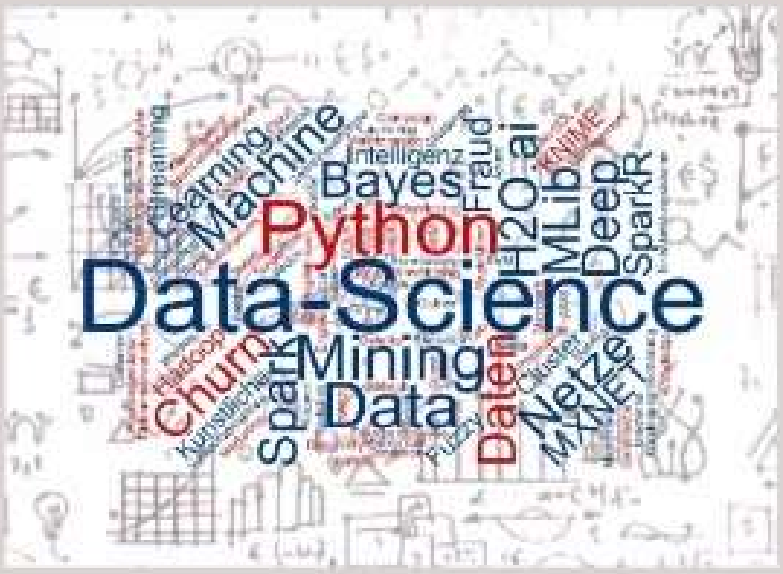
MICHAEL OETTINGER

# DATA SCIENCE UND AI

*Eine praxisorientierte Einführung im Umfeld von Machine Learning, künstlicher Intelligenz und Big Data*




3. erweiterte Auflage

Michael Oettinger

# Data Science und AI

## Eine praxisorientierte Einführung im Umfeld von Machine Learning, künstlicher Intelligenz und Big Data

3. erweiterte Auflage

© 2024 Michael Oettinger

Druck und Distribution im Auftrag des Autors:
tredition GmbH, Heinz-Beusen-Stieg 5, 22926 Ahrensburg, Germany

Das Werk, einschließlich seiner Teile, ist urheberrechtlich geschützt. Für die Inhalte ist der Autor verantwortlich. Jede Verwertung ist ohne seine Zustimmung unzulässig. Die Publikation und Verbreitung erfolgen im Auftrag des Autors, zu erreichen unter: Michael Oettinger, Römerschanzenstr. 1, 82110 Germering, Germany.

# Inhalt


1 Einleitung ..... 7

2 Daten bereitstellen ..... 10

2.1 Flatfiles ..... 10

2.2 Relationale Datenbanksysteme ..... 11

2.3 Data-Warehouse ..... 13

2.4 NoSQL ..... 16

2.5 Hadoop/Spark ..... 17

2.6 Cloud-Computing ..... 23

3 Datenanalyse ..... 28

3.1 Programmiersprachen ..... 28

3.2 Data-Science-Plattformen ..... 36

3.3 Machine Learning-Bibliotheken ..... 53

3.4 Cloud-Angebote ..... 59

3.5 Entscheidungshilfe für die Softwareauswahl ..... 64

4 Verfahren der Datenanalyse ..... 67

4.1 Künstliche Intelligenz ..... 67

4.2 Weitere Begriffe im Rahmen der Datenanalyse ..... 83

4.3 Datentypen und Skalentypen ..... 87

4.4 Einordnung der Verfahren ..... 89

4.5 Analyseverfahren – Machine Learning-Algorithmen ..... 96

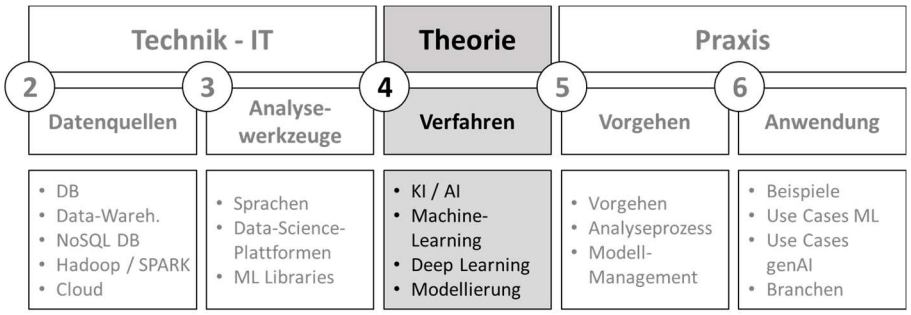
4.6 Auswahl des richtigen Verfahrens ..... 153

