinucleotide phosphate (NADPH) and adenosine triphosphate (ATP), the \"energy currency\" of cells.\n\nIn plants, algae and cyanobacteria, sugars are synthesized by a subsequent sequence of light-independent reactions called the Calvin cycle. In the Calvin cycle, atmospheric carbon dioxide is incorporated into already existing organic carbon compounds, such as ribulose biphosphate (RuBP).[5] Using the ATP and NADPH produced by the light-dependent reactions, the resulting compounds are then reduced and removed to form further carbohydrates, such as glucose. In other bacteria, different mechanisms such as the reverse Krebs cycle are used to achieve the same end.\n\nThe first photosynthetic organisms probably evolved early in the evolutionary history of life and most likely used reducing agents such as hydrogen or hydrogen sulfide, rather than water, as sources of electrons.[6] Cyanobacteria appeared later; the excess oxygen they produced contributed directly to the oxygenation of the Earth,[7] which rendered the evolution of complex life possible. Today, the average rate of energy capture by photosynthesis globally is approximately 130 terawatts,[8][9][10] which is about eight times the current power consumption of human civilization.[11] Photosynthetic organisms also convert around 100-115 billion tons (91-104 Pg petagrams, or billion metric tons), of carbon into biomass per year.[12][13] That plants receive some energy from light - in addition to air, soil, and water - was first discovered in 1779 by Jan Ingenhousz.\n\nPhotosynthesis is vital for climate processes, as it captures carbon dioxide from the air and then binds carbon in plants and further in soils and harvested products. Cereals alone are estimated to bind 3,825 Tg (teragrams) or 3.825 Pg (petagrams) of carbon dioxide every year, i.e. 3.825 billion metric tons.[14] \n\nPhotosynthesis is a process that is used by plants and other organisms to convert light energy into chemical energy. This chemical energy can later be used to fuel the organism's activities. Some of the chemical energy is stored in carbohydrate molecules, such as sugars and starches. Photosynthesis is responsible for producing and maintaining the oxygen content of the Earth's atmosphere, and supplies most of the",  
temperature=0.7,  
max\_tokens=256,  
top\_p=1,  
frequency\_penalty=0,  
presence\_penalty=0  
)

## Schlagworterkennung

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="Extract keywords from this text:\n\nSmall and medium-sized enterprises (SMEs) dominate the market in terms of sheer number of organisations; yet, scholarly resource materials to assist in honing skills and competencies have not kept pace. This well renowned textbook guides students through the complexities of entrepreneurship from the unique perspective of marketing in SME contexts, providing a clear grounding in the principles, practices, strategies, challenges, and opportunities faced by businesses today.\n\nSMEs now need to step up to the terrain of mobile marketing and consumer-generated marketing and utilise social media marketing tools. Similarly, the activities of various stakeholders in SME businesses like start-up accelerators, business incubators, and crowdfunding have now gained more prominence in SME activities. This second edition advances grounds covered in the earlier edition and has been fully updated to reflect this new, dynamic business landscape. Updates include:\n\nA consideration of social media imperatives on SME marketing;\n\nDiscussion of forms of capital formation and deployment for marketing effectiveness, including crowdfunding;\n\nUpdated international case studies drawn from diverse backgrounds;\n\nHands-on practical explorations based on real-life tasks to encourage deeper understanding.\n\nThis book is perfect for students studying SMEs, Marketing and Enterprise at both advanced undergraduate and postgraduate levels, as well as professionals looking to obtain the required knowledge to operate their businesses in this increasingly complex and turbulent marketing environment.\n\n-Small and medium-sized enterprises \n-SMEs \n-marketing \n-consumer-generated marketing \n-social media marketing \n-business incubators \n-crowdfunding \n-case studies \n-undergraduate \n-post",
    temperature=0.3,
    max_tokens=100,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0
)
```

## 8.6 Programmcode

### SQL-Übersetzer

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="### Postgres SQL tables, with their properties:\n#\n# Employee(id, name, department_id)\n# Department(id, name, address)\n# Salary_Payments(id, employee_id, amount, date)\n#\n### A query to list the names of the departments which employed more
```

```

than 10 employees in the last 3 months\nSELECT d.name\nFROM
Department d\nJOIN Employee e ON d.id = e.department_id\nJOIN
Salary_Payments sp ON sp.employee_id = e.id\nWHERE sp.date >=
DATE_SUB(CURDATE(), INTERVAL 3 MONTH)\nGROUP BY d.name\nHAVING
COUNT(*) > 10",
    temperature=0,
    max_tokens=250,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0,
    stop=[";"]
)

```

## Übersetzer Programmiersprachen

```

import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
    model="text-Davinci-002",
    prompt="##### Translate this function from Java into Python\n###
Java\n    \npublic class HelloWorld{\npublic static void main
(String [] args){\nSystem.out.println(\"Hallo Welt\");\n}\n}\n
\n### Python\n\ndef main():\n    print(\"Hello World\")\n\n\nif
__name__ == \"__main__\":\n    main()",
    temperature=0,
    max_tokens=100,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0,
    stop=[";"]
)

```

## A.2 Chatbot-Prototyp zur Testung von GPT-3 im Live-Einsatz

In diesem Anhang werden die Einstellungen des Prototyps dargestellt.

