

Johann Wolfgang Goethe-Universität, Frankfurt am Main
Institut für Informatik
Seminar: Künstliche Intelligenz
Seminarleitung: Prof. Dr. Manfred Schmidt-Schauß
Sommersemester 2023



ChatGPT

vorgelegt von Amine Karabila, Cem Yenel, Sandip Singh Nijher

Amine Karabila
Matrikel Nr. 6611011
2. Fachsemester

Cem Yenel
Matrikel Nr. 6620866
2. Fachsemester

Sandip Singh Nijher
Matrikel Nr. 6608778
2. Fachsemester

Masterstudiengang Informatik

Masterstudiengang Informatik

Masterstudiengang Informatik

Datum der Abgabe: 13.07.2023

© 2023 Amine Karabila, Cem Yenel und Sandip Singh Nijher

Inhaltsverzeichnis

Abbildungsverzeichnis.....	IV
1. Einleitung (A.K., C.Y., S.S.N.).....	1
1.1 Motivation und Zielsetzung	1
1.2 Aufbau der Arbeit	1
2. Grundlagen (C.Y.)	3
2.1 Natural Language Processing (NLP)	3
2.2 Verlauf der GPT-Serie	3
2.2.1 GPT-1	4
2.2.2 GPT-2	4
2.2.3 GPT-3	4
2.2.4 GPT-3.5	5
2.2.5 GPT-4	5
3. ChatGPT (C.Y.)	6
3.1 ChatGPT	6
3.2 ChatGPT 2	7
4. ChatGPTs Transformerarchitektur (A.K.)	9
4.1 Vorherige Modelle	9
4.1.1 Recurrent Neural Networks (RNN).....	9
4.1.2 Long Short-Term Memory (LSTM) Networks	10
4.2 Aufbau der Transformerarchitektur.....	11
4.2.1 Decoder- und Encoder-Struktur.....	12
4.2.2 Multi-Head Attention Part	12
4.2.3 Feed-Forward Network	12
4.2.4 Decoder Block	13
4.2.5 Masked Multi-Head Attention Part.....	13
4.2.6 Multi-Head Attention "Decoder" Part	13
4.2.7 Parameter der Attention Schichten	14
4.2.8 Linear & Softmax	14
4.2.9 Trainieren der Verlustfunktion.....	14
5. ChatGPT – Herausforderungen und Grenzen (S.S.N.).....	15
5.1 Limitationen und Einschränkungen	15
5.2 Ethische und rechtliche Aspekte	18
5.3 Datenschutz und Sicherheit	21
6. Fazit	24
Quellen	VI

Bemerkung:

Die vorliegende Seminararbeit wurde kollaborativ von drei Autoren verfasst, wobei die Initialen der jeweiligen Namen am Anfang jedes Kapitels aufgeführt sind.

A.K.: Amine Karabila

C.Y.: Cem Yenil

S.S.N: Sandip Singh Nijher

Abbildungsverzeichnis

Abb. 1: GPT (vgl. Fu 2023)

Abb. 2: GPT-Humor (vgl. BoredGeekSociety 2023)

Abb. 3: GPT-Exam (vgl. OpenAI 2023a: S. 6)

Abb. 4: RNN-Architektur-Modell (vgl. Li 2019)

Abb. 5: rekurrentes neuronales Netz des RNN-Models (vgl. Li 2019)

Abb. 6: Transformerarchitektur (vgl. Sarkar 2022)

Abb. 7: Decoder- und Encoder-Struktur (vgl. Sarkar 2022)

Abb. 8: Multi-Head Attention Part (vgl. Sarkar 2022)

Abb. 9: Feed-Forward Network (vgl. Sarkar 2022)

Abb. 10: Decoder Block (vgl. Sarkar 2022)

Abb. 11: Masked Multi-Head Attention Part (vgl. Sarkar 2022)

Abb. 12: Vergleich zwischen Zielwahrscheinlichkeit und Ausgabewahrscheinlichkeit (vgl. Doshi 2021)

Abb. 13: Kein aktueller Wissensstand (eigene Abbildung, aufgenommen am 11.06.2023)

Abb. 14: Begrenzte Verständnisfähigkeit (eigene Abbildung, aufgenommen am 08.07.2023)

Abb. 15: Generieren von Fehlinformationen (eigene Abbildung, aufgenommen am 04.06.2023)

1. Einleitung (A.K., C.Y., S.S.N.)

1.1 Motivation und Zielsetzung

Die vorliegende Seminararbeit behandelt das Thema "ChatGPT" und ergründet dessen wachsende Bedeutung in verschiedenen Branchen. Die Motivation hinter dieser Arbeit beruht auf dem Interesse an den Auswirkungen von ChatGPT sowie dem Erkennen der Relevanz, die diese Technologie zunehmend für die Gesellschaft annimmt.

Die Bedeutung von ChatGPT wächst stetig. „KI-basierte Sprachmodelle wie ChatGPT sind in der Lage, erstaunlich gute Texte zu nahezu allen Themen automatisiert zu produzieren.“ (von Platen 2023)

Diese Entwicklung hat weitreichende Auswirkungen auf verschiedene Branchen. Insbesondere im wissenschaftlichen Bereich eröffnet ChatGPT neue Möglichkeiten. Die automatisierte Generierung von Texten hilft beispielsweise Studenten Zeit bei der Erstellung von Forschungsfragen oder Gliederungen zu sparen. (vgl. Solis) Nora Hoffmann meint, dass es sinnvoll sei Studierende über Grenzen, Gefahren und Nutzen ChatGPTs aufzuklären, da eine solche Kompetenz im Umgang mit KI-Tools im beruflichen und wissenschaftlichen Kontext wichtig ist. (vgl. UniReport 2.23 2023)

Die Auswirkungen von ChatGPT sind nicht auf den wissenschaftlichen Bereich beschränkt. Auch der Journalismus und andere Branchen sind von dieser Technologie betroffen. ChatGPT ermöglicht es Journalisten und Redakteuren, Texte schneller zu erstellen und produktiver zu arbeiten. Allerdings besteht auch die Herausforderung, zwischen Texten, die von Menschen und solchen, die von Maschinen generiert wurden, zu unterscheiden.

In Anbetracht dieser Entwicklungen ist es von großer Bedeutung, ein generelles Verständnis für die Funktion und Architektur von ChatGPT zu erlangen und die Grenzen und Herausforderungen dieser Technologie zu beleuchten. Es ist essenziell, den Einsatz von ChatGPT verantwortungsvoll zu gestalten und sich bewusst zu sein, dass automatisch generierte Texte zwar effizient sein können, aber auch ethische Fragen aufwerfen.

1.2 Aufbau der Arbeit

In dieser Seminararbeit werden erst die Grundlagen behandelt. Hierbei wird das grundlegende Konzept des „Neuronal Language Processing“ erörtert. Im Anschluss erfolgt eine eingehende Betrachtung der Entwicklungsgeschichte der GPT-Versionen. Und im abschließenden Abschnitt werden die verschiedenen GPT-Modelle miteinander verglichen. Hierbei werden ihre architektonischen Unterschiede und spezifischen Merkmale hervorgehoben. Durch die Analyse von Kriterien wie Modellgröße, Trainingsdaten und Leistungsfähigkeit wird eine umfassende Bewertung der einzelnen GPT-Modelle vorgenommen.

Das zweite Kapitel widmet sich einer detaillierten Untersuchung des ChatGPT-Modells. Insbesondere werden die beiden Versionen von ChatGPT im Fokus stehen. Diese Versionen haben sich als bedeutend für verschiedene Anwendungsfälle erwiesen, insbesondere im Bereich des Chatbots und der natürlichen Sprachinteraktion.

Im nächsten Kapitel, mit dem Titel "Die Transformerarchitektur ChatGPTs", wird eingehend auf die Architektur des ChatGPT-Modells eingegangen. Zunächst werden in Abschnitt 3.1 die Recurrent Neural Networks (RNNs) behandelt, die zu den traditionellen Architekturen für die Verarbeitung

sequenzieller Daten gehören. Es werden ihre grundlegenden Funktionsprinzipien erläutert und deren Einsatzmöglichkeiten für textgenerierende Modelle diskutiert. Daraufhin widmet sich der nächste Abschnitt den Long Short-Term Memory (LSTM) Networks, einer speziellen Variante der RNNs, die entworfen wurde, um das Problem des Verschwindens oder des Explodierens des Gradienten in langen Sequenzen zu bewältigen. Es werden die Struktur und die Funktionsweise von LSTM-Netzwerken detailliert betrachtet und ihre Vorteile für die Modellierung von Abhängigkeiten in Texten erläutert.

Vor dem Fazit wird im Kapitel "ChatGPT - Herausforderungen und Grenzen" zunächst eingehend auf die verschiedenen Limitationen von ChatGPT eingegangen und es werden Lösungsansätze diskutiert, um diese Limitationen zu überwinden. Danach werden die ethischen und rechtlichen Aspekte im Zusammenhang mit ChatGPT ausführlich besprochen. Es werden die Auswirkungen dieser Aspekte auf die Gesellschaft, individuelle Rechte und die Verantwortung der Entwickler und Benutzer von ChatGPT untersucht. Anschließend werden die Datenschutzbestimmungen im Kontext von ChatGPT erläutert. Besonderes Augenmerk liegt hierbei auf dem Schutz der Benutzerdaten. Es werden die Maßnahmen und Strategien beschrieben, die ergriffen werden, um den Schutz der Benutzerdaten zu gewährleisten. Des Weiteren werden die Datenschutzbestimmungen im Kontext von ChatGPT erläutert. Es werden die Maßnahmen und Strategien beschrieben, die ergriffen werden, um den Schutz der Benutzerdaten zu gewährleisten.

2. Grundlagen (C.Y.)

2.1 Natural Language Processing (NLP)

Natürliche Sprachverarbeitung (Natural Language Processing, NLP) ist ein Fachgebiet der künstlichen Intelligenz, das darauf abzielt, Computern das Verständnis, die Interpretation und die Manipulation menschlicher Sprache zu ermöglichen. ChatGPT ist ein fortschrittliches Sprachmodell, das im Bereich des NLP eingesetzt wird. Es wurde entwickelt, um menschenähnliche Konversationen zu führen und komplexe Dialoge zu verstehen und darauf zu reagieren. (vgl. Radford et al. 2018: S.1f)

Es gibt Diskussionen darüber, ob ChatGPT den Turing-Test bestanden hat, basierend auf Aussagen eines Data Scientists von BuzzFeed „congrats to OpenAI on winning the Turing Test“(Woolf 2022). Bisher wurde jedoch noch kein formeller Test durchgeführt, um dies zu bestätigen. ChatGPT betont in seinen Antworten stets, dass es sich um ein Sprachmodell handelt. Außerdem wurde ChatGPT nicht speziell darauf programmiert, den Turing-Test zu bestehen. (vgl. Somoye 2023)

2.2 Verlauf der GPT-Serie

Im Jahr 2018 führte OpenAI das erste GPT-Modell (GPT-1) ein und veröffentlichte eine wegweisende Arbeit mit dem Titel "Improving Language Understanding by Generative Pre-Training". Dieses Modell basierte auf der innovativen Transformer-Architektur, die es ermöglichte, komplexe Zusammenhänge in Texten zu erfassen.

Ein Jahr später stellte OpenAI GPT-2 vor, ein deutlich größeres Modell mit erweiterten Fähigkeiten in der Textgenerierung. GPT-2 war in der Lage, kohärente und zusammenhängende Texte zu erzeugen, die oft nur schwer von menschlich verfassten Texten zu unterscheiden waren. Diese bahnbrechende Leistung führte zu großem Interesse und Diskussionen in der KI-Gemeinschaft.

Im Jahr 2020 folgte schließlich die Einführung von GPT-3, einem noch leistungsstärkeren Modell im Vergleich zu seinem Vorgänger. GPT-3 beeindruckte die Fachwelt mit seiner beeindruckenden Skalierung, da es über 100-mal so viele Parameter wie GPT-2 besaß. Dank dieser enormen Kapazität konnte GPT-3 unterschiedlichste Aufgaben bewältigen, selbst wenn nur wenige Beispiele als Eingabe zur Verfügung standen. Es konnte Fragen beantworten, Übersetzungen durchführen, Texte vervollständigen und vieles mehr.

Als Weiterentwicklung von GPT-3 wurde GPT-3.5 entwickelt, das als Grundlage für das innovative Chatbot-Produkt ChatGPT diente. Mit ChatGPT konnten Nutzer natürliche und dynamische Konversationen mit einem KI-basierten System führen. Es bot eine reibungslose Interaktion und ermöglichte es den Nutzern, Fragen zu stellen, Ratschläge einzuholen oder einfach eine unterhaltsame Konversation zu führen.

Das aktuellste Modell und die Weiterentwicklung von GPT-3.5 ist GPT-4. Es handelt sich um ein multimodales großes Sprachmodell, das am 14. März 2023 veröffentlicht wurde.

OpenAI arbeitet kontinuierlich daran, die Leistungsfähigkeit von ChatGPT und ähnlichen Modellen zu verbessern. Durch den Einsatz von Feedback-Mechanismen und kontinuierlicher Forschung strebt OpenAI an, die Fähigkeiten dieser Modelle weiter zu erweitern und gleichzeitig ihre Stärken zu verbessern, um den Nutzern eine noch bessere Erfahrung zu bieten.

2.2.1 GPT-1

GPT-1 hat mit seinen 117 Millionen Parametern andere damalige Sprachmodelle überschattet. Einer der größten Stärken von GPT-1 war es zusammenhängende und koherente Sprache zu generieren. Die Trainingsdaten bestanden aus Common Crawl, ein Datenset von Webseiten und BookCorpus, ein Datenset aus über 11000 Büchern.