5 Vorgehensmodell für ML-Projekte ..... 156

5.1 Vorgehensweise – Methode ..... 156

5.2 Modell-Management ..... 165
5.3 Deployment ..... 166
5.4 Cheat-Sheet zu SQL, Python und PySpark ..... 172
5.5 Cheat-Sheet Machine Learning im Python-Notebook ..... 176
6 Anwendungsfälle – Use-Cases ..... 186
6.1 Use Cases nach Branchen ..... 186
6.2 Beschreibung einzelner Use Cases ..... 199
6.3 Use Cases für genAI ..... 221
7 Abschluss ..... 231
8 Informationsquellen ..... 236
Autor ..... 238
Literaturverzeichnis ..... 240
Stichwortverzeichnis ..... 244

# 4 Verfahren der Datenanalyse




Der Begriff Data-Science enthält den Anspruch, dass es sich bei diesem Fachgebiet um eine *Wissenschaft* handelt. Die in den vorangegangenen Kapiteln beschriebenen Aspekte waren jedoch eher praxisorientiert. Im Fokus stand die Technik, insbesondere die Software, die Data-Scientists bei ihrer Arbeit einsetzen bzw. von denen sie abhängig sind (Datenlieferanten).

Im folgenden Kapitel wird nun der wissenschaftliche Kern betrachtet und die wichtigsten Verfahren im Rahmen der maschinellen Datenanalyse werden vorgestellt. Dabei ist das Ziel, beim Aufbau eines Verständnisses für den ‘Geist‘ der Verfahren zu helfen. Die statistischen und mathematischen Grundlagen und besonders die Feinheiten sowie die Ausprägungen der Methoden können im Rahmen dieses Buches nicht ausführlich behandelt werden. Bevor die Verfahren in Kapitel 4.4 dargestellt werden, erfolgt zuerst deren Einordnung in den Gesamtzusammenhang der künstlichen Intelligenz. Außerdem sollen zentrale Begriffe und Grundlagen erläutert werden.

## 4.1 Künstliche Intelligenz

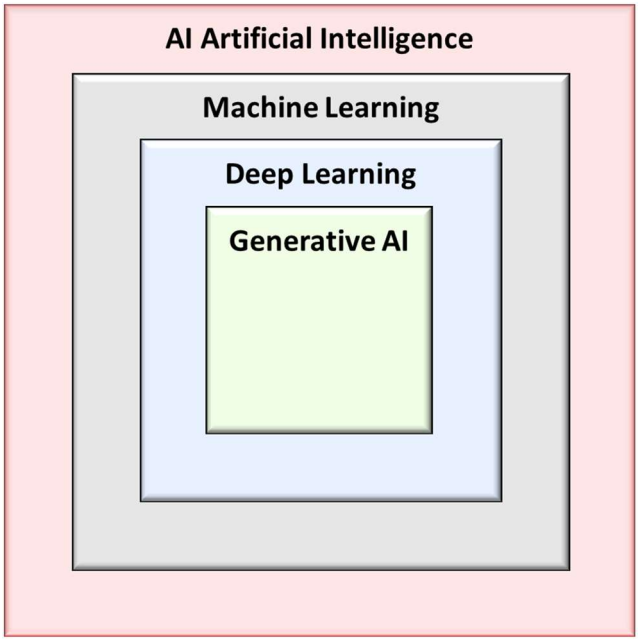
Seit ChatGPT im November 2022 wie aus dem Nichts in den Fokus der Öffentlichkeit gerückt ist, ist der Begriff künstliche Intelligenz bzw. Artificial Intelligence in aller Munde. Die Diskussion nimmt fast schon hysterische Formen an, in der entweder die baldige Weltherrschaft der künstlichen Intelligenz

bzw. der Verlust aller Arbeitsplätze prophezeit wird oder aber sehr lautsprechende ‘Crypto-Bros‘ auf meiner LinkedIn-Timeline erklären, wie ich dank künstlicher innerhalb von einem Monat zum Millionär werde.

Eine sachliche Beschreibung des Konzeptes der künstlichen Intelligenz ist notwendig, um die Chancen und Gefahren fundierter einschätzen zu können und darüber hinaus zu verstehen, wie die Aufgaben eines Data-Scientists und somit der Inhalt dieses Buches in den Gesamtkontext einzuordnen sind.

### 4.1.1 Abgrenzung künstlicher Intelligenz

Es bestehen unterschiedliche Ansätze, wie die Konzepte zu künstlicher Intelligenz abgegrenzt werden, wobei mir die folgende hierarchische Darstellung am meisten zusagt.



**Künstliche Intelligenz** (englisch: Artificial Intelligence, AI) ist ein Teilgebiet der Computer-Science bzw. Informatik und ein Oberbegriff, der sich auf die Fähigkeit von Maschinen bezieht, menschenähnliche Aufgaben wie Sprach-erkennung, Bilderkennung und Entscheidungsfindung auszuführen.

