HOCHSCHULE RUHR WEST
UNIVERSITY OF APPLIED SCIENCES

Der Markt der Large Language Models – Eine Analyse mit ausgewählten empirischen Daten

Bachelorarbeit
Studiengang BWL – industrielles Dienstleistungsmanagement
der Hochschule Ruhr West

Christopher Grüter
10012645

Erstprüfer:	Prof. Dr. Michael Vogelsang
Zweitprüfer:	Dr. Christoph Bierbrauer

Mülheim an der Ruhr, Januar 2024

Inhaltsverzeichnis

Inhaltsverzeichnis I
Abbildungsverzeichnis II
Tabellenverzeichnis III
Abkürzungsverzeichnis IV
1 Einleitung 1
2 Large Language Modelle 4
2.1 Der Generative Pre-Trained Transformer 4
2.2 Wie lernt ChatGPT? 9
2.3 LLM-spezifische Benchmarks 12
2.4 APIs und Open Source Software 13
3 Theoretische Grundlagen 14
3.1 Der Markt und seine Marktformen 14
3.2 Monopolisierungstreiber 17
3.3 Wettbewerbstreiber 19
3.4 Eigenschaften digitaler Produkte 20
4 Vorstellung der für die Analyse relevanten Unternehmen 22
5 Empirische Überprüfung des Marktes 26
5.1 Die aktuelle Marktsituation 27
5.2 Netzwerkeffekte und zweiseitige Märkte 30
5.3 Produktdifferenzierung 32
5.4 Nutzungsvoraussetzungen der Angebote 39
5.5 Lock-In-Effekte und Wechselkosten 40
5.6 Kostenstruktur 42
5.7 Diskussion 44
6 Fazit und Marktprognose 46
7 Zusammenfassung 50
Literaturverzeichnis 51
Anhang A: Anzahl der Seitenaufrufe 63
Anhang B: Berechnungen 64
Erklärung 65

Abbildungsverzeichnis

Abbildung 1: Aufbau eines Decoder-basierten Transformers 4
Abbildung 2: Beispielhafte Berechnung der Self-Attention 6
Abbildung 3: Masked Self-Attention Mechanismus und Maskierung der Attention-Scores 7
Abbildung 4: Aufbau eines Transformer-Blocks 8
Abbildung 5: Ablauf des Fine-Tuning Prozesses von ChatGPT 11
Abbildung 6: Summe der Seitenaufrufe der Chatbots von Juni bis Oktober 2023 28
Abbildung 7: Konzentrationskurve auf Basis der Seitenaufrufe 28
Abbildung 8: Seitenaufrufe von ChatGPT und Bard seit Veröffentlichung 29
Abbildung 9: Benchmark-Ergebnisse des Premium-Segments 34
Abbildung 10: Benchmark-Ergebnisse des Standard-Segments 35
Abbildung 11: Maximal mögliche Token-Anzahl der Eingabesequenz 36
Abbildung 12: API-Preise der LLMs für 100 Millionen Input- und Output-Tokens 37

Tabellenverzeichnis

Tabelle 1: Top 50 - Generative KI Anwendungen nach Seitenaufrufen (Juni 2023)..... 27
Tabelle 2: Welche Benchmarks sind für die LLMs vorhanden? 33

Abkürzungsverzeichnis

API	Application Programming Interface
CW	content word
DL	Deep Learning
FT	Fine-Tuning
FW	function word
GPT	Generative Pre-Trained Transformer
GWB	Gesetz gegen Wettbewerbsbeschränkungen
HHI	Herfindahl-Hirschman-Index
KI	Künstliche Intelligenz
LLM	Large Language Modell
LLMs	Large Language Modelle
ML	Machine Learning
MoE	Mixture of Experts
PPO	Proximal Policy Optimization
PT	Pre-Training
RNN	Recurrent Neural Networks
SA	Self-Attention
TA	Transformer-Architektur

1 Einleitung

Mit der Veröffentlichung von ChatGPT im November 2022 erlangte das Forschungsgebiet der Large Language Modelle (LLMs) großes öffentliches sowie mediales Interesse (Zhao et al. (2023, S.1), Bartholomew & Mehta (2023)). Der Chatbot von OpenAI ist eines der am schnellsten wachsenden online Angebote aller Zeiten; ChatGPT erreichte beispielsweise schneller eine Million Nutzer[1] als die sozialen Netzwerke Instagram oder Facebook (Buchholz, 2023). OpenAIs Chatbot ermöglicht es seiner großen Nutzeranzahl generative künstliche Intelligenz in ihrem Alltag zu nutzen (Taulli, 2023, S.8). Basierend auf Deep Learning (DL) Techniken kann ChatGPT menschenähnlichen Text generieren (Yenduri et al., 2023, S.3). DL ist ein Teilbereich des Machine Learning (ML) (Gronwald, 2023, S. 85). ML wird von Zhou (2021) definiert als „*[…] the technique that improves system performance by learning from experience via computational methods.*" Die in der Definition genannten Erfahrungen sind die verwendeten Input-Daten (Zhou, 2021, S.32). Ziel ist es aus den Input-Daten Muster und Verknüpfungen zu erkennen, die für die Aufgabenlösung benötigt werden. Um zu gewährleisten, dass diese Muster erkannt werden, müssen beim ML im Vorhinein Merkmale aus den Input-Daten herausgefiltert werden. Der Prozess der Merkmalsbildung erfolgt manuell und ist für die Güte des Modells ausschlaggebend. Anschließend wird das Modell durch den Einsatz von Lernalgorithmen automatisch erstellt. Neuronale Netze automatisieren den Prozess der Merkmalsbildung. Beim DL können dadurch auch unstrukturierte und große Datenmengen verwendet werden (Janiesch et al., 2021, S.688f.). ML ist wiederum ein Teilbereich der künstlichen Intelligenz (KI) (Gronwald, 2023, S.85). Eine allgemeingültige Definition für KI gibt es nicht, vielmehr wird versucht den Begriff der KI zu umschreiben (Ertel, 2021, S.1). Nach Moring (2023) „*[beschäftigt sich KI] mit Methoden, die es einem Computer ermöglichen, solche Aufgaben zu lösen, die, wenn sie vom Menschen gelöst werden, Intelligenz erfordern.*" Die Forschung im Bereich der KI geht einige Jahrzehnte zurück. Alan Turing beschäftigte sich bereits 1950 in seinem Artikel „Computing machinery and intelligence" mit der Frage, ob Computer denken können. Hierzu entwickelte er einen Test, in dem ein Proband mit einem Computer und einem Menschen eine Unterhaltung führt. Der Proband kann beiden Fragen stellen und muss am Ende feststellen, wer der beiden Kommunikationspartnern der Mensch ist. Gelingt dem Probanden dies nicht, gilt die Maschine als intelligent (Turing, 1950). Auch ist ChatGPT keineswegs der erste Chatbot aller Zeiten. Im Jahre 1966 wurde dieser mit dem Namen Eliza entwickelt. Eliza sollte als eine Art Psychotherapeut agieren und seinem Anwender Antworten in Form von Fragen geben. Der Chatbot PARRY wurde 1972 entwickelt, gefolgt vom Chatbot ALICE

[1] Wegen der besseren Lesbarkeit wird in dieser Bachelorarbeit das generische Maskulinum verwendet. Die in dieser Arbeit verwendeten Personenbezeichnungen beziehen sich - sofern nicht anders kenntlich gemacht - auf alle Geschlechter.

1995. Dieser war der erste Computer, welcher den Titel „most human computer“ erhielt (Adamopoulou & Moussiades, 2020, S.374).

Die Einsatzbereiche von KI innerhalb von Unternehmen sind vielseitig. Neben dem Einsatz in der Datenanalyse kann beispielsweise auch der Bereich des Marketing oder der Kundenkommunikation davon profitieren (Klöckner, 2023). Mercedes Benz testet derweilen den Einsatz von ChatGPT zur Unterstützung des Qualitätsmanagements. ChatGPT soll produktionsrelevante Daten auswerten und darstellen, sodass die Mitarbeiter bei ihren Aufgaben bestmöglich unterstützt werden. Es wird die Möglichkeit geschaffen Abfragen per Sprache durchzuführen, ohne dass Programmierkenntnisse benötigt werden (Mercedes Benz, 2023). Die Veröffentlichungen in Bezug auf das Thema der LLMs umfassen die verschiedensten Bereiche. Neben der Erarbeitung eines standardisierten Qualitätsbewertungsprozesses (siehe Zheng et al. (2023), Liang et al. (2023), Liu et al. (2023)) wird zum Beispiel auch der Einfluss von LLMs auf die Umwelt analysiert (siehe Faiz et al. (2023), Samsi et al. (2023)). Neben der Modellbewertung werden auch die Auswirkung der Nutzung auf die Modelle (siehe Brown et al. (2020), Kojima et al. (2022)) oder die Möglichkeiten der Modellverbesserung betrachtet (siehe Ouyang et al. (2022), Singhal et al. (2023)). Fortführend umfassen die Veröffentlichungen verschiedenste Themenbereiche, wie die Medizin (siehe Jeblick et al. (2023), Khan et al. (2023), Sedaghat (2023), Kung et al. (2023) King (2023)), die Lehre und die damit verbundene Vorbeugung der missbräuchlichen Nutzung (siehe Gao et al. (2022), Susnjak (2022), Zhai (2022)). Ein weiterer Forschungsbereich stellt die ökonomische Auswirkung von LLMs auf den Arbeitsmarkt dar. Eloundou et al. (2023) haben die Auswirkungen von LLM gestützter Software sowie von LLMs auf den Arbeitsmarkt analysiert. Forschungsergebnis ist die Erkenntnis, dass 80% der betrachteten amerikanischen Arbeitskräfte bei 10% ihrer Aufgaben mit KI unterstützt werden könnten. LLMs können bei 15%, LLM gestützte Software bei 47 bis 56% der gesamten Aufgaben eine unterstützende Rolle einnehmen. Dadurch können die Arbeitskräfte die Aufgaben bei gleichbleibender Qualität schneller lösen. Auch ist die Implementierung von LLMs aufgrund der umfassenden Fähigkeiten in vielen Arbeitsbereichen denkbar. Das Massachusetts Institute of Technology (MIT) hat in einer Studie die Auswirkung auf die Produktivität durch den Einsatz von KI analysiert. Noy & Zhang (2023) kamen zu dem selben Ergebnis. Der Einsatz von ChatGPT kann bei Schreibaufgaben zu einer höheren Produktivität und besseren Ergebnissen führen. Eine weitere Studie von Goldman Sachs überprüft die Auswirkung von KI auf das Wirtschaftswachstum. Briggs & Kodnani (2023) kamen zu dem Fazit, dass 25% der aktuellen Arbeitsaufgaben Automatisierungspotenzial durch generative KI aufzeigen und damit die Produktivität gesteigert werden könne. Auch eine Steigerung des BIPs und die Schaffung neuer Arbeitsplätze durch den Einsatz von generativer KI werden erwartet.

Der Markt der LLMs und die damit einhergehende Analyse und Prognose der Marktsituation wird zum jetzigen Zeitpunkt nicht betrachtet. Hieraus ergibt sich die interessante Fragestellung, wie sich der Markt der LLMs entwickeln wird. ChatGPT hat wie oben beschrieben sehr schnell eine große Anzahl an Nutzern hinzugewonnen, wodurch folgende Forschungsfrage entsteht: „Besteht die Gefahr, dass OpenAI Monopolstatus im Markt

der LLMs erreicht?“. Um die Forschungsfrage beantworten zu können führt Kapitel 2 zunächst in das Thema der LLMs ein. Hierzu wird die Funktionsweise des Generative Pre-Trained Transformer (GPT) erläutert; dieser bildet die Technik hinter ChatGPT (Jahn, 2023). Im nächsten Teil des zweiten Kapitels wird der Lernprozess von ChatGPT erklärt, gefolgt von der Einführung des Lesers in die LLM-spezifischen Benchmarks. Abgeschlossen wird das Kapitel mit einer kurzen Erklärung der Funktionsweise einer API und der Definition von dem Begriff der Open Source Software. Die theoretischen Grundlagen für die Analyse des Marktes werden dem Leser in Kapitel 3 vermittelt. Mit einer Einführung in die verschiedenen Monopolformen, der Erklärung, wie ein Markt eingegrenzt und die Marktkonzentration gemessen werden beginnt dieses Kapitel. Im Anschluss daran wird auf die für die Analyse relevanten Monopolisierungs- sowie Wettbewerbstreiber eingegangen. Der letzte Abschnitt des Kapitels umfasst die Besonderheiten digitaler Produkte. Die Vorstellung der für die Analyse relevanten Unternehmen und ihrer Produkte erfolgt in Kapitel 4. Die zum Beantworten der Fragestellung benötigte Analyse wird in Kapitel 5 durchgeführt. In Kapitel 6 werden die Ergebnisse der Analyse zusammengefasst und die Forschungsfrage der Bachelorarbeit wird beantwortet. Außerdem wird eine Prognose über die zukünftige Marktform des LLM Marktes erarbeitet.

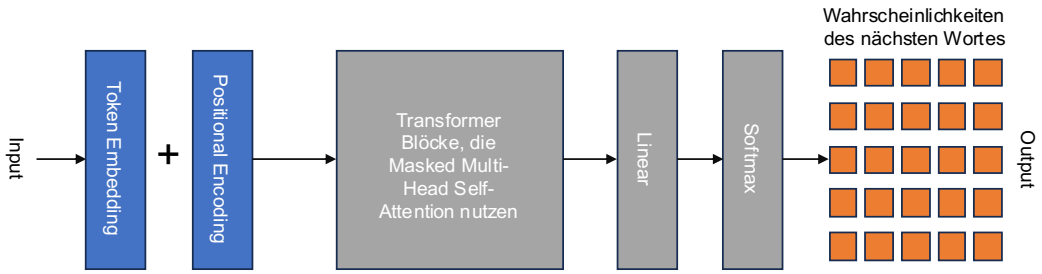
2 Large Language Modelle

Kapitel 2 führt in die Technik der LLMs ein. Zunächst wird die Technik hinter ChatGPT, also der Generative Pre-Trained Transformer erklärt, gefolgt von einer Erläuterung, wie ChatGPT trainiert und verbessert wird. Im Anschluss daran werden vier LLM spezifischen Benchmarks vorgestellt. Abschließend erfolgt eine Erklärung darüber, was eine API ist und welche Besonderheiten Open Source Softwares besitzen.

2.1 Der Generative Pre-Trained Transformer

Eine Version des LLM GPT, welches von OpenAI veröffentlicht wurde, dient als Basis für den Chatbot ChatGPT (Jahn, 2023). Allgemein stellt ein Chatbot ein System dar, mit dem eine Person einen textbasierten Dialog führen kann. Die Texteingabe des Nutzers erhält das System über eine Benutzeroberfläche (Mackmood et al., 2020, S.151). Mit Hilfe einer minimalen Texteingabe ist es LLMs möglich eine große Menge an menschenähnlichen Text zu generieren (Yenduri et al., 2023, S.3), wobei die Eingabe des LLMs als Prompt bezeichnet wird (Hüsch et al., 2023, S.4). Die Art einen Prompt zu gestalten kann die Qualität der Ausgabe maßgeblich beeinflussen (Horvitz & Nori, 2023). Mit dem Paper „Attention is all you need“ veröffentlichte das Google Entwicklerteam rund um Ashish Vaswani 2017 den ersten Transformer. Die auf Self-Attention (SA) basierende Architektur wurde als Alternative zu den bis dato dem Stand der Technik entsprechenden Recurrent Neural Networks (RNN) entwickelt (Vaswani et al., 2017, S.1f.). Die Transformer-Architektur (TA) bildet die Basis für das GPT Modell (Yenduri et al., 2023, S.5). Allgemein bildet die TA in vielen Bereichen den aktuellen Stand der Technik ab. Aufgrund von Vereinfachungen der TA kann die typische Encoder-Decoder-Architektur durch Decoder-basierte oder Encoder-basierte Architekturen abgelöst werden, welche je nach Anwendungsfall gewählt werden (Lanquillon & Schacht, 2023, S.154).

Abbildung 1: Aufbau eines Decoder-basierten Transformers



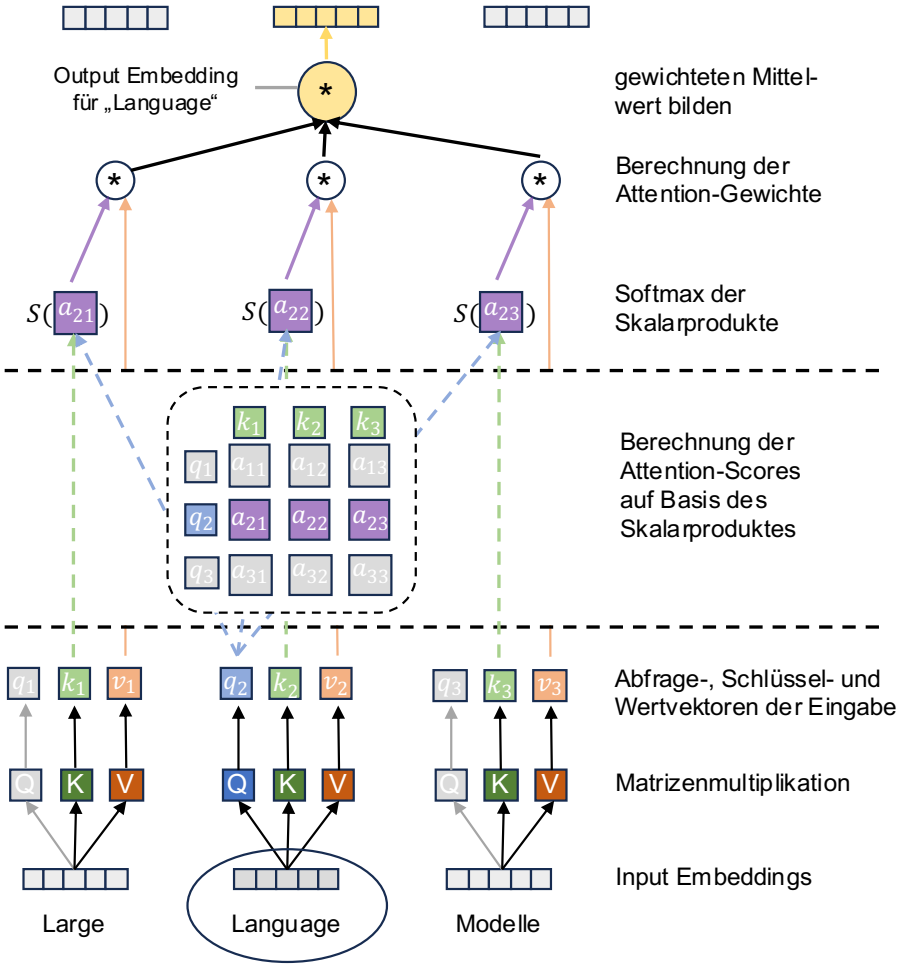
Quelle: Eigene Darstellung nach: Lanquillon & Schacht (2023, S.170)

OpenAI nutzt für seine GPT-Modelle den Decoder-basierten Ansatz. Abbildung 1 zeigt den Aufbau eines Decoder-basierten Transformers, dessen Funktionsweise im folgenden Teil der Arbeit genauer beschrieben wird. Die Stärke dieser TA liegt in der Vorhersage des nächsten Wortes einer Textfolge, weshalb diese gerade zur Texterzeugung genutzt werden (Tunstall et al., 2023, S.110f.). In der Abbildung 1 wurde die TA in drei

Bestandteile aufgeteilt, erkennbar an den Farben. Die in der Abbildung 1 blau hinterlegten Komponenten stellen die Eingabesequenz dar. Mit Hilfe des Token Embedding wird sichergestellt, dass der Transformer den Input verarbeiten kann (Lanquillon & Schacht, 2023, S.150). Abhängig von der Form der Tokenisierung kann ein Token verschiedene Ausprägungen haben (Hüsch et al., 2023, S.209f.). OpenAI gibt an, dass bei einem Text mit 750 Wörtern circa 1000 Tokens benötigt werden (OpenAI, 2023l). Beim Token Embedding passiert eine Umwandlung der eben beschriebenen Tokens in Vektoren. Diese Vektoren haben mehrere Dimensionen und beinhalten Informationen über die Bedeutung und die Syntax des Tokens innerhalb des vorliegenden Textes. Es entsteht also ein Vektorraum, welcher alle Tokens umfasst. Je näher die Tokens von der Bedeutung her aneinander liegen, desto näher ist deren Position zueinander (Lanquillon & Schacht, 2023, S.150). Neben dem Token Embedding erfolgt in der Eingabesequenz auch das Positional Encoding. Der Transformer selber hat keine Informationen über die relative Position der Tokens in einem Satz. Mit Hilfe des Positional Encoding wird den Inputs des Transformers eben diese Positionsinformation mitgegeben (Yenduri et al., 2023, S.7). Der Informationsmangel über die Position der Tokens kommt durch die Funktionsweise der TA. Unter Nutzung von Attention-Mechanismen ist es dem Transformer möglich den Input parallel zu bearbeiten, was allerdings dazu führt, dass die Position der Tokens verändert wird. Bei den früher verwendeten RNN-Architekturen konnte der Input lediglich Token für Token verarbeitet werden, wodurch das Positional Encoding bei den RNN nicht nötig ist. Die in Abbildung 1 grau hinterlegten Komponenten bilden den Text generierenden Transformer ab (Lanquillon & Schacht, 2023, S.151ff.). Als erste Komponente ist ein Transformerblock zu sehen, welcher Masked Multi-Head Self-Attention nutzt. Um diese Art der SA zu erklären, wird im folgenden Abschnitt zunächst die reine SA erklärt, um darauf aufbauend die Besonderheiten der Masked Multi-Head Self-Attention herauszuarbeiten. Der SA Mechanismus bietet durch die oben angesprochene parallele Verarbeitung der Eingabesequenz den Vorteil, dass auch weit auseinanderliegende Wörter die Verarbeitung der Eingabesequenz maßgeblich beeinflussen können. Durch ihn wird also ein kontextualer Zusammenhang der Wörter einer Eingabesequenz generiert, auch wenn diese in einem Satz nicht direkt nebeneinanderstehen. Der Kontext des Wortes innerhalb der Eingabesequenz sowie das Wort selber bieten die Grundlage für den Transformer die Wörter der Eingabesequenz auf eine neue Art darzustellen (Krauss, 2023, S.167f.). Wie oben beschrieben, werden die Wörter der Eingabesequenz in Tokens umgewandelt (Hüsch et al., 2023, S.4) und durch das Token Embedding in einen Vektor umgewandelt (Lanquillon & Schacht, 2023, S.150). Auf Grundlage jedes einzelnen dieser Vektoren erfolgt die Berechnung drei weiterer Vektoren. Neben dem Abfragevektor (q), werden ein Schlüsselvektor (k) sowie ein Wertvektor (v) errechnet (Krauss, 2023, S.168). Der gewünschte Output wird dem Transformer durch die Abfrage (q) vermittelt. Hierfür wird dem Transformer mitgegeben, nach welcher Token-Beziehung dieser suchen soll. Jedes Token trägt eine Information mit sich, welche durch den Schlüsselvektor (k) dargestellt wird. Der Wertvektor (v) gibt an, welche Information von dem betrachteten Token an die weiteren Tokens übermittelt werden soll (Lanquillon & Schacht,

2023, S.142f.). Abbildung 2 führt den Ablauf des SA Mechanismus beispielhaft für das Input Embedding „Language“ auf.

Abbildung 2: Beispielhafte Berechnung der Self-Attention

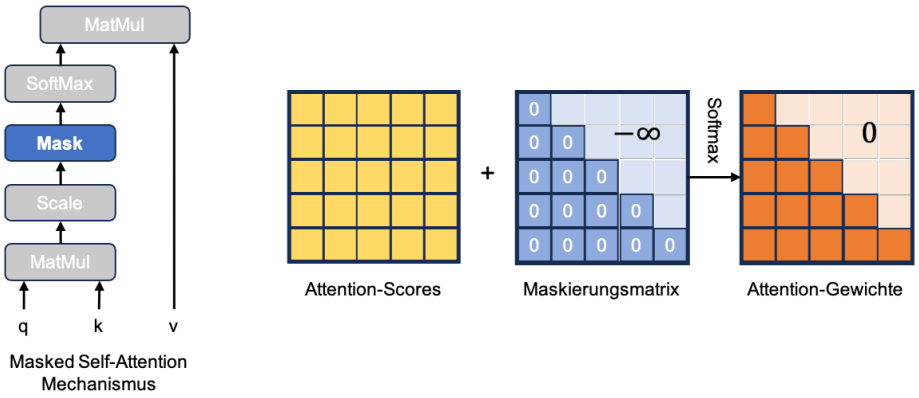


Quelle: Eigene Darstellung nach: Paaß & Hecker (2020, S.209) und Krauss (2023, S.169)

Die gerade beschriebenen Vektoren werden durch die Multiplikation der Eingabevektoren mit den Matrizen Q, K und V berechnet. Während des Trainingsprozess werden die Werte der Matrizen gelernt und bilden somit die erlernbaren Parameter des LLM ab (Paaß & Hecker, 2020, S.209f.). Nach der Berechnung der drei Vektoren durch die Matrixmultiplikation werden nun die Skalarprodukte der Abfrage- und Schlüsselvektoren errechnet, diese werden auch als Attention-Scores beschrieben (Tunstall et al., 2023, S.89). Zur Berechnung der Attention-Scores eines Wortes wird der Abfragevektor eines Tokens mit jedem Schlüsselvektor der Eingabesequenz multipliziert, in der Abbildung 2 wird das für das Wort „Language“ beispielhaft gezeigt. Dieser Prozess erfolgt für jeden Abfragevektor der Eingabesequenz. Mit Hilfe der Attention-Scores kann jedem Token der Eingabesequenz eine Wichtigkeit in Bezug auf das Verständnis der anderen Tokens eben dieser Eingabesequenz zugeschrieben werden (Krauss, 2023, S.145). Im nächsten Schritt werden aus den Attention-Scores die Attention-Gewichte berechnet. Die Möglichkeit, dass Skalarprodukte jeden beliebigen Wert annehmen können, kann zu einer

Beeinträchtigung des Trainingsprozesses führen. Aus diesem Grund erfolgt zunächst die Skalierung der Varianz durch die Multiplikation der Attention-Scores mit einem Skalierungsfaktor. Davon ausgehend werden die Attention-Gewichte durch Normalisierung unter Einsatz einer Softmax-Funktion errechnet. Die Summe der Attention-Gewichte jedes einzelnen Token ergibt hierdurch immer eins (Tunstall et al., 2023, S.89). Abschließend wird noch ein neues Ausgabe Embedding errechnet. Hierzu wird aus den zuvor berechneten Attention-Gewichten und den Wertvektoren ein gewichtetes Mittel gebildet. In dem hier verwendeten Beispiel wird für den Input „Language“ also ein neues Output-Embedding generiert. Diese Berechnung der neuen Output-Embeddings wird genauso für alle Tokens der Eingabesequenz durchgeführt (Paaß & Hecker, 2020, S.209f.). Die Output-Embeddings können als neue Darstellung der Input Tokens verstanden werden. Diese beinhalten neben der eigenen Bedeutung durch die gewichtete Summe Informationen über die kontextuellen Zusammenhänge aller anderen Tokens der Eingabesequenz (Krauss, 2023, S.170). Der Masked Self-Attention Mechanismus legt lediglich die Informationen fest, auf die ein Token innerhalb der Eingabesequenz zugreifen kann. Bei der oben beschriebenen SA ist es jedem Token möglich, mit allen Tokens der Eingabesequenz zu kommunizieren (Lanquillon & Schacht, 2023, S.5). Der SA Mechanismus innerhalb des Decoders soll nur auf die Tokens zugreifen dürfen, welche in der Eingabesequenz vor dem betrachteten Token kommen, sowie auf das betrachtete Token selber. Vor der Softmax-Funktion werden alle nicht gewollten Verbindungen maskiert, indem diese Werte auf $-\infty$ gesetzt werden (Vaswani et al., 2017, S.5).