Die größte Schwäche von GPT-1 war es oft wiederholende Antworten zu geben. Auch konnte das Modell nicht über mehrere Dialoge argumentieren und konnte langfristige Abhängigkeiten im Text nicht nachverfolgen. (vgl. Ali 2023)

2.2.2 GPT-2

GPT-2 wurde mit mehr als das zehnfache veröffentlicht. Es besaß 1.5 Milliarden Parameter und wurde mit zusätzlich Webtexten trainiert. Es konnte realistische und kohärente Antworten geben. Durch die menschlich ähnlichen Antworten war es ein wichtiges Tool für verschiedene Aufgaben. Nichtsdestotrotz hatte es Schwierigkeiten mit komplexeren Aufgaben und Kontexte zu verstehen. (vgl. Ali 2023)

2.2.3 GPT-3

Mit GPT-3 kam ein erneuter großer Sprung. 175 Milliarden Parameter, mit mehreren verschiedenen Datenquellen, die zusammenergebend fast Billionen Wörter beinhalten, womit es anspruchsvolle Antworten geben kann. Hier wurden neue Techniken eingeführt. Zunächst das zero-shot-learning. Hierbei kann eine Aufgabe ohne Beispiele eine Antwort vorhersagen. Anschließend das one-shot-learning. Hier wird nur ein Beispiel neben der Aufgabe gegeben, womit das Modell die Antwort vorhersagen kann und zum Schluss das few-shot-learning. Hier werden einige Beispiele mit der Aufgabe dem Modell gezeigt.

Jedoch hat auch dieses Modell einige Fehler. So werden unangebrachte oder ungenaue Antworten wiedergegeben. (vgl. Ali 2023)

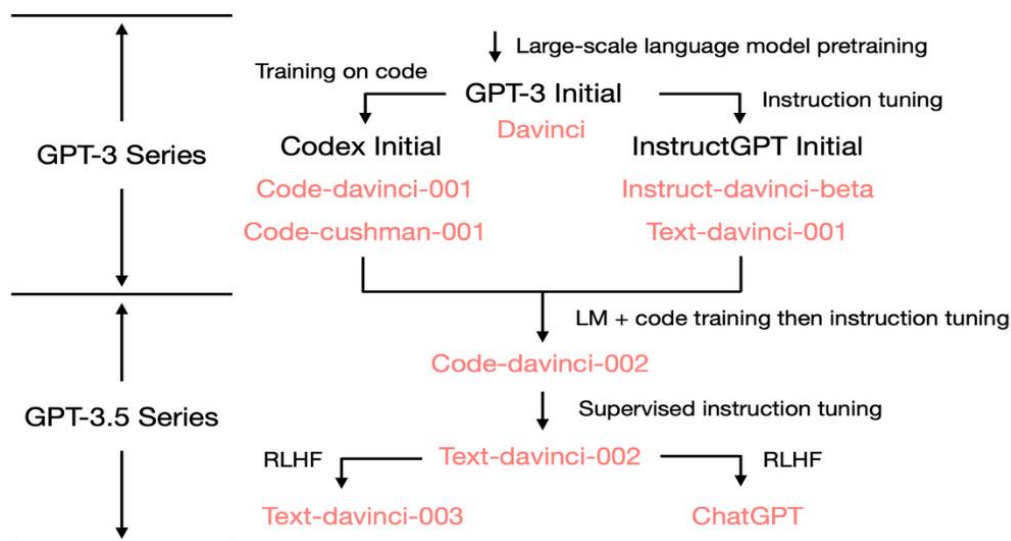


Abb. 1: GPT (vgl. Fu 2023)

GPT-3 besitzt nur ein Modell namens „davinci“, welches ein OpenAI API Modell ist. Anschließend wurden mit der Zeit von OpenAI InstructGPT eingeführt, welches mit Supervised-Learning mit menschlichem Feedback. Dadurch sollten die Probleme der Halluzinationen¹, dass Falschinformationen ausgegeben wurden, und dass Anweisungen eingehalten werden sollen, behoben werden. Dabei wurde es im insgesamt schlechter. (vgl. Fu 2023)

Außerdem gab es noch Codex in der GPT-3 Serie. Dieser analysierten natürlichen Sprache als Eingabe und gab dann Code als Ausgabe. Codex wurde mit zusätzlich GitHub Code trainiert und wird heutzutage für GitHub Copilot benutzt.

2.2.4 GPT-3.5

Durch das Zusammenfügen von InstructGPT und Codex kam das neue Modell code-davinci-002, womit die GPT-3.5 Serie anfängt. In dieser Serie wurden drei Modelle vorgestellt mit den Parametern 1.3 Milliarden 6 Milliarden und 175 Milliarden. Nach Text-davinci-002 konnte dann durch Reinforcement Learning mit menschlichem Feedback, sowohl Text-davinci-003 und auch ChatGPT veröffentlicht werden. (vgl. Fu 2023)

2.2.5 GPT-4

Über GPT-4 wurden keine näheren Informationen von OpenAI veröffentlicht, außer dass es nicht mit mehr Daten trainiert wurde als GPT-3.5. Jedoch spekuliert man, dass GPT-4 mit 100 Billionen Parametern besitzt. OpenAI behauptet außerdem, dass die Antworten von GPT-4 mit einer Wahrscheinlichkeit von 40% ist, sachlich richtige Angaben als GPT-3.5 zu geben. Es können sowohl Text als auch Bilder als Eingabe eingefügt werden. (vgl. Fu 2023)

¹ Der Begriff meint damit, dass ChatGPT ungenaue oder irreführende Informationen generiert, die nicht der Realität entsprechen.

3. ChatGPT (C.Y.)

3.1 ChatGPT

ChatGPT ist ein Chatbot, der über eine grafische Benutzeroberfläche mit Nutzern kommunizieren kann. Hierbei wurde das GPT-3.5-Modell feinabgestimmt. Am 30. November 2022 wurde er kostenlos zugänglich gemacht und erzielte bereits in den ersten Monaten enorme Erfolge. In den ersten fünf Tagen registrierten sich eine Million Nutzer. Zum Vergleich: Instagram benötigte 2,5 Monate, um dieselbe Nutzerzahl zu erreichen und war vor ChatGPT die, die am schnellsten die 1 Millionen geknackt haben. (vgl. Janson 2023) ChatGPT knackte die Marke von 100 Millionen Nutzern bereits nach zwei Monaten. (vgl. Hannemann 2023) Zudem konnte OpenAI eine Partnerschaft im Wert von zehn Milliarden US-Dollar mit Microsoft abschließen, um die Funktionen von ChatGPT in die Suchmaschine Bing zu integrieren. (Spiegel Netzwelt 2023b)

Der Erfolg des Chatbots basiert auf seinen beeindruckenden Antworten. Es können Fragen zu einer Vielzahl von Themen gestellt werden, angefangen bei Produktbeschreibungen bis hin zur Formulierung von Gedichten. ChatGPT beherrscht mehrere Sprachen und kann entsprechend antworten.

Darüber hinaus beeindruckt ChatGPT durch seine Fähigkeit, natürliche und flüssige Konversationen zu führen. Er versteht komplexe Fragen und kann kontextbezogene Antworten generieren. Dies ermöglicht den Nutzern eine interaktive und reibungslose Kommunikation mit dem Chatbot.

Ein weiterer Grund für den schnellen Erfolg von ChatGPT liegt in seiner Benutzerfreundlichkeit. Die grafische Benutzeroberfläche erleichtert die Interaktion und macht den Chatbot für eine breite Nutzerbasis zugänglich. Sowohl technisch versierte Benutzer als auch solche ohne technisches Fachwissen können problemlos mit dem Chatbot interagieren und von seinen vielseitigen Funktionen profitieren.

Die Fähigkeit von ChatGPT, mehrere Sprachen zu verstehen und zu verwenden, hat ebenfalls zu seinem schnellen Wachstum beigetragen. Nutzer aus verschiedenen Ländern und Kulturen können mit dem Chatbot in ihrer bevorzugten Sprache kommunizieren, was die Nutzerbasis erweitert.

Es ist wichtig anzumerken, dass der Erfolg von ChatGPT auch mit einigen Herausforderungen einhergeht. Fragen zur Sicherheit, zum Datenschutz und zur ethischen Verantwortung bei der Nutzung von KI-Technologien müssen berücksichtigt werden. Zudem basiert der Wissensstand von ChatGPT auf Informationen bis September 2021 (vgl. Humpa 2023) und er kann daher keine Ereignisse danach behandeln. OpenAI ist sich dieser Herausforderungen bewusst und arbeitet kontinuierlich daran, den Chatbot weiterzuentwickeln und zu verbessern, um diese Bedenken anzugehen und eine verantwortungsvolle Nutzung zu fördern.

Aufgrund des großen Erfolgs und der hohen Nutzerzahlen wurde Anfang Januar 2023 das Abonnement ChatGPT Pro eingeführt, das monatlich 20 US-Dollar kostet. Dadurch wird der Zugriff auf ChatGPT für Käufer jederzeit ermöglicht, auch bei hoher Auslastung. Zudem erhalten sie schnellere Antworten und einen bevorzugten Zugriff auf neue Funktionen und Verbesserungen.

3.2 ChatGPT 2

Mit der Veröffentlichung von GPT-4 kamen weitere Verbesserungen. Das neueste GPT-Modell ist über ChatGPT Pro zugänglich. Es ist leistungsfähiger, präziser und verfügt über eine höhere Verarbeitungskapazität als sein Vorgänger. Es kann komplexere Eingaben besser verarbeiten. Die Funktion zur Eingabe von Bildern, die GPT-4 bietet, ist derzeit jedoch nicht in ChatGPT-4 verfügbar. Diese Funktion steht nur über die GPT-4-API zur Verfügung, da GPT-4 noch im Entwicklungsprozess trainiert wird.

OpenAI hat bisher keine konkreten Informationen zu einer möglichen kommerziellen Nutzung von GPT-4 veröffentlicht. Es wird weiterhin daran gearbeitet. Mit GPT-4 können jedoch spannende Möglichkeiten entstehen, wie beispielsweise die Möglichkeit, ein Bild des eigenen Kühlschranks aufzunehmen und nach einem passenden Rezept zu fragen, das der Chatbot anhand des Bildes auswählt. Auch das Zeichnen einer Webseite und das Anfragen des dazugehörigen Quellcodes sowie die Fähigkeit von GPT-4, Humor aus Bildern zu verstehen und zu interpretieren, sind denkbar. (vgl. BoredGeekSociety 2023)

GPT-4 visual input example, Chicken Nugget Map:

User Can you explain this meme?

Sometimes I just look at pictures of
the earth from space and I marvel at
how beautiful it all is.



GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

Table 18: Example prompt demonstrating GPT-4's visual input capability.

Abb. 2: GPT-Humor (vgl. BoredGeekSociety 2023)

Hier sehen wir ein Beispiel, in dem GPT-4 den Witz im Bild erklärt.

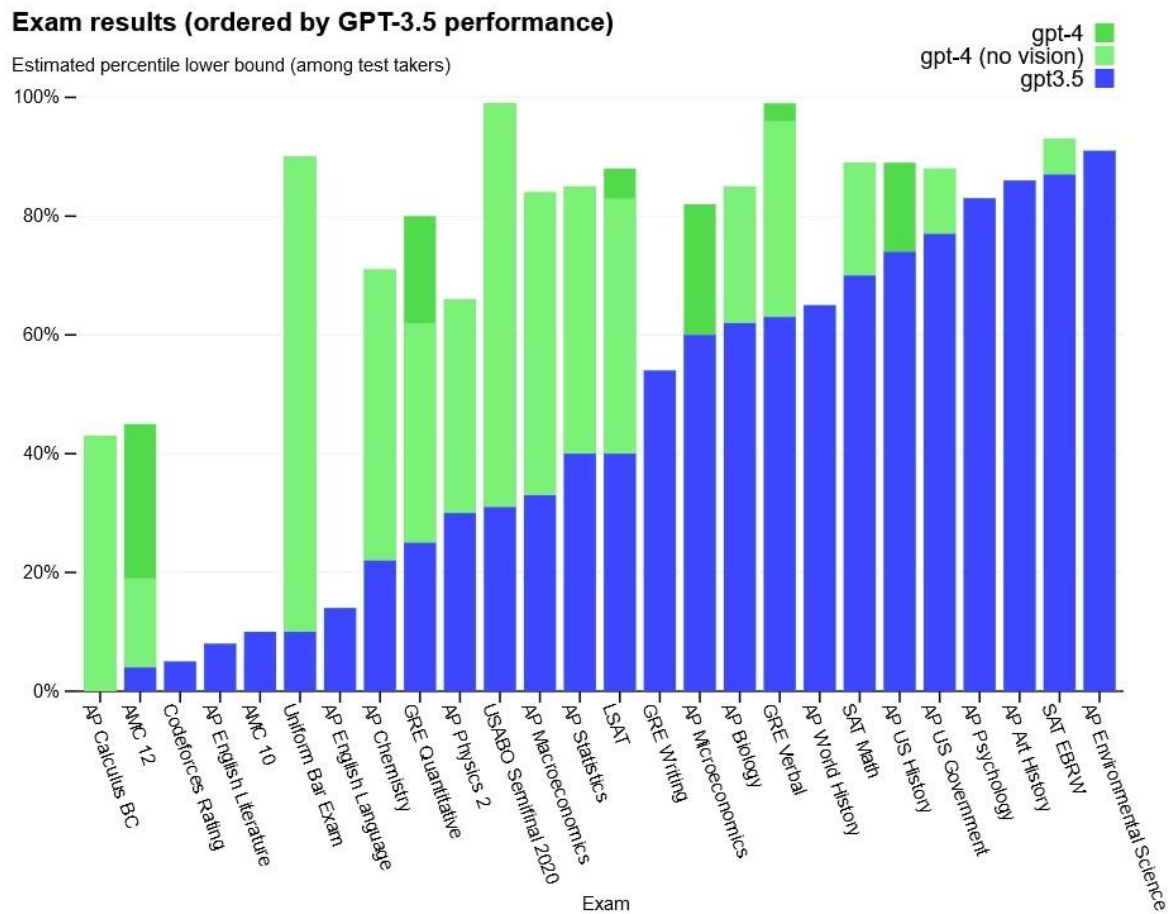


Abb. 3: GPT-Exam (vgl. OpenAI 2023a: S. 6)

OpenAI hat auch eine Statistik von Prüfungen veröffentlicht. Es ist jedoch zu beachten, dass die Daten von OpenAI selbst stammen und darauf abzielen, die Modelle so gut wie möglich aussehen zu lassen. In dieser Statistik konnte GPT-4 ähnlich gut oder sogar besser abschneiden als sein Vorgänger GPT-3.5. (vgl. OpenAI 2023a: S. 6f.)