Hinweis: In der Wissensbasis in Quickchat werden die einzelnen Informationen nicht nummeriert. Diese werden als Stichpunktliste aufgeführt. Die Nummerierung wurde eingefügt, um Referenzen aus dem Auswertungs-Kapitel zu ermöglichen.

### **Settings:**

Chat bot language: German/English

Chat bot creativity level: Low/normal/high

\*Einstellungen wurden für die Tests jeweils angepasst

### **Knowledge Base:**

Product/service/project name: ZBW Chatbot

Short description: This chatbot answers all questions about the library.

Knowledge Base:

- (1) ZBW - Leibniz Information Centre for Economics is a research infrastructure facility for the economic sciences. Two locations: Hamburg and Kiel. Its goal is to support economists and students: various research teams conduct research on the topics of open access, research data management, long-term archiving and automated indexing of media with artificial intelligence. Another service is the library, which supports students and economists on site as well as worldwide users in their search for literature.
- (2) Library opening hours: Monday and Thursday 10 a.m. to 6 p.m. Tuesday, Wednesday and Friday 10 a.m. to 4:30 p.m. The library is closed on Saturdays, Sundays and holidays.
- (3) During library opening hours books can be borrowed, the workstations in the reading room can be used and library cards can be issued.
- (4) Contact options: By telephone in Kiel at 0431/8814-555 and in Hamburg at 040/42834-219 during opening hours. Outside opening hours, an e-mail can be written: [info@zbw.eu](mailto:info@zbw.eu).
- (5) To use the ZBW library and borrow books and magazines, a free library card is required. online-registration:  
<https://www.zbw.eu/de/recherchieren/ausleihe/nutzerAusweis/online-anmeldung/>. card must be collected within five opening days during opening hours. After the registration

is processed, the user will receive an e-mail notification that the card can be picked up. This e-mail will also contain preliminary login information that can already be used to order books and periodicals for loan. On-site registration is also possible. Valid ID card and, if available, a student ID card are required for pickup.

- (6) ZBW card is valid for one year at a time. Can be renewed every year. For a renewal, the presentation of the identity card and, if applicable, a student ID is required.
- (7) The account will be blocked if a book has reached the fourth reminder level or if the account has more than 25 € in charges. If the account is blocked, no borrowing or renewal can take place. To unblock the account, the overdue book must be returned and the outstanding charges must be paid. Payment can be made in cash at the service counter or by EC card at the automat.
- (8) Printers, copiers and scanners are available in the ZBW. For printing and copying, a ZBW card is required, which contains credit. One page costs 0,05 €. Scanning is free of charge. A USB stick is required for this, onto which the scan is saved as a PDF.
- (9) Individual workstations are located in the quiet reading room, where individuals can work on their own. Listening to music only with headphones. Reservation of seats not possible. ZBW card is required for use.
- (10) Group rooms for learning and working together. In the group rooms, small groups can meet to work and talk to each other at room volume. The rooms are designed for four to eight people. Reservation of rooms is not possible. ZBW card is required for use.
- (11) EconBiz is an economics subject portal and library catalog. It contains over 11 million entries: books, articles and journals. Can be used nationally as a bibliographic database to search for economic literature: <https://www.econbiz.de/>.
- (12) To check whether a book is available in the ZBW, this can be searched in EconBiz: <https://www.econbiz.de/>.
- (13) Publications in economics are collected from the whole world. These include monographs, journals, working papers, statistics, and publications from ministries and research institutions.
- (14) EconBiz allows the export of bibliographic data into a literature management program. The following steps are required: 1) Select "Export selected results" at the bottom of the hit display or the individual hit. 2) Select the desired format. 3) You will then receive a file with the exported data, which can be transferred to your literature management program.