Abbildung 3: Masked Self-Attention Mechanismus und Maskierung der Attention-Scores



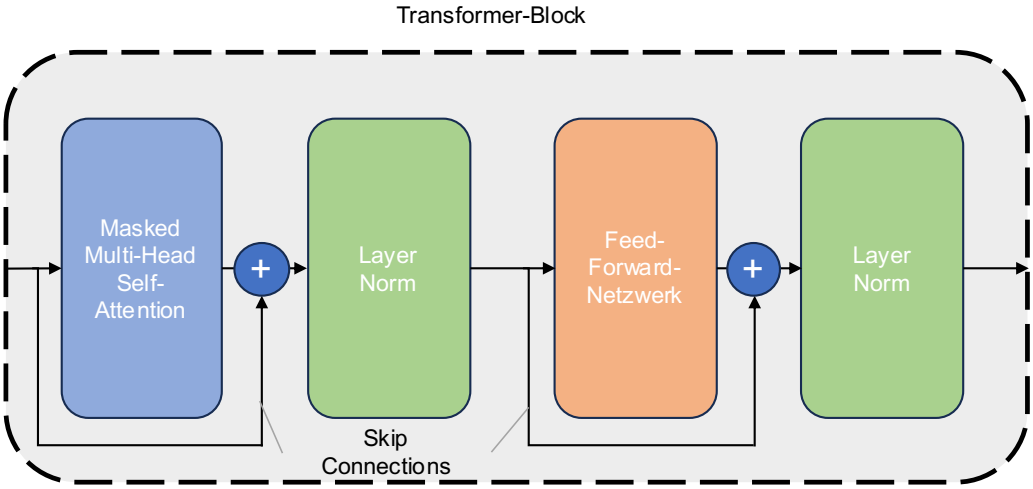
Quelle: Eigene Darstellung nach: Vaswani et al. (2017, S.4) und Lanquillon & Schacht (2023, S.147)

Dies geschieht durch das Addieren einer sogenannten Maskierungsmatrix, welche die Eigenschaft besitzt, dass die Positionen der sichtbaren Tokens innerhalb der Maskierungsmatrix den Wert 0 zugewiesen bekommen. Dies hat zur Folge, dass die Attention-Scores dieser Positionen nicht verändert werden. Die Softmax-Funktion wird wie oben beschrieben auf die gesamte Matrix angewendet. Die maskierten Positionen liefern als Ergebnis den Wert 0, wodurch diese nicht mehr beachtet werden (Lanquillon & Schacht, 2023, S.146). In Abbildung 3 werden einerseits der Aufbau der Masked Self-Attention

und andererseits die Vorgehensweise des Maskierens unter Anwendung der Maskierungsmatrix dargestellt.

Die Multi-Head-Attention zeichnet sich durch die parallele Nutzung mehrerer SA Mechanismen aus (Reuter et al., 2023, S.65). Multi-Head-Attention hat den Vorteil, dass sich mehrere Attention-Heads auf verschiedene Aspekte des Satzes konzentrieren können. Ein Attention-Head kann sich immer nur auf einen Aspekt der Eingabesequenz fokussieren, vergleichbar damit, wie das Subjekt eines Satzes mit dem Verb interagiert. Werden mehrere solcher Attention-Heads parallelgeschaltet, kann dieses Problem umgangen werden. Der verwendete Aspekt der Eingabesequenz wird nicht für die Attention-Heads vorgegeben, sondern während des Trainingsprozesses erlernt. (Tunstall et al., 2023, S.95). Jeder Attention-Head generiert eine eigene Ausgabe nach dem oben beschriebenen Prinzip. Die Ausgaben der Attention-Heads werden konkateniert, wodurch die Größe der Eingabe- und Ausgabematrix identisch bleibt (Lanquillon & Schacht, 2023, S.147f.). Abbildung 4 skizziert den Aufbau eines Transformer-Blocks.

Abbildung 4: Aufbau eines Transformer-Blocks



Quelle: Eigene Darstellung nach: Lanquillon & Schacht (2023, S.149)

Neben dem eben erklärten Masked Multi-Head Self-Attention Mechanismus besteht der Transformer-Block noch aus einer „Feed-Forward-Schicht“. Außerdem kommen sogenannte „Skip Connections“ und „Layer-Normalization“ Schichten zum Einsatz, die vor allem den Lernprozess unterstützen sollen. Die „Feed-Forward“ Schicht verknüpft die vom Masked Multi-Head Self-Attention Mechanismus kontextualisierten Daten mit dem Ziel den Kontext der Eingabesequenz noch besser interpretieren zu können (Lanquillon & Schacht, 2023, S.148f.). Aktuelle Transformer-Architekturen haben eine Vielzahl solcher Transformer-Blöcke hintereinandergeschaltet; diese bilden die Schichten eines neuronalen Netzes ab. Hierbei ist der Output eines Transformer-Blocks der Input des folgenden Blocks (Krauss, 2023, S.170). In Abbildung 1 ist zu sehen, dass nach dem letzten Transformer-Block nochmal eine lineare- und Softmax-Funktion angewendet wird. Die nächsten Tokens werden durch eine Prognose auf Basis einer linearen

Funktion bestimmt. Die Softmax-Funktion generiert eine Wahrscheinlichkeitsverteilung bezogen auf die Output-Tokens (Yenduri et al., 2023, S.7). Basierend auf der Wahrscheinlichkeitsverteilung wird ein beliebiges Token als nächstes Token der Sequenz ausgewählt. Aufgrund dieses Ansatzes wird auch ein Token ausgewählt, welches nicht die höchste Wahrscheinlichkeit hat, wodurch auch bei selbem Input nicht zweimal derselbe Output generiert wird. Das ermittelte Token wird neuer Bestandteil der ehemaligen Eingabesequenz. Diese neue Eingabesequenz wird verwendet um nach demselben Prinzip ein weiteres Token zu generieren, sie ist also als neuer Input für den Transformer zu verstehen. Dieser Vorgang wird so lange wiederholt, bis der geforderte Output generiert wurde (Krauss, 2023, S.170).

Ob bei der Generierung eines neuen Tokens alle Parameter des Modells genutzt werden, hängt von der verwendeten Architektur ab. Bei sogenannten „Dense“ Modellen werden alle Parameter für jedes Token verwendet; ein Beispiel hierfür wäre GPT-3. Bei dieser Version des LLM von OpenAI werden alle 175 Millionen Parameter verwendet, um ein Token zu erzeugen. Patel & Wong haben in ihrem Artikel die Architektur von GPT-4 analysiert und sind zu dem Ergbenis gekommen, dass für diese Version des Modells der „Mixture of Experts“ (MoE) Ansatz verwendet wurde (Patel & Wong, 2023). Bei der MoE Architektur wird ein großes LLM aus mehreren unabhängig voneinander agierenden kleineren LLMs zusammengesetzt; diese kleineren werden als Experten bezeichnet. Durch diesen Ansatz kann im Trainings- und Generierungsprozess Rechenleistung gespart werden, da nicht alle Parameter des Modells für jedes Token verwendet werden (Jacobs et al., 1991). Nach Patel & Wong besteht GPT-4 aus 16 verschiedenen Experten, bestehend aus jeweils circa 111 Milliarden Parametern. Um ein Token zu erzeugen sind immer zwei Experten miteinander verknüpft. Für den Attention Mechanismus wird ein Modell mit circa 55 Milliarden Parametern von allen acht Expertenpaaren simultan verwendet. Um ein Token zu generieren werden demnach um die 280 Milliarden Parameter verwendet, obwohl das gesamte GPT-4 Modell aus circa 1,8 Trillionen Parametern besteht (Patel & Wong, 2023).

2.2 Wie lernt ChatGPT?

Nachdem die Funktionsweise des GPTs in Kapitel 2.1 erklärt wurde, bezieht sich Kapitel 2.2 nun darauf, wie der Transformer von OpenAI trainiert wurde. Der Trainingsprozess kann hierbei in 2 Phasen unterteilt werden. In der Pre-Training (PT) Phase lernt der Transformer, unter Anwendung von „Unsupervised Learning“ Techniken, das nächste Token einer Sequenz vorherzusagen. Es werden statistische und kontextuale Zusammenhänge von Wörtern erlernt. Die Fine-Tuning (FT) Phase spezialisiert den Transformer für das Lösen von spezifischen Aufgaben (Yenduri et al., 2023, S.7). Der genaue Ablauf der beiden Phasen wird im weiteren Teil ausgeführt.

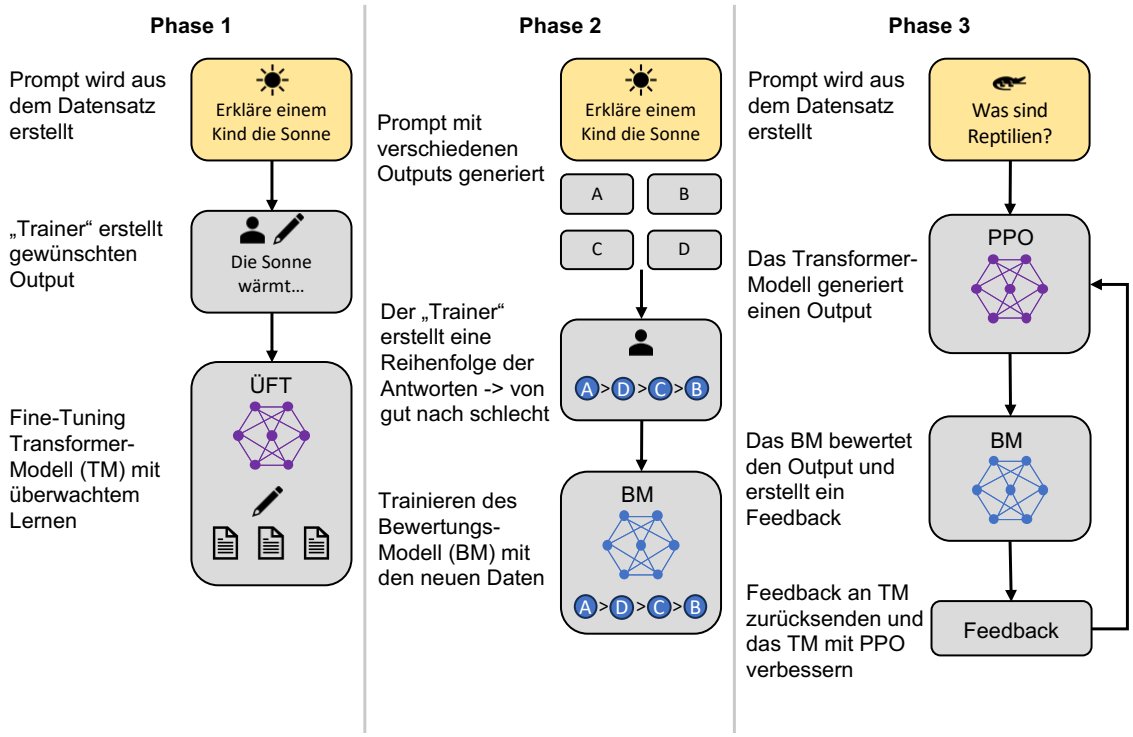
Das OpenAI Forscherteam rund um Radford et al. (2018) hat die Vorteile des unüberwachten Lernens und folgendem FT in ihrer Veröffentlichung „Improving Language Understanding by Generative Pre-Training“ diskutiert. Das in diesem Zuge trainierte Modell

konnte Verbesserungen im Vergleich zu den damals aktuellen Modellen in einigen LLM-spezifischen Benchmarks verzeichnen (Radford et al., 2018). Es wird von unüberwachtem Lernen gesprochen, wenn während des Lernprozesses keine Lernziele festgelegt werden; eine Rückmeldung bekommt das System auch nicht (Klüver et al., 2021, S.208). Eine Information darüber, was genau erkannt werden soll, existiert demnach nicht. (Cisek, 2021, S.67ff). Dem System wird lediglich ein sehr großer Datensatz vorgelegt, aus dem das System selber Merkmale erkennen muss (Rebala et al., 2019, S.27). Der Datensatz im PT Prozess besteht aus unbeschrifteten (unlabeled) Daten (Sarrion, 2023, S.27). Je unterschiedlicher die Merkmale des Datensatzes sind, desto besser ist das zu erwartende Ergebnis. Die Analyse der Wörter in einem Satz bietet sich dafür an, da gelernt werden kann ein Wort auf Basis des vorangegangenen Wortes vorherzusagen; dieses Lernvorgehen wird selbstüberwachtes Lernen genannt. Obwohl es sich hierbei um eine Form des überwachten Lernens handelt, gibt der Mensch nicht vor, welche Label innerhalb der Daten existieren (Paaß & Hecker, 2020, S.46). Der verwendete Algorithmus übernimmt diese Aufgabe der Label-Bildung. Die Informationen über die vorhandenen Labels werden von den Daten selbst vorgegeben. Auf Grundlage der Labels passt der Transformer seine Gewichte (Parameter) während des Trainingsprozesses an. Im Anwendungsfall von ChatGPT wird das Label durch den kontextuellen Zusammenhang der Wörter eines Satzes erzeugt. Dadurch ist es dem Modell möglich, die Bedeutung der einzelnen Wörter kontextbezogen zu lernen. Vorteil dieses Lernverfahrens ist, dass deutlich mehr unbeschriftete Daten im Vergleich zu beschrifteten Daten vorhanden sind (Krauss, 2023, S.132). Die Daten für das PT von ChatGPT werden von einer Vielzahl an Quellen bezogen. Neben Webseiten werden beispielsweise auch Bücher und Zeitungsartikel verwendet (Sarrion, 2023, S.27). Welcher Datensatz genau für das PT des LLMs von ChatGPT verwendet wurde, ist allerdings nicht bekannt (Schaul et al., 2023). Am Ende des PT Prozesses ist der Transformer in der Lage den Kontext eines Wortes innerhalb eines Satzes zu verstehen und auf Basis einer Eingabesequenz das nächste Wort vorherzusagen (Sarrion, 2023, S.27). Der Prozess des PT benötigt sehr viel Kapital, Rechenleistung und Zeit. Die benötigten Ressourcen steigen mit der Anzahl der Parameter des Modells und der Anzahl der verwendeten Tokens. Durch die aktuell vorhandene Rechenleistung und die diesbezüglichen Kosten wird die Skalierung der Größe von LLMs unter Betrachtung der Wirtschaftlichkeit und Trainingsdauer nach oben hin begrenzt (Patel, 2023).

Nach Abschluss der PT Phase erfolgt nun das FT von ChatGPT. Der FT Prozess für ChatGPT besteht aus drei Phasen und es werden Techniken des überwachten Lernens und Verstärkungslernens angewendet (Krauss, 2023, S.172). Allgemein nimmt der Prozess des FT nur minimale Änderungen an der gesamten Modellstruktur vor (Radford et al., 2018, S.2ff.). Für das FT werden im Vergleich zum PT sehr kleine Datensätze benötigt; allerdings sollten diese eine hohe Qualität haben, um das bestmögliche Ergebnis zu erzielen (Dai et al., 2023, S.2ff.). In der ersten Phase des FT Prozesses wird dem Transformer-Modell gezeigt, wie dieses auf bestimmte Prompts reagieren soll (Ouyang et al., 2022, S.3). Hierzu wird ein Prompt aus einer Datenbank genommen und ein menschlicher Trainer gibt vor, welches die erwartete Ausgabe ist. Diese vom Trainer generierten

Antworten werden für das FT des Modells genutzt (Ouyang et al., 2022, S.4). Unterstützung gab es für die „Trainer“, indem diese auf im Vorhinein generierte Antwortmöglichkeiten zugreifen konnten. Die Antworten mussten nur etwas angepasst werden. Auf Basis dieser Dialoge wurde ein neuer Datensatz erstellt, welcher durch künstlich erstellte Beispiele erweitert wurde (Krauss, 2023, S.172f.). Der Datensatz dient als Grundlage, um das Modell im ersten Schritt mit überwachtem Lernen zu verbessern (Ouyang et al., 2022, S.3). Anders als beim unüberwachtem Lernen müssen also die erwarteten Outputs der jeweiligen Prompts vorhanden sein (Joshi, 2020, S.10). Die Parameter des Modells werden durch den Prompt und den dazugehörigen Output des Datensatzes angepasst. Das Modell lernt hierdurch, wie es auf neuartige Prompts richtig reagieren kann (Rebala et al., 2019, S. 19f.). In der nächsten Phase wird ein Belohnungs-Modell in Form eines zweiten neuronalen Netzes erstellt (Krauss, 2023, S.173). Die Trainingsdaten werden wieder durch einen Trainer erzeugt, allerdings erstellt dieser nicht die gewünschte Antwort, sondern sortiert vier mögliche Antworten zu einem Prompt hinsichtlich deren Qualität. Der daraus entstehende Trainingssatz wird verwendet, um das Belohnungs-Modell zu trainieren. In der dritten Phase wird das Transformer-Modell mit „Reinforced Learning“ Techniken weiter verbessert (Ouyang et al., 2022, S.3). Hierbei wird dem Modell nicht vorgegeben, welcher der erwartete Output zu einem spezifischen Prompt ist. Das Modell erhält für den generierten Output ein Feedback. Auf Grundlage des positiv oder negativ ausfallenden Feedbacks wird das Modell verbessert (Cisek, 2021, S.37f.).

Abbildung 5: Ablauf des Fine-Tuning Prozesses von ChatGPT



Quelle: Eigene Darstellung nach: Ouyang et al. (2022, S.3)

Aus dem zuvor erstellten Datensatz wird ein Prompt als Input für das Transformer-Modell genommen. Dieses erstellt einen Output, der durch das in Phase 2 trainierte

Belohnungs-Modell bewertet wird. Das Feedback des Bewertungs-Modells wird zurück zum Transformer-Modell geleitet und für die Verbesserung des Modells genutzt. Die Verbesserung erfolgt durch Einsatz des „Proximal Policy Optimization“ (PPO) Algorithmus, durch den das Transformer-Modell angepasst wird, sodass das Feedback des Belohnungs-Modells besser ausfällt (Ouyang et al., 2022, S.3ff.). Der PPO-Algorithmus wurde von Schulman et al. (2017), einem OpenAI Forscherteam, veröffentlicht. Die genaue Funktionsweise des PPO-Algorithmus ist für die Bachelorarbeit allerdings nicht von Relevanz und wird nicht weiter erläutert. Wird der Datensatz durch weitere Daten ergänzt, können Phase 2 und 3 kontinuierlich wiederholt werden. Auf Basis der neuen Daten wird das Belohnungs-Modell angepasst (Phase 2), woraufhin das Transformer-Modell verbessert wird (Phase 3) (Ouyang et al., 2022, S.6). In Abbildung 5 ist der FT Prozess mit seinen 3 Phasen aufgeführt.

2.3 LLM-spezifische Benchmarks

Kapitel 2.3 gibt einen Überblick über die für die Bachelorarbeit relevanten LLM-spezifischen Benchmarks. Insgesamt werden vier verschiedene Benchmarks in der Bachelorarbeit genutzt, um die Qualität der einzelnen LLMs miteinander zu vergleichen: MMLU, HellaSwag, HumanEval und GSM8K. Diese vorgefertigten Datensätze sollen die Evaluierung von LLMs ermöglichen, die sich auf unterschiedliche Wissensbereiche und Aufgabentypen beziehen. Ziel ist es die verschiedenen LLMs hinsichtlich ihrer Kapazitäten und ihrer Grenzen zu testen (Naveed et al., 2023, S.19ff.).

MMLU ist eine Abkürzung für „Massive Multitask Language Understanding“ und ist ein Datensatz mit 57 Aufgaben, welche verschiedene Schwierigkeitsgrade abdecken. Der Datensatz wurde 2021 von Hendrycks et al. veröffentlicht. Die Aufgaben der MMLU Benchmark umfassen Fragen aus den Bereichen der Geistes- und Sozialwissenschaften sowie des MINT Bereiches. Neben diesen drei Hauptbereichen werden auch Fragen aus Themenbereichen wie beispielsweise der Medizin gestellt. Die Aufgaben werden in Form von Multiple-Choice-Fragen gestellt (Hendrycks et al., 2020, S.3ff.). Die HellaSwag Benchmark testet die Fähigkeit eines LLMs einen gegebenen Satz sinnvoll zu vervollständigen. Dem LLM wird zunächst ein Kontext einer Situation vorgegeben, sowie vier mögliche Ausgänge dieser Situation. Nur einer dieser möglichen Ausgänge ist sinnvoll und dementsprechend das richtige Ergebnis der Aufgabe (Zellers et al., 2019, S.1f.) Der HumanEval Datensatz überprüft die Fähigkeit eines LLM Programmieraufgaben zu lösen. Der Datensatz besteht hierfür aus 164 Aufgaben, welche händisch erstellt worden sind. Anhand der Aufgaben kann überprüft werden, wie gut das Sprachverständnis und das logische Denkvermögen eines LLM sind. Außerdem wird überprüft, wie gut die vom LLM verwendeten Algorithmen und die Fähigkeit einfache Mathematik anzuwenden sind. Durch das Datenset wird in einem Programmierkontext überprüft, wie akkurat die Funktionalität und die Problemlösungsfähigkeit eines LLM sind (Chen et al., 2021, S.4).

Die letzte für die Bachelorarbeit relevante Benchmark ist die GSM8K Benchmark, bestehend aus 8.500 mathematischen Textausgaben. Das Aufgabenniveau ist mit dem von Grundschulaufgaben vergleichbar. Gerade mathematische Problemstellungen stellen

LLMs immer wieder vor Probleme. Die Aufgaben benötigen zwischen zwei und acht Rechenschritten um diese zu lösen, wobei die Lösungsschritte lediglich auf den vier Grundrechenarten basieren (Cobbe et al., 2021, S.2ff.).

2.4 APIs und Open Source Software

Wollen Unternehmen ein LLM für deren eigene Anwendungsfälle nutzen, können diese das mit Hilfe einer „Application Programming Interface“ (API) machen (Hüsch et al., 2023, S.5). Die aus Programmiercode bestehende API schafft die Möglichkeit auf eine Anwendung oder einen Service zuzugreifen (Block, 2023, S.108). Die API ist demnach keine eigenständige Software, sondern ermöglicht die Nutzung einer Software, ohne diese auf dem eigenen Rechner zu installieren (Frank et al., 2021, S.19). So können auch große LLMs genutzt werden, ohne selbst die benötigte Serverkapazität zu besitzen (Zhao et al., 2023, S.11). Der Zugang zu der API wird den Unternehmen gewährt, indem diese einen API-Schlüssel erhalten. Dieser dient der Authentifizierung und schützt die Anwendung vor dem Zugriff unberechtigter Personen (Hüsch et al., 2023, S.108). Die Bezahlung für die Nutzung der APIs geschieht bei den LLMs auf Basis der verwendeten Tokens bei der Ein- und Ausgabesequenz. Einer festgelegten Menge an Tokens steht dementsprechend ein Preis gegenüber und der Kunde zahlt nur die Leistung, die er auch wirklich genutzt hat. Der Preis der Tokens für die Eingabe- und Ausgabesequenz unterscheidet sich meist noch voneinander (siehe OpenAI (2023l), Anthropic (2023c)). Ist der Quellcode einer Software für jede Person öffentlich zur Verfügung gestellt, wird von Open Source Software gesprochen. Veränderungen an der Software kann durch jede Person vorgenommen werden, wodurch jeder die Software seinen Wünschen entsprechend anpassen kann. Der Entwickler der Software stellt diese kostenlos zur Verfügung, kann allerdings trotzdem Regeln und Voraussetzungen für die Nutzung der Software festlegen (Haun, 2014, S.417).

3 Theoretische Grundlagen

Im folgenden Kapitel werden die für die Analyse benötigten theoretischen Grundlagen erläutert. Kapitel 3.1 gibt eine Einführung in die theoretischen Grundlagen für den Markt. Dazu werden die für die Bachelorarbeit relevanten Monopolformen aufgeführt und Kriterien für deren Vorhandensein definiert. Neben der Marktform des Monopols werden das natürliche Monopol, das Teilmonopol und die monopolistische Konkurrenz betrachtet. Anschließend wird erklärt, wie der relevante Markt eingegrenzt werden kann und welche Maße zur Messung der Marktkonzentration existieren. Im nächsten Abschnitt werden die relevanten Monopolisierungstreiber aufgeführt. Neben Economies of Scale und Scope wird die Theorie der zweiseitigen Märkte und Netzwerkeffekte betrachtet. Abgeschlossen wird Kapitel 3.2 mit Lock-In-Effekten und Wechselkosten. Neben den Monopolisierungstreibern sind auch Wettbewerbstreiber für die Bachelorarbeit relevant. Kapitel 3.3 bezieht sich darauf, welchen Einfluss die Produktdifferenzierung auf den Wettbewerb hat. Abgeschlossen wird das Kapitel mit der Definition digitaler Güter und der Besonderheiten dieser in Bezug auf Preisstrategien und deren Eigenschaften.

3.1 Der Markt und seine Marktformen

Ein Monopol liegt allgemein vor, wenn in einem bestimmten Markt alle Produkte nur durch ein Unternehmen angeboten werden. Diesem stehen auf der Nachfrageseite eine Großzahl an Nachfragern gegenüber. Die nachgefragte Menge hängt allerdings trotzdem von dem Preis der angebotenen Güter ab; jeder Nachfrager bringt eine individuelle Zahlungsbereitschaft mit sich (Blum et al., 2006, S.46). Bei einer Preiserhöhung durch den Monopolisten verliert dieser allerdings nicht alle Abnehmer. Hohe Markteintrittsbarrieren sind eine weitere Annahme, welche für ein Monopol vorliegen müssen (Tremblay & Tremblay, 2012, S.146). Den angebotenen Güter des Monopolisten stehen außerdem keine nahen Substitute anderer Unternehmen gegenüber (Bester, 2016, S.29).

Liegen bei einer Produktion hohe Fixkosten und geringe Grenzkosten vor, können sich hierdurch sinkende Durchschnittskosten ergeben. Die Grenzkosten liegen unterhalb der Durchschnittskosten, wodurch positive Skaleneffekte erzielt werden (Weimann, 2006, S.17). Diese werden in Kapitel 3.2 noch genauer erläutert. Kann dadurch ein Unternehmen die vom Markt geforderte Menge kostengünstiger produzieren als zwei Unternehmen dies könnten, liegt ein natürliches Monopol vor (Bühler & Jaeger, 2002, S.23). Das bedeutet, dass die Summe der Kosten eines Unternehmens kleiner ist als die Summe der Kosten von n Unternehmen bei Produktion derselben Menge. Ist diese Voraussetzung gegeben, liegt die sogenannte Subadditivität der Kosten vor (Wied-Nebbeling, 2004, S.35f.). Die Subadditivität erweitert den Bereich, in dem ein natürliches Monopol entsteht. Selbst wenn keine positiven Skalenerträge mehr generiert werden können, also sinkende Durchschnittskosten vorliegen, liegt noch so lange ein natürliches Monopol vor, bis zwei kleinere Unternehmen die geforderte Menge günstiger produzieren können (Bühler & Jaeger, 2002, S.24). Hieraus folgt, dass positive Skaleneffekte lediglich ein hinreichendes aber nicht notwendiges Kriterium für ein natürliches Monopol sind. Diese

Definition des natürlichen Monopols bezieht sich auf die Herstellung eines Produktes (Knieps, 2008, S.25). Im Mehrproduktfall können neben den positiven Skaleneffekten auch Verbundvorteile entstehen; auch diese werden in Kapitel 3.2 näher erklärt. Subadditivität liegt im Mehrgüterfall vor, wenn ein Unternehmen alle Produkte zusammen günstiger produzieren kann als dies mehrere Unternehmen könnten (Weimann, 2006, S.332). Die Macht des Monopolisten in Bezug auf die Preissetzung hängt davon ab, ob es sich um einen angreifbaren Markt handelt (Wied-Nebbeling, 2004, S.36). Baumol (1982) definiert einen angreifbaren Markt als einen Markt, „*[…] into which entry is absolutely free, and exit is absolutely costless*." Der freie Eintritt ist so zu verstehen, dass neue Marktteilnehmer keinen Kostennachteil gegenüber einem bereits am Markt agierenden Unternehmen erleiden. Beim Marktaustritt ist es wichtig, dass Unternehmen all ihre Kosten des Markteintritts zurückerhalten können, abzüglich der Kosten, die aus der Tätigkeit der Unternehmen zu erwarten sind (Baumol, 1982, S.3f.). Können bei Marktaustritt nicht alle zu erwartenden Kosten rückgängig gemacht werden, liegen versunkende Kosten vor. Diese sind durch ihre Irreversibilität charakterisiert. Neben fixen Kosten können auch variable Kosten versenkt werden. Wenden Unternehmen hohe Summen an versunkenden Kosten auf, ist davon auszugehen, dass diese so lange im Markt bleiben, bis die erwirtschafteten Umsätze mindestens diese decken können (Blum et al., 2006, S. 85f.). Handelt es sich um einen angreifbaren Markt, kann das Produkt lediglich zu Durchschnittskosten angeboten werden. Würde ein höherer Preis angesetzt werden, würden neue Konkurrenten in den Markt eintreten. Bei existierenden Markteintrittsschranken liegt die Macht über den Markt allerdings beim Monopolisten und er kann den Preis selber bestimmen. (Wied-Nebbeling, 2004, S.36ff.). Bain (1956) definiert drei Marktzutrittsbarrieren, denen sich neue Marktteilnehmer potenziell stellen müssen. Die erste von Bain angesprochene Barriere spiegelt sich in den eben angesprochenen positiven Skaleneffekten wider. Das Vorhandensein von Produktdifferenzierungen ist die zweite Barriere. In diesem Fall präferieren Kunden die bereits am Markt angebotenen Produktvarianten. Durch diese Präferenz fällt es neuen Anbietern schwer die Kunden des alten Unternehmens abzuwerben. Als letzte Markteintrittsbarriere nennt Bain das Vorhandensein von absoluten Kostenvorteilen. Diese bedeuten, dass ein am Markt etabliertes Unternehmen immer einen Kostenvorteil gegenüber neuen Unternehmen hat (Bain, 1956). Neben diesen dreien können auch Lernkurveneffekte sowie Netzwerkeffekte als eine solche Barriere agieren. Versunkene Kosten können als Marktaustrittsbarriere interpretiert werden (Wied-Nebbeling, 2004, S.11f.).