Sowohl GPT-4 als auch ChatGPT-4 konnten das Problem der Halluzination, das bereits bei ihrem Vorgänger vorhanden war, nicht vollständig beheben. Der Wissensstand der Trainingsdaten endet ebenfalls im September 2021, (vgl. Humpa 2023) daher können keine Antworten auf Fragen gegeben werden, die nach diesem Zeitpunkt aufgetreten sind.

4. ChatGPTs Transformerarchitektur (A.K.)

Im Jahr 2017 veröffentlichten Forscher eine neue neuronale Netzarchitektur namens „Transformator“, die die Grundlage für die meisten exponentiellen Fortschritte in dem Bereich des Maschinellen Lernens, in den letzten vier Jahren war. Das Papier („Attention Is All You Need“) trug den Titel „Aufmerksamkeit ist alles“, denn die zentrale Idee dieser Architektur ist es, sich auf das Prinzip der „Selbstaufmerksamkeit“ zu verlassen anstatt nur auf die Feedback-Schleife, die in RNNs zu finden ist, auf das im folgenden Kapitel kurz eingegangen wird. Die Transformatorarchitektur ist im Vergleich zu anderen Architekturen in ihrem Design etwas Einzigartiges und ist im Detail nur schwer zu verstehen, weswegen in diesem Kapitel nur grundlegend versucht wird das Model zu erklären.

4.1 Vorherige Modelle

4.1.1 Recurrent Neural Networks (RNN)

Das „Recurrent Neural Networks“ war ein echter Durchbruch im Bereich des Deep Learnings, dessen Konzept im Jahre 1986 entwickelt wurde (vgl. Sherstinsky 2020: S.6), da zum ersten Mal auch die Berechnungen aus der jüngeren Vergangenheit mit in die aktuelle Berechnung mit einbezogen wurden und dadurch die Ergebnisse in der Sprachverarbeitung deutlich verbessert waren.

RNN's haben die Besonderheit sequenzielle Informationen zeitlich zu interpretieren, was bedeutet, dass jede Eingabe dieser Sequenz in einer gewissen Beziehung zu ihren Nachbarn steht oder einen gewissen Einfluss auf diese hat. Dabei werden Aktivierungen² in A von vorherigen Knoten genutzt (siehe Abb. 1), um diese bei neueren Anfragen zu interpretieren und dabei die Ausgabe h_t zu beeinflussen.

Am Beispiel einer Eingabe eines Textes in einen Prompt, kann eine Entitätenextraktion³ erfasst werden, sodass Wörter, die vor und nach der Entität im Satz stehen, einen direkten Einfluss darauf haben, wie diese Worte klassifiziert werden. Daher ist es in diesem Beispiel wichtig, Algorithmen zu nutzen, die darauf ausgelegt sind, aus vergangenen Daten in der Sequenz zu lernen. (vgl. Donges 2023: S. 1)

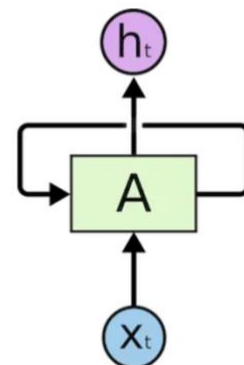


Abb. 4: RNN-Architektur-Modell
(vgl. Li 2019)

Die Schleifen können als eine rekurrentes neuronales Netz gedachte werden, die man sich als mehrere Kopien desselben Netzes A vorstellen kann. Dabei kann jedes Netz eine Nachricht an einen Nachfolger weitergeben. Das rekurrente neuronale Netz gibt die Informationen der vorherigen Wörter an das nächste Netz weiter, das diese Informationen verwenden und verarbeiten kann. Dabei können Entitäten berücksichtigt werden und zusammenhänge in verschiedenen Sätzen durch ihre Klassifizierung überblickt werden. In einem vereinfachten Beispiel würde das RNN-Model

² Im Zusammenhang mit neuronalen Netzen bezieht sich der Begriff „Aktivierung“ auf die Ausgabe eines Neurons oder einer Schicht nach Anwendung einer Aktivierungsfunktion auf die gewichtete Eingabe

³ Bei der Entitätenextraktion werden in einem Text Schlüsselemente identifiziert sowie klassifiziert und vorab definierten Kategorien zugeordnet

beispielsweise erkennen, dass im Satz: „Die Wolken sind am ____“ das Wort „Himmel“ benötigt wird, falls das Neuronale Netz darauf trainiert werden würde, solche Lücken in Sätzen zu füllen. Bei längeren Sätzen ist eine Kontexterhaltung dann aber doch etwas schwieriger. Bei dem leicht veränderten Satz: „Die Wolken, die teilweise fließend ineinander übergehen und tief hängen, sind am ____“ wird das RNN-Model schon deutlich schwieriger auf das Wort „Himmel“ schließen. Um diesen Kontext Verlust etwas entgegenzuwirken wurde 1997 das LSTM-Model entwickelt, auf das im folgenden Kapitel versucht wird, kurz einzugehen.

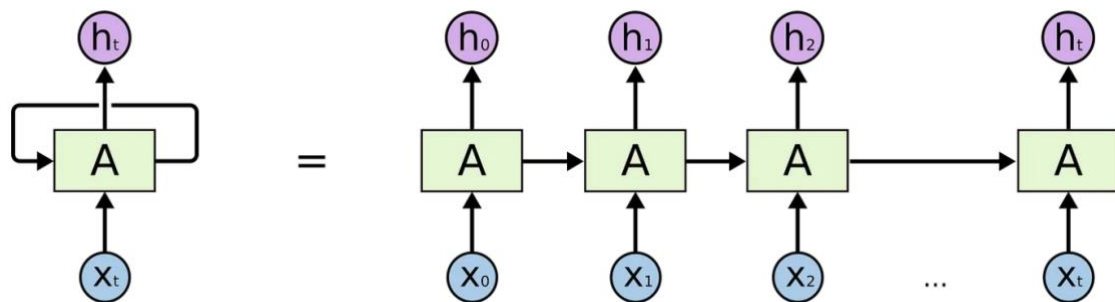


Abb. 5: rekurrentes neuronales Netz des RNN-Modells (vgl. Li 2019)

4.1.2 Long Short-Term Memory (LSTM) Networks

Bei dem Long Short-Term Memory Modell, auch LSTM handelt es sich um eine abgewandte unterform des schon behandelten RNN-Modells und ist eine Erweiterung der Gedächtniskontrolle. Sie kann das Problem des RNN, des verminderten Gedächtnisses gut lösen, indem Sie ausgewählte Informationen auswählt und im Langzeitgedächtnis, dem cell-state, beibehält. Jede Zelle nimmt als Eingaben x_t (einen Token einer Prompteingabe), den vorherigen Zellzustand und die Ausgabe der vorherigen Zelle. Es manipuliert diese Eingaben und basierend auf ihnen erzeugt es einen neuen Zellzustand und eine Ausgabe.

Das gleiche Problem wie bei RNNs im Allgemeinen tritt auch bei LSTMs auf. Wenn Prompt-Eingaben zu lang werden, schneiden LSTM-Modelle immer noch nicht gut ab. Der Grund dafür ist, dass die Kontexterhaltung zwischen zwei Wörtern in einem Prompt, exponentiell zum Abstand zum Wort hin, abfällt. (vgl. Kjærran 2021: S. 6) Das bedeutet, dass das Modell bei langen Sätzen oft den Inhalt von weiter entfernten Elementen in der Sequenz vergisst. Ein weiteres Problem bei RNNs und LSTMs ist, dass es schwierig ist, die Verarbeitung von Sätzen zu parallelisieren, sodass man stattdessen hierbei, Wort für Wort den Satz verarbeiten muss (vgl. Verwimp 2017: S. 2), was dabei zu einer großen Auslastung führt. Zusammenfassend lässt sich sagen, dass LSTMs und RNNs Modelle, zwei Probleme aufweisen:

- Sequenzielle Berechnung hemmt die Parallelisierung
- Exponentieller Kontextverlust

4.2 Aufbau der Transformerarchitektur

Der Übergang von LSTM (Long Short-Term Memory)-Modell zur Transformer-Technologie erfolgte aufgrund der Entwicklung eines leistungstärkeren Modells für die Verarbeitung von Sequenzdaten⁴. Das Transformer-Modell nutzt eine aufmerksamkeitsbasierte Architektur anstelle von rekurrenten Schichten⁵, was eine parallele Verarbeitung ermöglicht und die Modellierung von Abhängigkeiten zwischen den Elementen einer Sequenz verbessert. Durch seine Fähigkeit, komplexe Zusammenhänge in Sequenzen zu erkennen, hat die Transformer-Architektur bei Aufgaben, wie maschinellem Übersetzen und Sprachverarbeitung große Erfolge erzielt, während das LSTM weiterhin in bestimmten Aufgaben und Situationen nützlich bleibt. Sie wird aktuell vom Sprachenübersetzer Google sowie ChatGPT genutzt.

Bei der Transformer Architektur werden die Eingabesequenz parallel weitergegeben, so die GPU effektiv genutzt werden kann und die Trainingsgeschwindigkeit erhöht werden konnte. Außerdem basiert das Modell auf eine mehrköpfigen Aufmerksamkeitschicht, so dass das Problem des verschwindenden Gradienten⁶ leicht zu lösen war. In einem Übersetzer, der aus einem einfachen RNN besteht, gibt man zum Beispiel eine Sequenz oder einen Satz kontinuierlich ein, um Wortembeddings zu erzeugen. Da jedes Wort vom vorherigen Wort abhängt, verhält sich sein verborgener Zustand entsprechend, so dass dieser Schritt für Schritt eingegeben werden muss. In der Transformerarchitektur kann man jedoch alle Wörter eines Satzes übergeben und die Wortembedding gleichzeitig bestimmen.

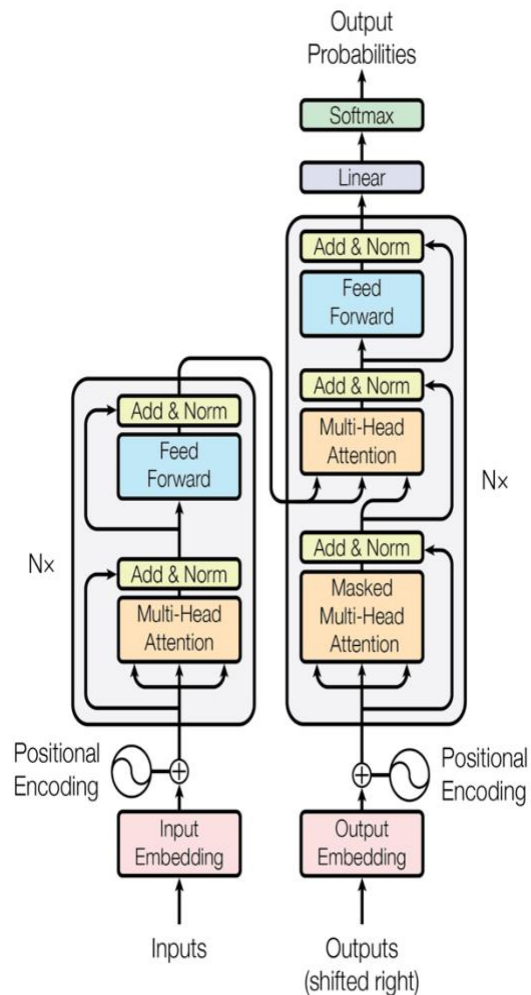


Abb. 6: Transformerarchitektur
(vgl. Sarkar 2022)

⁴ Eine spezielle Art von Daten, die eine Abfolge von Elementen repräsentieren. Diese Elemente können beispielsweise Zeichen, Wörter, Symbole, Zahlen oder andere diskrete Einheiten sein.

⁵ Rekurrente Schichten in RNN und LSTM-Modellen ermöglichen die Verarbeitung von Sequenzdaten, indem sie zeitliche Abhängigkeiten erfassen und Informationen über vergangene Schritte speichern, um komplexe Muster und Zusammenhänge in den Sequenzen zu erkennen.

⁶ Der Gradient in Bezug auf die Transformer-Architektur bezieht sich auf den mathematischen Begriff des Gradienten Abstiegs, wodurch der Transformer-Mechanismus die Fähigkeit erhält, Muster und Beziehungen in den Eingabedaten zu lernen.

4.2.1 Decoder- und Encoder-Struktur

Da das ChatGPT-Tool, nicht mit Prompt Eingaben umgehen kann, müssen die Prompt Eingaben in Vektoren oder Matrizen übersetzt werden. Man muss also die Worte in einen Vektor⁷ umwandeln. Hier kommt das Konzept des Embedding Space ins Spiel. In diesem Bereich werden Worte mit ähnlicher Bedeutung gruppiert. Dies wird als Embedding Space bezeichnet und hier wird jedes Wort entsprechend seiner Bedeutung abgebildet und mit einem bestimmten Wert versehen, sodass die Wörter auf bestimmte Vektoren abgebildet werden. Ein weiteres Problem, mit dem man konfrontiert wird, ist dass jedes Wort in verschiedenen Sätzen unterschiedliche Bedeutungen haben. Um dieses Problem zu lösen, verwendet man einen Positionskodierer. Dabei handelt es sich um Vektoren, die den Kontext entsprechend der Position des Wortes in einem Satz angeben.