- (15) EconBiz offers a calendar of events, in which advanced training courses, meetings and conferences in economics are listed. The calendar can be accessed at this link: <https://www.econbiz.de/Events/Results>. To add your event to the calendar, please use the contact form: <https://www.econbiz.de/eb/de/kontakt/veranstaltung-vorschlagen>.
- (16) To order books and journals, you must log in to EconBiz with your ZBW account.
- (17) User ID: is the eight-digit number on the ZBW card.
- (18) Password: by default the date of birth with six digits in the format: DDMMYY. If you have forgotten your password, contact us by phone or mail.
- (19) If you do not have access to the library in Hamburg or Kiel, you can create an EconBiz account. With this account only favorite-lists can be created. No books can be ordered or borrowed. You assign the login data yourself. If you have forgotten your password, please contact [info@econbiz.de](mailto:info@econbiz.de).
- (20) To borrow books from the ZBW, they must be ordered in advance via EconBiz. The titles will then be searched for from the stacks and made available for pickup. To order a book, proceed as follows: 1) Log in to EconBiz with your ZBW account. 2) Search for the book you want. 3) Click on "Request" next to the signature. 4) Select the pick-up location. 5) The order is completed. You can track the current status in your account summary in EconBiz. As soon as the book is ready for you, the status is "Checked out". Books can be borrowed for three weeks and renewed up to four times if there is no reservation.
- (21) To borrow journals from the ZBW, they must be ordered in advance via EconBiz. The titles are then searched for from the magazine and made available for pickup. To order a journal, proceed as follows: 1) Log in to EconBiz with your ZBW account. 2) Search for the journal you want. 3) Select the correct volume. 3) Click on "Request" next to the signature. 4) Select the pick-up location. 5) The order is completed. You can track the current status in your account overview in EconBiz. As soon as the journal is ready for you, the status is "Pickup shelf". Journals cannot be borrowed, they can only be used in the reading room. It is possible to make copies or scans.
- (22) To order printed articles, please follow these steps: 1) Log in to EconBiz with your ZBW account. 2) Search for the desired article. 3) Note the information under "Published in", there you will find the information in which volume the article was published. 4) Select the appropriate volume from the list. 5) If you order the whole volume: Select the pick-up location. The order is completed. You can track the current status in your account overview in EconBiz. As soon as the journal is ready for you, the status is "Pickup

shelf". Journals cannot be borrowed, they can only be used in the reading room. It is possible to make copies or scans. 6) When you fill in the order form, you will receive an e-mail with the access options to the article.

- (23) An article is available either printed in the ZBW or digitally. A printed article must be ordered for viewing via EconBiz. A digital article is either freely available or accessible via remote access, this is indicated in EconBiz. Or a license is required, in which case please visit the ZBW on site to access the article.
- (24) Remote Access is only available to selected individuals for licensing reasons. You can then access licensed articles from home. You are eligible if your primary residence is in Hamburg or Kiel. In addition, persons are entitled who study at a university that the ZBW has listed as a cooperation partner. If you have any questions about this, please contact us.
- (25) Access to licensed articles and e-books is only possible if you are logged in to EconBiz. If the access does not work, please check these requirements: 1) Are you logged in with your ZBW account (data from ZBW card)? 2) Is your main residence in Hamburg or Kiel? 3) Do you study at a university in Hamburg or Kiel that counts as a cooperation partner? 4) Have you tried all the links displayed? 5) Are you connected to a VPN access of your university? If you still have problems, please contact our staff.
- (26) The location of the books can be viewed in EconBiz. Books and journals can be ordered regardless of the desired pick-up location. Ordered books from Hamburg can be brought to Kiel by a transport service and vice versa.
- (27) After ordering, it takes a certain time until the books and journals are available. A book or magazine from the same location is available after about one hour during opening hours. A book or magazine from the other location is usually available the next business day if ordered by 4pm.
- (28) There is no limit to how many items can be borrowed at the same time.
- (29) If a book or magazine has already been borrowed by another user, it can be reserved.
- (30) Ordered books are available in the collection shelves. You have five opening days to pick them up during opening hours. You will find the books sorted by name in the shelves. For your compartment number, take the last two numbers of your ZBW ID and swap them. So for the final numbers 23, the shelf is 32. For Hamburg users, the H at the end is ignored.