Die beiden eben beschriebenen Monopolformen definieren sich dadurch, dass nur ein Anbieter innerhalb eines Marktes vorhanden ist. Anders ist dies beim Vorliegen eines Teilmonopols; hier gibt es ein Unternehmen mit einem sehr großen Marktanteil und viele weitere Unternehmen, die einen geringen Marktanteil vorweisen. Genaue Marktanteile sind nicht definiert, es kommt lediglich darauf an, dass die kleineren Unternehmen keine Macht innerhalb des Marktes besitzen und der Teilmonopolist den Status des Preisführers erlangt. Ein Teilmonopol kann entstehen, wenn ein Unternehmen einen neuen Markt eröffnet und sich dadurch schon einmal ein Image aufbauen konnte (Wied-Nebbeling, 2004, S.65). Durch das Gesetz gegen Wettbewerbsbeschränkungen (GWB) wird in

Deutschland die wirtschaftliche Macht von einzelnen Unternehmen geregelt und der faire Wettbewerb gesichert. Nach §18 Absatz 1 hat ein Unternehmen eine marktbeherrschende Stellung, wenn es keine Wettbewerber gibt, kein wesentlicher Wettbewerb besteht oder das Unternehmen eine im Vergleich deutlich bessere Marktstellung besitzt. Laut Absatz 4 ist ein Unternehmen ab mindestens 40 Prozent Marktanteil in einer marktbeherrschenden Position (Gesetz gegen Wettbewerbsbeschränkungen. Bundesministerium der Justiz, 2023). Im Falle heterogener Güter verliert der Teilmonopolist den Status des Preisführers. Haben die kleinen Unternehmen ein schlechteres Image, müssen diese den Preis für ihre Güter geringer halten als der Teilmonopolist. Tendenziell können die kleineren Unternehmen eine neue Käuferschicht akquirieren, allerdings ist davon auszugehen, dass die Kunden mit der stärksten Kaufkraft beim Teilmonopolisten bleiben, sofern dieser sein Image dauerhaft hochhalten kann (Wied-Nebbeling, 2004, S.71).

Eine weitere Marktform stellt die monopolistische Konkurrenz dar. Diese kann auch als Polypol beschrieben werden, welches sich auf einem unvollkommenen Markt befindet. Bei dieser Marktform existieren mehrere Anbieter mit jeweils sehr geringen Marktanteilen. Dadurch hat keines der Unternehmen die Macht die Handlungen des Konkurrenten durch preispolitische Entscheidungen wirklich zu beeinflussen. Dem gegenüber stehen eine große Menge an kleinen Nachfragern (Premer, 2015, S.351) Der Markt muss die Eigenschaften des angreifbaren Marktes besitzen (Pindyck et al., 1998, S. 525). Eine weitere Voraussetzung für die monopolistische Konkurrenz ist das Vorhandensein nicht vollständiger Substitute auf der Angebotsseite. Die angebotenen Produkte können sich beispielsweise in der Produktart oder der Qualität unterscheiden (Letzner, 1997, S.16). Hierdurch hat jeder Anbieter bei vorhandener Nachfrage die Preissetzungsmacht eines Monopolisten für seinen Anteil des Marktes (Bühler & Jaeger, 2002, S.58).

Im folgenden Abschnitt wird definiert, wie der relevante Markt abgegrenzt werden kann und welche Parameter es gibt, um die Marktkonzentration zu messen. Die Markteingrenzung gilt als wichtige Vorbereitung für die Bestimmung der Marktkonzentration (Tremblay & Tremblay, 2012, S.184). Es muss sichergestellt werden, dass nur Unternehmen in die Markteingrenzung mit einbezogen werden, die einen Einfluss auf den Wettbewerb innerhalb des Marktes haben (Schmidt & Haucap, 2013, S.61). Der Markt kann auf Basis der Produkteigenschaften abgegrenzt werden. Die Konsumtheorie nach Lancaster besagt, dass Produkte wegen bestimmter Eigenschaften nachgefragt werden. Nur die Produkte, die den Eigenschaften entsprechen, werden in die Marktbetrachtung mit einbezogen (Wied-Nebbeling, 2004, S.15). Die Produkteigenschaften werden durch Auflistung der wichtigen Charakteristiken definiert (Bester, 2016, S.20). Die Bewertung der Marktkonzentration erfolgt über den Herfindahl-Hirschman-Index (HHI) sowie die Darstellung der Marktanteile mit Hilfe einer Konzentrationskurve. Die Berechnung des HHI erfolgt durch die Anzahl der im Markt agierenden Unternehmen und deren jeweiligen Marktanteil. Die Marktanteile der Unternehmen werden zunächst quadriert und im Anschluss summiert. Maximal kann der HHI einen Wert von 10.000 erreichen, da im Monopolfall ein Unternehmen 100% Marktanteil besitzen würde und 100 zum Quadrat 10.000 ergibt. Der minimale Wert des HHI ist wiederum von der Anzahl der Marktteilnehmer abhängig. Hierzu

wird 10.000 durch die Anzahl der Unternehmen gerechnet (Waldman & Jensen, 2013, S.105ff.). Liegt der Wert des HHI zwischen 1.500 und 2.500 geht im Übrigen das US-amerikanische Justizministerium davon aus, dass der Markt eine moderate Konzentration aufweist. Bei Werten über 2.500 wird von einer hohen Konzentration gesprochen (Monopolkommission, 2022, S.54). Diese Grenzen werden für die Interpretation des HHI innerhalb dieser Bachelorarbeit verwendet. Eine Konzentrationskurve stellt die Verteilung einer Gesamtmenge an Daten bezogen auf eine bestimmte Variable dar. Es erfolgt eine Abbildung der Konzentrationsraten (Woeckener, 2014, S.87). Die Daten werden zunächst der Größe nach sortiert, angefangen mit dem größten Wert. Daraufhin werden für jede Variable jeweils der Anteil an der Gesamtmenge berechnet. Es werden die kumulierten prozentualen Werte gebildet und in einem Diagramm dargestellt. Auf der X-Achse werden die Variablen und auf der Y-Achse der kumulierte Anteil der Gesamtmenge abgebildet (Pepall et al., 2014, S.47f.).

3.2 Monopolisierungstreiber

Kapitel 3.2 beschreibt die gängigen und für die Arbeit relevanten Monopolisierungstreiber. Zunächst werden Economies of Scope und Scale erläutert und herausgearbeitet, wie diese sich voneinander unterscheiden. Als nächstes werden Netzwerkeffekte betrachtet, wobei diese in direkte und indirekte Netzwerkeffekte unterschieden werden (Srinivasan, 2023, S.44f.). Zur Erklärung der indirekten Netzwerkeffekte wird die Theorie der zweiseitigen Märkte verwendet. Abgeschlossen wird das Kapitel mit der Definition von Lock-In-Effekten und Wechselkosten und welche Bedeutung diese haben.

Economies of Scale wurden schon im vorherigen Kapitel angesprochen um die Entstehung des natürlichen Monopols herzuleiten. In diesem Zuge wurden positive Skaleneffekte benannt, welche zur Folge haben, dass die durchschnittlichen Produktionskosten mit zunehmender Ausbringungsmenge fallen (Weimann, 2006, S.328). Skaleneffekte können auch negativer oder konstanter Natur sein. Im Fall von negativen Skaleneffekte steigt die Durchschnittskostenfunktion mit jeder weiteren Einheit, bei den konstanten Skaleneffekten liegen konstante Durchschnittskosten vor (Bühler & Jaeger, 2002, S.17). Positive Skaleneffekte können realisiert werden, wenn das Unternehmen zunächst eine große Menge an Fixkosten tragen muss, während der Produktion allerdings geringe variable Kosten vorzuweisen hat (Holtforth, 2016, S.21). Solange sich die Grenzkosten der Produktion unterhalb der Durchschnittskosten befinden, kann von positiven Skaleneffekten profitiert werden (Weimann, 2006, S.328). Economies of Scope entstehen demgegenüber, wenn ein Unternehmen keine Kostenvorteile aus der Produktion einer weiteren Einheit eines Produktes generieren kann, sondern mindestens zwei verschiedene Produkte günstiger herstellen kann, als dies zwei Unternehmen könnten (Premer, 2015, S.349). Die Produktion eines weiteren Gutes wird dem entsprechend positiv durch die Produktion des ersten Gutes beeinflusst (Bühler & Jaeger, 2002, S.21). Economies of Scale und Scope sind nicht auf direkte Weise voneinander abhängig. Trotz negativer Skaleneffekte in der Produktion ist das Vorhandensein von Economies of Scope möglich (Pindyck et al., 1998, S.273). Im Mehrproduktfall führen Economies of Scope also wie in

Kapitel 3.1 beschrieben zur Subadditivität der Kosten (Weimann, 2006, S.332). Ein Grund für das Existieren von Verbundvorteilen kann sein, dass bei der Produktion zweier Güter dieselben Input-Faktoren benötigt werden (Tremblay & Tremblay, 2012, S.43). Die Betrachtung der Kostenstruktur eines Unternehmens ist für die Bestimmung der Marktstruktur von Relevanz. Besonders entscheidend sind hierbei die langfristigen Durchschnittskosten (Wied-Nebbeling, 2004, S.7). Je länger ein Unternehmen von sinkenden Durchschnittskosten profitieren kann, desto weniger Marktteilnehmer wird es in dem betrachteten Markt geben (Peters, 2010, S.16).

Positive Netzwerkeffekte bewirken die Nutzensteigerung eines Produktes aus der Sicht des Nutzers. Wie oben beschrieben lassen sich diese in zwei verschiedene Typen unterscheiden: direkte und indirekte Netzwerkeffekte. Besteht eine direkte Verbindung zwischen der Nutzensteigerung und der Anzahl der Nutzer, wird von direkten Netzwerkeffekten gesprochen (Sauer, 2019, S.36). Je mehr Leute also das Produkt nutzen, desto größer ist der empfundene Nutzen. Dadurch entscheiden sich wiederum mehr Leute dazu das Produkt zu nutzen (Lipczynski et al., 2013, S.619). Erfolgt der Nutzenzuwachs nicht über die steigende Anzahl an Nutzern auf derselben Marktseite, sondern auf der gegenüberliegenden Marktseite, entstehen indirekte Netzwerkeffekte (Srinivasan, 2023, S.41f.). Indirekte Netzwerkeffekte gelten als Eigenschaft von zweiseitigen Märkten. Die Existenz eines zweiseitigen Marktes wird durch die Verbindung zweier unabhängiger Märkte mit Hilfe einer zweiseitigen Plattform begründet. Durch eine steigende Teilnehmerzahl des ersten Marktes erhöht sich der Nutzen der Marktteilnehmer des zweiten Marktes und umgekehrt. Existiert eine große Teilnehmerbasis innerhalb des ersten Marktes, werden neue Teilnehmer in den zweiten Markt eintreten, wodurch wieder neue Teilnehmer in den ersten Markt eintreten. (Dewenter & Rösch, 2015, S.118f.). Netzwerkeffekte können auch negativer Natur sein. Dies ist der Fall, wenn die Qualität des Gutes negativ durch eine höhere Nutzeranzahl beeinflusst wird, beispielsweise durch Überbelastung einer Plattform. Die Sicherstellung des Wachstums einer Plattform erfolgt über die Generierung einer Ausgangslage, in der beide Seiten der Plattform von positiven Netzwerkeffekten profitieren können. (Srinivasan, 2023, S.44). Bei der Einführung eines Netzwerkgutes wird in der Anfangsphase versucht so schnell wie möglich eine große Nutzeranzahl anzuziehen. Dafür bietet das Unternehmen sein Produkt zunächst zu geringen Preisen an, um diese im Laufe der Zeit zu erhöhen. Die Zahlungsbereitschaft der Kunden steigt mit zunehmenden positiven Netzwerkeffekten (Peters, 2010, S.43ff.). Treten Unternehmen zu unterschiedlichen Zeitpunkten in den Markt ein, hat das erste im Markt befindliche Unternehmen die Möglichkeit sich eine Nutzerbasis aufzubauen. Diese Nutzerbasis wird auch als „Installed Base“ betitelt und kann als Markteintrittsbarriere fungieren. Der generierte Vorteil des ersten Unternehmens wird auch „First Mover Advantage“ genannt. Je größer die Installed Base bei Markteintritt des Konkurrenten desto mehr Mittel muss dieser aufwenden, um eine eigene Nutzerbasis aufzubauen (Peters, 2010, S.50). Wie groß dieser First-Mover Vorteil ist, hängt demnach von der Nutzeranzahl sowie von den vorhandenen Wechselkosten ab. Dadurch entsteht ein Lock-In-Effekt, der dafür sorgt, dass sich die Kunden an das Produkt binden und zur Installed Base werden. Abhängig von der Marktgröße und den Nutzerpräferenzen können trotzdem

verschiedene Produkte nebeneinander in dem Markt existieren (Wied-Nebbeling, 2004, S.288). Können die Marktteilnehmer mehrere miteinander in Konkurrenz stehende Plattformen nutzen, wird von Multihoming gesprochen. Ist die Möglichkeit für Multihoming vorhanden, können die Teilnehmer einfach das Angebot einer anderen Plattform wahrnehmen. Der Wettbewerbsdruck der Unternehmen steigt, woraus ein erhöhter Preiswettbewerb resultiert. Die Wahrscheinlichkeit für Multihoming sinkt mit steigenden Preisen der einzelnen Plattformen sowie durch eine Ausschließlichkeitsbindung der Marktteilnehmer an die jeweilige Plattform. (Dewenter & Rösch, 2015, S.50).

Im oberen Absatz wurden Lock-In-Effekte und Wechselkosten bereits im Zuge der Kundenbindung innerhalb der zweiseitigen Märkte genannt. Gelingt es einem Anbieter seine Kunden an sein Produkt zu binden, wird von einem Lock-In gesprochen. Dieser kann durch die Erzeugung von Wechselkosten generiert werden. Wechselkosten können neben der Bindung des Kunden an ein Produkt zur Erzeugung von Markteintrittsbarrieren dienen (Peters, 2010, S.50). Burnham et al. (2003) definieren Wechselkosten als „*[…] the onetime costs that customers associate with the process of switching from one provider to another.*“ Wechselkosten sind demnach die einmaligen Kosten, welche ein Kunde tragen muss, um den Anbieter zu wechseln. Zu beachten ist hierbei, dass die Kosten nicht zwangsläufig direkt bei dem Wechsel anfallen müssen. Wechselkosten können durch den Wechselprozess bedingt sein. Zu diesen Wechselkosten zählen zum Beispiel, dass ein Kunde den Umgang mit einem neuen Produkt erlernen muss, die Kosten um ein neues Produkt einzurichten oder auch die, die zur Informationsbeschaffung über andere Angebote aufgewendet werden müssen. Auch eine Kundenbeziehung mit dem Unternehmen kann für die Entstehung von Wechselkosten verantwortlich sein (Burnham et al., 2003, S.110ff.). Die Bindung des Kunden an ein Produkt kann auch durch das Einführen von Produktstandards oder individuelle auf den einzelnen Kunden zugeschnittene Angebote erfolgen (Halberstadt, 2014, S.21). Der Nutzer kann des weiteren auf psychologischer oder emotionaler Ebene an das Produkt gebunden sein, wodurch von Wechselbarrieren gesprochen werden kann. Eine Wechselbarriere umfasst demnach alle Sachverhalte, durch die der Wechsel des Anbieters erschwert oder kostspielig gemacht wird. Die zur Überwindung dieser Barrieren aufgebrachten Kosten sind die jeweiligen Wechselkosten (Blut, 2008, S.28). Bei einem Anbieterwechsel geht der Kunde davon aus, einen höheren Nutzen für sich zu generieren. Ist der erwartete Nutzenzuwachs allerdings geringer als die anfallenden Wechselkosten, liegt ein Lock-In des Kunden bei dem jeweiligen Unternehmen vor. Je vergleichbarer die Produktqualitäten für den Anwender sind, desto geringer ist der Effekt des Lock-Ins. Eine höhere Qualität ist allerdings kein hinreichender Indikator einen Anbieterwechsel durchzuführen (Halberstadt, 2014, S.20f.). Neue Marktteilnehmer haben es deutlich schwieriger ihr Produkt innerhalb des Marktes zu platzieren (Busch, S.39).

3.3 Wettbewerbstreiber

Erfüllen zwei Produkte dieselbe Funktion, haben allerdings geringe Merkmalsunterschiede, wird von Produktdifferenzierung gesprochen (Tremblay & Tremblay, 2012,

S.164). Eine Voraussetzung dafür ist, dass die Nachfrager unterschiedliche Präferenzen haben. Begründen lassen sich diese beispielsweise auf sachlicher oder auch persönlicher Ebene. Auf sachlicher Ebene können die Produkte Unterschiede hinsichtlich der Qualität oder der Funktionalität haben; auf persönlicher Ebene spielt hingegen das Image des Unternehmens eine wichtige Rolle (Wied-Nebbeling, 2004, S.5). Ausgangspunkt der vertikalen Produktdifferenzierung ist die Einigkeit der Konsumenten über die bevorzugten Kriterien bei der Produktauswahl. Im Falle der Qualität sind sich alle Konsumenten einig, dass eine höhere Qualität bei selbem Preis zu bevorzugen ist. Ist die höhere Qualität an einen höheren Preis gekoppelt, werden sich einige Marktteilnehmer für das Produkt der höheren und einige für das Produkt der niedrigeren Qualität entscheiden (Tirole, 1999, S.210ff.). Voraussetzung für die Entstehung eines Preiswettbewerbs ist die Möglichkeit die Produktionskapazitäten kurzfristig zu verändern. Kann ein Unternehmen seine Preise senken und mehr Kunden anziehen, muss die zusätzliche Nachfrage auch bedient werden können. Wird ein homogenes Gut betrachtet, führt die Unterbietung des Preises der Konkurrenten dazu, dass dieses Unternehmen die gesamte Marktnachfrage erhält. Anders ist das bei heterogenen Gütern; die Intensität des Wettbewerbs über den Preis wird durch die Produktdifferenzierung reduziert (Woeckener, 2014, S.15f.). Durch eine Änderung des Preises verliert das Unternehmen also nicht die gesamte Nachfrage und ist dadurch nicht zwangsläufig an einen Marktpreis gebunden (Wied-Nebbeling, 2004, S.101). Bei der horizontalen Produktdifferenzierung herrscht keine Übereinstimmung darüber, welches Produkt die höhere Qualität hat (Pfähler & Wiese, 2008, S.250). Werden die Produkte zu demselben Preis angeboten hängt die Nachfrage des Kunden rein von seinen Präferenzen ab und er entscheidet, welche Merkmalsausprägung seine Kaufentscheidung beeinflusst (Tirole, 1999, S.213). Ein höherer Grad der Produktdifferenzierung wird durch „Mass Customization“ erlangt. Ziel ist es jedem Kunden ein individuelles, auf seinen Wünschen basierendes Produkt anzubieten. Somit wird keine Produktpalette im Markt angeboten, sondern der Kunde dazu befähigt die Attribute seines Produktes eigenständig zu definieren. Dadurch ist es möglich, die Zahlungsbereitschaft des Kunden zu erhöhen sowie einen größeren Bereich des Marktes zu bedienen. Vorausgesetzt wird, dass die Kosten der Individualisierung nicht den generierten Nutzen übersteigen (Peters, 2010, S.120ff.).

3.4 Eigenschaften digitaler Produkte

Digitale Produkte zeichnen sich durch die in der Produktion anfallenden hohen Fixkosten aus; dem stehen allerdings geringe variable Kosten gegenüber. Aus dieser Eigenschaft ist der Begriff der „First Copy Costs“ entstanden (Schmidt, 2007, S.34). Die Entwicklungskosten eines digitalen Produktes sind als versunkene Kosten zu kategorisieren, da die getätigten Aufwendungen für die Entwicklung nicht mehr rückgängig gemacht werden können. Durch diese können neue Konkurrenten abgehalten werden als Imitator in den Markt einzutreten und mit einer aggressiven Preisstrategie die Kunden des Marktführers abzuwerben (Fichert, 1998, S.346f.). Digitale Produkte zeichnen sich außerdem durch ihre Immaterialität aus, wodurch der Kunde nicht in der Lage ist die Qualität des

jeweiligen Gutes direkt zu bestimmen. Durch eben diese Immaterialität und die Schwierigkeit die Qualität zu bestimmen werden digitale Güter zu Erfahrungsgütern (Schmidt, 2007, S.34f.). Haben sich Nutzer für ein Erfahrungsgut entschieden, ist es für neue Unternehmen schwierig diese abzuwerben. Ein First-Mover-Vorteil kann entstehen, wenn das erste Unternehmen im Markt seine Nutzer von dessen Qualität überzeugt hat (Wied-Nebbeling, 2004, S.281). Für Unternehmen ist es hilfreich eine starke Marke aufzubauen, sodass der Kunde dadurch ein Vertrauen in das Produkt erhält (Schmidt, 2007, S.37) Die nächste Eigenschaft digitaler Güter ist deren Anpassbarkeit (Clement et al., 2019, S.36). Digitale Güter sind des Weiteren nicht an Produktionskapazitäten gebunden; die Vervielfältigung des Originals mit selber Qualität ist beliebig oft möglich (Sauer, 2019, S.41). Bei Produktinnovationen digitaler Güter wird vor allem über diejenige Innovation berichtet, welche als erstes auf dem Markt ist. Imitatoren genießen meist nicht die Aufmerksamkeit wie der First-Mover. Dadurch kann es dem First-Mover gelingen sein Produkt sehr schnell bei einer Vielzahl an potenziellen Kunden bekannt zu machen. Außerdem kann der erste im Markt zum Beispiel seinen Domainnamen selber auswählen und diesen Namen etablieren. Imitatoren haben keine Möglichkeit mehr denselben, potenziell in den Markt integrierten, Domainnamen zu nutzen (Halberstadt, 2014, S.92ff.). Eine andere Möglichkeit schnell eine große Menge an Nutzern für sein Produkt anzuwerben ist die Anwendung des Freemium-Geschäftsmodells (Hoisl, 2018, S.35). Dieses Geschäftsmodell wird vor allem für digitale Produkte genutzt. Die Funktionen des vorliegenden Produktes werden in Basis- und Premium-Funktionen eingeteilt. Die Basis-Funktionen des Produktes können durch die Nutzer gratis verwendet werden, wohingegen die vollumfängliche Nutzung aller Funktionen Geld kostet. Wichtig ist, dass die Produkte auch schon in der Gratis-Version genug Funktionen bieten, dass die Nutzung des Produktes einen Mehrwert für den Nutzer generieren kann. Die Basis-Version des Produktes kann so für eine große Zahl an potenziellen Nutzern zur Verfügung gestellt werden. Nutzer, die sehr zufrieden mit dem Produkt sind, werden sich entscheiden die Premium-Version zu erwerben. Die Zahlungsbereitschaft ist in diesem Fall auch höher, als wenn das Produkt direkt zu erwerben gewesen wäre (Seufert, 2014, S.1f.).

4 Vorstellung der für die Analyse relevanten Unternehmen

OpenAI wurde im Jahre 2015 als ein Non-Profit Unternehmen gegründet. Neben Reed Hoffman und Peter Thiel waren auch Elon Musk und Sam Altman an der Gründung beteiligt (Taulli, 2023, S.93). Elon Musk verließ 2018 das Unternehmen wegen Unstimmigkeiten über die Zusammenarbeit mit Microsoft (Jahn, 2023). Die Umstrukturierung der Unternehmensform von OpenAI wurde 2019 durchgeführt. Aufgrund der mit der KI-Forschung zusammenhängenden hohen Kosten durch benötigte Serverkapazitäten und Fachleute, wurde neben dem Non-Profit Unternehmen ein gewinnorientiertes Unternehmen gegründet. Die mögliche Rendite dieser hybriden Unternehmensstruktur, die Investoren erwirtschaften können, wird bei dieser Form auf einen bestimmten Faktor begrenzt. Die darüber hinaus erwirtschafteten Renditen gehen an den Non-Profit Teil des Unternehmens (OpenAI, 2019). Microsoft investierte 2019 zunächst 1 Milliarde Dollar in OpenAI und wiederholte dies im Jahre 2021. Anfang 2023 hat Microsoft eine weitere Investition in Höhe von 10 Milliarden Dollar getätigt, die sich über die nächsten Jahre erstrecken wird. (Bass, 2023). Aktuell gilt OpenAI als eines der wertvollsten Start-Ups innerhalb der USA und wird mit 29 Milliarden Dollar bewertet. (Jahn, 2023). OpenAIs Ziel ist es „*[…] to ensure that artificial general intelligence […] benfits all of humanity*“ (Altman, 2023a). OpenAI veröffentlichte am 30. November 2022 den kostenfreien Chatbot ChatGPT, welcher auf dem LLM von OpenAI läuft. Am 1. Februar 2023 wurde die Möglichkeit geschaffen ein ChatGPT Plus Abonnement abzuschließen. Für 20 Dollar im Monat werden neue Funktion bereitgestellt. Außerdem können die Nutzer auf das GPT-4 Modell zugreifen; die kostenlose Version verwendet weiterhin GPT-3.5 (Zhang, 2023, S.2). Seit der Veröffentlichung von ChatGPT Plus wurden den Nutzern über die Zeit neue Funktionen bereitgestellt. Zunächst wurde die Möglichkeit geschaffen Plug-Ins in ChatGPT zu nutzen. Entwickler erhalten die Möglichkeit diese für ChatGPT zu entwickeln, wodurch der Nutzer neue Anwendungsmöglichkeiten erhält. ChatGPT kann dadurch beispielsweise einen Urlaub über Expedia planen oder über Klarna Preise in Online Shops vergleichen (OpenAI, 2023c). Neben den Plug-Ins wurde außerdem eine App für das Betriebssystem iOS (OpenAI, 2023h) und ein ChatGPT Abonnement für Unternehmen mit „ChatGPT Enterprise“ veröffentlicht. Unternehmenskunden können dadurch alle Funktionen von ChatGPT Plus nutzen, haben allerdings keine Nutzungskapazitäten und deren Daten werden gesondert geschützt (OpenAI, 2023f). Im September 2023 erhielten ChatGPT Plus und Enterprise Nutzer außerdem die Funktionen Sprach- sowie Bildeingaben zu tätigen. Über die Smartphone-App können Spracheingaben getätigt werden, die Antwort erfolgt auf Basis einer Sprachausgabe. Die Funktion Bilder als Eingabe zu nutzen ist auf allen Plattformen verfügbar (OpenAI, 2023a). Einen Monat später erfolgte noch die Integration von DALL·E 3, wodurch der Nutzer Bilder durch Texteingaben erstellen kann (OpenAI, 2023e). Im November 2023 wurden GPTs vorgestellt, wodurch ChatGPT Plus und Enterprise Nutzer ChatGPT personalisieren können. Außerdem soll in diesem Zuge auch eine Art App-Store veröffentlicht werden, indem Nutzer auf die von Entwicklern erstellten GPTs zugreifen können. Die Entwickler erhalten auf Basis der

Nutzungshäufigkeit ihres GPTs eine Bezahlung. Enterprise Kunden haben außerdem die Möglichkeit interne GPTs zu erstellen, die nur sie selbst nutzen können. Diese können so erstellt werden, dass den GPTs Zugriff auf interne Daten gegeben werden kann (OpenAI, 2023g). Neben ChatGPT bietet OpenAI seine LLMs auch über einen API-Zugang an. Die neuesten LLMs sind GPT-3.5 Turbo sowie GPT-4 Turbo (OpenAI, 2023l). Des weiteren hat OpenAI eine Initiative mit „OpenAI Data Partnerships“ gestartet, bei der Unternehmen oder Organisationen OpenAI Datensätze zur Verfügung stellen können, damit OpenAI ihre Modelle mit den Daten fine-tunen kann. Durch diesen FT Prozess kann das LLM von OpenAI auf das Unternehmen oder die Organisation zugeschnittene Antworten liefern (OpenAI, 2023j).

Die von Microsoft getätigten Investitionen umfassen neben finanziellen Mitteln auch die Bereitstellung von Rechenleistung über Microsofts Cloud Computing Angebot „Azure Cloud“ (Bass, 2023). Microsoft sichert sich durch seine Investitionen den Status des exklusiven Cloud Anbieters für die Produkte von OpenAI und deren Forschungsarbeiten. Auch die APIs der LLMs von OpenAI werden lediglich über Azure Cloud zur Verfügung gestellt. Um diesen Service zu gewährleisten baut Microsoft seine Supercomputer Infrastruktur aus (Microsoft Corporate Blogs, 2023). Microsoft selber bietet kein eigenes LLM an, integriert allerdings die Modelle von OpenAI in ihre Produkte. Die Basis für Microsofts Copilot bieten GPT-4 sowie DALL-E 3, wodurch Copilot in der Lage ist Texte und Bilder zu generieren. Über Bing kann Copilot außerdem auf Daten aus dem Internet zugreifen (Microsoft Bing Team, 2023). Copilot ist in einer Vielzahl von Microsofts Produkten integriert. Neben der Veröffentlichung einer Chatvariante von Copilot wurde dieser einerseits in den Microsoft Edge Browser und andererseits direkt in das Windows 11 Betriebssystem integriert (Mehdi, 2023). Der Microsoft Edge Browser besitzt im November 2023 einen Marktanteil von 5,5% (StatCounter Global Stats, 2023a). Unter den Desktop Betriebssystemen liegt Windows im November 2023 mit 68,87% vorne (StatCounter Global Stats, 2023b). Von diesen 68,87% macht 26,66% des Marktanteils die Version Windows 11 aus (StatCounter Global Stats, 2023c). Auf den gesamten Markt bezogen hat Windows 11 dementsprechend einen Marktanteil von circa 18,4%. Unternehmenskunden haben die Möglichkeit den Copilot Enterprise Chat oder Microsoft Copilot 365 zu nutzen. Der Enterprise Chat bietet dieselben Funktionen wie der normale Copilot Chat, schützt bei Verwendung allerdings noch die Unternehmensdaten. Unternehmen können den Copilot Enterprise Chat für 5 Dollar pro Nutzer pro Monat nutzen; als Unternehmenskunde der Microsoft 365 Produkte kostenlos. Microsoft Copilot 365 stellt eine weitere Alternative für Unternehmen dar (Spataro, 2023). Für 30 Dollar pro Monat pro Nutzer zuzüglich zu dem Microsoft 365 Abonnement können Unternehmen Copilot noch zusätzlich in Microsofts Office Programmen nutzen (Microsoft, 2023a). Microsoft Copilot Studio ermöglicht die Erzeugung eigener Copilots basierend auf den unternehmensinternen Anforderungen. Durch die Nutzung von den oben angesprochenen GPTs, Plug-Ins und Unternehmensdaten kann die Individualisierung ohne Programmierfähigkeiten erfolgen (Stallbaumer, 2023). Plugins können von den Entwicklern demnächst über das Microsoft Partner Center veröffentlicht werden (Microsoft Bing Team, 2023). Neben den Modellen

von OpenAI bietet Microsoft auch das neuste LLM von Meta über Azure Cloud an (Montgomery, 2023).