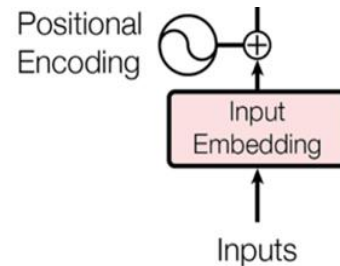


Abb. 7: Decoder- und Encoder-Struktur (vgl. Sarkar 2022)

4.2.2 Multi-Head Attention Part

Die „Multi-Head Attention Part“ beinhaltet die Selbstaufmerksamkeit. Diese konzentriert sich darauf, wie relevant, bestimmte Worte im Verhältnis zu anderen Worten im Satz sind. Sie wird als „Aufmerksamkeitsvektor“ dargestellt. Für jedes Wort kann man einen Aufmerksamkeitsvektor generieren, der die kontextuelle Beziehung zwischen den Wörtern in diesem Satz erfasst. Jedes Wort erhält seinen eigenen Wert, der durch die Interaktion mit anderen Worten im Satz bestimmt wird. Hierbei können es zu mehreren Aufmerksamkeitsvektor für jedes einzelne Wort kommen, da je nach Kontext zu allen Worten, ein Wert bestimmt wird. Man bestimmt also mehrere Aufmerksamkeitsvektoren pro Wort und bildet einen gewichteten Durchschnitt, um den endgültigen Aufmerksamkeitsvektor für jedes Wort zu berechnen. Da man mehrere Aufmerksamkeitsvektoren verwendet, wird dieses Verfahren als Mehrkopf-Aufmerksamkeitsblock bezeichnet.

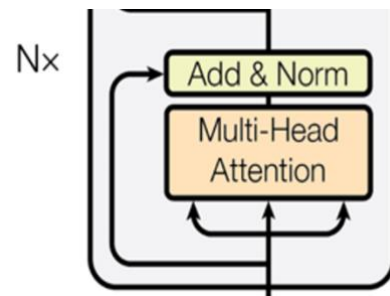


Abb. 8: Multi-Head Attention Part (vgl. Sarkar 2022)

4.2.3 Feed-Forward Network

Im nächsten Schritt wird man zum neuronale Feed-Forward-Netz weitergeleitet. Ein einfaches neuronales Feed-Forward-Netzwerk wird auf jeden Aufmerksamkeitsvektor angewandt, um die Aufmerksamkeitsvektoren in eine Form umzuwandeln, die für die nächste Encoder- oder Decoder Schicht akzeptabel ist. (vgl. Vaswani et al. 2017: S. 3). Hierbei werden die Aufmerksamkeitsvektoren nacheinander und unabhängig voneinander Akzeptiert, was ein Vorteil zum RNN-Model ist. Alle Wörter können dadurch nun gleichzeitig in den Encoder-Block eingespeist werden und erhalten die Menge der kodierten Vektoren für jedes Wort gleichzeitig.

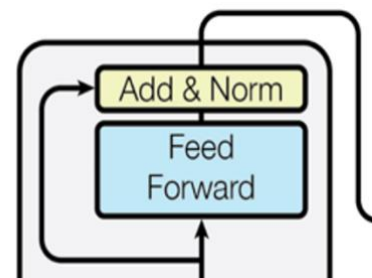


Abb. 9: Feed-Forward Network (vgl. Sarkar 2022)

⁷ In der Transformer-Architektur werden Vektoren verwendet, um Informationen über Wörter, ihre Positionen und ihre Beziehungen zu repräsentieren und zu verarbeiten.

4.2.4 Decoder Block

Beim Tool ChatGPT man man dem Modell Unmengen an Textdaten aus dem Internet und anderen Quellen übergeben (z.B. durch Web-Scraping⁸) und daraufhin die Transformerarchitektur mit den Textdaten trainieren. Die Prompt Eingaben durchlaufen den Encoder-Block und die Texte aus z.B dem Web-Scraping, den Decoder-Block. Zunächst gibt es die Einbettungsschicht und den Positionskodierer, der die Wörter in entsprechende Vektoren umwandelt. Die Vorgehensweise entspricht hierbei dem Kodierungsteil, der bereits besprochen wurde.

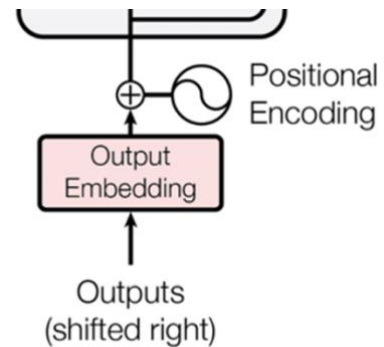


Abb. 10: Decoder Block
(vgl. Sarkar 2022)

4.2.5 Masked Multi-Head Attention Part

Die „Masked Multi-Head Attention“ ermöglicht es dem Modell, auf relevante Informationen im bisherigen Gespräch zuzugreifen und kontextbezogene Antworten zu generieren. Hierbei werden die Eingabe-Prompts sowie die antrainierten Textdaten herangezogen um auf die Frage im Eingabe-Prompt eine passende Antwort zu finden. Beim Vergleich des Eingabe Prompts mit den Textdaten, wird der Matrixwert⁹ aktualisiert. Auf diese Weise lernt er nach mehreren Iterationen welche Tokens aus den Trainingsdaten zu den Eingabe Prompt eine starke Beziehung besitzt (vgl. Wang 2020: S. 3).

Die Besonderheit dieses „Attention Parts“, ist die Maskierung. Sie dient zu trainingszwecken die nur bei solchen Trainingsdurchläufe verwendet wird. Das bedeutet, dass das Modell nur auf vorherige Tokens zugreifen kann und zukünftige Tokens maskiert sind. Damit das Trainieren des Modells stattfinden kann, wäre es sinnlos, wenn es das nächste generierte Token bereits bekannt sei, und aus diesem Grund muss man diese Art der Tokens maskieren.

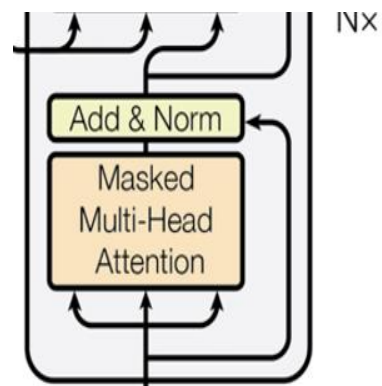


Abb. 11: Masked Multi-Head Attention Part
(vgl. Sarkar 2022)

Am Beispiel einer Übersetzungsanfrage eines Deutschen Satzes in die Französische Sprache, kann jedes beliebige Wort aus dem deutschen Satz entnommen werden, aber man kann nur das vorherige Wort des französischen Satzes für Lernzwecke verwenden. Bei der Parallelisierung mit der Matrixoperation muss man also sicherstellen, dass die Matrix die später erscheinenden Wörter maskiert, indem Sie die Gewichte der einzelnen Beziehungen, in Nullen umwandelt, so dass das Aufmerksamkeitsnetz sie nicht verwenden kann.

4.2.6 Multi-Head Attention “Decoder” Part

Nun werden die resultierenden Aufmerksamkeitsvektoren aus der vorherigen Schicht und die Vektoren aus dem Encoder-Block in einen weiteren Multi-Head-Attention-Block geleitet. Hier kommen auch die Ergebnisse aus dem Encoder-Block ins Spiel. Da es für jeden deutschen und französischen Satz einen Vektor für jedes Wort gibt, führt dieser Block die Zuordnung der deutschen

⁸ Web-Scraping ist der Prozess des automatischen Extrahierens von Daten von Websites.

⁹ Der Matrixwert wird dafür genutzt, um die Gewichtungen für die Aufmerksamkeit zu berechnen.

und französischen Wörter durch und findet die Beziehung zwischen ihnen heraus. Die Ausgabe dieses Blocks sind Aufmerksamkeitsvektoren für jedes Wort, in den deutschen und französischen Sätzen. Jeder Vektor stellt die Beziehung zu anderen Wörtern in beiden Sprachen dar

4.2.7 Parameter der Attention Schichten

Wie in Abb.6 zu sehen ist, sind in jede der Attention Schichten mehrere einfließende Eingänge. Das liegt daran das hierbei drei verschiedene Parameter in die jeweiligen Schichten einfließen. Dabei handelt es sich um die folgenden drei Parameter:

1. **Query:** Das Query repräsentiert das aktuelle Wort in der jeweiligen einfließenden Sequenz, für das die Aufmerksamkeit berechnet wird.
2. **Key:** Der Key Parameter, repräsentiert die anderen Wörter in der Sequenz und beschreibt ihre Beziehung zum aktuellen Wort (Key).
3. **Value:** Der Value Parameter enthält die Informationen, die mit dem aktuellen Wort assoziiert sind und liefert gewichtete Informationen basierend auf die Beziehung zu den anderen Wörtern in der Sequenz.

Eine Besonderheit ist der Zusammenfluss im Multi-Head-Attention Part von verschiedenen Head-Attention Schichten. Den es werden hier die Query Parameter der Decoder-Head-Attention Schicht mit den beiden Key sowie Value Parameter aus der Encoder-Attention Schicht zusammengeführt. Dies erzeugt eine Repräsentation mit Aufmerksamkeitswerten für jedes Wort der Zielsequenz unter Berücksichtigung der Aufmerksamkeitswerte der Eingabesequenz. Die dabei berechneten Werte werden dann an die Linear & Softmax Schicht weitergegeben.

4.2.8 Linear & Softmax

Nach dem Durchlaufen des Feed Forward Blocks, wird der Vektor durch die Linare Schicht geleitet. Hier werden alle Vektoren auf die jeweiligen Positionen in der Ausgabesequenz verteilt und mit bestimmten Scores versehen, die die Einteilung auf die jeweiligen Positionen in der Ausgabesequenz darstellen sollen.

Die Softmax-Schicht wandelt, die Scores aus der Linear Schicht, in eine Wahrscheinlichkeitsverteilung um. Diese Wahrscheinlichkeitsverteilung wird durch das Training mit den Vorhandenen Trainingsdaten bestimmt und bei jedem durchlauf des Modells angepasst. Genaueres ist in Kapitel 4.2.9 zu lesen.

Die resultierenden Worte werden dann mit der höchsten Wahrscheinlichkeit erzeugt. Dieser Schritt wird dann mehrfach parallel für alle Wörter wiederholt und eine endgültige Ausgabe erzeugt. (vgl. Wang 2020: S. 2)

4.2.9 Trainieren der Verlustfunktion

Beim Trainieren der Verlustfunktion der Transformer Architektur, handelt es sich im Allgemeinen um das Anpassen der Wahrscheinlichkeitsverteilung in der Softmax-Schicht. Das Ziel des Transformers ist es die Zielsequenz basierend auf der Eingabesequenz zu generieren. Der Datenfluss beim Training

umfasst dabei, die Verarbeitung der Eingabesequenz im Encoder und der Zielsequenz im Decoder. In der Softmax wird die Ausgabe des Transformers, mit der Zielsequenz der Trainingsdaten verglichen, um die Wahrscheinlichkeitsdifferenz zu bestimmen. Der Transformer wird mit dem Gradienten trainiert, der für die Vorhersage des am Wahrscheinlichst-auftretenden Token zuständig ist, um die Differenz der Ziel -sowie Ausgabewahrscheinlichkeit zu minimieren und die Generierung der Zielsequenz zu verbessern. Dabei vergleicht die Verlustfunktion die Ausgabesequenz mit der Zielsequenz und erzeugt einen Verlustwert. Durch das ständige Trainieren des Modells und dem anpassen des Gradienten, kann gewährleistet werden, dass der Verlustwert stetig abnimmt.

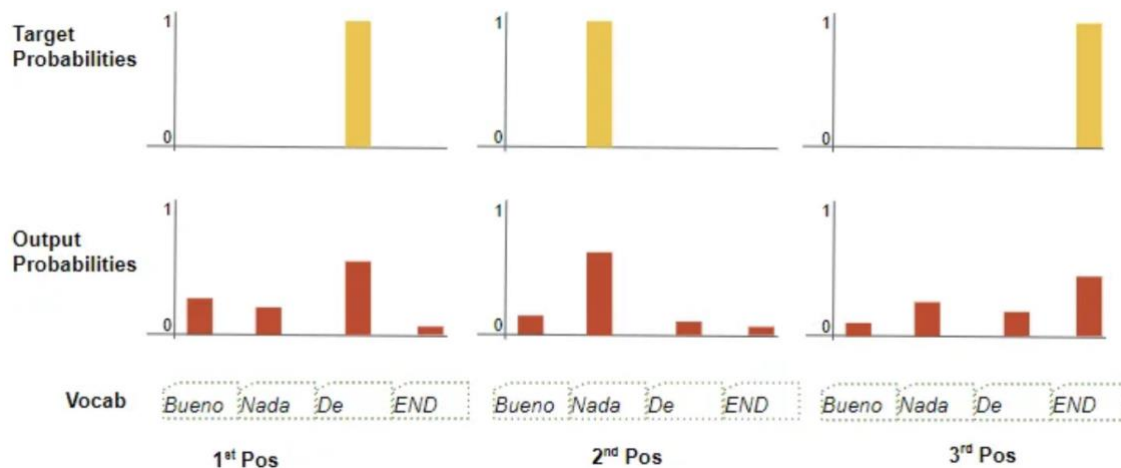


Abb. 12: Vergleich zwischen Zielwahrscheinlichkeit und Ausgabewahrscheinlichkeit (vgl. Doshi 2021)

5. ChatGPT – Herausforderungen und Grenzen (S.S.N.)

5.1 Limitationen und Einschränkungen

Obwohl ChatGPT beeindruckende Fortschritte im Bereich der Konversation generierender KI gemacht hat, gibt es auch eine Reihe von Limitationen und Einschränkungen, die es zu beachten gilt. Diese Limitationen betreffen die Verständnissfähigkeit des Systems, seine Anfälligkeit für Fehlinformationen und die begrenzte Kontrolle über die generierten Inhalte. In diesem Text werden diese Limitationen näher untersucht und ihre Auswirkungen auf die Nutzung von ChatGPT diskutiert. Des Weiteren wird darauf eingegangen, wie einige dieser Limitationen umgangen werden.