- (31) There is no separate notification when an ordered book is ready for pickup. The status can be tracked in the account. If a book or journal that has been reserved has been returned, you will be notified by mail that it is now available.
- (32) Borrowed books can be renewed in the account in EconBiz. The renewal counts from the day one renews the books. It is three weeks again each time. Four renewals are possible if there is no reservation. If the account is locked, renewal is no longer possible.
- (33) The return of borrowed books is possible during opening hours at the service desk. Outside opening hours, books can be placed in the return box in the entrance area in Hamburg from Monday to Friday between 6:30 am and 7:30 pm. In Kiel, a book drop box outside next to the entrance is accessible around the clock.
- (34) If a user has lost a book, a replacement must be provided. There are two possibilities: 1) The user procures a new copy himself. This can be used, but must be in good condition and have at least the same or a higher edition. Any reminder fees incurred and a processing fee of 20 € must then be paid. 2) The second possibility is that the ZBW buys the book new. Then the user will be charged the book price, a procurement fee of 10 € and the processing fee of 20 €.
- (35) If books are returned late, reminder fees will be charged. This is 1 € per book in the first week, 2,50 € in the second week, 5 € in the third week and 15 € in the fourth week. After the fourth week the account will be blocked. Overdue fines can be paid in cash at the service counter or by EC card at the automat. From a total amount of 10 € the fees can be invoiced and paid by bank transfer. To do so, contact us at: [info@zbw.eu](mailto:info@zbw.eu).
- (36) The ZBW has licensed access for over 400 databases. An overview can be found here: [https://dbis.ur.de/dbinfo/dbliste.php?bib\\_id=zbw&colors=127&ocolors=40&lett=f&gebiete=16](https://dbis.ur.de/dbinfo/dbliste.php?bib_id=zbw&colors=127&ocolors=40&lett=f&gebiete=16). The databases are accessible via the computers in the ZBW. Access from home is not possible.
- (37) If a book you are looking for is not in the ZBW's collection, you can submit an acquisition proposal. This form can be used for this purpose: <https://www.econbiz.de/eb/de/kontakt/buch-vorschlagen>. For journals, an acquisition proposal can be submitted using this form: <https://www.econbiz.de/eb/de/kontakt/zeitschrift-vorschlagen>.
- (38) The ResearchGuide EconDesk offers support for researching economic literature. Tips for research are given, e.g. search words and combinations as well as databases where information can be found. No literature lists with recommendations are provided. It is up to the inquirer to check whether a title is relevant for his/her own work. Personal



consultations with a subject specialist are also possible. Please use the contact options on this website: <https://www.zbw.eu/de/service/beratung>.

- (39) When help with research is requested, we need information about the topic, the type of work, and the type of information being sought. The more precisely this is described, the more specifically we can help. Contact at: <https://www.zbw.eu/de/service/beratung>.
- (40) Interlibrary loan involves ordering books or journals from another library. This happens when your own library does not have the titles you are looking for in its collection. A small fee of usually 1,50 € is charged for this service. The order takes a few days, as the volume is usually delivered by mail.
- (41) The ZBW does not offer interlibrary loan for users. We do not order books from other libraries. However, our own holdings are sent to other libraries through interlibrary loan.
- (42) The document delivery service subito can be used to order books from libraries directly to your home. This is more expensive than interlibrary loan, but often faster. More information and registration can be found at this website: <https://www.subito-doc.de/>.
- (43) With direct delivery, books or articles are sent from the ZBW to users. This costs 9 € per book and 5,50 € for an article. Payment is made by invoice. A special condition applies to ZBW users: if several books are ordered, the first book costs 9 € and each additional book is charged at 2 €. Here, too, payment is by invoice.
- (44) Found items will be collected at the service counter. Please contact us by phone to find out if your missing items have been found.
- (45) Snack and beverage vending machines are available at both locations.

Answer to "How to reach you?" / "Where can I get more info?" questions (optional):

Please contact us via mail: [info@zbw.eu](mailto:info@zbw.eu) or telephone: Kiel at 0431/8814-555 and in Hamburg at 040/42834-219.

### A.3 Dokumentation der Test-Durchläufe

Eine ausführliche Dokumentation der Durchläufe findet sich auf dem angelegten GitLab-Repository der TH Wildau:

<https://git.th-wildau.de/albe6395/evaluation-des-sprachmodells-gpt-3>

→ G2\_DokumentationTestDurchläufe.pdf

Dies beinhaltet die von den Teilnehmenden geführten Chatverläufe, ein Transkript des Gesprächs während der Durchführung sowie die Beantwortung der Fragen.

## A.4 Verzeichnis der auf der beiliegenden CD gespeicherten Dateien

Masterarbeit als PDF-Version

Vollständiger Programmcode aller Engines zu den Beispielen aus Kapitel 8

Vollständige Dokumentation der durchgeführten Tests

## A.5 Verzeichnis des GitLab-Repository

Link: <https://git.th-wildau.de/albe6395/evaluation-des-sprachmodells-gpt-3>

README.md

➔ Beschreibung des Repository

Masterarbeit\_GPT3\_Behrens.pdf

➔ Masterarbeit als PDF-Version

G1\_ProgrammcodeKapitel8.py

➔ Vollständiger Programmcode aller Engines zu den Beispielen aus Kapitel 8

G2\_DokumentationTestDurchläufe.pdf

➔ Vollständige Dokumentation der durchgeführten Tests

## Selbstständigkeitserklärung

Ich versichere, dass die vorliegende Arbeit von mir selbständig und ohne unerlaubte Hilfe angefertigt worden ist. Ich habe alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen sind, durch Zitate bzw. Literaturhinweise als solche kenntlich gemacht.

Hamburg, 09.09.2022

Alena Behrens

Alena Behrens