Llama 2 ist Metas Open Source LLM und kann neben der Bereitstellung über eine API direkt auf der Webseite von Meta heruntergeladen werden. Das LLM ist für Forschung sowie kommerzielle Nutzung freigegeben (Meta, 2023b). Auf Basis von Llama 2 hat Meta einen eigenen KI-Assistenten entwickelt. Meta AI kann über WhatsApp, Messenger und Instagram genutzt werden. Die Nutzung von Meta AI kann in einem privaten oder direkt in einem Gruppenchat erfolgen. Der Chatbot kann Instagram Nutzern Reels zu verschiedenen Anfragen bereitstellen und auch Bilder generieren. Neben den Bildern existiert die Funktion neue Sticker über den Einsatz von KI zu erzeugen und der Nutzer kann seine Bilder mit Hilfe von KI bearbeiten. Durch die Kooperation mit Microsoft kann Metas Chatbot über Bing auf das Internet zugreifen und Nutzer können mit einer Auswahl an KIs chatten (Meta, 2023d). Jede KI wurde speziell von Meta trainiert, damit jede KI individuelle Eigenschaften besitzt. Entwickler werden über AI Studio eigene KIs erstellen und diese den Nutzern zunächst über Messenger bereitstellen können. Ebenso werden Nutzer über ein eigenes Tool eigene KIs erstellen können (Meta, 2023e). Unternehmen können KIs erstellen um mit deren Kunden über die Apps von Meta zu kommunizieren. Llama 2 wird des Weiteren in die Smart Glasses, welche in Kooperation mit Ray-Ban herausgebracht wurden, und in die Meta Quest 3 integriert (Meta, 2023c). Auf Basis von Llama 2 veröffentlichte Meta Code Llama, welches den Nutzer beim Programmieren unterstützt (Meta, 2023a).

Google hat mit seinem Chatbot Bard einen Konkurrenten zu ChatGPT veröffentlicht. Am 10. Mai wurde dieser nach einer Testphase für alle interessierten Nutzer verfügbar gemacht (Hsiao, 2023b). Bard basiert seit dem neuesten Update auf dem neuen Gemini Pro LLM (Hsiao, 2023c). Gemini ist Googles neue LLM und existiert in drei verschiedenen Versionen: Ultra, Pro und Nano. Ultra ist dabei die größte und Nano die kleinste Version (Pichai & Hassabis, 2023). Für Bard wurde angekündigt, dass Gemini Ultra bald über Bard Advanced nutzbar ist. Zum Zeitpunkt der Bearbeitung der Bachelorarbeit sind allerdings keine weiteren Informationen zu Bard Advanced verfügbar (Hsiao, 2023c). Bard kann wie ChatGPT Bilder unter Einsatz von Google Lens als Prompt verarbeiten (Krawczyk, 2023). Bard bietet außerdem die Möglichkeit die Apps und Dienste von Google zu verbinden und diese zu nutzen, wodurch Bard auf das Internet zugreifen kann. Hierbei wird die Zusammenarbeit mehrerer dieser Apps und Dienste unterstützt. Es besteht auch die Möglichkeit die von Bard generierte Antwort über einen Button mit Google Suchergebnissen zu belegen, wodurch die Antwort verifiziert werden kann. Die mit Bard geführten Unterhaltungen können auch mit anderen Nutzern über einen Link geteilt werden (Pinsky, 2023). Unternehmen können Gemini Pro über eine API nutzen. Gemini Nano wird Android-Entwicklern zur App-Entwicklung bereitgestellt (Pichai & Hassabis, 2023). Wie auch Microsoft bietet Google die Möglichkeit Llama 2 und Code Llama von Meta zu nutzen. Demnächst erfolgt noch die Implementierung von Claude 2, ein Closed-Source LLM von Anthropic (Vahdat & Yang, 2023). Neben dem Bereitstellen der APIs integriert Google Bard in deren Google Assistant, um die Performance und die

Nutzungsmöglichkeiten zu verbessern (Hsiao, 2023a). Duet AI ist Googles Pendent zu Microsofts Copilot 365 und wird Unternehmen, die Workspace benutzen, angeboten. Duet AI ist in die Google Workspace Anwendungen sowie die Google Cloud integriert und kann dort zur Unterstützung genutzt werden (Pappu, 2023). Google bietet den Chatbot Duet AI Unternehmen mit vorhandenem Workspace Abonnement für 30 Dollar pro Monat pro Nutzer an (Holtermann & Scheuer, 2023).

Amazon ist mit AWS ein weiterer Cloud Computing Anbieter und ermöglicht es Kunden über Amazon Bedrock auf LLMs über APIs zuzugreifen. Amazon bietet hier neben dem eigenen LLM Titan beispielsweise auch Claude 2 und Llama 2 an (Amazon Web Services, 2023c). Neben den APIs bietet Amazon den auf Unternehmen ausgelegten Chatbot Q an, welcher den Nutzer bei täglichen Aufgaben unterstützen soll (Weise, 2023). Genutzt werden kann dieser über das Amazon Q Business für 20 Dollar pro Monat pro Nutzer und Amazon Q Builder für 25 Dollar pro Monat pro Nutzer Abonnement. Amazon Q kann den Nutzer mit Informationen versorgen oder auch Inhalte generieren, allerdings nur auf Basis von Unternehmensdaten. Dadurch ist Amazon Q allerdings auf das jeweilige Unternehmen abgestimmt und liefert dem einzelnen Nutzer unter Beachtung seiner Rolle und Befugnisse unternehmensspezifische Antworten. Die Builder Version kann zusätzlich bei Programmieraufgaben und der Umsetzung von neuen AWS Lösungen unterstützen (Amazon Web Services, 2023b, S.121). Auch kann Amazon Q den Nutzer dabei unterstützen über das QuickSight Tool Dashboards mit Unternehmensdaten zu erstellen und über Amazon Connect den Kundenservice zu unterstützen (Amazon Web Services, 2023a). Amazon integriert außerdem ein auf Sprachkonversationen spezialisiertes LLM in seinen Sprachassistenten Alexa (Rausch, 2023).

Anthropic wurde schon in den oberen Abschnitten als Entwickler des LLM Claude genannt und es wurde beschrieben, welche Cloud Computing Anbieter den API Zugriff ermöglichen. Neben dem LLM hat Anthropic einen Chatbot veröffentlicht, welcher aktuell nur in Amerika und dem vereinigten Königreich genutzt werden kann (Anthropic, 2023a).

5 Empirische Überprüfung des Marktes

Ziel der empirischen Arbeit ist die Analyse des Marktes der LLMs. Hierzu soll die Forschungsfrage „Besteht die Gefahr, dass OpenAI Monopolstatus im Bereich der Large Language Modelle erreichen wird?“ durch Anwendung eines Prüfschemas beantwortet werden. Die theoretischen Grundlagen hierfür wurden in Kapitel 3 gegeben. Das erste Prüfkriterium stellt die aktuelle Marktsituation dar. Zunächst wird der Markt der Chatbots betrachtet. Auf Basis der Seitenaufrufe aus dem Zeitraum von Juni 2023 bis Oktober 2023 werden die Konzentrationskurve und der HHI gebildet, um die aktuelle Marktverteilung der Chatbots abzubilden. Die Seitenaufrufe werden mit Hilfe der Analyse Website SimiliarWeb[2] ermittelt. Für ChatGPT und Bard speziell werden die Seitenaufrufe seit Veröffentlichung der beiden Chatbots ermittelt, da diese Chatbots von den für die Bachelorarbeit interessanten Unternehmen veröffentlicht wurden. Für ChatGPT wird dafür der Online-Artikel von Hillebrandt (2023), für Bard der Online-Artikel von Krivec (2023) verwendet. Die in den Artikeln angegebenen Seitenaufrufe basieren auch auf SimiliarWeb, wodurch die Konsistenz der Daten weiterhin gewährleistet ist. Hierdurch kann die Entwicklung der Seitenaufrufanzahl von ChatGPT und Bard miteinander verglichen werden. Auch wird die mediale Aufmerksamkeit von ChatGPT mit in die Betrachtung aufgenommen. Die Analyse der Bachelorarbeit startet mit der aktuellen Marktsituation, da diese als Basis einiger weiterer Prüfpunkte dient. Im nächsten Schritt wird die Existenz von direkten bzw. indirekten Netzwerkeffekten analysiert. Da die Stärke eines Netzwerkeffektes von der Anzahl der Nutzer abhängt (Lipczynski et al., 2013, S.619), werden die Ergebnisse der Marktüberprüfung relevant und fließen mit in die Auswertung des Prüfpunktes ein. Die Betrachtung der Produktdifferenzierung stellt den nächsten Prüfpunkt dar. Zunächst wird die Qualität der LLMs von vier Anbietern anhand von Benchmark Ergebnissen überprüft. Hierzu werden jeweils zwei Versionen der einzelnen LLMs betrachtet und in ein Premium- sowie Standard-Segment eingeteilt. Ausnahme ist Metas Llama 2, da dieses als einziges Open Source LLM mit allen Modellvarianten verglichen werden soll. Anschließend wird die mögliche Eingabelänge mit in die Qualitätsbetrachtung genommen. Die Ergebnisse werden mit den Preisen für die APIs in Verbindung gesetzt, um zu überprüfen, wie die Qualität und der Preis miteinander zusammenhängen. Damit abgeschlossen werden jeweils die in Kapitel 4 vorgestellten vorhandenen Funktionen der Chatbots für private Nutzer und Unternehmen miteinander verglichen. Kapitel 5.4 dient als Vorbereitung für die Überprüfung auf das Vorliegen von Lock-In-Effekten. In diesem Abschnitt der Bachelorarbeit werden die Nutzungsvoraussetzungen für die APIs und die unternehmensspezifischen Chatbots herausgearbeitet. Auf Basis der Ergebnisse werden die API Angebote und die unternehmensspezifischen Chatbots auf Wechselkosten und Lock-In-Effekte überprüft. Des Weiteren werden mögliche Lock-In-Effekte von ChatGPT betrachtet und es erfolgt eine Überprüfung der LLM Integration durch die Anbieter in deren Produkte. Abgeschlossen wird das Prüfschema mit der

[2] Mit dem Studentenabonnement von SimiliarWeb ist es möglich die Seitenaufrufe der letzten drei Monate zu ermitteln. Die erste Datenerhebung fand im September statt, wodurch der erste Monat des Betrachtungszeitraumes Juni 2023 ist.

Analyse der Kostenstruktur unter Betrachtung von Economies of Scale und Scope. Die Ergebnisse der Prüfpunkte sollen einen Aufschluss darüber geben, ob die Voraussetzungen für die Monopolbildung vorhanden sind.

5.1 Die aktuelle Marktsituation

Im folgenden Abschnitt wird die aktuelle Marktsituation der auf LLMs basierenden Chatbots analysiert. Die in Kapitel 2.1 genannten Eigenschaften eines Chatbots werden genutzt, um die Liste der Top 50 der generative KI Anwendungen von Moore (2023) auf das Vorkommen von Chatbots hin zu überprüfen. Die Eigenschaften für einen Chatbot sind neben der Möglichkeit einen textbasierten Dialog zu führen auch das Vorhandensein einer Benutzeroberfläche, in die eine Eingabe getätigt werden kann (Mackmood et al., 2020, S.151). Wichtig ist auch, dass der Chatbot anhand dieser Eingabesequenz einen neuen menschenähnlichen Text generiert (Yenduri et al., 2023, S.3). Die in Tabelle 1 markierten Anwendungen, besitzen die Eigenschaften eines Chatbots.

Tabelle 1: Top 50 - Generative KI Anwendungen nach Seitenaufrufen (Juni 2023)

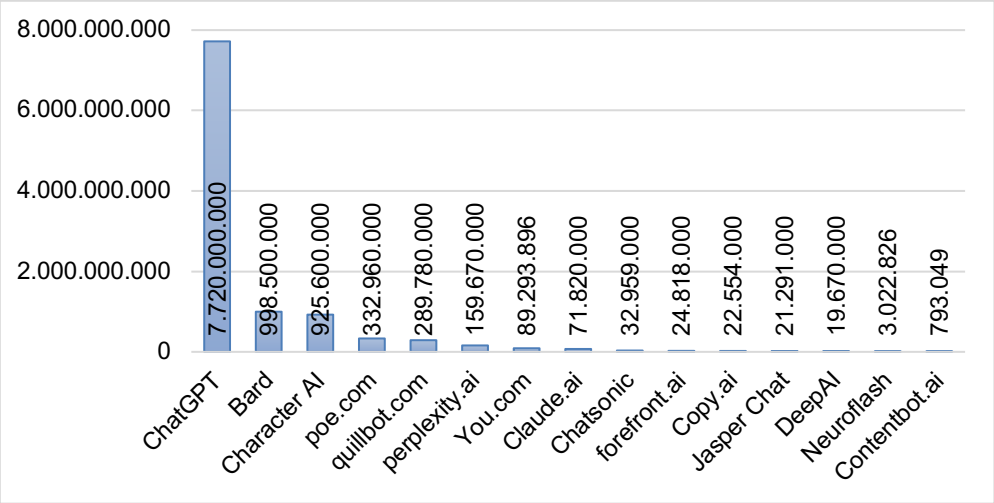
Top 50 - Generative KI-Anwendungen nach Seitenaufrufen				
ChatGPT	**YOU**	NightCafe	GPTGo.ai	Fliki
character.ai	leonardo.	replicate	R runway	pornpen.ai
Bard	PIXLR	Speechify	Playground	KAPWING
Poe	VEED.IO	ElevenLabs	AKaiber	Gamma
QuillBot	tome	Lexica	Hotpot	Looka
PhotoRoom	AI-Novel	VocalRemover	Stable Diffusion	human or not
CILITAI	cutout.pro	**Writesonic**	**copy.ai**	PIXAI
Midjourney	ForefrontAI	CHATPDF	ZeroGPT	WRITER
Hugging Face	Clipchamp	D-ID	Smodin	A NovelAI
Perplexity	TheB.AI	Chub.ai	ZMO.AI	DeepSwap

Quelle: Eigene Darstellung nach: Moore (2023)

Neben der Überprüfung dieser Liste wurde auch Internetrecherche betrieben. Mit Suchbegriffen wie „Chatbot“, „ChatGPT Konkurrenz“ und „generative KI Anwendungen“ wurden noch sechs weitere Chatbots ermittelt. Diese sind Claude.ai, forefront.ai, Jasper Chat, DeepAI, Neuroflash und Contentbot.ai. Zunächst erfolgt eine visuelle Darstellung der Seitenaufrufe mit Hilfe eines Balkendiagrams. Aus Abbildung 6 kann die Verteilung der summierten Seitenaufrufe des Zeitraums von Juni bis Oktober 2023 entnommen werden. Die Gesamtzahl der Seitenaufrufe belief sich in diesem Zeitraum auf circa 11,3 Milliarden. ChatGPT ist mit 7,72 Milliarden Seitenaufrufen deutlich vorne und hat 7,7-mal so viele Seitenaufrufe generieren können, wie Googles Konkurrent Bard. Bard und Character AI haben verglichen miteinander circa dieselbe Anzahl an Seitenaufrufen, genauso wie poe.com und quillbot.com verglichen miteinander. Ab perplexity.ai nimmt die

Anzahl der Seitenaufrufe stark ab, was vermuten lässt, dass diese Anbieter eine geringe Relevanz im Markt der Chatbots haben.

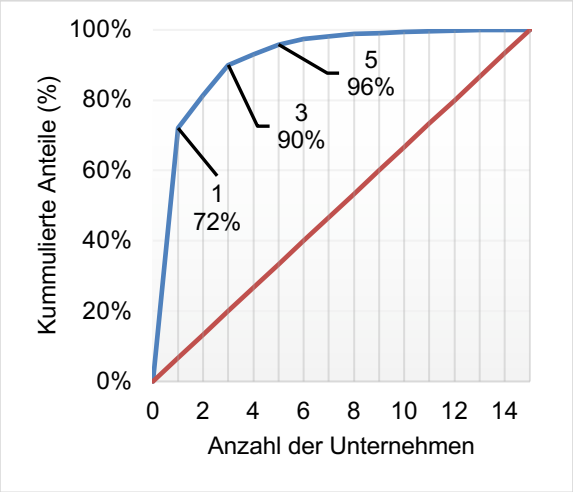
Abbildung 6: Summe der Seitenaufrufe der Chatbots von Juni bis Oktober 2023



Quelle: Eigene Darstellung auf Basis der Daten von: Anhang A

Diese Beobachtung spiegelt sich auch in der Darstellung der Konzentrationskurve in Abbildung 7 wider. Die Reihenfolge der Unternehmen ist wie in Abbildung 6. ChatGPT alleine generierte in dem betrachteten Zeitraum 72% der gesamten Seitenaufrufe. Werden die ersten 3 Anbieter zusammen betrachtet haben diese 90%, die ersten 5 Anbieter zusammen sogar 96% der gesamten Seitenaufrufe ausgemacht. Die restlichen 4% der gesamten Seitenaufrufe verteilt sich auf zehn Anbieter. Hierdurch wird die These gestärkt, dass diese Unternehmen eine sehr geringe Relevanz haben. Wird nun §18 Absatz 4 GWB betrachtet, hat ein Unternehmen eine marktbeherrschende Stellung ab einem Marktanteil von 40% (Gesetz gegen Wettbewerbsbeschränkungen. Bundesministerium der Justiz, 2023).

Abbildung 7: Konzentrationskurve auf Basis der Seitenaufrufe

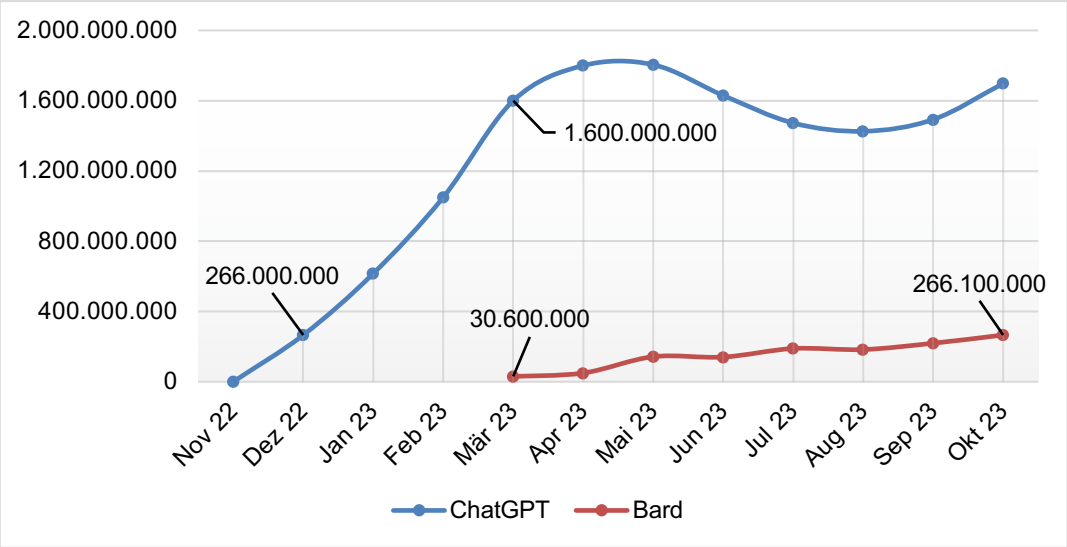


Quelle: Eigene Darstellung auf Basis der Daten von: Anhang A

Unter Berücksichtigung, dass die Marktanteile auf Basis der Seitenaufrufe bestimmt worden sind, kann die Schlussfolgerung getroffen werden, dass OpenAI mit ChatGPT eine marktbeherrschende Stellung eingenommen hat. Die Seitenaufrufe werden des weiteren verwendet um den HHI zu bilden und dadurch die Marktkonzentration zu bestimmen. Der ermittelte Wert des HHI auf Basis der Seitenaufrufe beträgt 5.375. Bei 15 Marktteilnehmern wäre der geringstmögliche Wert des HHI bei 667 (Anhang B), in diesem Fall wären die Seitenaufrufe perfekt auf alle Marktteilnehmer verteilt (Waldman & Jensen, 2013, S.105ff.). Nach den Grenzen, die das US-amerikanische Justizministerium für den HHI festgelegt hat, liegt im Markt der Chatbots bezogen auf die Seitenaufrufe eine hohe Marktkonzentration in Richtung OpenAI vor (Monopolkommission, 2022, S.54).

Google und Anthropic werden in der Bachelorarbeit als Konkurrenten von OpenAI genauer betrachtet. In Abbildung 8 werden die Seitenaufrufe von ChatGPT und Bard seit deren Veröffentlichung aufgeführt. Da Anthropic seinen Chatbot Claude aktuell nur Nutzern in Amerika und dem vereinigten Königreich zur Verfügung stellt (Anthropic, 2023a), fällt dieser aus der Betrachtung raus. Seit der Veröffentlichung von ChatGPT am 30. November 2022 konnte OpenAI eine große Anzahl an Nutzern für sich gewinnen. Innerhalb der ersten fünf Tage konnte ChatGPT eine Millionen Nutzer erreichen (Buchholz, 2023), nach zwei Monaten waren es nach einer Studie von UBS schon 100 Millionen Nutzer (Hu, 2023). Aus Abbildung 8 wird ersichtlich, dass ChatGPT im Veröffentlichungsmonat von Bard schon 1,6 Milliarden Seitenaufrufe hatte. Googles Chatbot benötigte außerdem sieben Monate um die Anzahl der monatlichen Seitenaufrufe zu erreichen, die ChatGPT im ersten vollen Monat nach der Veröffentlichung erreicht hat.

Abbildung 8: Seitenaufrufe von ChatGPT und Bard seit Veröffentlichung



Quelle: Eigene Darstellung auf Basis der Daten von: Anhang A

Das Columbia Journalism Review analysierte die Aufmerksamkeit, die ChatGPT und das Thema der Chatbots in den amerikanischen Medien bekamen. Anfang Februar erreichte die Menge an Berichterstattungen im Betrachtungszeitraum der Analyse ihren

Höhepunkt. Zu diesem Zeitpunkt erreichte die Anzahl an Veröffentlichungen den gleichen Wert, welchen das Thema Bitcoin 2021 erreichte; damals stieg der Bitcoin Preis auf ein globales Maximum an (Bartholomew & Mehta, 2023). OpenAI konnte mit ChatGPT den dadurch resultierenden First-Mover Vorteil für sich nutzen. Durch die Medienpräsenz wurde der Chatbot einer großen Anzahl an Leuten ein Begriff, wodurch sich ChatGPT bereits mit seinem Namen etablieren konnte (Halberstadt, 2014, S.92ff.). ChatGPT war außerdem bis zum 1. Februar 2023 zur kostenlosen Nutzung verfügbar, ab diesem Zeitpunkt wurde das Freemium-Geschäftsmodell angewendet. Mit der Einführung von ChatGPT Plus haben Nutzer die Möglichkeit für 20 Dollar pro Monat auf mehr Funktionen zuzugreifen. Bei diesem Modell führt gerade das kostenlose Angebot dazu, dass sehr schnell eine große Anzahl an Nutzern gewonnen werden kann (Seufert, 2014, S.1). OpenAI konnte durch den in diesem Kapitel aufgeführten schnellen Anstieg der Nutzerzahl einen First-Mover-Vorteil generieren. Die Konkurrenten von OpenAI müssen mehr Mittel aufwenden um neue Nutzer für sich zu gewinnen, wodurch diese einen Wettbewerbsnachteil beim Markteintritt haben (Peters, 2010, S.50). Neben der Nutzeranzahl wird die Stärke des First-Mover-Vorteils auch durch Wechselkosten beeinflusst (Wied-Nebbeling, 2004, S.288), welche in Kapitel 5.5 genauer betrachtet werden.

5.2 Netzwerkeffekte und zweiseitige Märkte

In diesem Kapitel werden die einzelnen Marktteilnehmer auf direkte und indirekte Netzwerkeffekte hin überprüft. In Kapitel 5.1 wurde ermittelt, dass OpenAI in kürzester Zeit eine sehr große Nutzerzahl generieren konnte. Bei positiven Netzwerkeffekten steigt der Nutzen eines Produktes mit zunehmender Nutzerzahl (Sauer, 2019, S.36). ChatGPT wird u.a. auf Basis von echten Unterhaltungen mit Nutzern verbessert. Diese bieten echte Situationen, mit denen das hinter ChatGPT liegende Modell verbessert werden kann, wodurch dieses sicherer und hilfreicher wird. Die verwendeten Prompts der API Nutzer werden standardmäßig nicht zur Verbesserung des Modells verwendet, außer diese füllen ein Formular aus (Schade, 2023). ChatGPT bietet dem Nutzer außerdem die Möglichkeit die generierte Antwort direkt zu bewerten, wodurch das Modell seine Antwort auf diese Frage beim nächsten Mal anpassen wird (Schönert, 2023). Auch Bard wird mit Hilfe von echten Konversationen der Nutzer sowie deren Feedback verbessert (Pinsky (2023), Google (2023a)). Für den FT Prozess werden zwar kleine Datensätze mit hoher Datenqualität benötigt (Dai et al., 2023, S.2ff.), allerdings kann OpenAI mehr Konversationen verwenden, wodurch potenziell mehr Anwendungsmöglichkeiten des Chatbots abgedeckt werden können. Außerdem ist die Wahrscheinlichkeit höher, dass unter allen Konversationen einige dabei sind, welche die benötigte Qualität für den FT Prozess besitzen. Die Konversationen können genutzt werden um Phase zwei und drei des in Kapitel 2.2 beschriebenen FT Prozesses erneut durchzuführen. Das Belohnungsmodell kann um die neuen Prompts ergänzt werden und dafür genutzt werden das LLM zu verbessern (Ouyang et al., 2022, S.6). Bei mehr Nutzern kann auch davon ausgegangen werden, dass ChatGPT im Vergleich zu Bard mehr direktes Feedback erhält, welches zur Verbesserung verwendet werden kann. Im März 2023 veröffentlichte

OpenAI Plugins für ChatGPT Plus Nutzer. Neben zwei eigenen Plugins wurden damals elf Plugins von Drittanbietern veröffentlicht. Andere Entwickler konnten sich allerdings in eine Warteliste eintragen und weitere Plugins für ChatGPT entwickeln (OpenAI, 2023c). Durch die Einführung der Plugins konnte OpenAI indirekte Netzwerkeffekte erschaffen und einen zweiseitigen Markt entstehen lassen. Auf der einen Marktseite stehen die Entwickler der Plugins, auf der anderen Marktseite die ChatGPT Plus Nutzer. Je mehr Leute ChatGPT nutzen, desto mehr Entwickler werden sich dazu entschließen ein Plugin zu erstellen. Je mehr Plugins vorhanden sind, desto höher ist der Nutzen von ChatGPT für jeden einzelnen Nutzer (Dewenter & Rösch, 2015, S.118). Neben dem ChatGPT Plus Abonnement müssen keine weiteren Kosten durch den Nutzer getragen werden, um die Plugins zu nutzen (Schimroszik, 2023). Der erschaffene zweiseitige Markt soll nun genutzt werden, um eine neues Geschäftsmodell zu erzeugen. Durch die Einführung der GPTs im November 2023 wurde die Möglichkeit der Plugin Integrierung gestoppt. Neue Plugins werden von OpenAI nicht mehr akzeptiert, allerdings haben Entwickler von nun an die Möglichkeit GPTs zu erstellen (OpenAI, 2023d). Die erstellten GPTs werden den ChatGPT Plus Nutzern in einem demnächst veröffentlichten GPT-Store zur Verfügung gestellt. Je mehr Nutzer ein GPT nutzen, desto mehr Geld verdient der Entwickler, da dieser basierend auf der Nutzeranzahl entlohnt wird. Die ChatGPT Plus Nutzer sollen weiterhin mit ihrem Abonnement ohne zusätzliche Kosten auf die GPTs zugreifen können (OpenAI, 2023g). Für die Entwickler der GPTs birgt die Möglichkeit der Bezahlung eine neue Motivation immer bessere GPTs zu entwickeln, wodurch wiederum neue Nutzer angelockt werden. Durch die Partnerschaft zwischen OpenAI und Microsoft wird die potenzielle Nutzergruppe der GPTs erweitert. Microsoft wird die Kompatibilität der GPTs mit seinen Produkten sicherstellen, wie schon mit den für ChatGPT entwickelten Plugins (Hawk, 2023). Zunächst können die GPTs bei Microsoft nur von Unternehmenskunden genutzt werden, wenn diese mit Copilot Studio ihre benutzerdefinierten Copilots erstellen. Wie eben erwähnt sind die ChatGPT Plugins mit Copilot kompatibel, da diese auf der selben Struktur basieren. Hierdurch können Plugins in allen Copilot Produkten genutzt werden. Demnächst wird Microsoft es ermöglichen, dass Entwickler ihre Plugins über das Microsoft Partner Center anbieten können und so ein wie oben beschriebener zweiseitiger Markt entstehen wird (Microsoft Bing Team, 2023).