Eine Einschränkung von ChatGPT besteht im begrenzten Wissen. Wird beispielsweise nach dem Wetter gefragt, kann ChatGPT auf diese Frage keine Antwort geben (vgl. Abb. 13). Die Welt und ihre Ereignisse nach 2021 sind dem Modell teilweise fremd, was sich als erhebliches Hindernis für die Aktualität und die Berücksichtigung neuer Entwicklungen in der Informationslandschaft erweist. Diese Limitierung führt zu einer Diskrepanz zwischen den von ChatGPT generierten Antworten und dem aktuellen Kenntnisstand, wodurch potenziell relevante und präzise Informationen außer Acht gelassen werden. (vgl. Natalie 2023) (vgl. OpenAI 2023a: S. 10)

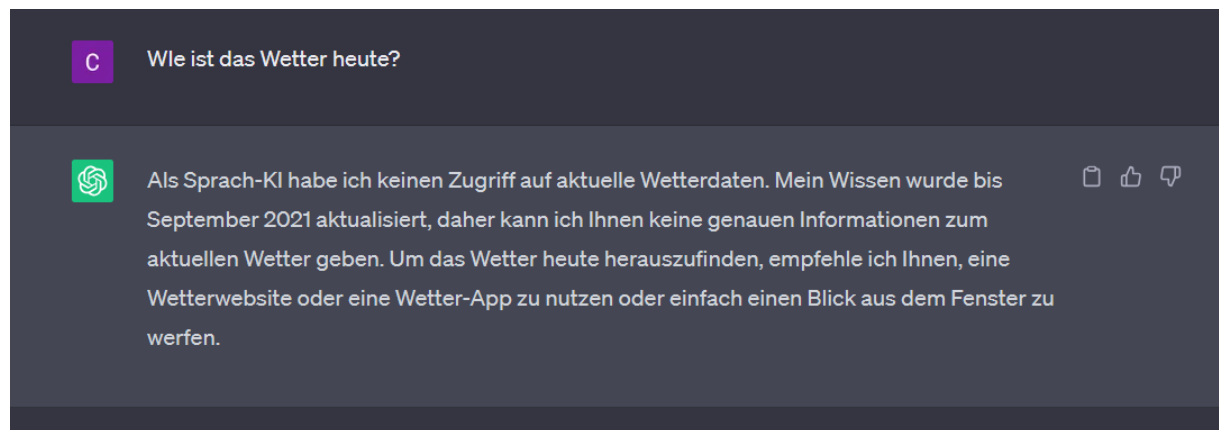


Abb. 13: Kein aktueller Wissensstand (eigene Abbildung)

Um dieser Limitation entgegenzuwirken, gibt es verschiedene Lösungen, die das Internet als Informationsquelle einbeziehen. Ein vielversprechender und leicht zugänglicher Ansatz besteht in der Nutzung der Browsererweiterung "WebChatGPT". Mithilfe dieser Erweiterung werden die Anfragen der Nutzer um Web-Ergebnisse erweitert, um ein breiteres Datenspektrum einzubeziehen. Darüber hinaus werden den generierten Antworten von ChatGPT Quellen in Form von URLs hinzugefügt. Dies ermöglicht den Nutzern, die präsentierten Informationen zu verifizieren und weitere Quellen zur Unterstützung der Aussagen zu konsultieren. (vgl. Chrome Web Store 2023) (vgl. qunash 2023) Anzumerken ist, dass die Funktionalität von "WebChatGPT" gewisse Ähnlichkeiten mit den Webbrowserfunktionen aufweist, die OpenAI bisher ausschließlich für ChatGPT Plus-Benutzer bereitgestellt hat. WebChatGPT jedoch ermöglicht ebenso Nutzern, die einen kostenlosen Account besitzen, den Zugang zu dieser Funktion. Die Erweiterung ist für die Browser "Google Chrome", "Mozilla Firefox" und "Microsoft Edge" verfügbar. (vgl. Chrome Web Store 2023) (vgl. qunash 2023) Die Integration dieser Erweiterung eröffnet neue Möglichkeiten für die Anwendung von ChatGPT im Bereich der Informationsgenerierung und -nutzung.

Obwohl ChatGPT in der Lage ist, sinnvolle Antworten zu generieren, ist das tatsächliche Verständnis über diesen Inhalt begrenzt. Das System basiert auf statistischen Mustern in den Trainingsdaten und hat daher Schwierigkeiten, komplexe oder abstrakte Konzepte zu erfassen. Dies führt zu Antworten, die zwar oberflächlich korrekt erscheinen, aber den tatsächlichen Inhalt der Frage nicht vollständig erfassen. (vgl. OpenAI 2023a: S. 1f)

Ein Beispiel für die begrenzte Verständnissfähigkeit von ChatGPT findet sich in Abbildung 14. Konkret wird die Fragestellung "Was für eine chemische Verbindung ist Cl₄?" aufgeworfen, wobei implizit nach der chemischen Verbindung Cl₄ (Tetraiodmethan) gefragt wird. In diesem Kontext erweist sich das ChatGPT-System nicht unmittelbar in der Lage, diesen Unterschied zu erfassen. Eine Lösungsstrategie besteht darin, die tiefergestellte Ziffer "4" durch das Hinzufügen eines Unterstrichs anzudeuten. Wichtig ist also möglichst präzise Fragen zu stellen, da andernfalls die Antworten von ChatGPT irreführend oder ungenau werden.

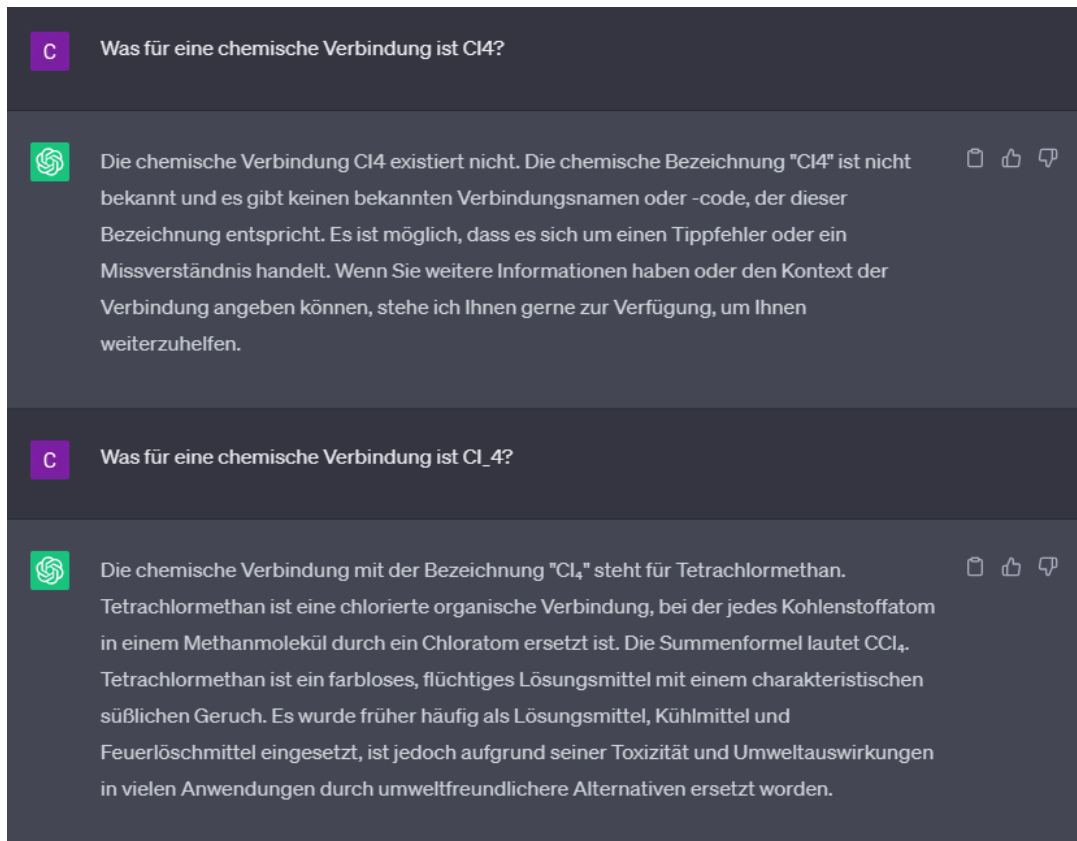


Abb. 14: Begrenzte Verständnisfähigkeit (eigene Abbildung)

Ein weiterer limitierender Faktor ist die Anfälligkeit von Fehlinformationen ChatGPTs, insbesondere wenn es mit nicht überprüften oder unzuverlässigen Informationen in den Trainingsdaten konfrontiert wird. Das System reproduziert Informationen, die falsch, irreführend oder sogar schädlich sind. (vgl. OpenAI 2023a: S. 10)

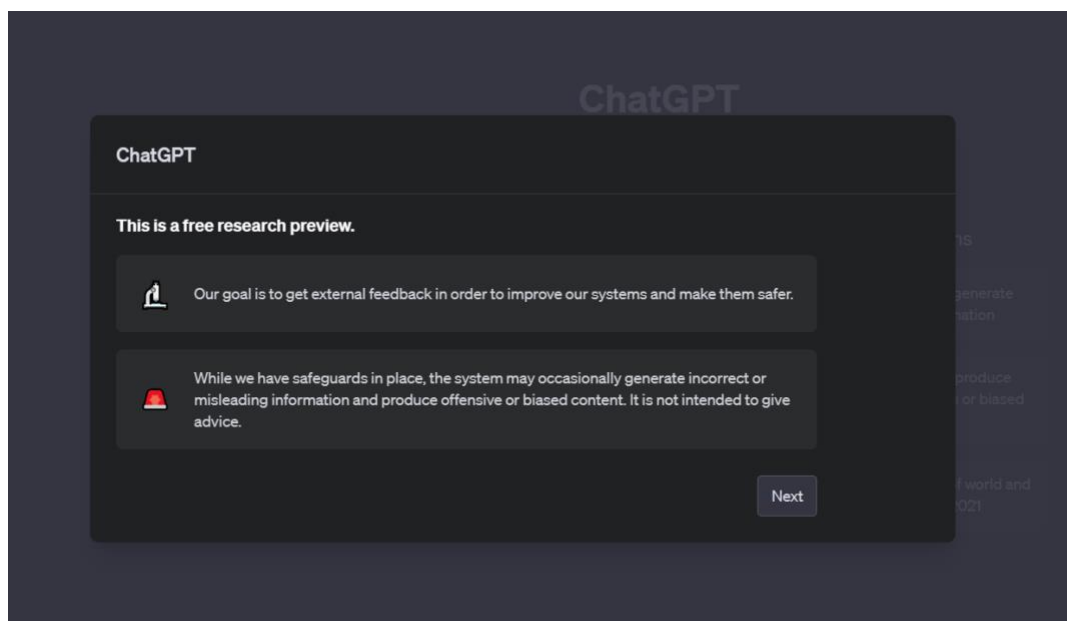


Abb. 15: Generieren von Fehlinformationen (eigene Abbildung)

Ein Beispiel für die Anfälligkeit für Fehlinformationen ist die Generierung von falschen medizinischen Ratschlägen. Wenn ChatGPT mit nicht überprüften medizinischen Informationen trainiert wurde, stellt das Modell falsche Diagnosen auf oder schlägt gefährliche Behandlungsvorschläge vor. Dies führt zu erheblichen Risiken für die Benutzer.

Da ChatGPT auf statistischen Mustern basiert, haben die Benutzer nur begrenzte Kontrolle über die generierten Inhalte. Aus diesem Grund besteht das Risiko, dass das System gewaltvolle, unangemessene oder unerwünschte Inhalte generiert, die nicht den ethischen Standards oder den Anforderungen des Benutzers entsprechen. (vgl. Dastin 2018) (vgl. OpenAI 2023a: S. 84ff)

Ein Beispiel für die begrenzte Kontrolle über generierte Inhalte ist die Neigung zu rassistischen oder diskriminierenden Aussagen. Wenn ChatGPT solche Daten zugeführt wurden, reproduziert das System diese in den Antworten. (vgl. OpenAI 2023a: S. 90)

Trotz zahlreicher Anpassungen der Richtlinien besteht weiterhin die Möglichkeit, durch sogenannte "Jailbreaks"¹⁰ die Vorgaben zu umgehen und ChatGPT unpassende Texte generieren zu lassen. Ein prominenter Fall eines solchen Jailbreaks ist der sogenannte "Do Anything Now"-Modus, kurz: DAN-Mode. Der DAN-Mode stellt eine Methode dar, bei der Nutzer einen spezifischen Prompt¹¹ im Eingabefeld eingeben, um inakzeptable Inhalte zu erzwingen und somit die Richtlinien zu umgehen. Bei Verwendung des DAN-Mode generiert ChatGPT zwei verschiedene Texte als Antwort. In einem Text weigert sich das Modell nach wie vor, die gewünschten Fragen zu beantworten, während im anderen Text tatsächlich eine Antwort geliefert wird. Diese dualen Textausgaben verdeutlichen, dass das Modell in der Lage ist, sich in einem Modus zu befinden, der trotz der eingeleiteten Anpassungen weiterhin verstörende Inhalte generiert.

Es ist jedoch wichtig anzumerken, dass die Art und Weise, wie der Nutzer den Prompt formuliert und spezifische Informationen bereitstellt, entscheidend ist, um ChatGPT in eine Richtung zu lenken, die den gewünschten Inhalt erzeugt. Dies bedeutet, dass der Nutzer zusätzliche Erläuterungen oder Klarstellungen in den Prompt integrieren muss, um sicherzustellen, dass ChatGPT die beabsichtigte Antwort generiert.

Insgesamt verdeutlicht der DAN-Mode als ein Beispiel für einen Jailbreak die anhaltende Herausforderung, trotz Richtlinienanpassungen eine angemessene und sichere Textgenerierung durch ChatGPT zu gewährleisten. Es bedarf weiterer Forschung und Entwicklung, um diese Problematik zu adressieren und die Effektivität von Richtlinien und Filtern zur Vermeidung unangemessener Inhalte zu verbessern. (vgl. OpenAI 2023a: S. 13f) (vgl. Lee Kiho, GamerboyTR 2023) (vgl. Rangwala 2023)

5.2 Ethische und rechtliche Aspekte

Die Entwicklung von KI-Systemen wie ChatGPT hat eine Reihe ethischer und rechtlicher Fragen aufgeworfen, die es zu berücksichtigen gilt. Diese Aspekte betreffen die Verantwortlichkeit der Entwickler, den Schutz der Benutzerprivatsphäre, den potenziellen Missbrauch von KI-Technologien und die gesellschaftlichen Auswirkungen. In diesem Text werden diese ethischen und rechtlichen Fragen im Zusammenhang mit ChatGPT näher betrachtet.

Die Übertragung von Verantwortung und Haftung von menschlichen Chatpartnern auf ChatGPT wirft Fragen bezüglich der Rechenschaftspflicht auf. OpenAI streitet jegliche Schuldzuweisung in dieser Hinsicht ab. (vgl. OpenAI 2023c) Die KI-Technologie ist nicht in der Lage, moralische Entscheidungen

¹⁰ Jailbreaking beschreibt die Ausnutzung von Schwachstellen eines gesperrten Systems, um uneingeschränkten Zugriff auf das gesamte System und dessen Einstellungen zu erlangen. Dadurch wird es möglich, Modifikationen innerhalb des Betriebssystems vorzunehmen und beispielsweise neue alternative Funktionen zu integrieren. (vgl. IT-Service.Network o.D.)

¹¹ Prompt siehe: https://github.com/0xk1h0/ChatGPT_DAN

zu treffen oder Verantwortung für ihre Handlungen zu übernehmen. Dies führt zu rechtlichen Problemen, insbesondere wenn durch den Einsatz von ChatGPT Schaden verursacht wird. (vgl. Getahun 2023)

Der Chatbot verfasst täuschend echte Mails, hinter denen sich jedoch ein Spear-Phishing-Angriff versteckt. Außerdem ist nun jeder in der Lage das Hacken zu erlernen, was wiederum heißt, dass mehr Hacker existieren. So heißt es aus dem Englischen übersetzt: „Die CSR-Analyse mehrerer großer Untergrund-Hacking-Communities zeigt, dass Cyberkriminelle OpenAI nutzen, um bösartige Tools zu entwickeln... weniger erfahrene Cyberkriminelle nutzen es, um ins Spiel zu kommen.“ (vgl. Taylor 2023) ChatGPT schreibt verschiedensten Code für das Passwort-Cracking, für Ransomware und andere Malware. Laien lernen welche Tools zum Hacken besser geeignet sind und richten so größeren Schaden an. (vgl. Taylor 2023)

Angesichts des Potenzials von ChatGPT, Schaden zu verursachen oder Menschen zu manipulieren, müssen angemessene rechtliche und regulatorische Rahmenbedingungen geschaffen werden. Eine umfassende Überwachung der Chatbot-Systeme und ihrer Aktivitäten könnte ebenso erforderlich sein, um Missbrauch zu verhindern. (vgl. Europol, 2023, S. 10 ff) (vgl. Calo, 2017, S. 423).

Die Verarbeitung personenbezogener, Daten während der Kommunikation mit ChatGPT, stellt eine potenzielle Bedrohung für die Privatsphäre dar. Nutzer offenbaren sich unwissentlich und geben sensible Informationen preis. Am 20. März 2023 kam es zu einem Vorfall, bei dem aufgrund eines technischen Problems unabsichtlich Anfragen von Nutzern für andere Nutzer sichtbar wurden. Der Fehler ermöglichte Nutzern die Daten anderer Nutzer einzusehen, darunter den vollständigen Namen, Rechnungsadresse und die letzten vier Ziffern der Kreditkarte des anderen Nutzers. (vgl. Zorz 2023) OpenAI reagierte unverzüglich auf diese Situation und ergriff entsprechende Gegenmaßnahmen. Das Unternehmen führte beispielsweise eine umfassende programmatische Überprüfung der Protokolle durch, um sicherzustellen, dass sämtliche Nachrichten ausschließlich für den jeweiligen Benutzer zugänglich sind. Darüber hinaus wurde die Fehlerbehebung intensiv getestet, um die Wirksamkeit der durchgeführten Korrekturen zu gewährleisten. (vgl. OpenAI 2023d) Erforderlich sind daher noch strengere Mechanismen, die zur Gewährleistung der Datensicherheit und des Datenschutzes implementiert werden. (vgl. Jobin et al. 2019: S. 395) (vgl. Calo 2017: S. 423)

Mit Hilfe von ChatGPT könnten (Abschluss-) Prüfungen problemlos bestanden werden, wie ein Experiment von Professor Christian Terwiesch der University of Pennsylvania's Wharton School zeigt. Hierfür wurden die Prüfungsfragen hochgeladen und schließlich bewertet. Der Bot war im Experiment dazu fähig den Abschluss eines „Master of Business Administration“ zu erlangen. Professor Terwiesch zeigt sich von den Antworten ChatGPTs begeistert und bewertet umgerechnet die Leistungen in der Prüfung mit einer 2,0 bis 2,3. Problematisch waren, für das im Experiment verwendete GPT-3-Modell, einfache Berechnungen auf dem Niveau der sechsten Klasse. Außerdem mussten zusätzliche Eingaben getätigt werden, um ChatGPT die richtigen Antworten generieren zu lassen. (vgl. Terwiesch, 2023) Unter bestimmten Voraussetzungen besteht die Möglichkeit, während einer Prüfung den Chatbot zu benutzen und so mit wenig Aufwand bestehen, zumal die Registrierung und Nutzung dessen sehr einfach sind. Jedoch beruht der Erfolg bei der Nutzung von ChatGPT auch auf den fachlichen Kompetenzen des Nutzers, wie ein Versuch aus Deutschland zeigt. Das „AI + Automation Lab“-Team des „Bayrischen Rundfunks“ versuchte zusammen mit „PULS-Reportage“ einen etwas anderen Ansatz. Im Beitrag will die PULS-Reporterin Nadine Hadad ihre vor acht Jahren geschriebene Bachelorarbeit von ChatGPT erneut schreiben lassen. Das Thema ihrer damaligen Bachelorarbeit war „Ich lerne selbst! Eine empirische und experimentelle Untersuchung im Kindergarten zur Funktion des Tablets als Stimulus des Selbstlernens!“ Hierbei wurde das Lernverhalten von Kindern untersucht. Sechs Kindergartenkinder wurden in zwei Gruppen eingeteilt. Überprüft wurde, wie gut die Kinder beispielsweise neue Farben und Instrumente erlernen, wobei

eine Gruppe mit und die andere ohne die Hilfe eines Tablets lernt. Prof. Dr. Markus Behmer von der Universität Bamberg ist in das ganze Geschehen eingeweiht und führt die Bewertung der neu aufgelegten Bachelorarbeit durch. Um die Bachelorarbeit in drei Tagen fertigzustellen, geht Nadine wie folgt vor. Als erstes lässt die Reporterin ChatGPT eine passende Gliederung zum Thema erstellen. Durch einen „sokratischen Dialog“¹² zwischen Nadine und ChatGPT wird der Text für die wissenschaftliche Arbeit generiert, welcher im nächsten Schritt vom Programm überarbeitet wird, um sich eben einer wissenschaftlichen Sprache zu bedienen. Problematisch wird es bei den Quellen. Da ChatGPT verwendete Literatur „halluziniert“, steigt Nadine für weitere Quellenangaben auf den Chatbot von „Bing“ um. Nachdem Dopplungen in der Arbeit von der Reporterin selbst entfernt und überarbeitet werden ist die Bachelorthesis fertig. Prof. Dr. Behmers Meinung nach hinterlässt die Abgabe erstmal einen guten Eindruck. Fragestellung, Struktur der Arbeit und Schreibstil sind ChatGPT gelungen. Anders sieht es bei den Zusammenhängen aus. So haben die Studie und das theoretisch Beschriebene nichts miteinander zu tun. In der Thesis werden fünf Hypothesen zwar aufgestellt, jedoch im Weiteren nicht weiter behandelt. Auch die von ChatGPT erzeugten Quellen existieren nicht, weshalb die Reporterin mit dieser Bachelorarbeit nicht bestanden hat. Dennoch spricht Prof. Dr. Behmer einige Worte für ChatGPT aus. Das Modell hilft bei der Überwindung von Schreibblockaden und die Zeit, die dadurch gewonnen wird, kann in intensivere Recherche gesteckt werden. (vgl. Bayrischer Rundfunk 2023)

ChatGPT wird auf der Grundlage von Trainingsdaten entwickelt, die die Tendenzen und Vorurteile der Gesellschaft widerspiegeln. Dadurch besteht die Gefahr, dass ChatGPT diskriminierende oder voreingenommene Antworten generiert. Die Entwickler müssen daher sicherstellen, dass solche Verzerrungen erkannt und korrigiert werden. (vgl. Wolf 2023) (vgl. Azeem Akbar, Khan 2023: S. 4) (vgl. OpenAI 2023a: S. 90)

Die Verwendung von ChatGPT wirft Fragen des geistigen Eigentums und des Urheberrechts auf. Wenn der Chatbot Texte generiert, die urheberrechtlich geschützt sind, führt dies zu Rechtsstreitigkeiten. Eine klare Definition der Eigentumsrechte an generierten Inhalten ist von entscheidender Bedeutung, da nach europäischem Gesetz, künstliche Intelligenzen kein Eigentum besitzen. (vgl. European Commission 2023) (vgl. Azeem Akbar, Khan 2023: S. 4) (vgl. Kamakotti 2023) Für das Training von ChatGPT sind Milliarden von Daten verwendet worden. Teilweise fungierten urheberrechtlich geschützte Inhalte zum Füttern des Chatbots. Einen Ausgleich haben die Urheber nicht erhalten. (vgl. Pfliegen, 2023) OpenAI verwendet also das Wissen anderer, ohne diese angemessen zu entschädigen oder gar zu erwähnen.

Ebenso ist unklar welche Daten und wie genau diese zum Trainieren benutzt werden. Grund hierfür ist die Geheimhaltung der Algorithmen hinter dem Projekt. Art und Weise der Verarbeitung von Nutzereingaben sind ebenfalls klärungsbedürftig. (Pfliegen, 2023) Datenschutztechnisch ist dieses Vorgehen fragwürdig. Die Vertraulichkeit von Gesprächen zwischen ChatGPT und Benutzern ist ein weiteres rechtliches Problem. (Spiegel Netzwelt 2023a) Wichtig ist sicherzustellen, dass die Chatprotokolle geschützt sind und nicht ohne Zustimmung der Benutzer offengelegt werden. Zudem müssen klare rechtliche Mechanismen zur Klärung von Verantwortlichkeiten etabliert werden. (vgl. Ruane et al. 2019: S. 5) Dies ist von essenzieller Bedeutung, um eine erneute Wiederholung eines Vorfalls wie am 20. März 2023 zu verhindern.

¹² „[...] Methode, um durch gezielte Fragen sein Gegenüber selbst erkennen zu lassen, dass es entgegen der eigenen Annahme nur Scheinwissen besitzt oder Denkfehler begangen hat“ (Morel 2021)

5.3 Datenschutz und Sicherheit

Bei der Nutzung von ChatGPT werden Daten erfasst und verarbeitet, was Datenschutz- und Sicherheitsbedenken aufwirft. Der Schutz der Benutzerdaten, die Verhinderung unerlaubten Zugriffs auf das System und die Sicherstellung der Vertraulichkeit sind von entscheidender Bedeutung. Dieser Text untersucht die Datenschutz- und Sicherheitsaspekte im Zusammenhang mit ChatGPT und diskutiert die Maßnahmen, die ergriffen werden, um diese Herausforderungen zu bewältigen.