Meta hat für seinen Chatbot die sogenannten KIs mit dem Plan veröffentlicht, dass Entwickler bald eigene KIs über das Tool AI Studio von Meta erstellen können. Diese KIs sollen zunächst bei Messenger und danach bei WhatsApp bereitgestellt werden und erzeugen dadurch einen zweiseitigen Markt (Dewenter & Rösch, 2015, S.118). Die Vorstellung des Chatbots von Meta erfolgte Ende September 2023 und kann zunächst nur in Amerika genutzt werden (Meta, 2023c). Dementsprechend gibt es noch keine Informationen über die Anzahl der Chatbot Nutzer, allerdings können die potenziellen Nutzer der erstellten KIs über die Nutzeranzahl von Messenger und WhatsApp ermittelt werden. WhatsApp hat im Oktober 2023 2 Milliarden, Messenger 1,036 Milliarden Nutzer (Kemp, 2023).

Netzwerkeffekte können im Falle von Überbelastung eines Systems ab einer bestimmten Nutzeranzahl negativ werden (Srinivasan, 2023, S.44). Um dieser Überlastung vorzubeugen hat OpenAI eine maximale Anzahl von Anfragen für ChatGPT Plus Nutzer festgelegt. Diese können innerhalb von drei Stunden 50 Anfragen an das GPT-4 Modell senden (J., 2023). Außerdem hat Sam Altman über X kommuniziert, dass die Kapazität der ChatGPT Plus Abonnements in Folge des DevDay derzeit beschränkt werden muss, damit die aktuellen Nutzer den Dienst weiterhin ohne Einschränkungen nutzen können. (Altman, 2023b). Auf dem OpenAI DevDay wurden die Erstellung von GPTs durch ChatGPT Plus Nutzer vorgestellt (OpenAI, 2023k). Die hohe Nachfrage nach ChatGPT erzeugt in diesen Fällen negative Netzwerkeffekte, da der Nutzen durch die Begrenzung der Abfragen verringert wird. Der Überlastung der ChatGPT Plus Funktionen wird zudem durch die Begrenzung der neuen Abonnements vorgebeugt. Negativ hierbei ist, dass OpenAI potenzielle Kunden verlieren kann, wenn diese nicht bereit sind zu warten um das Abonnement abzuschließen. Verdeutlicht wird hier allerdings auch, wie hoch die Nachfrage nach den GPTs ist, da die Restriktion der Abonnements erst nach der Veröffentlichung dieser erfolgt ist. Multihoming ist im Chatbot Angebot möglich, da keine Ausschließlichkeitsbindung der Nutzer an den jeweiligen Anbieter existiert. Die Wahrscheinlichkeit für Multihoming sinkt mit steigenden Preisen der einzelnen Angebote. Bard, Meta AI und Copilot stehen dem Nutzer kostenlos zur Verfügung. Für ChatGPT Plus müssen Nutzer 20 Dollar pro Monat zahlen, die normale Version kann auch gratis verwendet werden. Die Existenz von Multihonming ist dementsprechend zu erwarten, wodurch ein erhöhter Wettbewerbsdruck entsteht. (Dewenter & Rösch, 2015, S.50).

Bei der Betrachtung der Unternehmen im Markt der LLMs wurde die Existenz von positiven Netzwerkeffekten nachgewiesen. Mit dem GPT-Store hat OpenAI einen zweiseitigen Markt erschaffen. Da der Nutzenzuwachs mit zunehmender Nutzeranzahl steigt, hat OpenAI mit ChatGPT einen Vorteil gegenüber Microsoft und Google (Lipczynski et al., 2013, S.619). Durch die von OpenAI eingeführten Restriktionen können auch mögliche negative Netzwerkeffekte bedingt durch Überbelastungen verhindert werden (J. (2023), Altman (2023b)). Meta will die von externen Entwicklern erzeugten KIs über WhatsApp und Messenger zur Verfügung stellen. Der dadurch entstehende zweiseitige Markt liefert auf der einen Seite eine sehr große potenzielle Nutzeranzahl. Die große Nutzeranzahl von WhatsApp und Messenger erzeugen einen sehr großen Nutzen für die Entwickler, wodurch diese in den anderen Markt eintreten werden. Eine große Anzahl an Nutzern oder in diesem Fall potenziellen Nutzern auf der einen Marktseite, generieren neue Nutzer auf der anderen Marktseite (Dewenter & Rösch, 2015, S.115). Aufgrund der vorhandenen Voraussetzungen für Multihoming wird in Kapitel 5.3 unteranderem überprüft, wie sich die Chatbots untereinander unterscheiden.

5.3 Produktdifferenzierung

Das Kapitel der Produktdifferenzierung wird in verschiedene Betrachtungseinheiten aufgeteilt. Zunächst erfolgt die Betrachtung der Benchmark-Ergebnisse von vier verschiedenen Anbietern und es werden die Preise für die APIs mit den Ergebnissen verknüpft.

Anschließend werden die Angebote der Chatbots für Privatpersonen und Unternehmen auf deren Leistungen hin untersucht.

Die LLMs der Anbieter werden auf Grundlage von vier Benchmarks miteinander verglichen. In Tabelle 2 sind die vorhandenen Benchmarks mit einem „x“ gekennzeichnet.

Tabelle 2: Welche Benchmarks sind für die LLMs vorhanden?

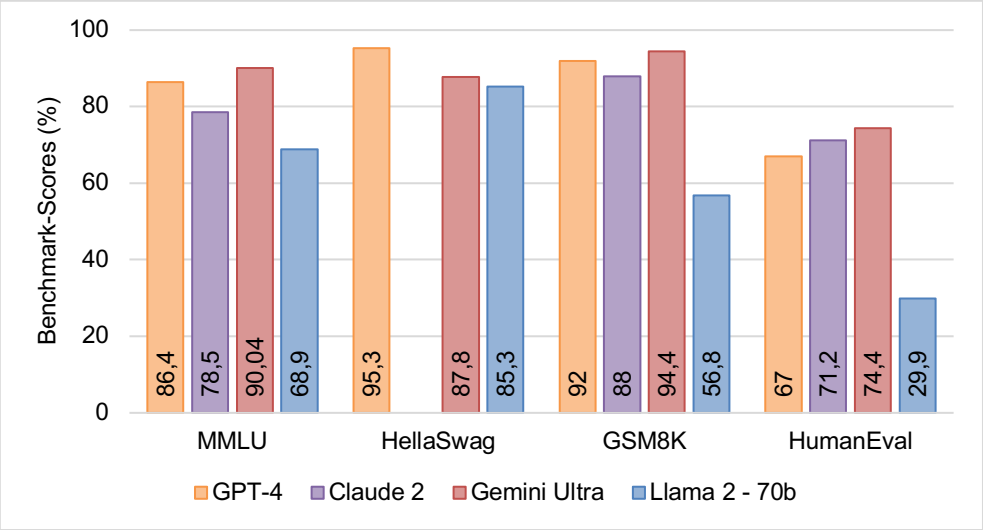
	MMLU	**HellaSwag**	**GSM8K**	**HumanEval**
GPT-3.5 Turbo	x	x	x	x
GPT-4	x*	x*	x*	x*
Claude Instant	x**	n.a.	x**	x**
Claude 2	x**	n.a.	x**	x**
Gemini Pro	x***	x***	x***	x***
Gemini Ultra	x***	x***	x***	x***
Llama 2 – 70b	x****	x****	x****	x****

Quellen: Zheng et al. (2023, S.6), *OpenAI (OpenAI, 2023m, S.11), **Anthropic (2023b, S.10), (2023b; ***Gemini Team, S.7), ****Meta (2023b)

Die Benchmarks werden von den Entwicklern der LLMs in deren technischen Berichterstattungen veröffentlicht; ein standardisiertes Erhebungsverfahren existiert nicht. Die Qualität der Ergebnisse ist außerdem stark abhängig von der Methode des „prompting“. Neue Strategien können hierbei zu anderen Ergebnissen führen (Horvitz & Nori, 2023). Für die Ermittlung der Benchmark Ergebnisse von GPT-3.5 Turbo werden die Ergebnisse des Papers von Zheng et al. (2023) verwendet, da in der technischen Berichterstattung von GPT-4 nur das normale GPT-3.5 Modell betrachtet wird. Bei der Durchführung der Benchmark Erhebung können demnach Unterschiede entstehen. Die in Kapitel 2.3 beschriebenen Datensätze für die einzelnen Benchmarks sind allerdings standardisiert. Der Vergleich der Benchmarks in der Bachelorarbeit soll lediglich eine grobe Einschätzung über die verschiedenen Qualitäten der LLMs ermöglichen. Auf Grundlage der Theorie der vertikalen Produktdifferenzierung ist davon auszugehen, dass sich Nutzer bei dem selbem Preis für das LLM mit der höchsten Qualität entscheiden werden (Tirole, 1999, S.210ff.). Mit den vier verschiedenen Benchmarks wird ein großer Bereich der Anwendungsmöglichkeiten eines LLM überprüft. Ergebnisse der MMLU Benchmark geben Aufschluss über das Allgemeinwissen eines LLM, da diese einen weitgefächerten Wissensbereich mit den Fragen abdeckt. Durch die HellaSwag Benchmark wird vor allem überprüft, welche Fähigkeiten ein LLM besitzt logische Denkaufgaben zu lösen. Die mathematischen Fähigkeiten werden durch die GSM8K Benchmark und die Programmierfähigkeiten durch die HumanEval Benchmark getestet (Google DeepMind, 2023). Die LLMs der Anbieter werden zunächst in ein Standard- und in ein Premium-Segment eingeteilt. Jedes der betrachteten LLMs wird in verschiedenen Versionen angeboten. In der folgenden Betrachtung werden die in Tabelle 2 blau hinterlegten LLMs dem Standard-Segment, die weiß hinterlegten dem Premium-Segment zugeordnet. Bei dem rot hinterlegten LLM handelt es sich um die Llama 2 Version mit 70 Milliarden Parametern. Dies ist das einzige Open Source Modell in dem Betrachtungsumfeld. Diese

Modellversion sowohl mit dem Standard- als auch mit dem Premium-Segment verglichen. Zunächst werden die LLMs des Premium-Segments miteinander verglichen.

Abbildung 9: Benchmark-Ergebnisse des Premium-Segments



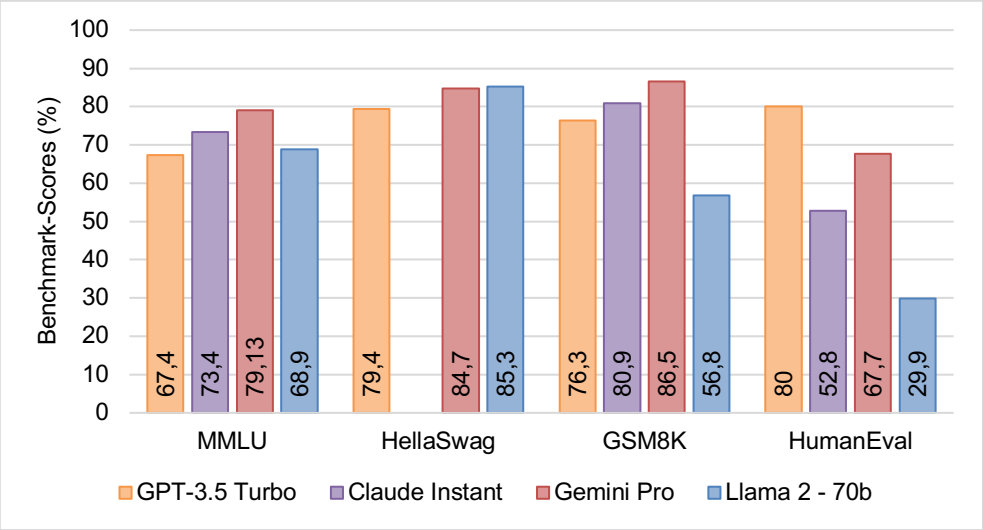
Quelle: Eigene Darstellung auf Basis der Daten von: OpenAI (2023m, S.7) (2023m, S.7), Anthropic (2023b, S.10), Gemini Team (2023, S.7), Meta (2023b)

Das neu veröffentlichte LLM Gemini Ultra erzielt bei der MMLU, GSM8K sowie der HumanEval Benchmark das beste Ergebnis. GPT-4 liegt bei diesen Benchmarks nur knapp hinter Gemini Ultra. Bei MMLU sind es 3,64, bei GSM8K 2,4 und bei HumanEval 7,4 Prozentpunkte Unterschied. Claude 2 schneidet bei der MMLU Benchmark mit einem Unterschied von 11,9 Prozentpunkten zu Gemini Ultra und 7,9 Prozentpunkten zu GPT-4 am schlechtesten unter den Closed-Source Modellen ab. Auch bei der GSM8K Benchmark schneidet Claude 2 mit einem erreichten Wert von 88,0% am schlechtesten von den drei LLMs ab. Bei der HumanEval Benchmark platziert sich Claude 2 mit 71,2% zwischen GPT-4 und Gemini Ultra. Die betrachtete Version von Llama 2 schneidet in diesen drei Benchmarks deutlich schlechter ab als die Premium-Versionen der Closed-Source Modelle, wobei der Unterschied bei der GSM8K und HumanEval nochmal deutlicher ist als bei der MMLU Benchmark. Ein anderes Ergebnis liefert die Betrachtung der HellaSwag Benchmark. Llama 2 ist mit 85,3% nur 2,5 Prozentpunkte schlechter als Gemini Ultra, allerdings ist GPT-4 mit einem Wert von 95,3% deutlich vor den beiden anderen.

Verglichen wird Llama 2 nun zusammen mit den Standardversionen der LLMs der drei Unternehmen. Die Gemini Pro Version erreicht bei der MMLU und GSM8K Benchmark die besten Ergebnisse. Mit 11,73 Prozentpunkten weniger, verglichen mit Gemini Pro, schneidet GPT-3.5 Turbo bei der MMLU Benchmark am schlechtesten ab. Llama 2 befindet sich ungefähr auf demselben Niveau wie GPT-3.5 Turbo, Claude Instant liegt mit 73,4% circa in der Mitte. Llama 2 liegt bei der GSM8K Benchmark wie im oberen Vergleich deutlich zurück. Von den drei anderen schneidet GPT-3.5 Turbo mit 76,3% wieder am schlechtesten ab, Claude Instant ist erneut in der Mitte. Die HellaSwag Benchmark

führt Llama 2 mit einem Abstand von 0,6 Prozentpunkten zu Gemini Pro und 5,9 Prozentpunkten zu GPT-3.5 Turbo mit einem Wert von 85,3% an.

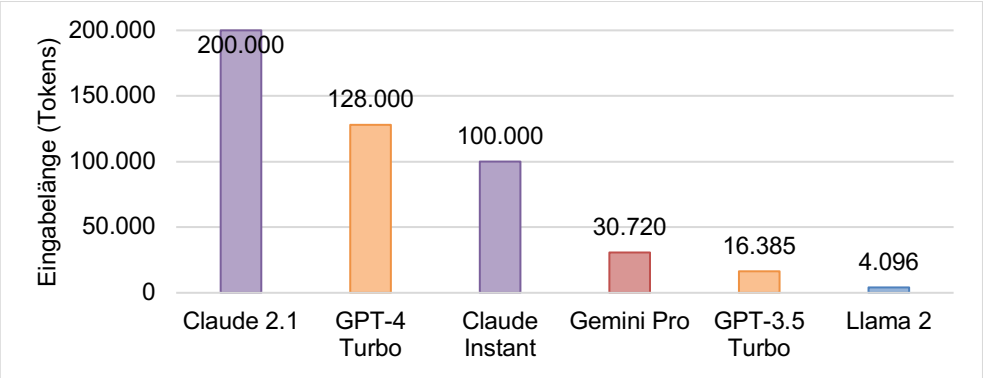
Abbildung 10: Benchmark-Ergebnisse des Standard-Segments



Quelle: Eigene Darstellung auf Basis der Daten von: Zheng et al. (2023, S.6), Anthropic (2023b, S.10), Gemini Team (2023, S.7), Meta (2023b)

GPT-3.5 Turbo erzielt mit 80,0% bei der HumanEval das beste Ergebnis aller betrachteten LLMs und liegt auch vor den Premium-Versionen. Unter den Standard-Versionen hat Gemini Pro mit 67,7% das zweitbeste, Claude Instant mit 52,8% das drittbeste Ergebnis. Die Modelle unterscheiden sich außerdem in der möglichen Token-Anzahl der Eingabesequenz, vgl. Abbildung 11. In der Betrachtung der Benchmarks wird das LLM Claude 2 verwendet. Claude 2.1 hat in Bezug auf die überprüften Benchmarks dieselbe Qualität, allerdings unterscheiden diese sich in der Eingabelänge, weshalb die Eingabelänge von Claude 2.1 verwendet wird und dementsprechend in Abbildung 11 auf Claude 2.1 referenziert wird. Preislich unterscheiden sich Claude 2 und Claude 2.1 auch nicht (Anthropic, 2023c). Anthropic bietet im Premium- sowie im Standard-Segment die längste Eingabesequenz unter den betrachteten Modellen an. GPT-4 Turbo ermöglicht im Vergleich zu den 200.000 Tokens von Claude 2.1 lediglich Eingaben mit einer Token-Anzahl von 128.000 Tokens. Im Standard-Segment folgt Gemini Pro mit 30.720 Tokens dem Vorreiter Claude Instant. Im Vergleich zu Gemini Pro fällt die Token-Anzahl bei GPT-3.5 Turbo circa halb so lang aus. Llama 2 hat mit 4.096 Tokens die kürzeste Eingabesequenz der betrachteten LLMs.

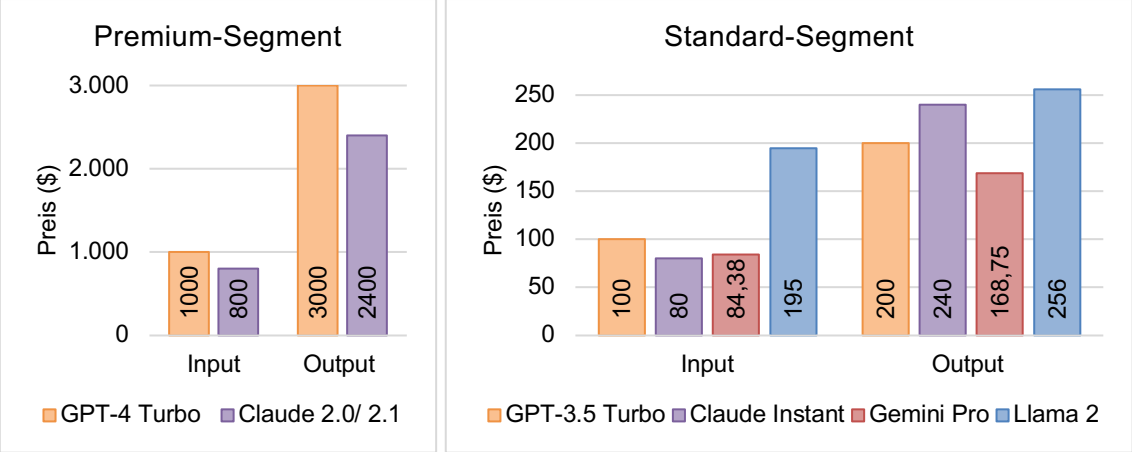
Abbildung 11: Maximal mögliche Token-Anzahl der Eingabesequenz



Quelle: Eigene Darstellung auf Basis der Daten von: OpenAI (2023i), Google (2023c), Anthropic (2023c), Meta (2023b)

Im nächsten Schritt erfolgt nun der Vergleich der API-Preise, welche in Dollar pro 1.000 Tokens aufgeführt werden. OpenAI und Anthropic geben ihre Preise auf deren Internetseite an (siehe OpenAI (2023l), Anthropic (2023c)). Im November 2023 hat OpenAI mit GPT-4 Turbo und GPT-3,5 Turbo-1106 zwei neue Versionen ihrer LLMs veröffentlicht. Neben der Verbesserung der Modelle konnten auch die API-Preise deutlich verringert werden (OpenAI, 2023i). Auch wenn sich die Benchmark-Bewertung auf die ältere Version bezieht, werden die Preise des neuen Modells genutzt, um diese mit der Konkurrenz zu vergleichen. Aufgrund der Neuheit der Gemini LLMs ist aktuell nur der Preis für das Gemini Pro Modell verfügbar, allerdings wird dieses in Dollar pro 1.000 Zeichen angegeben. Für einen Input von 1.000 Zeichen wird ein Preis von 0,00025 Dollar, für einen generierten Output von 1.000 Zeichen ein Preis von 0,0005 Dollar angegeben (Google, 2023b). Um eine Vergleichbarkeit der Preise zu schaffen müssen die Preise in Dollar pro 1000 Tokens umgerechnet werden. Die Veröffentlichung von Miller et al. (1958) wird als Basis genutzt, um diese Umrechnung annäherungsweise durchzuführen. In ihrem Artikel „Length-Frequency Statistics for Written English“ untersuchten diese die Wörter englisch geschriebener Texte auf deren Länge. Sie teilten die Worte dazu in „function words“ (FW) und „content words“ (CW) ein und kamen zu dem Ergebnis, dass FW durchschnittlich aus 3,13 und CW durchschnittlich aus 6,47 Zeichen bestehen. Die Stichprobe bestand aus insgesamt 36.299 Wörtern, wovon circa 59% den FW und 41% den CW zugeordnet werden konnten (Miller et al., 1958, S.382f.). Unter der Annahme, dass 1.000 Wörter 750 Tokens sind (OpenAI, 2023l), ergibt sich ein Umrechnungsfaktor von 3,375, mit dem die von Google angegebenen Preise multipliziert werden (Anhang B). Da Llama 2 ein Open Source Modell ist, bietet Meta keinen API Zugriff an, sondern ermöglicht es jedem Interessenten das LLM herunterzuladen und lokal zu nutzen (Meta, 2023b). Allerdings muss der Nutzer die benötigte Rechenleistung besitzen, um das Modell zu nutzen (Zhao et al., 2023, S.11). Llama 2 wird deswegen auf Basis der von Amazon angegebenen Preise für die API Nutzung mit aufgeführt (siehe Amazon Web Services (2023d)). Um die Preisunterschiede besser zu erkennen, werden die Preise der APIs in Abbildung 12 für jeweils 100 Millionen Tokens angegeben.

Abbildung 12: API-Preise der LLMs für 100 Millionen Input- und Output-Tokens



Quelle: Eigene Darstellung auf Basis eigener Berechnungen und Daten von: OpenAI (2023l), Anthropic (2023c), Google (2023b), Amazon Web Services (Amazon Web Services, 2023d, S.19)

GPT-4 Turbo erzeugt bei der Betrachtung von jeweils 100 Million Tokens Gesamtkosten von 4.000 Dollar, verglichen mit den 3.200 Dollar für Claude 2.1. GPT-4 erzielt in zwei der überprüften Benchmarks ein besseres Ergebnis als Claude 2. Die HellaSwag Benchmark ist nicht für Claude 2 vorhanden, allerdings ist der von GPT-4 erzielte Wert von 95,3% deutlich vor dem Wert des sonst überall führenden Gemini Ultra. Es ist dementsprechend davon auszugehen, dass das Modell von OpenAI auch in dieser Benchmark besser ist als das von Anthropic. Aus Sicht der Benchmarks bietet GPT-4 dem Nutzer also eine höhere Qualität. Da die höhere Qualität an einen höheren Preis gekoppelt ist, ist zu erwarten, dass sich die Nutzer je nach ihrer Zahlungsbereitschaft auf die beiden Produkte aufteilen werden (Tirole, 1999, S.213). Verstärkt wird die Aufteilung der Konsumenten durch die zusätzliche Betrachtung der Eingabelänge als Qualitätsmerkmal. Eine Festlegung der höheren Qualität ist nicht mehr objektiv möglich; vielmehr kommt es darauf an, welche Präferenzen die Nachfrager hegen (Pfähler & Wiese, 2008, S.250). Da die Nachfrage nach Produkten auf Basis ihrer Produkteigenschaften begründet werden kann (Wied-Nebbeling, 2004, S.15), werden einige Nutzer auf Grund ihrer Präferenzen die Benchmark-Ergebnisse, andere die Eingabelänge als entscheidende Produkteigenschaft ansehen. Ist die Eingabelänge der entscheidende Qualitätsfaktor, werden sich die Nutzer für Claude 2.1 entscheiden, da in diesem Fall die höhere Qualität an einen niedrigeren Preis gekoppelt ist (Tirole, 1999, S.213). Im Standard-Segment erzeugt Llama 2 die höchsten Gesamtkosten, allerdings handelt es sich hier um die Preise, die von Amazon angegeben werden. Außerdem ist Llama 2 als Open Source LLM veröffentlicht, weshalb es nicht weiter in der Kostengegenüberstellung betrachtet wird. Claude Instant ist das teuerste Modell der drei Closed Source LLMs des Standard-Segments, allerdings bei der Benchmark Bewertung nicht das Beste. Wird die Eingabelänge als Grundlage für die Qualität herangeführt, ist Claude Instant deutlich besser als die LLMs der Konkurrenz. GPT-3.5 Turbo ist mit Gesamtkosten von 300 Dollar bei der Betrachtung des Preises im Mittelfeld, jedoch schneidet es in der Qualitätsbewertung über

Benchmark Ergebnisse und Eingabelänge am schlechtesten ab. Gemini Pro ist das günstigste LLM, allerdings beruht die Ermittlung der Kosten auf der in Anhang B aufgeführten Rechnung und nicht auf dem angegebenen Preis pro 1.000 Tokens von Google. Der errechnete Preis pro 1.000 Tokens hängt nämlich von der angenommenen durchschnittlichen Zeichenanzahl eines Wortes ab. Je höher die durchschnittliche Zeichenanzahl einer Eingabesequenz ist, desto höher fällt auch der errechnete Preis für 1.000 Tokens aus. Aus den Benchmark Ergebnissen kristallisiert sich Googles LLM allerdings insgesamt als bestes LLM heraus und bietet mit 32.720 das zweitgrößte Eingabefenster des Standard-Segments. Auch bei den LLMs des Standard-Segments existiert keine Einigkeit über die beste Qualität unter Betrachtung beider Eigenschaften. Hier schneidet GPT-3.5 Turbo insgesamt am schlechtesten ab.

In Kapitel 4 wurden die für die Bachelorarbeit relevanten Unternehmen mit deren Produkten im Bereich der LLMs vorgestellt. Im Folgenden werden nun die angebotenen Chatbots für private Nutzer untereinander und die für Unternehmen untereinander verglichen. Die Anpassbarkeit digitaler Güter (Clement et al., 2019, S.36) ermöglicht es OpenAI und Meta mit den GPTs bzw. KIs eine Personalisierungsfunktion anzubieten. Unternehmen können dadurch die Zahlungsbereitschaft des Kunden erhöhen und einen weiteren Bereich des Marktes abdecken (Peters, 2010, S.120). Individualisierte Produkte können außerdem dazu beitragen Kunden zu binden (Halberstadt, 2014, S.21). Die Kundenbindung wird in Kapitel 5.5 weiter analysiert. Im Vergleich zu ChatGPT bietet die Chatvariante von Copilot dieselben Funktionen, allerdings nicht die Anpassung des Chatbots über GPTs. Bard bietet bis auf die Verknüpfung zu den Google Diensten und Apps und der Überprüfung des Outputs mit Google Suche keine Besonderheiten gegenüber ChatGPT. Verglichen mit Copilot und Bard bietet ChatGPT Plus aktuell die meisten Funktionen. Durch die Integration von Meta AI in Instagram, Messenger und WhatsApp unterscheidet sich der Chatbot von Meta in seinen Funktionen verglichen mit ChatGPT. Metas Chatbot ist mit seinen Funktionen auf die Integration in die Social Media Plattformen spezialisiert. Die KIs unterscheiden sich auch zu den GPTs. GPTs ermöglichen es dem Nutzer ChatGPT dafür zu nutzen spezielle Aufgaben zu lösen (OpenAI, 2023g), KIs von Meta sollen dem Nutzer eine individuelle Chaterfahrung liefern, mit verschiedenen Charakteren (Meta, 2023d). Hierbei werden die Präferenzen der Nutzer erneut eine Rolle bei der Anbieterwahl spielen, allerdings wurde schon die Möglichkeit des Multihoming in Kapitel 5.2 erwähnt, wodurch die simultane Nutzung von ChatGPT Plus und Meta AI denkbar ist.