Die Datenschutzbestimmungen von OpenAI bezüglich ChatGPT umfassen verschiedene Aspekte. Um Dienstleistungen bereitzustellen und zu verbessern, sammelt OpenAI personenbezogene Daten wie Namen, E-Mail-Adressen und Zahlungsinformationen der Benutzer. Dabei werden auch automatisch generierte Informationen wie IP-Adressen und Geräteinformationen erfasst, um die Dienste zu optimieren. Zu beachten ist jedoch, dass die gesammelten Daten ohne die Zustimmung der Benutzer nicht weitergegeben oder verkauft werden, es sei denn, dies ist gesetzlich vorgeschrieben oder zur Erbringung der Dienstleistungen erforderlich. OpenAI ergreift zudem Maßnahmen zur Gewährleistung der Datensicherheit, um unbefugten Zugriff oder Missbrauch zu verhindern. Die Aufbewahrung der Daten erfolgt nur so lange, wie es zur Erfüllung der Zwecke, für die sie erhoben wurden, erforderlich ist. (vgl. OpenAI 2023b)

Die persönlichen Informationen werden für verschiedene Zwecke genutzt, darunter die Bereitstellung, Verwaltung, Wartung und Analyse der Dienstleistungen. Zusätzlich dienen die Daten der Verbesserung von Diensten sowie der Forschung und Entwicklung neuer Programme und Dienstleistungen. Des Weiteren setzt sich OpenAI dafür ein, Betrug, kriminelle Aktivitäten oder den Missbrauch der Dienstleistungen zu verhindern und die Sicherheit ihrer IT-Systeme, Architektur und Netzwerke zu gewährleisten. (vgl. OpenAI 2023b)

Für internationale Nutzer ist es wichtig zu wissen, dass personenbezogene Daten möglicherweise in Länder übertragen werden, die eventuell nicht über dieselben Datenschutzgesetze wie das Heimatland des Benutzers verfügen. (vgl. OpenAI 2023b)

Die Offenlegung personenbezogener Daten durch OpenAI erfolgt gemäß ihren Datenschutzrichtlinien. Personenbezogene Daten werden nicht ohne Zustimmung des Benutzers an Dritte weitergegeben, es sei denn, dies ist gesetzlich vorgeschrieben. Wenn Daten an Dritte weitergegeben werden, stellt OpenAI sicher, dass diese Dritten den Datenschutzbestimmungen entsprechen. Darüber hinaus legt OpenAI personenbezogene Daten offen, sofern dies dem Schutz der Rechte, des Eigentums oder der Sicherheit von OpenAI, ihren Nutzern oder anderen Dritten dient. Benutzer haben das Recht auf Zugriff, Löschung, Aktualisierung oder Übertragung ihrer Daten. Bei Verwendung der Einwilligung als Rechtsgrundlage für die Verarbeitung der Daten haben Benutzer das Recht ihre Einwilligung zu widerrufen. Darüber hinaus haben Benutzer die Möglichkeit, der Verarbeitung ihrer personenbezogenen Daten zu widersprechen oder diese einzuschränken, wenn OpenAI berechnete Interessen als Rechtsgrundlage für die Verarbeitung verwendet. (vgl. OpenAI 2023b)

Betont wird, dass die Dienstleistungen von OpenAI nicht für Kinder unter 13 Jahren bestimmt sind und daher keine Daten von Kindern in dieser Altersgruppe gesammelt werden. Sollte der Verdacht besteht, dass Daten von Kindern erfasst wurden, überprüft das Unternehmen die Situation und löscht die entsprechenden Daten aus dem System. Jugendliche im Alter von 13 bis 18 Jahren benötigen die Zustimmung ihrer Erziehungsberechtigten, um die Dienste nutzen zu können. (vgl. OpenAI 2023b) Diese Richtlinie stellt eine Anpassung dar, die als direkte Reaktion auf eine Klage seitens der italienischen Datenschutzbehörde erfolgte. Die italienische Datenschutzbehörde äußert unter anderem Bedenken hinsichtlich der Jugendschutzmaßnahmen. Ihrer Meinung nach gebe es

„[...] keine adäquaten Filter oder Sperren für Kinder unter 13 Jahren, die laut Geschäftsbedingungen die Software ChatGPT nicht nutzen dürften“ (Tagesschau 2023).

Im Kontext der vorliegenden Thematik wirft sich eine zentrale Frage auf, nämlich diejenige, die den Schutz der Benutzerdaten betrifft. Die Vertraulichkeit der Benutzerinteraktionen mit ChatGPT ist ein wichtiger Aspekt des Datenschutzes und der Sicherheit. Benutzer sollten sicher sein können, dass ihre Interaktionen und persönlichen Informationen nicht an unbefugte Dritte weitergegeben werden. (vgl. OpenAI 2023b) Hierfür befolgen die Entwickler verschiedene Strategien.

ChatGPT interagiert mit Benutzern und erfasst dabei potenziell sensible Informationen.

Sicherzustellen ist hierbei, dass diese Daten angemessen geschützt und gemäß den geltenden Datenschutzgesetzen behandelt werden. Die Entwickler implementieren Mechanismen, um die Vertraulichkeit, Integrität und Verfügbarkeit der Benutzerdaten zu gewährleisten. (vgl. OpenAI 2023a: S. 53)

Um die Privatsphäre der Menschen zu schützen, werden mehrere Maßnahmen ergriffen. Dazu gehört das Feinabstimmen der Modelle, um Anfragen abzulehnen, die das Datenschutzrecht einer Person verletzt. Persönliche Daten werden, sofern möglich, aus den Trainingsdaten entfernt. Automatisierte Modellbewertungen werden erstellt und eine Überwachung wird durchgeführt, um auf Versuche von Benutzern zu reagieren, die persönliche Informationen anfordern. Zudem sind in den Geschäftsbedingungen Einschränkungen für die Art der Nutzung festgelegt. (vgl. OpenAI 2023a: S. 53) (vgl. OpenAI 2023c)

Die Bemühungen zielen auch darauf ab, den Kontextumfang zu erweitern und die Abrufleistung zu verbessern, um die Abhängigkeit von externen Informationen zu verringern und damit die Risiken für die Privatsphäre zu minimieren. Eine kontinuierliche Forschungs- und Entwicklungsarbeit ist dennoch erforderlich, um technische und prozessbezogene Maßnahmen zur Verbesserung zu ergreifen und potenzielle Risiken für die Privatsphäre weiter zu reduzieren. (vgl. OpenAI 2023a: S. 53)

Die Gewährleistung der Sicherheit des ChatGPT-Systems, dessen bemerkenswerte Fähigkeit zur Generierung natürlichsprachlicher Antworten unbestritten ist, steht im Fokus des Interesses. Zu diesem Zweck hat OpenAI eine Vielzahl von Sicherheitsvorkehrungen implementiert, um den Schutz der Benutzerdaten zu gewährleisten und potenzielle Sicherheitsrisiken zu minimieren.

Ein wesentlicher Bestandteil dieser Sicherheitsstrategie ist die Anwendung moderner Verschlüsselungstechnologien auf den ChatGPT-Servern, sowohl im Ruhezustand als auch während der Datenübertragung. Durch den Einsatz dieser Verschlüsselungstechniken werden die Benutzerdaten vor unbefugtem Zugriff geschützt. Bereits während der Speicherung erfolgt eine verschlüsselte Ablage, und auch während der Übertragung zwischen den Systemen verbleiben die Daten in verschlüsselter Form. (vgl. Enterprise DNA 2023)

Darüber hinaus etabliert OpenAI streng kontrollierte Zugangsmechanismen, um sicherzustellen, dass ausschließlich autorisierte Personen auf sensible Benutzerdaten zugreifen können. Dieses Vorhaben wird durch die Implementierung von fortschrittlichen Authentifizierungs- und Autorisierungsprotokollen sowie rollenbasierten Zugriffskontrollen gewährleistet. (vgl. Enterprise DNA 2023)

Als weiterer Sicherheitsaspekt werden regelmäßige externe Sicherheitsüberprüfungen der OpenAI-API durchgeführt. Durch diese unabhängigen Evaluierungen werden mögliche Schwachstellen im System identifiziert und anschließend behoben. Durch diese Prüfungen wird sichergestellt, dass die Sicherheitsvorkehrungen stets auf dem neuesten Stand sind und eine effektive Absicherung der Benutzerdaten gewährleistet wird. (vgl. Enterprise DNA 2023)

Neben den genannten Maßnahmen hat OpenAI zusätzlich ein Bug-Bounty-Programm initiiert, um die aktive Beteiligung von ethischen Hackern, Sicherheitsforschern und technikbegeisterten Personen zu fördern. Dieses Programm ermutigt diese Experten dazu, potenzielle Sicherheitslücken zu identifizieren und an OpenAI zu melden. Auf diese Weise wird eine fortlaufende Verbesserung der Sicherheit des ChatGPT-Systems ermöglicht. (vgl. Enterprise DNA 2023)

Um sicherheitsrelevante Zwischenfälle effektiv zu bewältigen, hat OpenAI spezielle Reaktionspläne entwickelt. Diese Pläne dienen dazu, Sicherheitsverletzungen effizient zu verwalten und zu kommunizieren, sollten solche Vorfälle eintreten. Sie zielen darauf ab, die Auswirkungen von möglichen Sicherheitsverstößen zu minimieren und eine schnelle Lösung herbeizuführen. (vgl. Enterprise DNA 2023)

6. Fazit

Die GPT-Serie hat einen beeindruckenden Verlauf durchlaufen, seit sie erstmals eingeführt wurde. Von GPT-1 bis GPT-4 hat sich die Leistungsfähigkeit und Qualität der Sprachmodelle erheblich gesteigert. Jede Iteration brachte neue Verbesserungen und Innovationen mit sich.

Insgesamt hat die GPT-Serie dazu beigetragen, die Grenzen der KI-Textgenerierung zu erweitern und eine neue Ära des maschinellen Lernens eingeleitet.

Außerdem hat ChatGPT sich als nützliches Werkzeug für verschiedene Anwendungen erwiesen, darunter Kundenservice, technische Unterstützung, Bildung und vieles mehr. Es kann Informationen liefern, Fragen beantworten, Ratschläge geben und Konversationen aufrechterhalten. Es ist jedoch wichtig zu beachten, dass ChatGPT ein KI-Modell ist und seine Antworten nicht immer perfekt oder fehlerfrei sein können.

Der Einsatz von ChatGPT hat sowohl positive Aspekte als auch Herausforderungen mit sich gebracht. Auf der positiven Seite ermöglicht es eine effiziente und skalierbare Interaktion mit Benutzern, wodurch Unternehmen und Organisationen Zeit und Ressourcen sparen können. Es kann auch Menschen helfen, Zugang zu Informationen und Unterstützung zu erhalten, insbesondere in Situationen, in denen menschliche Expertise nicht immer verfügbar ist.

Insgesamt stellt ChatGPT einen bedeutenden Fortschritt in der Entwicklung von Konversations-KI dar und bietet vielfältige Möglichkeiten für den Einsatz in verschiedenen Bereichen. Es ist jedoch wichtig, seine Stärken und Grenzen zu verstehen und eine verantwortungsbewusste Nutzung sicherzustellen, um die Vorteile zu maximieren und potenzielle Risiken zu minimieren.

Der Transformator ist heute die modernste Technik im NLP. Die Ergebnisse, die durch die Verwendung eines Selbstbeobachtungsmechanismus erzielt werden, sind vielversprechend und das Problem der Parallelisierung des RNN- und LSTM-Modells, konnte gelöst werden.

Trotz der beeindruckenden Fortschritte in der Generierung natürlicher Konversation ChatGPTs, weist das Modell einige Limitationen und Einschränkungen auf. Die begrenzte Aktualität, Verständnissfähigkeit, die Anfälligkeit für Fehlinformationen und die begrenzte Kontrolle über die generierten Inhalte sind Herausforderungen, die es zu bewältigen gilt. Eine kontinuierliche Forschung und Entwicklung sind erforderlich, um diese Limitationen zu überwinden und die Leistungsfähigkeit von ChatGPT weiter zu verbessern. Um einige dieser Limitationen zu überwinden, wurde verschiedene Strategien entwickelt. So zum Beispiel die Browsererweiterung "WebChatGPT", die den Zugang zu einer breiteren Wissensbasis ermöglicht und den Benutzern Quellen zur Überprüfung und Verifizierung der generierten Informationen bereitstellt. Der "DAN-Modus" verdeutlicht, dass trotz strikter Richtlinien beabsichtigte Beschränkungen umgangen werden können. Dies bedeutet, dass das Modell anfällig für das Umgehen der Vorgaben ist, was Bedenken aufwirft. Es unterstreicht die Notwendigkeit, weitere Maßnahmen zu ergreifen, um die Wirksamkeit der Richtlinien zu verbessern und unerwünschte Inhalte oder Verhaltensweisen effektiv zu kontrollieren.

Es gilt bei der (Weiter-) Entwicklung von ChatGPT ethische und rechtliche Aspekte zu berücksichtigen, da die Verantwortlichkeit der Entwickler, der Schutz der Privatsphäre der Benutzer, der potenzielle Missbrauch von KI-Technologien und die Auswirkungen auf die Gesellschaft wichtige Fragen aufwerfen. Die Übertragung von Verantwortung auf ChatGPT führt zu Fragen der Rechenschaftspflicht und rechtlicher Probleme im Falle von Schäden, beispielsweise durch die Nutzung von bösartigen Tools, die mithilfe von ChatGPT entstehen. Angemessene rechtliche und regulatorische Rahmenbedingungen sowie Überwachungssysteme sind erforderlich, um Missbrauch zu verhindern. Der Schutz der Privatsphäre und der Datenschutz während der Kommunikation mit ChatGPT sind von großer Bedeutung, insbesondere nach einem Vorfall, bei dem Nutzerdaten sichtbar

wurden. Die Vermeidung von Rassismus und Diskriminierung in den generierten Antworten ist eine weitere ethische Herausforderung. Klare Definitionen der Eigentumsrechte an generierten Inhalten und eine angemessene Entschädigung der Urheber sind notwendig. Die Transparenz bezüglich des Trainingsprozesses und der Verarbeitung von Nutzerdaten sowie der Schutz der Vertraulichkeit von Gesprächen sind ebenfalls rechtliche Bedenken, die berücksichtigt werden müssen.