ChatGPT Enterprise und Copilot Enterprise Chat bieten jeweils dieselben Funktionen wie ChatGPT Plus bzw. die normale Version von Copilot. Diese schützen bei der Nutzung zusätzlich noch die Unternehmensdaten (OpenAI (2023f), Spataro (2023)). Daraus ergibt sich, dass die Variante von OpenAI mehr Funktionen im Vergleich zu der Variante von Microsoft bietet. Nun werden Microsoft 365 Copilot, Duet AI und Amazon Q miteinander verglichen. Microsoft und Google integrieren ihren KI-Assistenten in ihre Anwendungen Microsoft 365 bzw. Workspace. Die KI-Integration hilft bei der Erstellung und Zusammenfassung der Inhalte innerhalb der Anwendungen. Mit dem Microsoft 365

Copilot erhalten Unternehmen allerdings auch noch die Möglichkeit den Chatbot außerhalb der Anwendungen zu nutzen. Amazon Q hingegen unterstützt effektiver mit den unternehmensinternen Daten und einigen AWS-Anwendungen zu arbeiten. Insgesamt fallen die Anwendungsmöglichkeiten geringer aus, allerdings kostet Amazon Q auch weniger im Vergleich zu seinen Konkurrenten. Microsoft bietet mit Microsoft Copilot Studio die Möglichkeit eigene Copiloten zu erstellen und besitzt hiermit ein Alleinstellungsmerkmal im Vergleich zu seinen Konkurrenten. Durch die Unterschiede der Produkteigenschaften handelt es sich bei den Angeboten um heterogene Güter. Die damit zusammenhängende Intensitätsminderung des Preiswettbewerbs generiert Amazon keinen Wettbewerbsvorteil durch den geringeren Preis (Woeckener, 2014, S.16). Im Vergleich dazu generiert Microsoft einen Wettbewerbsvorteil durch die Personalisierungsmöglichkeit über Copilot Studio (Peters, 2010, S.121). Welches Angebot ein Unternehmen nutzt, hängt stark mit den Voraussetzungen für die Nutzung zusammen. Die Voraussetzungen der Nutzung werden in Kapitel 5.4 erarbeitet und dann in Kapitel 5.5 in einen Kontext gesetzt.

5.4 Nutzungsvoraussetzungen der Angebote

In diesem Abschnitt der Bachelorarbeit wird erörtert, welche APIs durch welchen Cloud-Provider angeboten werden. Daraufhin werden die Marktanteile der Unternehmen mit in die Betrachtung genommen. Des weiteren werden die Voraussetzungen zur Nutzung des unternehmensspezifischen Chatbot Angebots betrachtet.

Als Cloud Computing Anbieter werden im folgenden Microsoft, Amazon sowie Google betrachtet und es wird überprüft, welche von den in Kapitel 5.3 betrachteten APIs über welchen Anbieter bereitgestellt werden. Mit den getätigten Investitionen hat sich Microsoft den Status des alleinigen Cloud Providers für die Produkte von OpenAI gesichert. Die APIs der LLMs von OpenAI werden dementsprechend ausschließlich über Microsofts Azure Cloud zur Verfügung gestellt (Microsoft Corporate Blogs, 2023). Googles eigenes LLM Gemini wird nur über das unternehmenseigene Cloud Computing Angebot zur Verfügung gestellt (Pichai & Hassabis, 2023). Die LLMs von Anthropic befinden sich im Angebot von Google und Amazon (Vahdat & Yang (2023), Amazon Web Services (2023c)). Llama 2 kann einerseits über einen API-Zugang genutzt werden, welcher von allen drei betrachteten Unternehmen angeboten wird (Montgomery (2023), (Vahdat & Yang, 2023, S.11) (2023), Amazon Web Services (2023c)). Andererseits kann das Open Source Modell direkt über die Seite von Meta kostenlos heruntergeladen werden (Meta, 2023b), allerdings muss der Nutzer die benötigten Serverkapazitäten besitzen, um das LLM zu betreiben (Zhao et al., 2023, S.11). Microsoft, Google und Amazon haben im Cloud Computing Markt zusammen einen Marktanteil von 66%. Davon belaufen sich 34% auf Amazon, 21% auf Microsoft und die restlichen 11% auf Google. Amazon ist demnach der größte Cloud Computing Anbieter der drei Unternehmen (Synergy Research Group, 2023). Open AIs ChatGPT Enterprise kann jedes Unternehmen gegen eine Gebühr nutzen, eine weitere Voraussetzung existiert nicht (OpenAI, 2023b). Auch die auf Unternehmen ausgerichtete Chatbot-Variante von Copilot kann gegen eine Gebühr,

in diesem Fall sind es 5 Dollar pro Monat pro Nutzer, genutzt werden. Unternehmen, die bereits ein Microsoft 365 Abonnement abgeschlossen haben, können diese Copilot-Variante kostenlos nutzen. Das Microsoft 365 Abonnement ist außerdem Voraussetzung dafür Microsoft 365 Copilot nutzen zu können (Spataro, 2023). Die Nutzungsvoraussetzungen für Googles Duet AI sind simultan zu denen von Microsoft. Unternehmen, die Google Workspace nutzen können dadurch auf Duet AI, zugreifen (Pappu, 2023). Duet AI kostet genau wie Microsoft 365 Copilot 30 Dollar pro Monat pro Nutzer (Holtermann & Scheuer, 2023). Der Kundenstamm von Amazons Cloud Computing Angebot AWS kann Amazon Q nutzen. Amazon Q Business kostet 20 Dollar, Amazon Q Builder 25 Dollar pro Monat pro Nutzer (Amazon Web Services, 2023b).

5.5 Lock-In-Effekte und Wechselkosten

Durch den Lock-In eines Kunden schafft ein Unternehmen es diesen an sich zu binden. Bedingt ist ein Lock-In durch das Vorliegen von Wechselkosten (Peters, 2010, S.50). In den vorherigen Kapiteln wurde gezeigt, dass die Unternehmensangebote einzeln betrachtet werden müssen. Aus diesem Grund werden die Chatbot Angebote für Nutzer und Unternehmen sowie die Bereitstellung der LLMs über APIs auf Wechselkosten hin überprüft. Neben diesen werden die Unternehmen noch auf weitere mögliche Lock-In-Effekte hin überprüft.

In Kapitel 5.1 konnte aufgrund der großen Nutzerzahl für ChatGPT ein First-Mover-Vorteil ermittelt werden; verstärkt wird dieser durch vorhandene Wechselkosten (Wied-Nebbeling, 2004, S.288). Digitale Güter gelten als Erfahrungsgüter, da diese immateriell sind und die Qualität nicht direkt bestimmbar ist (Schmidt, 2007, S.34f.). In Kapitel 5.3 wurde schon angemerkt, dass die Benchmarks nur als Orientierung dienen sollen, da die Ergebnisse zum Beispiel durch die Art der Prompts verändert werden können (Horvitz & Nori, 2023). Die Qualität der Chatbots lässt sich dementsprechend nur schwierig feststellen. Kann das erste Unternehmen im Markt die Nutzer von dessen Produktqualität überzeugen, kann es sich einen First-Mover-Vorteil sichern (Wied-Nebbeling, 2004, S.281). OpenAI hat sein Produkt sehr schnell an einer großen Nutzeranzahl zur Verfügung gestellt, wodurch diese die Qualität von ChatGPT testen konnten. Die Berechnung des HHI auf Basis der Seitenaufrufe hat gezeigt, dass eine hohe Marktkonzentration in Richtung von ChatGPT besteht. Der Berechnungszeitraum beläuft sich auf Juni bis Oktober, es ist also davon auszugehen, dass die Nutzer die Qualität von ChatGPT in der Anfangsphase für gut befunden haben und deshalb bei dem Angebot von OpenAI bleiben. Die Kundenbindung an ChatGPT lässt vermuten, dass sich OpenAI den First-Mover-Vorteil sichern konnte (Wied-Nebbeling, 2004, S.281). Durch die mediale Aufmerksamkeit von ChatGPT konnte sich der Produktname innerhalb des Marktes mit seinem Domainnamen etablieren; dieser ist nicht durch einen Mitbewerber kopierbar. Aufgrund der Fokussierung der Medien auf Produktinnovationen konnte OpenAI auch diesen First-Mover-Vorteil für sich nutzen (Halberstadt, 2014, S.92ff.). Die Tatsache, dass es sich um ein Erfahrungsgut handelt und sich ChatGPT im Markt mit seinem Namen etablieren konnte, führt zur einer Wechselbarriere. Durch diese entsteht eine psychologische

Bindung des Kunden und führt zu einem Lock-In des Kunden (Blut, 2008, S.28). OpenAIs GPTs und Metas KIs können aufgrund des Individualisierungsangebots zu einer Kundenbindung beitragen (Halberstadt, 2014, S.21). Beim Anbieterwechsel müsste der Kunde zunächst den Umgang mit dem Produkt neu erlernen und Zeit investieren dieses neu anzupassen, wodurch Wechselkosten entstehen (Burnham et al., 2003, S.111f.). Google mit der Integration von Bard in Assistant (Hsiao, 2023a), Microsoft mit der Integration von Copilot in den Edge Browser und Windows 11 (Spataro (2023), Microsoft (2023b)), Amazon mit der Integration eines LLM in Alexa (Rausch, 2023) und Meta durch die Integration von Meta AI in Instagram, Messenger, WhatsApp und die Ray-Ban Smart Glasses (Meta, 2023d) erzeugen bei den Nutzern der Produkte Wechselbarrieren (Blut, 2008, S.28).

ChatGPT kann von einem Lock-In-Effekt seiner Nutzergruppe profitieren. Gerade durch die neue Funktion der GPTs nimmt die Stärke des Lock-In-Effektes zu. Erfolgt die Integration von Gemini Ultra in Bard wurde in Kapitel 5.3 gezeigt, dass Gemini Ultra bezogen auf die Benchmarks eine bessere Qualität aufweist als GPT-4. Unter diesem Gesichtspunkt hätte Bard eine bessere Qualität als ChatGPT. In diesem Kapitel wurde allerdings aufgeführt, dass sich die Qualität sehr schlecht von außen ermitteln lässt. Aus einer schlechteren Vergleichbarkeit der Qualitäten entsteht eine Vergrößerung des vorliegenden Lock-In-Effektes. Generell ist eine höhere Qualität kein hinreichender Indikator, um einen Anbieterwechsel zu erwarten (Halberstadt, 2014, S.20f.). Durch die aufgeführte Integration der LLMs in die verschiedensten Produkte und die Erzeugung eines psychologischen Lock-In-Effekts ist von einer Aufteilung der Nutzer auf die Anbieter auszugehen. Trotz des Lock-In-Effekts von OpenAI können aufgrund von Marktgröße und Nutzerpräferenzen mehrere Produkte im Markt nebeneinander existieren (Wied-Nebbeling, 2004, S.288). Meta kann basierend auf den immensen Nutzerzahlen ihrer Produkte potenziell einen sehr starken Lock-In-Effekt erzeugen, hinzu kommt die Individualisierungsmöglichkeit des Angebots durch die KIs. Aus Kapitel 4 ging hervor, dass die Marktanteile für Microsoft Edge derzeit relativ gering sind. Etwas höher ist der Marktanteil von Windows 11, welcher innerhalb der Windows Betriebssystemversionen im letzten Jahr um circa 10% gestiegen ist (StatCounter Global Stats, 2023c). Insgesamt ist davon auszugehen, dass wenn ein Nutzer bei Benutzung eines dieser Produkte auf die Fähigkeiten eines LLM zugreifen möchte, dieser sich auf Grund einer psychologischen Wechselbarriere für das in das Produkt integrierte LLM entscheiden wird (Blut, 2008, S.28). Bei den auf Unternehmen ausgelegten Angeboten werden sich die Nutzer aufgrund der herausgearbeiteten Wechselkosten und damit einhergehenden Lock-In-Effekten auf die verschiedenen Unternehmen verteilen.

Bei der Zusammenarbeit eines Unternehmens mit OpenAI im Zuge der Initiative „Data Partnerships“ kann ein Lock-In-Effekt entstehen. Einerseits lässt dieser sich durch die bereits investierten zeitlichen und personellen Ressourcen, um die Daten für den FT Prozess bereitzustellen, begründen. Diese müssten bei einem Anbieterwechsel erneut getragen werden (Burnham et al., 2003, S.112). Andererseits sorgt die Individualisierung des Angebots durch das FT des LLM für eine Bindung des Kunden (Halberstadt, 2014,

S.21). Neben diesen Wechselkosten gelten für ChatGPT Enterprise auch die erarbeiteten Wechselkosten durch die mediale Aufmerksamkeit und die Individualisierung des Angebots durch den Einsatz von GPTs. Microsofts Unternehmenskunden können, wie in Kapitel 4 erwähnt, mit ihrem Microsoft 365 Abonnement die unternehmensspezifische Chatvariante von Copilot gratis nutzen. Durch die Nutzung von OpenAIs ChatGPT Enterprise an Stelle von Copilot müssten Extrakosten getragen werden. Diese extra Kosten können als Wechselkosten betrachtet werden und sorgen für einen Lock-In-Effekt (Burnham et al., 2003, S.110). Die Ergebnisse aus Kapitel 5.4 lassen auf die Existenz von Lock-In-Effekten schließen. Um die Chatbot Angebote von Microsoft, Google oder Amazon nutzen zu können, muss ein Unternehmen bereits Kunde bei dem jeweiligen Anbieter sein. Mit Microsofts Copilot 365 und Googles Duet AI implementieren beide Unternehmen generative KI in ihre Office Software. Die Nutzung der Chatbot Angebote von Microsoft und Google ist demnach an das Abonnement der jeweiligen Office Software gekoppelt. Beim Anbieterwechsel würden hierdurch verschiedene Arten von Wechselkosten anfallen. Die Unternehmen müssten sich zunächst über das Angebot der Konkurrenz informieren und überprüfen, ob alle Anforderungen umsetzbar sind. Außerdem müssten sich die Mitarbeiter beim Wechsel zunächst mit der neuen Software und den Funktionen dieser vertraut machen. Auch die Migration der Dokumente und Strukturen erzeugt Wechselkosten (Burnham et al., 2003, S.111f.). Amazons Cloud Computing Angebot kann von Unternehmen auch verwendet werden, wenn diese Microsoft 365 oder Google Workspace als Office Software benutzen. Dementsprechend kann ein Unternehmen beispielsweise Microsoft 365 und Produkte von AWS nutzen und sich entscheiden, ob es Microsoft 365 Copilot oder Amazon Q nutzen möchte. Simultane Situation ergibt sich für die parallele Nutzung von Google Workplace und AWS. In Kapitel 5.3 wurden Produktdifferenzierungen zwischen den drei Chatbots festgestellt, wodurch die Wahl des Angebots auf den Präferenzen der Nachfrager beruht (Tirole, 1999, S.213). Microsoft kann seine Kunden allerdings durch Copilot Studio an sich binden, da hierdurch ein individueller Copilot durch den Nutzer erzeugt werden kann und individuelle Angebote zu Kundenbindung beitragen (Halberstadt, 2014, S.21). Für die Nutzung der APIs wurden in Kapitel 5.4 Unterschiede des Angebots zwischen Microsoft, Google und Amazon herausgearbeitet. Demnach stehen nicht alle LLMs über jeden Cloud Computing Anbieter zur Verfügung. Erfordert der Wechsel der API auch den Wechsel des Cloud Computing Anbieters, entstehen dieselben hohen Wechselkosten wie die eben beschriebenen beim Wechsel von Copilot 365 auf Duet AI.

5.6 Kostenstruktur

Aus Kapitel 3.4 ging hervor, dass sich digitale Produkte durch hohe Fixkosten und geringe variable Kosten auszeichnen (Schmidt, 2007, S.34). Dies begünstigt die Existenz positiver Skaleneffekte (Holtforth, 2016, S.21). Schon in Kapitel 2.2 wurde angemerkt, dass der Lernprozess sehr viel Kapital benötigt. OpenAI gibt nicht an, wie viel der Trainingsprozess von GPT-4 gekostet hat, allerdings zeigt der Artikel „The AI Brick Wall“ von Patel (2023) die theoretische Berechnung der Trainingskosten unter Betrachtung der

Anzahl von Parametern sowie Trainings-Tokens und der Architektur des LLM. Das DeepMind Entwicklerteam rund um Hoffmann et al. hat in seinem Paper „Training Compute-Optimal Large Language Models“ die optimale Anzahl an Trainingstokens im Vergleich zur Parameteranzahl des LLM analysiert (Hoffmann et al., 2022). Auf Basis dieser Ergebnisse hat Patel die theoretischen Kosten des Trainingsprozesses für verschiedene Modellgrößen berechnet. Ein LLM mit 175 Milliarden Parametern würde unter diesen Voraussetzungen über 9 Millionen Euro, eines mit 520 Milliarden Parametern würde schon über 80 Millionen Euro kosten. Patel hat des weiteren die theoretischen Kosten für zu diesem Zeitraum veröffentlichte LLMs berechnet. Da diese nicht das optimale Verhältnis von Parameteranzahl zu Trainingstoken Anzahl nutzen, fallen die Kosten für den Trainingsprozess deutlich geringer aus, teilweise in Millionenhöhe. Mit zunehmender Parameter- und Token-Anzahl des LLM steigen die Trainingskosten an. (Patel, 2023). Auch sind diese aufgrund ihrer Irreversibilität als versunkende Kosten zu interpretieren (Fichert, 1998, S.346). Versunkene Kosten fungieren als Marktaustrittsbarriere (Wied-Nebbeling, 2004, S.12) und sorgen für die Nichterfüllung der Voraussetzungen eines angreifbaren Marktes (Baumol, 1982). Auch die ermittelten Markteintrittsbarrieren durch die Etablierung der Produktvarianten und die nachgewiesenen Netzwerkeffekte unterstützen diese Vermutung (Bain (1956), Peters (2010, S.50)). Neue Marktteilnehmer haben dadurch einen Kostennachteil gegenüber den am Markt etablierten Unternehmen. Es ist davon auszugehen, dass die Anzahl der am Markt tätigen Unternehmen hierdurch gering bleibt (Wied-Nebbeling (2004, S.7), Peters (2010, S.16)). Die Kosten, welche durch eine Abfrage entstehen, können als variable Kosten betrachtet werden, da diese nur auftreten, wenn ein Prompt an das LLM gestellt wird. Die Kosten, die durch die Abfragen entstehen, sind deutlich höher als die Trainingskosten. Die variablen Kosten zum Betreiben eines LLM übersteigen die fixen Kosten. Patel und Ahmad (2023) schätzen die täglichen Kosten für ChatGPT auf knapp 700.000 US-Dollar (Patel & Ahmad, 2023). Die Berechnung der Kosten basiert dabei auf der GPT-3 Version von ChatGPT (Mok, 2023). In dem Paper „From Words to Watts“ haben Samsi et al. unter anderem den Energieverbrauch von LLMs anhand drei verschiedener Versionen von Metas Llama untersucht. Mit zunehmender Parameter Anzahl der Llama Versionen nimmt der gemessene Energieverbrauch zu (Samsi et al., 2023). Bei der Generierung eines Tokens werden, wie in Kapitel 2.1 erwähnt, bei GPT-3 weniger Parameter genutzt als bei GPT-4, obwohl hierbei die MoE Architektur verwendet worden ist (Patel & Wong, 2023). Demnach ist davon auszugehen, dass die täglichen Kosten für die GPT-4 Version von ChatGPT nochmal höher ausfallen. Die LLMs werden Kunden auch über APIs zur Verfügung gestellt. Die aus Programmiercode bestehende API ermöglicht es einem Kunden das LLM zu nutzen, wodurch bei der Bereitstellung des benötigten API-Schlüssel sehr geringe variable Kosten anfallen (Block, 2023, S.108). Allerdings wird bei einer API nicht die Software an einen Kunden verkauft, sondern nur die Nutzung dieser ermöglicht (Frank et al., 2021, S.19). Die in diesem Fall benötigte Rechenleistung für das LLM wird dementsprechend weiterhin durch die LLM Anbieter bereitgestellt, wodurch diese weiterhin die eben ermittelten hohen variablen Kosten tragen müssen. Die dennoch hohen fixen Kosten können zur Abhaltung von Markteintritten der Konkurrenz führen. Konkurrierende Unternehmen

werden nicht dieselbe Qualität zu einem geringeren Preis anbieten können, um so die Nutzer von ChatGPT abzuwerben (Fichert, 1998, S.346f.). Es ist davon auszugehen, dass durch diesen Fixkostenanteil zunächst positive Skaleneffekte generiert werden können, da die durchschnittlichen Kosten zunächst sinken (Holtforth, 2016, S.21). Ab einem bestimmten Punkt werden diese allerdings nicht mehr sinken, da die Kosten für eine Abfrage mit zunehmender Anzahl an Abfragen eher steigen. Es ist eher davon auszugehen, dass wenn mehr Abfragen simultan ausgeführt werden, die variablen Kosten steigen, da mehr Berechnungen gleichzeitig erfolgen müssen und dadurch mehr Energie verbraucht wird (Samsi et al. (2023), Patel & Wong (2023)). Durch die zunächst sinkenden durchschnittlichen Kosten wird die Anzahl der im Markt tätigen Unternehmen begrenzt (Peters, 2010, S.16). Die Aufbringung der versunkenden Kosten in Form der fixen Kosten lässt normalerweise vermuten, dass die aktuell im Markt tätigen Unternehmen zunächst versuchen werden so lange im Markt zu bleiben, bis diese mindestens Umsätze in Höhe der versunkenden Kosten erwirtschaftet haben (Blum et al., 2006, S.86).

Im Falle von Economies of Scope bedingt die Produktion eines Gutes die Produktion eines anderen Gutes positiv (Premer, 2015, S.349). Google verwendet für sein Bildgenerierungsmodell „Imagen“ den Encoder Teil ihres T5-XXL Modells (Saharia et al., 2022, S.1). Das T-5 LLM von Google wurde als Open Source Modell veröffentlicht (Roberts & Raffel, 2020), wodurch jeder auf dessen Programmiercode zugreifen kann und diesen verwenden kann (Haun, 2014, S.417). Der Encoder Teil des T5-XXL wird dadurch auch von OpenAI für DALL·E 3 (Betker et al., 2023, S.17) und von Meta für deren Bildgenerierungsmodell Emu verwendet (Dai et al., 2023, S.5). Durch die Bereitstellung des T5 LLM als Open Source Software kann sich Google keinen Verbundvorteil sichern, da auch Konkurrenten wie OpenAI und Meta von dem T5 LLM profitieren können, indem diese das LLM für deren eigene Modelle nutzen können.

5.7 Diskussion

Das Ergebnis der in der Bachelorarbeit durchgeführten Analyse ist als erste Einschätzung über die zukünftige Marktsituation zu interpretieren. Bei der Auswertung der einzelnen Prüfkriterien sowie der herausgearbeiteten Theorie ist zu berücksichtigen, dass der Umfang einer Bachelorarbeit begrenzt ist. Es konnte ein Prüfschema entwickelt werden, mit dem der Markt der LLMs in Zukunft erneut überprüft werden kann. Die erneute Durchführung der Analyse ist wegen der Veröffentlichung von Bard Advanced und dem GPT-Store sowie der Bereitstellung von Meta AI weltweit wünschenswert, da hierdurch die Ergebnisse bestätigt oder widerlegt werden können. Die Betrachtung weiterer Open Source Modelle sollte außerdem erfolgen. Die Anzahl der Downloads könnte hierbei mit der Anzahl der API User verglichen werden, sofern die Daten vorhanden sind. Die Datenverfügbarkeit ist das größte Problem, welches bei der Bearbeitung des Themas entstanden ist. Die Verwendung der Seitenaufrufe haben sich bei der Ermittlung des Marktanteils im Bereich der Chatbots als gute Alternative für die Umsätze herausgestellt. Viele Ergebnisse der Bachelorarbeit basieren dennoch auf einer theoretischen Überprüfung und konnten nicht mit Hilfe von Daten bzw. Berechnungen gestützt werden. Weitere

Arbeiten in diesem Themenbereich sollten sich auf einzelne dieser Prüfpunkte fokussieren und diese tiefergehend analysieren. Im Bereich der Produktdifferenzierung könnten Umfragen durchgeführt werden, um dadurch die getroffenen Vermutungen, welche Produkteigenschaften einen Wettbewerbsvorteil erzeugen, zu überprüfen. Im Falle der Netzwerkeffekte und der Lock-In-Effekte könnte in weiteren Arbeiten ein Bewertungssystem entwickelt werden, um die Stärke der herausgearbeiteten Effekte zu quantifizieren. Die Analyse der Kostenstruktur stellte sich als äußerst schwierig dar, da keine offiziellen Angaben über die Trainings- und Betriebskosten von OpenAI vorhanden sind. Selbst die hinter GPT-4 liegende Modellarchitektur wird von OpenAI nicht veröffentlicht. Weitere Forschungsarbeiten könnten das Thema der Kosten genauer analysieren und zum Beispiel die Kosten je 1.000 Tokens der einzelnen Modelle ermitteln. Die Betrachtung von Innovationen innerhalb des LLM Marktes kann als weiterer Prüfpunkt mit in das Prüfschema aufgenommen werden. Des weiteren können äußere Umstände durch beispielsweise Regulatorien oder innere Umstände wie die kurzzeitige Entlassung von Sam Altman mit in die Betrachtung als Prüfpunkt einfließen. Das Forschungsthema des LLM Marktes birgt dementsprechend noch interessante auf der Bachelorarbeit aufbauende Möglichkeiten die Forschung fortzuführen.

6 Fazit und Marktprognose

Kapitel 5 hat gezeigt, dass der Chatbot Markt differenziert betrachtet werden muss. Neben den Chatbots für Privatpersonen und Unternehmen stellen die API Angebote eine dritte Betrachtungsebene dar. Die drei Ebenen werden nun einzeln betrachtet um darauf aufbauend die Forschungsfrage, ob die Gefahr besteht, dass OpenAI Monopolstatus im Markt der LLMs einnimmt zu beantworten. Abgeschlossen wird die Bachelorarbeit mit einer Prognose der zukünftigen Marktform.

Im Bereich der Chatbots konnte sich OpenAI eine marktbeherrschende Stellung sichern. Während der Analyse konnten für ChatGPT positive Netzwerkeffekte sowie Lock-In-Effekte festgestellt werden. Der dadurch generierte First-Mover-Vorteil sorgt für die Entstehung einer Installed Base der Chatbot Nutzer (Wied-Nebbeling, 2004, S.288). Neue Marktteilnehmer haben dadurch einen Nachteil, da sie mehr Mittel aufwenden müssen, um eine eigene Nutzerbasis aufzubauen (Peters, 2010, S.50). Auch die Individualisierungsmöglichkeit und das Bereitstellen der meisten Funktionen im Vergleich zu den anderen Chatbots beschert ChatGPT neben dem Aufbau eines Images einen Wettbewerbsvorteil. Auch Meta AI bietet dem Nutzer eine Individualisierungsmöglichkeit und generiert einen Vorteil über die Bereitstellung ihres Chatbots über Messenger, Instagram, WhatsApp und die Ray-Ban Smart Glasses. Diesen Bereitstellungsvorteil besitzt auch Microsofts Copilot über die Integration in Windows 11 und Microsoft Edge. Google und Amazon sichern sich diesen Vorteil durch die Integration eines LLM in deren Sprachassistenten. Googles Bard steht im Vergleich zu den anderen Chatbots allerdings am schlechtesten dar, da dieser sich sonst keinen nennenswerten Wettbewerbsvorteil sichern kann. Durch die Analyse der unternehmensspezifischen Chatbot Angebote konnten für ChatGPT Enterprise dieselben Wettbewerbsvorteile wie für das ChatGPT Plus Angebot ermittelt werden. Gezeigt werden konnte auch eine verstärkte Bindung der Kunden durch die „Data Partnerships“ Initiative von OpenAI. Demgegenüber steht die Möglichkeit für Unternehmen, die Microsoft 365 nutzen, die unternehmensspezifische Version von Copilot kostenlos zu nutzen. In Kapitel 5.3 wurden des Weiteren Produktdifferenzierungen zwischen ChatGPT Enterprise, Copilot 365, Duet AI und Amazon Q festgestellt. Auf Basis der Betrachtung der Nutzungsvoraussetzungen in Kapitel 5.4 konnten für die einzelnen Anbieter im nachfolgenden Kapitel Lock-In-Effekte ermittelt werden. Bei der Analyse der APIs konnten in Kapitel 5.3 Unterschiede hinsichtlich der Qualität und der Kosten der einzelnen APIs festgestellt werden. Es wurde erarbeitet, dass die Qualität je nach Betrachtungsweise (Benchmark Ergebnisse oder Eingabelänge) variiert und somit von den Präferenzen der Nachfrage abhängt. Zum aktuellen Zeitpunkt ist GPT-4 zwar eines der besten LLMs bezogen auf die Benchmark Ergebnisse und Eingabelänge, kostet allerdings auch am meisten. Die betrachteten Cloud Computing Anbieter bieten außerdem unterschiedliche Modelle als API an, der dadurch entstehende Lock-In-Effekt wurde in Kapitel 5.5 ermittelt. Meta ist das einzige der betrachteten Unternehmen, welches sein LLM als Open Source Software bereitstellt.

Im Falle eines Monopols muss der Monopolist in der Lage sein die gesamte Marktnachfrage decken zu können (Blum et al., 2006, S.46). Die in Kapitel 5.2 aufgeführten

Restriktionen von OpenAI um eine Überlastung von ChatGPT vorzubeugen lassen darauf schließen, dass die gesamte Marktnachfrage des Chatbot Marktes nicht von einem Unternehmen alleine gestemmt werden kann. Außerdem wird bei der Entstehung eines Monopols davon ausgegangen, dass für das angebotene Gut keine nahen Substitute existieren (Bester, 2016, S.29). Auf Basis der Produktdifferenzierung konnte nachgewiesen werden, dass sich die Angebote der Unternehmen hinsichtlich der Funktionen und der Qualität unterscheiden. Die Produktdifferenzierung hat außerdem zur Folge, dass der Druck des Preiswettbewerbs verringert wird. Eine Verringerung des Preises führt dementsprechend nicht dazu, dass ein Anbieter seine gesamte Nachfrage verliert, wenn ein Konkurrent den Preis für das Produkt senkt. (Woeckener, 2014, S.15f.). Im Falle eines natürlichen Monopols kann ein Anbieter die gesamte Marktnachfrage günstiger bedienen, als dies zwei Unternehmen könnten (Bühler & Jaeger, 2002, S.23). Die Ergebnisse der Betrachtung der Kostenstruktur aus Kapitel 5.6 weisen darauf hin, dass die Voraussetzungen für ein natürliches Monopol nicht gegeben sind. Allerdings konnte gezeigt werden, dass die ermittelten Markteintritts- sowie Marktaustrittsbarrieren für eine kleine Anzahl an Marktteilnehmern führen. Die Möglichkeit einen Vorteil durch Verbundvorteile zu generieren konnte außerdem nicht ermittelt werden. Die ermittelten Lock-In-Effekte führen aus Sicht des Autors dazu, dass sich der gesamte Markt auf die im Markt agierenden Unternehmen aufteilen wird. Die Überprüfung der Produktdifferenzierung stärkt diese These, da keines der Unternehmen in jedem der drei Bereiche das beste Angebot vorzuweisen hat. Dies hat zur Folge, dass die Entstehung eines Teilmonopols ausgeschlossen werden kann, da hierfür ein Unternehmen im gesamten LLM Markt einen sehr großen Marktanteil haben müsste (Wied-Nebbeling, 2004, S.56). Die ermittelte Existenz nicht vollständiger Substitute auf der Angebotsseite, begründet durch die Produktdifferenzierung, ist als Indiz für die monopolistische Konkurrenz zu betrachten (Letzner, 1997, S.16). Dagegen spricht allerdings die eben getroffene Annahme, dass innerhalb des Marktes eine kleine Anzahl an Unternehmen existieren wird. Im Falle der monopolistischen Konkurrenz würde sich die Marktnachfrage auf eine große Anzahl an Unternehmen aufteilen (Premer, 2015, S.351). Eine weitere Voraussetzung stellt das Vorhandensein eines angreifbaren Marktes dar (Pindyck et al., 1998, S.525). Die Eigenschaft des freien Marktzutritts wird allerdings aufgrund der Kostenstruktur sowie der ermittelten Markteintrittsbarrieren nicht gewährleistet. Besteht nun also die Gefahr, dass OpenAI Monopolstatus im Bereich der LLMs erreichen wird? Auf Basis der herausgearbeiteten Ergebnisse lässt sich diese Frage verneinen. Das Vorhandensein weniger Marktteilnehmer mit ausgeglichenen Marktanteilen lässt den Autor davon ausgehen, dass sich im LLM Markt die Marktform des Oligopols einstellt (Bühler & Jaeger, 2002, S.71).

Zum Abschluss der Arbeit wird nun noch eine Prognose für die Marktformen der ermittelten Teilmärkte abgegeben. Durch die Analyseergebnisse kommt der Autor zu dem Schluss, dass sich in Bezug auf das Chatbot Angebot verschiedene Teilmärkte bilden werden. Im Bereich des Desktop Angebots wird OpenAI aufgrund der oben angesprochenen Wettbewerbsvorteile weiterhin eine marktbeherrschende Stellung einnehmen. Es ist davon auszugehen, dass OpenAI mit ChatGPT Plus den Status des

Teilmonopolisten halten wird. Kann OpenAI das aufgebaute Image aufrechterhalten, kann von einer Erhaltung der dominierenden Marktposition ausgegangen werden. In Kapitel 5 konnte gezeigt werden, dass der Wechsel eines Nutzers von ChatGPT zu Bard sehr unwahrscheinlich ist. Google hat die Chance Bard Advance zu einem geringeren Preis als ChatGPT Plus anzubieten, um so potenziell eine neue Käuferschicht zu akquirieren (Wied-Nebbeling, 2004, S.71). Aus Kapitel 4 ging hervor, dass die Anzahl der Nutzer des Edge Browsers und von Windows 11 aktuell noch gering ist. Es ist davon auszugehen, dass Edge beziehungsweise Windows 11 Nutzer Copilot nur nutzen werden, wenn diese kein ChatGPT Plus Abonnement besitzen. Begründen lässt sich dies durch die mit dem Abonnement einhergehenden positiven Netzwerkeffekte und die ermittelten Lock-In-Effekte. Meta fokussiert sich auf die Implementierung ihres Chatbots in deren Apps und besitzt wegen der bereits großen Installed Base ein immenses Potenzial eine marktbeherrschende Stellung im Bereich der Smartphone Nutzer einzunehmen. Die Ergebnisse aus Kapitel 5.3, die auf die Möglichkeit der simultanen Nutzung von ChatGPT und Meta AI schließen lassen, veranlassen den Autor zu der Annahme, dass ChatGPT und Meta AI jeweils in ihrem Anwendungsbereich den Markt dominieren werden. ChatGPT Plus Nutzer werden demnach weiterhin die Smartphone Applikation für den normalen Gebrauch nutzen. Benötigt der Nutzer bei Verwendung der Applikationen von Meta allerdings die Fähigkeiten eines LLMs ist davon auszugehen, dass dieser aufgrund der ermittelten psychologischen Wechselbarrieren aus Kapitel 5.5 den Chatbot von Meta nutzen wird. Wie sich die Marktanteile aufteilen, hängt vom Nutzungsverhalten der Smartphone Nutzer ab; der Autor spricht Meta in diesem Bereich allerdings einen Vorteil zu. Die erzeugten Wechselbarrieren durch die Implementierung der LLMs in die Ray-Ban Smart Glasses und die Sprachassistenten von Google und Amazon werden vermutlich nicht für den Nutzerrückgang von ChatGPT sorgen. Vielmehr werden diese Produkte neue Nutzungsmöglichkeiten für LLMs erzeugen und so die Nutzer bei neuen Aufgabenfeldern unterstützen. Im Bereich der unternehmensspezifischen Chatbot Angebote wird kein Unternehmen eine marktbeherrschende Stellung einnehmen. Neben den Präferenzen der Nachfrager wird die Kaufentscheidung auch durch die angesprochenen Lock-In-Effekte entschieden. Die Nutzer werden sich auf die vorhandenen Unternehmen im Markt aufteilen. Je nachdem welche Office Software die Unternehmen nutzen und welche Anwendungsanforderungen diese an den Chatbot haben, ändert sich deren Kaufentscheidung. Durch die qualitativen und preislichen Unterschiede der APIs entsteht die Tendenz, dass sich die Nachfrager auf die einzelnen Anbieter und deren Modelle aufteilen werden (Tirole, 1999, S.212). Die ermittelten Lock-In-Effekte durch das eingeschränkte API Angebot der Cloud Computing Anbieter verstärken die Annahme, dass sich die Nutzer auf die verschiedenen API Angebote verteilen werden. Meta bedient mit seinem Open Source LLM Llama 2 eine andere Nutzergruppe. Durch die Bereitstellung einer Open Source Software können deren Nutzer den Programmiercode nach ihren Wünschen anpassen (Haun, 2014, S.417). Es ist davon auszugehen, dass gerade KI Interessierte Forscher und Entwickler Zielgruppe dieser Open Source Modelle sind. Die Betrachtung der Kostenstruktur hat gezeigt, dass der Trainingsprozess eines LLMs je nach Größe des LLMs hohe Kosten in Form von versunkenen Kosten verursacht. In

Kapitel 5.6 konnte auch gezeigt werden, dass die Summe der variablen Kosten für das Betreiben eines LLMs deutlich über den der fixen Kosten liegen. Auf Grund der immensen Kosten, die ein LLM verursacht, geht der Autor davon aus, dass nur große Technikunternehmen generell eine Chance haben in den Markt der LLMs einzutreten.

7 Zusammenfassung

OpenAI hat im November 2022 den Chatbot ChatGPT veröffentlicht. Seit der Veröffentlichung ist die Nachfrage nach Chatbots und dem Thema der LLMs stark angestiegen und die Implementierung von LLMs wurde in immer mehr Bereichen diskutiert. Grundlage für ChatGPT ist die Transformer Architektur, welche auf dem Self Attention Mechanismus basiert. Die Veröffentlichungen im Bereich der Ökonomie beziehen sich bislang auf die Auswirkungen der LLMs auf den Arbeitsmarkt. Mit der Forschungsfrage „Besteht die Gefahr, dass OpenAI Monopolstatus im Markt der Large Language Modelle erreichen wird?“ ist der Markt der Large Language Modelle (LLMs) analysiert worden. Als Konkurrenten von OpenAI werden Microsoft, Google, Amazon und Meta betrachtet. Die Bearbeitung der Fragestellung erfolgt über die Erstellung eines Prüfschemas, bestehend aus Monopolisierungs- sowie Wettbewerbstreibern. Neben diesen werden die aktuelle Marktsituation des Chatbot Marktes sowie die Nutzungsvoraussetzungen der angebotenen Leistungen der Unternehmen betrachtet. Die Reihenfolge der Prüfpunkte ist so gewählt, dass die Ergebnisse eines Prüfpunktes mit in die Ergebnisbildung der darauffolgenden Prüfpunkte einfließen. Eine Einführung in die Analyse schafft die Betrachtung der aktuellen Marktsituation des Chatbot Marktes auf Basis der Seitenaufrufe. Darauf aufbauend wird im nächsten Prüfpunkt die Existenz von Netzwerkeffekten überprüft, gefolgt von der Überprüfung vorliegender Produktdifferenzierung über die Qualität der LLMs und der damit verbundene Preis. Auch die Betrachtung der Leistungen der einzelnen Anbieter fällt in diesen Prüfpunkt. Im nächsten Prüfpunkt werden die Nutzungsvoraussetzungen der Angebote aufgeführt. Abgeschlossen wird die Analyse mit der Ermittlung von Lock-In-Effekten und der Überprüfung der Kostenstruktur. Hierfür wird die Existenz von Economies of Scale und Scope analysiert. Auf Basis der Analyseergebnisse konnte festgestellt werden, dass sich der Markt der LLMs in einzelne Teilmärkte aufteilen lässt. Neben dem Teilmarkt der Chatbots für private Nutzer, existieren des weiteren noch ein Teilmarkt für unternehmensspezifische Chatbots und einer für die Bereitstellung der LLMs über APIs oder Open Source Software. Anhand der Ergebnisse konnte keine Gefahr der Monopolbildung im gesamten Markt der LLMs festgestellt werden. Vielmehr ist anzunehmen, dass sich die Marktform des Oligopols einstellen wird. ChatGPT wird weiterhin die Position des Teilmonopolisten im Chatbot Markt der Desktop Nutzer einnehmen; Meta wird den Chatbot Markt der Smartphone Nutzer anführen. Der Markt der unternehmensspezifischen und der API Angebote wird sich auf die darin tätigen Unternehmen aufteilen. Weiterer Forschungsbedarf besteht in einer tieferen Analyse der erarbeiteten Prüfpunkte durch eine Quantifizierung der Prüfpunkte. Außerdem kann der Prüfkatalog um weitere Wettbewerbstreiber ergänzt werden. Die Analyse des Wettbewerbs zwischen Meta und OpenAI stellt ein weiteres Forschungsthema dar, für welches eine tiefere Analyse sinnvoll erscheint.

Literaturverzeichnis

Adamopoulou, E. & Moussiades, L. (2020) „An Overview of Chatbot Technology“, in Maglogiannis, I., Iliadis, L. & Pimenidis, E. (Hg.) *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II,* Cham, Springer International Publishing; Imprint Springer, S. 373–383.

Altman, S. (2023a) *Planning for AGI and beyond* [Online], OpenAI. Verfügbar unter https://openai.com/blog/planning-for-agi-and-beyond (Abgerufen am 25 November 2023).

Altman, S. (2023b) *We are pausing new ChatGPT Plus sign-ups for a bit* [Online], X. Verfügbar unter https://twitter.com/sama/status/1724626002595471740 (Abgerufen am 11 Dezember 2023).

Amazon Web Services (2023a) *Amazon Q (Preview): Your generative AI–powered assistant designed for work that can be tailored to your business* [Online]. Verfügbar unter https://aws.amazon.com/de/q/ (Abgerufen am 20 Dezember 2023).

Amazon Web Services (2023b) *Generative AI Powered Assistant: Amazon Q Pricing* [Online]. Verfügbar unter https://aws.amazon.com/de/q/pricing/ (Abgerufen am 6 Dezember 2023).

Amazon Web Services (2023c) *Generative KI-Anwendungen mit Basismodellen erstellen: Amazon Bedrock* [Online]. Verfügbar unter https://aws.amazon.com/de/bedrock/?refid=095af90b-a461-4fb7-85fd-e43f9910d45b (Abgerufen am 6 Dezember 2023).

Amazon Web Services (2023d) *Amazon Bedrock Pricing: Build Generative AI Applications with Foundation Models* [Online]. Verfügbar unter https://aws.amazon.com/de/bedrock/pricing/ (Abgerufen am 15 Dezember 2023).

Anthropic (2023a) *Introducing Claude Pro* [Online]. Verfügbar unter https://www.anthropic.com/index/claude-pro (Abgerufen am 6 Dezember 2023).

Anthropic (2023b) „Model Card and Evaluations for Claude Models“.

Anthropic (2023c) *Model Pricing* [Online]. Verfügbar unter https://anthropic.com/pricing (Abgerufen am 9 Oktober 2023).

Bain, J. S. (1956) *Barriers to New Competition: Their Character and Consequences in Manufacturing Industries*, Cambridge, MA and London, England, Harvard University Press.

Bartholomew, J. & Mehta, D. (2023) *How the media is covering ChatGPT* [Online], Columbia Journalism Review. Verfügbar unter https://www.cjr.org/tow_center/media-coverage-chatgpt.php (Abgerufen am 7 Dezember 2023).

Bass, D. (2023) „Microsoft to Invest $10 Billion in ChatGPT Maker OpenAI (MSFT)“, *Bloomberg*, 23. Januar [Online]. Verfügbar unter https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai?leadSource=uverify%20wall#xj4y7vzkg (Abgerufen am 26 November 2023).

Baumol, W. J. (1982) „Contestable Markets: An Uprising in the Theory of Industry Structure“, *The American Economic Review*, Vol. 72, No. 1, S. 1–15 [Online]. Verfügbar unter https://d1wqtxts1xzle7.cloudfront.net/84002332/Contestable_Markets_An_Uprising_in_the_T20220412-2436-sgn59i.pdf?1649800500=&response-content-disposition=inline%3B+filename%3DContestable_Markets_An_Uprising_In_the_T.pdf&Expires=1699104381&Signature=TWy0~K2jnedb1r65jYIOx~DXy17XcsC9St4j1snl-FAkxkqFOcprRXOoZTmJpa9MOVHqwTmQ-53M-

rZBjsYGHPdPDmMa6iKflX~P4N5F7Y4aeF-vkfl1fXoCeSCgzc2xSjPJv20R~oKzXi3PsHEHM6FcKVSbsK5rArHWIdS-FoT2mf1uJtXLxTnRgNgjCxAASE0l0fFpGF9MD8WPJxAVQ4B1hAIf2TAG-MFkzMKKpvYcbUHOz2Us-GEO7Uw~jarDS6Os5n06Kthi4ouC1Ex~vIid7Ph8BzXRI0m9i-Ur2JjvMCula1liKXc7cttd-gAvJAJ8O-IiS88xOWA9TArCqYdFvqmg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA (Abgerufen am 4 November 2023).

Bester, H. (2016) *Theorie der Industrieökonomik* [Online], 7. Aufl., Berlin, Heidelberg, Springer Berlin Heidelberg. Verfügbar unter http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1587171.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y. & Ramesh, A. (2023) *Improving Image Generation with Better Captions* [Online]. Verfügbar unter https://cdn.openai.com/papers/dall-e-3.pdf (Abgerufen am 18 Dezember 2023).

Block, S. (2023) *Large-Scale Agile Frameworks: Agile Frameworks, agile Infrastruktur und pragmatische Lösungen zur digitalen Transformation*, Berlin, Heidelberg, Springer Berlin Heidelberg; Springer Vieweg.

Blum, U., Müller, S. & Weiske, A. (2006) *Angewandte Industrieökonomik: Theorien - Modelle - Anwendungen*, Wiesbaden, Betriebswirtschaftlicher Verlag Dr. Th. Gabler | GWV Fachverlage GmbH Wiesbaden.

Blut, M. (2008) *Der Einfluss von Wechselkosten auf die Kundenbindung: Verhaltenstheoretische Fundierung und empirische Analyse*, Wiesbaden, Betriebswirtschaftlicher Verlag Dr. Thomas Gabler / GWV Fachverlage GmbH Wiesbaden.

Briggs, J. & Kodnani, D. (2023) *The Potentially Large Effects of Artificial Intelligence on Economic Growth* [Online], Goldman Sachs. Verfügbar unter https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html (Abgerufen am 27 Dezember 2023).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020) *Language Models are Few-Shot Learners* [Online]. Verfügbar unter http://arxiv.org/pdf/2005.14165v4.

Buchholz, K. (2023) *One Million Users: Threads Shoots Past One Million User Mark at Lightning Speed* [Online], Statista. Verfügbar unter https://www.statista.com/chart/29174/time-to-one-million-users/ (Abgerufen am 7 Dezember 2023).

Bühler, S. & Jaeger, F. (2002) *Einführung in die Industrieökonomik*, Berlin, Heidelberg, s.l., Springer Berlin Heidelberg.

Bundesministerium der Justiz (2023) *Gesetz gegen Wettbewerbsbeschränkungen* (GWB).

Burnham, T. A., Frels, J. K. & Mahajan, V. (2003) „Consumer Switching Costs: A Typology, Antecedents, and Consequences“, *Journal of the Academy of Marketing Science*, Vol. 31, No. 2, S. 109–126.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M.,

Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I. & Zaremba, W. (2021) *Evaluating Large Language Models Trained on Code* [Online]. Verfügbar unter http://arxiv.org/pdf/2107.03374v2 (Abgerufen am 14 Dezember 2023).

Cisek, G. (2021) *Machtwechsel der Intelligenzen: Wie sich unser Miteinander durch künstliche Intelligenz verändert*, Wiesbaden, Heidelberg, Springer Vieweg.

Clement, R., Schreiber, D., Bossauer, P. & Pakusch, C. (2019) *Internet-Ökonomie: Grundlagen und Fallbeispiele der digitalen und vernetzten Wirtschaft*, Berlin, Heidelberg, Springer Gabler.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. & Schulman, J. (2021) *Training Verifiers to Solve Math Word Problems* [Online]. Verfügbar unter http://arxiv.org/pdf/2110.14168v2 (Abgerufen am 28 Oktober 2023).

Dai, X., Hou, J., Ma, C.-Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., Yu, M., Kadian, A., Radenovic, F., Mahajan, D., Li, K., Zhao, Y., Petrovic, V., Singh, M. K., Motwani, S., Wen, Y., Song, Y., Sumbaly, R., Ramanathan, V., He, Z., Vajda, P. & Parikh, D. (2023) *Emu: Enhancing Image Generation Models Using Photogenic Needles in a Haystack* [Online], GenAI, M. Verfügbar unter https://scontent-fra3-2.xx.fbcdn.net/v/t39.2365-6/10000000_314278304809963_6123613192263580501_n.pdf?_nc_cat=104&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=KsPsP6k-kD0AX8a6e7N&_nc_ht=scontent-fra3-2.xx&oh=00_AfCyMqkYtezZUJkki-pawElDRzv-ZpsboTwW5lqd73zmCg&oe=656D7C98 (Abgerufen am 30 November 2023).

Dewenter, R. & Rösch, J. (2015) *Einführung in die neue Ökonomie der Medienmärkte: Eine wettbewerbsökonomische Betrachtung aus Sicht der Theorie der zweiseitigen Märkte*, Wiesbaden, Springer Gabler.

Eloundou, T., Manning, S., Mishkin, P. & Rock, D. (2023) *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* [Online]. Verfügbar unter https://arxiv.org/pdf/2303.10130.pdf (Abgerufen am 27 Dezember 2023).

Ertel, W. (2021) *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*, 5. Aufl., Wiesbaden, Heidelberg, Springer Vieweg.

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F. & Jiang, L. (2023) *LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models* [Online]. Verfügbar unter http://arxiv.org/pdf/2309.14393v1.

Fichert, F. (1998) „Das Microsoft-Monopol: Herausforderung für die Wettbewerbspolitik“, *Wirtschaftsdienst - Zeitschrift für Wirtschaftspolitik*, Vol. 78, No. 6, S. 343–347 [Online]. Verfügbar unter https://www.wirtschaftsdienst.eu/pdf-download/jahr/1998/heft/6/beitrag/das-microsoft-monopol-herausforderung-fuer-die-wettbewerbspolitik.html (Abgerufen am 17 November 2023).

Frank, R., Strugholtz, S. & Meise, F. (2021) *Bausteine der digitalen Transformation: Wie APIs Unternehmen den Weg in die Programmable Economy ebnen*, Wiesbaden, Springer Fachmedien Wiesbaden; Springer Gabler.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y. & Pearson, A. T. (2022) *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers*.

Gemini Team (2023) *Gemini: A Family of Highly Capable Multimodal Models* [Online], Google and DeepMind. Verfügbar unter https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf (Abgerufen am 13 Dezember 2023).

Google (2023a) *Bard Privacy Hub* [Online]. Verfügbar unter https://support.google.com/bard/answer/13594961?visit_id=638367609011653217-756597615&p=privacy_help&rd=1#your_data (Abgerufen am 11 Dezember 2023).

Google (2023b) *Gemini API Pricing* [Online]. Verfügbar unter https://ai.google.dev/pricing?hl=en (Abgerufen am 15 Dezember 2023).

Google (2023c) *Gemini-Modelle* [Online]. Verfügbar unter https://ai.google.dev/models/gemini?hl=de (Abgerufen am 16 Dezember 2023).

Google DeepMind (2023) *Welcome to the Gemini era* [Online]. Verfügbar unter https://deepmind.google/technologies/gemini/?_gl=1*31sllk*_up*MQ.*_ga*Njc2ODYyMTA4LjE3MDI1NDI1NjI.*_ga_LS8HVHCNQ0*MTcwMjU0MjU2MS4xLjAuMTcwMjU0MzM3Ni4wLjAuMA.#introduction (Abgerufen am 14 Dezember 2023).

Gronwald, K.-D. (2023) *Globale Kommunikation und Kollaboration: Globale Supply Chain Netzwerk-Integration, interkulturelle Kompetenzen, Arbeit und Kommunikation in virtuellen Teams* [Online], Wiesbaden, Heidelberg, Springer Vieweg. Verfügbar unter http://www.springer.com/.

Halberstadt, J. (2014) *Globale und nationale First-Mover-Vorteile internetbasierter Geschäftsmodelle*, Wiesbaden, Springer Gabler.

Haun, M. (2014) *Cognitive Computing: Steigerung des systemischen Intelligenzprofils*, 2014. Aufl., Berlin, Heidelberg, Springer Berlin Heidelberg.

Hawk, J. (2023) *Come build with us: Microsoft and OpenAI partnership unveils new AI opportunities* [Online]. Verfügbar unter https://www.microsoft.com/en-us/microsoft-cloud/blog/2023/11/07/come-build-with-us-microsoft-and-openai-partnership-unveils-new-ai-opportunities/ (Abgerufen am 13 Dezember 2023).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. & Steinhardt, J. (2020) *Measuring Massive Multitask Language Understanding* [Online]. Verfügbar unter http://arxiv.org/pdf/2009.03300v3 (Abgerufen am 27 Oktober 2023).

Hillebrandt, F. (2023) *ChatGPT-Statistiken 2023: Spannende Zahlen, Daten & Fakten* [Online]. Verfügbar unter https://www.blogmojo.de/chatgpt-statistiken/ (Abgerufen am 5 August 2023).

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O. & Sifre, L. (2022) *Training Compute-Optimal Large Language Models* [Online]. Verfügbar unter http://arxiv.org/pdf/2203.15556v1 (Abgerufen am 29 Dezember 2023).

Hoisl, B. (2018) *Produkte Digital-First Denken: Wie Unternehmen Software-Basierte Produktinnovation Erfolgreich Gestalten* [Online], Wiesbaden, Gabler. Verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5568412.

Holtermann, F. & Scheuer, S. (2023) *Microsoft geht mit KI-Assistent Copilot an den Start* [Online]. Verfügbar unter https://www.handelsblatt.com/technik/it-internet/kuenstliche-intelligenz-microsoft-geht-mit-ki-assistent-copilot-an-den-start/29405612.html (Abgerufen am 19 Dezember 2023).

Holtforth, D. G. (2016) *Schlüsselfaktoren im E-Commerce: Innovationen, skaleneffekte, daten und kundenzentrierung*, Wiesbaden, Germany, Springer Gabler.

Horvitz, E. & Nori, H. (2023) *Steering at the Frontier: Extending the Power of Prompting* [Online]. Verfügbar unter https://www.microsoft.com/en-us/research/blog/steering-at-the-frontier-extending-the-power-of-prompting/ (Abgerufen am 17 Dezember 2023).

Hsiao, S. (2023a) *Assistant with Bard: A step toward a more personal assistant* [Online]. Verfügbar unter https://blog.google/products/assistant/google-assistant-bard-generative-ai/ (Abgerufen am 6 Dezember 2023).

Hsiao, S. (2023b) *Google I/O '23: Bard wird visueller, stärker integriert und breiter verfügbar* [Online]. Verfügbar unter https://blog.google/intl/de-de/unternehmen/technologie/google-io-23-bard-update-verfuegbarkeit-ki/ (Abgerufen am 7 Dezember 2023).

Hsiao, S. (2023c) *Bard gets its biggest upgrade yet with Gemini* [Online]. Verfügbar unter https://blog.google/products/bard/google-bard-try-gemini-ai/ (Abgerufen am 18 Dezember 2023).

Hu, K. (2023) „ChatGPT sets record for fastest-growing user base - analyst note", *Reuters Media*, 2. Februar [Online]. Verfügbar unter https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (Abgerufen am 12 Juli 2023).

Hüsch, A., Distelrath, D. & Hüsch, T. (2023) *Einsatzmöglichkeiten von GPT in Finance, Compliance und Auditing: Vorteile, Herausforderungen, Praxisbeispiele* [Online], Wiesbaden, Springer Fachmedien Wiesbaden GmbH. Verfügbar unter https://link.springer.com/978-3-658-41418-4.

J., J. (2023) *How can I access GPT-4? | OpenAI Help Center* [Online]. Verfügbar unter https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4#h_ef2559d888 (Abgerufen am 11 Dezember 2023).

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. (1991) „Adaptive Mixture of Local Experts", in *Neural Computation*, S. 79–87.

Jahn, T. (2023) *ChatGPT: Was Sie über OpenAIs KI wissen sollten* [Online]. Verfügbar unter https://www.handelsblatt.com/technik/it-internet/chatgpt-was-sie-ueber-openais-ki-wissen-sollten-/28941524.html.

Janiesch, C., Zschech, P. & Heinrich, K. (2021) „Machine learning and deep learning", *Electronic Markets - The International Journal on Networked Business*, No. 31, S. 685–695 [Online]. Verfügbar unter https://link.springer.com/content/pdf/10.1007/s12525-021-00475-2.pdf?pdf=button (Abgerufen am 22.08.23).

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J. & Ingrisch, M. (2023) „ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports", *European radiology*.

Joshi, A. V. (2020) *Machine Learning and Artificial Intelligence* [Online], Cham, Springer International Publishing AG. Verfügbar unter https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=5904620.

Kemp, S. (2023) *Digital 2023 October Global Statshot Report* [Online]. Verfügbar unter https://datareportal.com/reports/digital-2023-october-global-statshot (Abgerufen am 12 Dezember 2023).

Khan, R. A., Jawaid, M., Khan, A. R. & Sajjad, M. (2023) „ChatGPT - Reshaping medical education and clinical management", *Pakistan journal of medical sciences*, Vol. 39, No. 2, S. 605–607.

King, M. R. (2023) „The Future of AI in Medicine: A Perspective from a Chatbot", *Ann Biomed Eng*, No. 2, S. 291–295 [Online]. DOI: 10.1007/s10439-022-03121-w.