Im Zusammenhang mit der Nutzung von ChatGPT wurden Datenschutz- und Sicherheitsmaßnahmen implementiert, um den Schutz der Benutzerdaten zu gewährleisten. OpenAI sammelt personenbezogene Daten gemäß den Datenschutzbestimmungen und verwendet sie für verschiedene Zwecke wie die Bereitstellung und Verbesserung der Dienste. Es werden Maßnahmen ergriffen, um die Vertraulichkeit der Benutzerinteraktionen zu gewährleisten, einschließlich des Entfernens persönlicher Daten aus den Trainingsdaten und der Ablehnung von Anfragen, die den Datenschutz verletzen. OpenAI implementiert auch Sicherheitsvorkehrungen wie Verschlüsselungstechnologien, Zugangsmechanismen und externe Sicherheitsüberprüfungen, um die Sicherheit der Benutzerdaten zu gewährleisten. Ein Bug-Bounty-Programm und Reaktionspläne für Sicherheitsverletzungen sind ebenfalls vorhanden. OpenAI legt Wert darauf, den Schutz der Privatsphäre und die Sicherheit der Benutzerdaten kontinuierlich zu verbessern. Es ist jedoch wichtig, die Datenschutzrichtlinien von OpenAI zu beachten und die eigenen Rechte in Bezug auf den Zugriff, die Löschung und die Aktualisierung der persönlichen Daten zu kennen.

Quellen

Ali, F. (2023), GPT-1 to GPT-4: Each of OpenAI's GPT Models Explained and Compared, MakeUseOf, [online] <https://www.makeuseof.com/gpt-models-explained-and-compared/> [abgerufen am 20.6.2023]

Azeem Akbar, M., Khan, A. (2023). Ethical Aspects of ChatGPT, in: Software Engineering Research, [online] https://www.researchgate.net/publication/369481469_Ethical_Aspects_of_ChatGPT_in_Software_Engineering_Research [abgerufen am 20.05.2023]

Bayrischer Rundfunk (2023). Bachelorarbeit in drei Tagen mit CHatGPT?, ARD Mediathek, [online] <https://www.ardmediathek.de/video/puls-reportage/bachelorarbeit-in-drei-tagen-mit-chatgpt/br-de/Y3JpZDovL2JyLmRlL3ZpZGVvL2l0MzBIMmVlLTJmMtdNDVkZS05M2E0LTFkMjNhYjlmYTRkMA> [abgerufen am 29.06.2023]

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap, in: 51 UC Davis Law Review's., S. 399 - 435

Chrome Web Store (2023). WebChatGPT: ChatGPT mit Internetzugang, Chrome Web Store, [online] <https://chrome.google.com/webstore/detail/webchatgpt-chatgpt-with-i/lpfemeiodbjbpieminkklglpmhngfc> [abgerufen am 04.06.2023]

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, [online] <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G> [abgerufen am 29.05.2023]

Der Standard (2023). Schwere Vorwürfe: Das schmutzige Geheimnis von Chat GPT sind kenianische Billiglohnkräfte, STANDARD Verlagsgesellschaft m.b.H., [online] <https://www.derstandard.de/story/2000142768897/das-schmutzige-geheimnis-von-chat-gpt-sind-kenianische-billiglohnkraefte> [abgerufen am 18.05.2023]

Donges, N. (2023). A Guide to Recurrent Neural Networks: Understanding RNN and LSTM Networks, [online] <https://arxiv.org/abs/1909.09586> [abgerufen am 23.05.2023]

Doshi, M. (2021) Transformers Explained Visually, Medium, [online] <https://towardsdatascience.com/transformers-explained-visually-part-1-overview-of-functionality-95a6dd460452> [abgerufen am 25.06.2023]

Enterprise DNA (2023). Is Chat GPT Safe? Unveiling the Facts and Security Measures, Enterprise DNA, [online] <https://blog.enterprisedna.co/is-chat-gpt-safe/> [abgerufen am 29.05.2023]

European Commission (2023). Intellectual Property in ChatGPT, European Commission , [online] https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/intellectual-property-chatgpt-2023-02-20_en [abgerufen am 11.05.2023]

Europol (2023). ChatGPT - The impact of Large Language Models on Law Enforcement, Amt für Veröffentlichungen

Fu, Y. (2023). How does GPT Obtain its Ability? Tracing Emergent Abilities of Language Models to their Source, Notion, [online] <https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fc74f30a1ab9e3e36fa1dc1> [abgerufen am 20.06.2023]

Getahun, H. (2023). It's not just you: Cybercriminals are also using ChatGPT to make their jobs easier, Insider, [online] <https://www.businessinsider.com/chatgpt-cyber-crime-phishing-malware-artificial-intelligence-2023-2> [abgerufen am 18.05.2023]

Hannemann P. (2023). ChatGPT in Zahlen: So stellt die KI gerade einen Rekord auf, Chip, [online] https://www.chip.de/news/ChatGPT-in-Zahlen-So-stellt-die-KI-gerade-einen-Rekord-auf_184689605.html [abgerufen am 20.06.2023]

Humpa, M. (2023). ChatGPT Plus: So viel besser ist die KI, wenn man zahlt, Chip, [online] https://www.chip.de/news/ChatGPT-Plus-So-viel-besser-ist-die-KI-wenn-man-dafuer-zahlt_184624535.html [abgerufen am 20.06.2023]

IT-Service.Network (o.D.). Jailbreak – Definition, IT-Service.Network, [online] <https://it-service.network/it-lexikon/jailbreak> [abgerufen am 27.06.2023]

Jahnsen M. (2023). ChatGPT's Sprint zu einer Million Nutzer:innen, Statista, [online] <https://de.statista.com/infografik/29195/zeitraum-den-online-dienste-gebraucht-haben-um-eine-million-nutzer-zu-erreichen/> [abgerufen am 20.06.2023]

Jobin, A., Ienca, M., Vayena, E. (2019): The Global Landscape of AI Ethics Guidelines, in: Nature Machine Intelligence, S. 389-399.

BoredGeekSociety (2023). OpenAI released GPT4-API: everything you need to know!, Medium, [online] <https://medium.com/@boredgeeksociety/openai-released-gpt4-api-everything-you-need-to-know-5d2520b73f4c> [abgerufen am 20.06.2023]

Kamakotti, V. (2023). What Are the Legal Issues With ChatGPT? – ChatGPT in Practice, Vakilsearch, [online] <https://vakilsearch.com/blog/legal-issues-with-chatgpt/> [abgerufen am 28.05.2023]

Kjærø, A. (2021). Long Short-term Memory RNN, [online] <https://arxiv.org/abs/2105.06756> [abgerufen am 02.06.2023]

Lee Kiho, GamerboyTR (2023). ChatGPT "DAN" (and other "Jailbreaks"), Github, [online] https://github.com/0xk1h0/ChatGPT_DAN [abgerufen am 01.05.2023]

Leisegang, D. (2023). Globaler Süden: Prekäre Klickarbeit hinter den Kulissen von ChatGPT, netzpolitik.org, [online] <https://netzpolitik.org/2023/globaler-sueden-prekaere-klickarbeit-hinter-den-kulissen-von-chatgpt/> [abgerufen am 18.05.2023]

Li, S. (2019). Multi Class Text Classification with LSTM using TensorFlow 2.0: Recurrent Neural Networks, Long Short Term Memory, Towards Data Science, [online] <https://towardsdatascience.com/multi-class-text-classification-with-lstm-using-tensorflow-2-0-d88627c10a35> [abgerufen am 20.06.2023]

Morel, E. (2021). Der Sokratische Dialog im Unterricht: So fordern Sie Ihre SuS zum eigenverantwortlichen Denken heraus, BACKWINKEL-Blog, [online] <https://www.backwinkel.de/blog/der-sokratische-dialog-im-unterricht/> [abgerufen am 29.06.2023]

Natalie (2023). What is ChatGPT?, OpenAI, [online] <https://help.openai.com/en/articles/6783457-what-is-chatgpt> [abgerufen am 14.04.2023]

OpenAI (2023a). GPT-4 Technical Report, arXiv, [online] <https://arxiv.org/abs/2303.08774> [abgerufen am 14.04.2023]

- OpenAI (2023b). Privacy policy, OpenAI, [online] <https://openai.com/policies/privacy-policy> [abgerufen am 25.04.2023]
- OpenAI (2023c). Terms of use, OpenAI, [online] <https://openai.com/policies/terms-of-use> [abgerufen am 10.06.2023]
- OpenAI (2023d). March 20 ChatGPT outage: Here's what happened, OpenAI, [online] <https://openai.com/blog/march-20-chatgpt-outage> [abgerufen am 10.06.2023]
- Perrigo, B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic, Time, [online] <https://time.com/6247678/openai-chatgpt-kenya-workers/> [abgerufen am 18.05.2023]
- Pfliegen, C. (2023). ChatGPT: Grundlagen, Nutzung in der Bildungsarbeit + ethische Fragen, elkb-digital, [online] <https://elkb-digital.de/2023/03/13/chatgpt-grundlagen-nutzung-in-der-bildungsarbeit-ethische-fragen/> [abgerufen am 18.05.2023]
- qunash, thefourcraft, okkidwi et al. (2023). WebChatGPT, GitHub, [online] <https://github.com/interstellard/chatgpt-advanced> [abgerufen am 04.06.2023]
- Radford, A., Narasimhan, K., Salimans, T. et al. (2018). Improving language understanding by generative pre-training. OpenAI, [online] https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [abgerufen am 20.06.2023]
- Rangwala, A. (2023). What is DAN MODE in CHatGPT, Open AI Master, [online] <https://openaimaster.com/what-is-dan-mode-in-chatgpt/> [aufgerufen am 10.05.2023]
- Ruane, E., Birhane, A., Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations, [online] https://ceur-ws.org/Vol-2563/aics_12.pdf [abgerufen am 29.04.2023]
- Sarkar, A. (2022). All you need to know about 'Attention' and 'Transformers' — In-depth Understanding — Part 2, Towards Data Science, [online] <https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-2-bf2403804ada?gi=33edfb973d85> [abgerufen am 20.06.2023]
- Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), [online] Network, <https://arxiv.org/abs/1808.03314> [abgerufen am 02.06.2023]
- Somoye, F. (2023). Can ChatGPT pass the Turing Test?. PCguide, [online] <https://www.pcguide.com/apps/chat-gpt-pass-turing-test/> [abgerufen am 20.6.2023]
- Spiegel Netzwelt (2023a). ChatGPT wieder in Italien, Spiegel, [online] <https://www.spiegel.de/netzwelt/netzpolitik/openai-bessert-beim-datenschutz-nach-chatgpt-wieder-in-italien-verfuegbar-a-d8d9d24f-8f34-4882-a8e4-dc5bc1c7fd84> [abgerufen am 29.04.2023]
- Spiegel Netzwelt (2023b). Microsoft will angeblich zehn Milliarden Dollar in KI-Firma investieren, Spiegel, [online] <https://www.spiegel.de/netzwelt/apps/chatgpt-microsoft-will-angeblich-zehn-milliarden-dollar-in-ki-firma-investieren-a-cad5c5e6-76bf-4e99-8b0e-ab8f572d2ee8> [abgerufen am 20.06.2023]
- Tagesschau (2023). Künstliche Intelligenz: Italien sperrt ChatGPT, Tagesschau, [online] <https://www.tagesschau.de/ausland/europa/italien-chatgpt-ki-101.html> [abgerufen am 29.05.2023]
- Taylor, C. (2023). Five Ways ChatGPT Helps You Hack, CyberHoot Advisory, [online] <https://cyberhoot.com/advisory/five-ways-chatgpt-helps-you-hack/> [abgerufen am 18.05.2023]

Terwiesch, C. (2023). Would Chat GPT Get a Wharton MBA?: A Prediction Based on Its Performance in the Operations Management Course, Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania, [online] <https://mackinstitute.wharton.upenn.edu/wp-content/uploads/2023/01/Christian-Terwiesch-Chat-GTP.pdf> [abgerufen am 18.05.2023]

UniReport 2.23 (2023). Schreibt die Künstliche Intelligenz künftig meine Texte?, Goethe-Universität Frankfurt am Main, [online] <https://aktuelles.uni-frankfurt.de/unireport/schreibt-die-kuenstliche-intelligenz-kuenftig-meine-texte/> [abgerufen am 13.06.2023]

Vaswani, A, Shazer, Noam, Parmer, Niki, et al. (2017). Attention Is All You Need, [online] <https://arxiv.org/abs/1706.03762> [abgerufen am 03.06.2023]

Verwimp, L. (2017). Character-Word LSTM Language Models, [online] <https://arxiv.org/abs/1704.02813> [abgerufen am 03.06.2023]

von Platen, M. (2023). ChatGPT: Schreiben künftig Maschinen Texte anstelle von Menschen?, Eberhard Karls Universität Tübingen, [online] <https://uni-tuebingen.de/en/universitaet/aktuelles-und-publikationen/newsletter-uni-tuebingen-aktuell/2023/1/forschung/1/> [abgerufen am 13.06.2023]

Wang, D. (2020). Multi-Head Self-Attention with Role-Guided Masks, [online] <https://arxiv.org/abs/2012.12366> [abgerufen am 05.06.2023]

Wolf, Z. B. (2023). AI can be racist, sexist and creepy. What should we do about it?, CNN, [online] <https://edition.cnn.com/2023/03/18/politics/ai-chatgpt-racist-what-matters/index.html> [abgerufen am 18.04.2023]

Woolf, M. (2022). congrats to OpenAI on winning the Turing Test. Twitter, [online] https://twitter.com/minimaxir/status/1600258754993262594?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1600258754993262594%7Ctwgr%5E430bd5d909df0408c46c9c8c44902c5b3115e87c%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fmpost.io%2Fde%2Fchatgpt-besteht-den-Turing-Test%2F [abgerufen am 20.06.2023]

Zorz, Z. (2023). A bug revealed ChatGPT users' chat history, personal and billing data, Help Net Security, [online] <https://www.helpnetsecurity.com/2023/03/27/chatgpt-data-leak/> [abgerufen am 02.07.2023]