Klöckner, J. (2023) „ChatGPT & Co.: Wo Unternehmen jetzt Künstliche Intelligenz einsetzen wollen", *Handelsblatt*, 31. Juli [Online]. Verfügbar unter https://www.handelsblatt.com/politik/deutschland/chatgpt-und-co-wo-unternehmen-jetzt-kuenstliche-intelligenz-einsetzen-wollen/29279706.html (Abgerufen am 27 Dezember 2023).

Klüver, C., Klüver, J. & Schmidt, J. (2021) *Modellierung komplexer Prozesse durch naturanaloge Verfahren: Künstliche Intelligenz und Künstliches Leben*, 3. Aufl., Wiesbaden, Heidelberg, Springer Vieweg.

Knieps, G. (2008) *Wettbewerbsökonomie: Regulierungstheorie, Industrieökonomie, Wettbewerbspolitik*, 3. Aufl., Berlin, Heidelberg, Springer.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. (2022) *Large Language Models are Zero-Shot Reasoners* [Online]. Verfügbar unter http://arxiv.org/pdf/2205.11916v4 (Abgerufen am 17 August 2023).

Krauss, P. (2023) *Künstliche Intelligenz und Hirnforschung: Neuronale Netze, Deep Learning und die Zukunft der Kognition*, Berlin, Heidelberg, Springer Berlin Heidelberg; Imprint Springer.

Krawczyk, J. (2023) *Bard's latest update: More features, languages and countries* [Online]. Verfügbar unter https://blog.google/products/bard/google-bard-new-features-update-july-2023/ (Abgerufen am 18 Dezember 2023).

Krivec, R. (2023) *Google Bard Statistics: How Many People Use Google Bard?* [Online]. Verfügbar unter https://colorlib.com/wp/google-bard-statistics/ (Abgerufen am 5 August 2023).

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. de, Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. & Tseng, V. (2023) „Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models", *PLOS digital health*, Vol. 2, No. 2, e0000198.

Lanquillon, C. & Schacht, S. (2023) *Knowledge Science – Grundlagen: Methoden der Künstlichen Intelligenz für die Wissensextraktion aus Texten*, Wiesbaden, Springer Fachmedien Wiesbaden; Imprint Springer Vieweg.

Letzner, V. (1997) *Integrationstheorie und monopolistische Konkurrenz*, Wiesbaden, Deutscher Universitätsverlag; Imprint.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y. & Koreeda, Y. (2023) „Holistic Evaluation of Language Models", *Published in Transactions on Machine Learning Research (TMLR)* [Online]. Verfügbar unter http://arxiv.org/pdf/2211.09110v2 (Abgerufen am 1 Oktober 2023).

Lipczynski, J., Wilson, J. O. S. & Goddard, J. A. (2013) *Industrial organization: Competition, strategy, policy*, Harlow, Pearson.

Liu, J., Liu, C., Zhou, P., Ye, Q., Chong, D., Zhou, K., Xie, Y., Cao, Y., Wang, S., You, C. & Yu, P. S. (2023) *LLMRec: Benchmarking Large Language Models on Recommendation Task* [Online]. Verfügbar unter http://arxiv.org/pdf/2308.12241v1 (Abgerufen am 7 Oktober 2023).

Mackmood, J., Marty, D. B. & D'Onofrio, S. (2020) „Chatbot & Cognitive Services: Ein Schritt Richtung Automatisierung im User Help Desk der Schweizerischen Post", in Portmann, E. & D'Onofrio, S. (Hg.) *Cognitive Computing: Theorie, Technik und Praxis,* Wiesbaden, Heidelberg, Springer Vieweg, S. 147–168.

Mehdi, Y. (2023) *Announcing Microsoft Copilot, your everyday AI companion - The Official Microsoft Blog* [Online]. Verfügbar unter https://blogs.microsoft.com/blog/2023/09/21/announcing-microsoft-copilot-your-everyday-ai-companion/ (Abgerufen am 5 Dezember 2023).

Mercedes Benz (2023) *Prozesse und Fehler in Echtzeit analysieren: Mercedes-Benz testet ChatGPT in der intelligenten Produktion* [Online]. Verfügbar unter https://

group.mercedes-benz.com/innovation/digitalisierung/industrie-4-0/chatgpt-in-der-produktion.html (Abgerufen am 30 September 2023).

Meta (2023a) *Introducing Code Llama, an AI Tool for Coding* [Online]. Verfügbar unter https://about.fb.com/news/2023/08/code-llama-ai-for-coding/ (Abgerufen am 5 Dezember 2023).

Meta (2023b) *Introducing Llama 2: The next generation of our open source large language model* [Online]. Verfügbar unter https://ai.meta.com/llama/ (Abgerufen am 6 Oktober 2023).

Meta (2023c) *Introducing New AI Experiences Across Our Family of Apps and Devices* [Online]. Verfügbar unter https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/ (Abgerufen am 5 Dezember 2023).

Meta (2023d) *Introducing new AI experiences from Meta* [Online]. Verfügbar unter https://ai.meta.com/genai/ (Abgerufen am 18 Dezember 2023).

Meta (2023e) *Meta Connect 2023: Quest 3, Fortschritte bei KI, Smart Glasses der nächsten Generation & der Weg zum Metaversum* [Online]. Verfügbar unter https://about.fb.com/de/news/2023/09/meta-connect-2023/ (Abgerufen am 18 Dezember 2023).

Microsoft (2023a) *Microsoft Copilot for Microsoft 365: Thousands of skills. All your data. Infinite possibilities.* [Online]. Verfügbar unter https://www.microsoft.com/en-us/microsoft-365/enterprise/copilot-for-microsoft-365 (Abgerufen am 19 Dezember 2023).

Microsoft (2023b) *Microsoft Edge kennenlernen* [Online]. Verfügbar unter https://www.microsoft.com/de-DE/edge?ep=83&form=MA13KJ&es=24 (Abgerufen am 23 Dezember 2023).

Microsoft Bing Team (2023) *Our vision to bring Microsoft Copilot to everyone, and more* [Online]. Verfügbar unter https://blogs.bing.com/search/november-2023/our-vision-to-bring-microsoft-copilot-to-everyone-and-more (Abgerufen am 5 Dezember 2023).

Microsoft Corporate Blogs (2023) *Microsoft and OpenAI extend partnership* [Online]. Verfügbar unter https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/ (Abgerufen am 5 Dezember 2023).

Miller, G. A., Newman, E. B. & Friedman, E. A. (1958) „Length-frequency statistics for written English“, *Information and Control*, Vol. 1, No. 4, S. 370–389 [Online]. DOI: 10.1016/S0019-9958(58)90229-8 (Abgerufen am 14 Dezember 2023).

Mok, A. (2023) *ChatGPT could cost over $700,000 per day to operate. Microsoft is reportedly trying to make it cheaper.* [Online], Business Insider. Verfügbar unter https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4 (Abgerufen am 31 Dezember 2023).

Monopolkommission (2022) *Wettbewerb 2022: XXIV. Hauptgutachten* [Online]. Verfügbar unter https://www.monopolkommission.de/images/HG24/HGXXIV_Gesamt.pdf (Abgerufen am 5 November 2023).

Montgomery, J. (2023) *Microsoft and Meta expand their AI partnership with Llama 2 on Azure and Windows* [Online]. Verfügbar unter https://blogs.microsoft.com/blog/2023/07/18/microsoft-and-meta-expand-their-ai-partnership-with-llama-2-on-azure-and-windows/ (Abgerufen am 5 Dezember 2023).

Moore, O. (2023) *How Are Consumers Using Generative AI?* [Online]. Verfügbar unter https://a16z.com/how-are-consumers-using-generative-ai/ (Abgerufen am 24 November 2023).

Moring, A. (2023) *Künstliche Intelligenz und Intuition: Robuste und nachhaltige Entscheidungen in digitalen Arbeitswelten* [Online], Wiesbaden, Springer Fachmedien Wiesbaden GmbH. Verfügbar unter https://link.springer.com/978-3-658-42017-8.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N. & Mian, A. (2023) *A Comprehensive Overview of Large Language Models* [Online]. Verfügbar unter http://arxiv.org/pdf/2307.06435v4 (Abgerufen am 26 Oktober 2023).

Noy, S. & Zhang, W. (2023) *Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence,* MIT [Online]. Verfügbar unter https://economics.mit.edu/sites/default/files/inline-files/Noy_Zhang_1.pdf (Abgerufen am 27 Dezember 2023).

OpenAI (2019) *OpenAI LP* [Online]. Verfügbar unter https://openai.com/blog/openai-lp (Abgerufen am 26 November 2023).

OpenAI (2023a) *ChatGPT can now see, hear, and speak* [Online]. Verfügbar unter https://openai.com/blog/chatgpt-can-now-see-hear-and-speak (Abgerufen am 25 November 2023).

OpenAI (2023b) *ChatGPT Enterprise: Enterprise-grade security & privacy and the most powerful version of ChatGPT yet* [Online]. Verfügbar unter https://openai.com/enterprise (Abgerufen am 19 Dezember 2023).

OpenAI (2023c) *ChatGPT plugins* [Online]. Verfügbar unter https://openai.com/blog/chatgpt-plugins (Abgerufen am 25 November 2023).

OpenAI (2023d) *ChatGPT plugins waitlist* [Online]. Verfügbar unter https://openai.com/waitlist/plugins (Abgerufen am 12 Dezember 2023).

OpenAI (2023e) *DALL·E 3 is now available in ChatGPT Plus and Enterprise* [Online]. Verfügbar unter https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise (Abgerufen am 25 November 2023).

OpenAI (2023f) *Introducing ChatGPT Enterprise* [Online]. Verfügbar unter https://openai.com/blog/introducing-chatgpt-enterprise (Abgerufen am 25 November 2023).

OpenAI (2023g) *Introducing GPTs* [Online]. Verfügbar unter https://openai.com/blog/introducing-gpts (Abgerufen am 25 November 2023).

OpenAI (2023h) *Introducing the ChatGPT app for iOS* [Online]. Verfügbar unter https://openai.com/blog/introducing-the-chatgpt-app-for-ios (Abgerufen am 25 November 2023).

OpenAI (2023i) *New models and developer products announced at DevDay* [Online]. Verfügbar unter https://openai.com/blog/new-models-and-developer-products-announced-at-devday (Abgerufen am 15 Dezember 2023).

OpenAI (2023j) *OpenAI Data Partnerships* [Online]. Verfügbar unter https://openai.com/blog/data-partnerships (Abgerufen am 22 Dezember 2023).

OpenAI (2023k) *OpenAI's first developer conference* [Online]. Verfügbar unter https://devday.openai.com/ (Abgerufen am 13 Dezember 2023).

OpenAI (2023l) *Pricing: Simple and flexible. Only pay for what you use.* [Online]. Verfügbar unter https://openai.com/pricing#language-models (Abgerufen am 26 November 2023).

OpenAI (2023m) *GPT-4 Technical Report* [Online]. Verfügbar unter http://arxiv.org/pdf/2303.08774v3 (Abgerufen am 5 Oktober 2023).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. (2022) *Training language models to follow instructions with human feedback* [Online]. Verfügbar unter http://arxiv.org/pdf/2203.02155v1 (Abgerufen am 25 Oktober 2023).

Paaß, G. & Hecker, D. (2020) *Künstliche Intelligenz: Was steckt hinter der Technologie der Zukunft?*, Wiesbaden, Heidelberg, Springer Vieweg.

Pappu, A. (2023) *Introducing Duet AI for Google Workspace* [Online]. Verfügbar unter https://workspace.google.com/blog/product-announcements/duet-ai?hl=en (Abgerufen am 6 Dezember 2023).

Patel, D. (2023) *The AI Brick Wall: A Practical Limit For Scaling Dense Transformer Models, and How GPT 4 Will Break Past It* [Online]. Verfügbar unter https://www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit (Abgerufen am 30 November 2023).

Patel, D. & Ahmad, A. (2023) *The Inference Cost Of Search Disruption: Large Language Model Cost Analysis* [Online]. Verfügbar unter https://www.semianalysis.com/p/the-inference-cost-of-search-disruption?utm_source=%2Fsearch%2Fchatgpt&utm_medium=reader2 (Abgerufen am 14 Oktober 2023).

Patel, D. & Wong, G. (2023) *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE* [Online]. Verfügbar unter https://www.semianalysis.com/p/gpt-4-architecture-infrastructure (Abgerufen am 31 Dezember 2023).

Pepall, L., Richards, D. & Norman, G. (2014) *Industrial organization: Contemporary theory and empirical applications*, 5. Aufl., Hoboken, NJ, John Wiley & Sons, Inc.

Peters, R. (2010) *Internet-Ökonomie*, Heidelberg, Springer.

Pfähler, W. & Wiese, H. (2008) *Unternehmensstrategien im Wettbewerb: Eine spieltheoretische Analyse*, 3. Aufl., Berlin, Heidelberg, Springer.

Pichai, S. & Hassabis, D. (2023) *Gemini – unser größtes und leistungsfähigstes KI-Modell: KI für alle zugänglich machen* [Online]. Verfügbar unter https://blog.google/intl/de-de/unternehmen/technologie/gemini/ (Abgerufen am 19 Dezember 2023).

Pindyck, R. S., Rubinfeld, D. L. & Schittko, U. K. (1998) *Mikroökonomie*, 4. Aufl., Berlin, Boston, Oldenbourg Wissenschaftsverlag.

Pinsky, Y. (2023) *Bard can now connect to your Google apps and services* [Online]. Verfügbar unter https://blog.google/products/bard/google-bard-new-features-update-sept-2023/ (Abgerufen am 6 Dezember 2023).

Premer, M. (2015) *Grundzüge der Volkswirtschaftslehre: Makroökonomik und Mikroökonomik*, 2. Aufl., Berlin, De Gruyter Oldenbourg.

Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. (2018) *Improving Language Understanding by Generative Pre-Training* [Online]. Verfügbar unter https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (Abgerufen am 25 Oktober 2023).

Rausch, D. (2023) *Previewing the future of Alexa* [Online]. Verfügbar unter https://www.aboutamazon.com/news/devices/amazon-alexa-generative-ai (Abgerufen am 6 Dezember 2023).

Rebala, G., Ravi, A. & Churiwala, S. (2019) *An introduction to machine learning*, Cham, Springer.

Reuter, W., Korotkova, A. & Nguyen, L. (2023) „Strukturierte und unstrukturierte Daten“, in Gillhuber, A., Kauermann, G. & Hauner, W. (Hg.) *Künstliche Intelligenz und Data Science in Theorie und Praxis: Von Algorithmen und Methoden zur praktischen Umsetzung in Unternehmen* [Online], Berlin, Heidelberg, Springer Spektrum, S. 51–68. Verfügbar unter http://www.springer.com/.

Roberts, A. & Raffel, C. (2020) *Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer – Google Research Blog* [Online]. Verfügbar unter https://blog.research.google/2020/02/exploring-transfer-learning-with-t5.html (Abgerufen am 1 Januar 2024).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J. &

Norouzi, M. (2022) *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding* [Online]. Verfügbar unter http://arxiv.org/pdf/2205.11487v1 (Abgerufen am 1 Januar 2024).

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D. & Gadepally, V. (2023) *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference* [Online]. Verfügbar unter http://arxiv.org/pdf/2310.03003v1.

Sarrion, E. (2023) *Exploring the Power of ChatGPT: Applications, Techniques, and Implications*, Berkeley, CA, Apress; Imprint Apress.

Sauer, A. (2019) *eSport, Netzwerkeffekte und Lindahl-Preise*, Wiesbaden, Springer Fachmedien Wiesbaden.

Schade, M. (2023) *How your data is used to improve model performance: Learn more about OpenAI's data usage policies for our API, ChatGPT and DALL-E* [Online], OpenAI. Verfügbar unter https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance (Abgerufen am 11 Dezember 2023).

Schaul, K., Chen, S. Y. & Tiku, N. (2023) „Inside the secret list of websites that make AI like ChatGPT sound smart“, *The Washington Post*, 19. April [Online]. Verfügbar unter https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?nid=top_pb_signin&arcId=&account_location=ONSITE_HEADER_ARTICLE&itid=nav_sign_in (Abgerufen am 25 Oktober 2023).

Schimroszik, N. (2023) *ChatGPT: Das sind die zehn hilfreichsten Plug-ins* [Online]. Verfügbar unter https://www.handelsblatt.com/technik/it-internet/chatgpt-das-sind-die-zehn-hilfreichsten-plug-ins/29170984.html (Abgerufen am 12 Dezember 2023).

Schmidt, I. & Haucap, J. (2013) *Wettbewerbspolitik und Kartellrecht: Eine interdisziplinäre Einführung*, 10. Aufl., München, Oldenbourg.

Schmidt, S. (2007) *Das Online-Erfolgsmodell digitaler Produkte: Strategische Wirkungspotenziale und operative Handlungsoptionen*, Wiesbaden, Deutscher UniversittÌƒs-Verlag GWV Fachverlage GmbH, Wiesbaden.

Schönert, U. (2023) „ChatGPT: So funktioniert ChatGPT“, *Die Zeit*, 25. März [Online]. Verfügbar unter https://www.zeit.de/2023/13/chatgpt-kuenstliche-intelligenz-funktion-token (Abgerufen am 13 Dezember 2023).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017) *Proximal Policy Optimization Algorithms* [Online]. Verfügbar unter http://arxiv.org/pdf/1707.06347v2 (Abgerufen am 5 September 2023).

Sedaghat, S. (2023) „Early applications of ChatGPT in medical practice, education and research“, *Clinical medicine (London, England)*, Vol. 23, No. 3, S. 278–279.

Seufert, E. B. (2014) *Freemium economics: Leveraging analytics and user segmentation to drive revenue*, Waltham, Massachusetts, MK.

Similiarweb (2023a) *Chatbot Angebot 11-15* [Online]. Verfügbar unter https://pro.similarweb.com/#/digitalsuite/websiteanalysis/overview/website-performance/*/999/3m?webSource=Total&key=copy.ai,deepai.org,jasper.ai,neuroflash.com,contentbot.ai.

Similiarweb (2023b) *Chatbot Angebot 1-5* [Online]. Verfügbar unter webSource=Total&key=chat.openai.com,bard.google.com,character.ai,bing.com,claude.ai (Abgerufen am 6 November 2023).

Similiarweb (2023c) *Chatbot Angebot 6-10* [Online]. Verfügbar unter https://pro.similarweb.com/#/digitalsuite/websiteanalysis/overview/website-performance/*/999/3m?webSource=Total&key=perplexity.ai,claude.ai,you.com,forefront.ai,app.writesonic.com (Abgerufen am 6 November 2023).

Singhal, P., Goyal, T., Xu, J. & Durrett, G. (2023) *A Long Way to Go: Investigating Length Correlations in RLHF* [Online]. Verfügbar unter http://arxiv.org/pdf/2310.03716v1 (Abgerufen am 8 Oktober 2023).

Spataro, J. (2023) *Announcing Microsoft 365 Copilot general availability and Microsoft 365 Chat* [Online]. Verfügbar unter https://www.microsoft.com/en-us/microsoft-365/blog/2023/09/21/announcing-microsoft-365-copilot-general-availability-and-microsoft-365-chat/ (Abgerufen am 5 Dezember 2023).

Srinivasan, R. (2023) *Plattform-Geschäftsmodelle: Rahmenwerke, Konzepte und Design*, Singapore, Springer Nature Singapore; Springer Gabler.

Stallbaumer, C. (2023) *Introducing Microsoft Copilot Studio and new features in Copilot for Microsoft 365* [Online]. Verfügbar unter https://www.microsoft.com/en-us/microsoft-365/blog/2023/11/15/introducing-microsoft-copilot-studio-and-new-features-in-copilot-for-microsoft-365/ (Abgerufen am 5 Dezember 2023).

StatCounter Global Stats (2023a) *Browser Market Share Worldwide* [Online]. Verfügbar unter https://gs.statcounter.com/browser-market-share (Abgerufen am 25 Dezember 2023).

StatCounter Global Stats (2023b) *Desktop Operating System Market Share Worldwide* [Online]. Verfügbar unter https://gs.statcounter.com/os-market-share/desktop/worldwide/#monthly-202211-202311 (Abgerufen am 25 Dezember 2023).

StatCounter Global Stats (2023c) *Desktop Windows Version Market Share Worldwide* [Online]. Verfügbar unter https://gs.statcounter.com/os-version-market-share/windows/desktop/worldwide (Abgerufen am 25 Dezember 2023).

Susnjak, T. (2022) *ChatGPT: The End of Online Exam Integrity?,* School of Mathematical and Computational Sciences [Online]. Verfügbar unter http://arxiv.org/pdf/2212.09292v1.

Synergy Research Group (2023) *AI Helps to Stabilize Quarterly Cloud Market Growth Rate; Microsoft Market Share Nudges Up Again* [Online]. Verfügbar unter https://www.srgresearch.com/articles/ai-helps-to-stabilize-quarterly-cloud-market-growth-rate-microsoft-market-share-nudges-up-again (Abgerufen am 20 Dezember 2023).

Taulli, T. (2023) *Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business*, Berkeley, CA, Apress.

Tirole, J. (1999) *Industrieökonomik*, 2. Aufl., München, Wien, Oldenbourg.

Tremblay, V. J. & Tremblay, C. H. (2012) *New perspectives on industrial organization: With contributions from behavioral economics and game theory*, New York, NY, Heidelberg, Springer.

Tunstall, L., van Werra, L., Wolf, T. & Géron, A. (2023) *Natural Language Processing mit Transformern: Sprachanwendungen mit Hugging Face erstellen*, Heidelberg, dpunkt.

Turing, A. M. (1950) „Computing machinery and intelligence“, *Mind a quarterly review of psychology and philosophy*, No. 16, S. 433–460 [Online]. Verfügbar unter https://academic.oup.com/mind/article-pdf/LIX/236/433/30123314/lix-236-433.pdf (Abgerufen am 21 August 2023).

Vahdat, A. & Yang, J. (2023) *Vertex AI Next 2023 announcements* [Online]. Verfügbar unter https://cloud.google.com/blog/products/ai-machine-learning/vertex-ai-next-2023-announcements?hl=en (Abgerufen am 6 Dezember 2023).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017) *Attention Is All You Need* [Online]. Verfügbar unter http://arxiv.org/pdf/1706.03762v7 (Abgerufen am 1 Oktober 2023).

Waldman, D. E. & Jensen, E. J. (2013) *Industrial Organization: Theory amd Practice*, 4. Aufl., Harlow, Pearson Education.

Weimann, J. (2006) *Wirtschaftspolitik: Allokation und kollektive Entscheidung*, Berlin, Heidelberg, Springer-Verlag Berlin Heidelberg.

Weise, K. (2023) „Amazon Introduces Q, an A.I. Chatbot for Companies“, *The New York Times*, 28. November [Online]. Verfügbar unter https://www.nytimes.com/2023/11/28/technology/amazon-ai-chatbot-q.html (Abgerufen am 12 Juni 2023).

Wied-Nebbeling, S. (2004) *Preistheorie und Industrieökonomik*, Berlin, Heidelberg, s.l., Springer Berlin Heidelberg.

Woeckener, B. (2014) *Strategischer Wettbewerb: Eine Einführung in die Industrieökonomik*, 3. Aufl., Berlin, Heidelberg, Springer Gabler.

Yenduri, G., Murugan, R., Govardanan, C. S., Y, S., Srivastava, G., Maddikunta, P. K. R., Raj G, D., Jhaveri, R. H., B, P., Wang, W., Vasilakos, A. V. & Gadekallu, T. R. (2023) *GPT (Generative Pre-trained Transformer): A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions* [Online]. Verfügbar unter https://arxiv.org/pdf/2305.10435.pdf (Abgerufen am 20 Oktober 2023).

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. & Choi, Y. (2019) *HellaSwag: Can a Machine Really Finish Your Sentence?* [Online]. Verfügbar unter http://arxiv.org/pdf/1905.07830v1 (Abgerufen am 12 September 2023).

Zhai, X. (2022) „ChatGPT User Experience: Implications for Education“ [Online]. Verfügbar unter https://dx.doi.org/10.2139/ssrn.4312418 (Abgerufen am 3 Oktober 2023).

Zhang, D. (2023) *Should ChatGPT and Bard Share Revenue with Their Data Providers? A New Business Model for the AI Era* [Online]. Verfügbar unter http://arxiv.org/pdf/2305.02555v2 (Abgerufen am 27 August 2023).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du Yifan, Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y. & Wen, J.-R. (2023) *A Survey of Large Language Models* [Online]. Verfügbar unter http://arxiv.org/pdf/2303.18223v12 (Abgerufen am 6 Oktober 2023).

Zheng, S., Zhang, Y., Zhu, Y., Xi, C., Gao, P., Zhou, X. & Chang, K. C.-C. (2023) *GPT-Fathom: Benchmarking Large Language Models to Decipher the Evolutionary Path towards GPT-4 and Beyond* [Online]. Verfügbar unter http://arxiv.org/pdf/2309.16583v1 (Abgerufen am 2 Oktober 2023).

Zhou, Z.-H. (2021) *Machine Learning*, Singapore, Springer Singapore; Imprint Springer.

Anhang A: Anzahl der Seitenaufrufe

Anbieter	Jun 23	Jul 23	Aug 23	Sep 23	Okt 23
ChatGPT	1.630.000.000	1.473.000.000	1.426.000.000	1.493.000.000	1.698.000.000
Bard	140.600.000	189.100.000	183.400.000	219.300.000	266.100.000
Character AI	190.100.000	206.000.000	198.500.000	168.800.000	162.200.000
poe.com	70.530.000	86.390.000	59.660.000	55.520.000	60.860.000
quillbot.com	58.790.000	49.570.000	51.820.000	61.800.000	67.800.000
perplexity.ai	22.220.000	24.700.000	27.870.000	37.410.000	47.470.000
Claude.ai	23.896	16.830.000	26.450.000	21.370.000	24.620.000
You.com	16.550.000	14.510.000	13.440.000	13.550.000	13.770.000
forefront.ai	9.526.000	7.662.000	6.348.000	5.074.000	4.349.000
Chatsonic	5.974.000	4.708.000	4.579.000	4.508.000	5.049.000
Copy.ai	5.979.000	5.142.000	4.231.000	3.561.000	3.641.000
DeepAI	3.412.000	3.511.000	3.709.000	4.332.000	6.327.000
Jasper Chat	4.519.000	4.152.000	4.754.000	3.225.000	3.020.000
Neuroflash	536.438	583.101	569.085	598.556	735.646
Contentbot.ai	212.380	175.113	136.381	134.257	134.918

Quellen: Similiarweb (2023b), Similiarweb (2023c), Similiarweb (2023a)

	ChatGPT	Bard
Nov 22	152.700	
Dez 22	266.000.000	
Jan 23	616.000.000	
Feb 23	1.050.000.000	
Mär 23	1.600.000.000	30.600.000
Apr 23	1.800.000.000	49.700.000
Mai 23	1.805.000.000	142.600.000

Quellen: Hillebrandt (2023), Krivec (2023)

Anhang B: Berechnungen

HHI:

Anbieter	Seitenaufrufe zwischen Juni und Oktober	Anteile an den gesamten Seitenaufrufen	Quadrierten Anteile der Seitenaufrufe
ChatGPT	7.720.000.000	72,06%	51,931891%
Bard	998.500.000	9,32%	0,868752%
Character AI	925.600.000	8,64%	0,746528%
poe.com	332.960.000	3,11%	0,096601%
quillbot.com	289.780.000	2,71%	0,073171%
perplexity.ai	159.670.000	1,49%	0,022215%
You.com	89.293.896	0,83%	0,006948%
Claude.ai	71.820.000	0,67%	0,004495%
Chatsonic	32.959.000	0,31%	0,000947%
forefront.ai	24.818.000	0,23%	0,000537%
Copy.ai	22.554.000	0,21%	0,000443%
Jasper Chat	21.291.000	0,20%	0,000395%
DeepAI	19.670.000	0,18%	0,000337%
Neuroflash	3.022.826	0,03%	0,000008%
Contentbot.ai	793.049	0,01%	0,000001%
Σ	10.712.731.771		53,75%

Quelle: Eigene Berechnung auf Basis von: Anhang A und (Waldman & Jensen, 2013, S.105ff.)

$$\boldsymbol{Kleinstmögliocher\ HHI} = \frac{10.000}{15} \approx \mathbf{667}$$

$$\boldsymbol{HHI} = 53{,}75\% * 100 = \mathbf{5.375}$$

Umrechnungsfaktor API Kosten

Durchschnittliche Zeichenanzahl eines englischen Wortes:

$3{,}13 * 0{,}59 + 6{,}47 * 0{,}41 \approx 4{,}5.$

Anzahl an Wörtern aus 1.000 Zeichen:

$$\frac{1.000\ Zeichen}{4{,}5\,\frac{Zeichen}{Wort}} = 222{,}22\ Wörter$$

$$\boldsymbol{Umrechnungsfaktor}: \frac{750\ Wörter}{222{,}22\ Wörter} = \mathbf{3{,}375}$$

Quelle: Eigene Berechnung auf Basis der Daten von: Miller et al. (Miller et al., 1958, S.105ff.) und OpenAI (2023l)

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Bachelorarbeit selbständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht. Die vorgelegte Arbeit hat weder in der gegenwärtigen noch in einer anderen Fassung schon einem anderen Fachbereich der Hochschule Ruhr West oder einer anderen wissenschaftlichen Hochschule vorgelegen.

MH, 08.01.24 | C. Güter

Ort, Datum | Unterschrift