**Machine Learning** ist ein Teilbereich von künstlicher Intelligenz. Unter Anwendung fortschrittlicher Algorithmen soll aus großen Datenmengen gelernt werden, indem Muster erkannt werden. Das Ziel ist es, Vorhersagen zu treffen und damit Entscheidungen zu unterstützen.

**Deep Learning** bezieht sich auf einen Ausschnitt der Machine Learning-Verfahren, insbesondere auf die neuronalen Netze mit mehreren Hidden Layers (s. Kapitel 4.4.7) und Graphenverfahren (z. B. Multi-Class Logistic Regression). Es sind Verfahren, die in der Grundidee die Arbeitsweise des menschlichen Gehirns simulieren und besonders in den Bereichen Bild- und Sprach-erkennung eingesetzt werden.

**Generative Artificial Intelligence** ist eine Teilmenge von Deep-Learning-Modellen, die dazu verwendet werden, eigenständig Content wie Texte, Bilder, Videos oder Programmiercode anhand eines gegebenen Inputs zu generieren. Sie werden an riesigen Datenmengen trainiert und bedürfen keiner besonderen Instruktion.

Wenn in der Presse von künstlicher Intelligenz die Rede ist, muss somit differenziert werden, ob damit die Gesamtheit der computergestützten Lernmethoden (das ‘Gesamtkästchen‘ in der obigen Abbildung) oder der Einsatz von Generative Artificial Intelligence gemeint ist (das innerste ‘Kästchen‘).

In den folgenden Abschnitten werden die Verfahren der Generative Artificial Intelligence (genAI) erläutert.

### 4.1.2 ChatGPT und LLMs

Der Auslöser für den Hype um künstliche Intelligenz war die öffentliche Verfügbarkeit von ChatGPT im November 2022. Die Qualität der Antworten auf frei formulierte Fragen bzw. Aufforderungen stellte alles in den Schatten, was vergleichbare Sprachmodelle zuvor leisten konnten.

ChatGPT ist eine von OpenAI entwickelte Chatbot-Software, die auf künstlicher Intelligenz basiert und mit Menschen kommunizieren kann. Sie beruht auf einem Modell für Machine Learning, das Texte verarbeitet und Vorhersagen macht. ChatGPT lernt aus den Gesprächen mit den Nutzern und nutzt bereits gelesene Texte, um Antworten zu generieren.

Um die Grundlagen von ChatGPT zu verstehen, ist der Begriff ‘Large Language-Model‘ (großes Sprachmodell) zu klären. ChatGPT beruht auf einem GPT (Generative Pre-trained Transformer), der die Fähigkeit aufweist, große Mengen an Textdaten zu verstehen und darauf basierend qualitativ hochwertige Texte zu generieren.

Transformer-Modelle ermöglichen es, komplexe Abhängigkeiten in Daten zu modellieren und sind besonders effektiv bei der Verarbeitung langer Sequenzen, wie sie in Texten vorkommen.

Wer die grundsätzliche Arbeitsweise von Transformer-Modellen detaillierter verstehen möchte, sei auf die ausführliche Webseite der Financial Times verwiesen: ig.ft.com/generative-ai.

*Abbildung 7: Auszug aus ig.ft.com/generative-ai.*

Dort werden die grundlegenden Konzepte der Transformer (wie Token, Embeddings, Vector und Self-Attention) anschaulich erklärt.

Ein entscheidender Aspekt von ChatGPT ist das Konzept des unüberwachten Lernens. Das Modell wird mit riesigen Mengen an Textdaten vortrainiert, ohne dabei spezifische Anweisungen für eine bestimmte Aufgabe zu erhalten. Durch dieses Verfahren kann ChatGPT eine breite Vielfalt von Sprachmustern und Kontexten erlernen.

ChatGPT zeichnet sich durch seine Fähigkeit aus, kontextsensitive Vorhersagen zu treffen. Das bedeutet, dass es nicht nur auf den vorherigen Satz eines Textes reagiert, sondern auf den gesamten bisherigen Kontext. Dadurch kann das Modell präzisere und kohärentere Antworten generieren und Informationen über lange Textabschnitte hinweg berücksichtigen. Somit werden komplexe Zusammenhänge verstanden und darauf basierend adäquate Antworten generiert.

Durch adaptives Lernen kann ChatGPT während der Interaktion mit Benutzern neue Informationen aufnehmen und in den Kontext einbinden. Auf diese Weise ist das Modell flexibel und anpassungsfähig.

### Weitere Large Language-Models

ChatGPT ist der prominenteste Vertreter der Transformer-based Large Language-Models, aber nicht der einzige. Der zeitliche Verlauf der Entwicklung und die Verwandtschaft der bedeutenden Modelle sind nachfolgend in einem Stammbaum dargestellt.

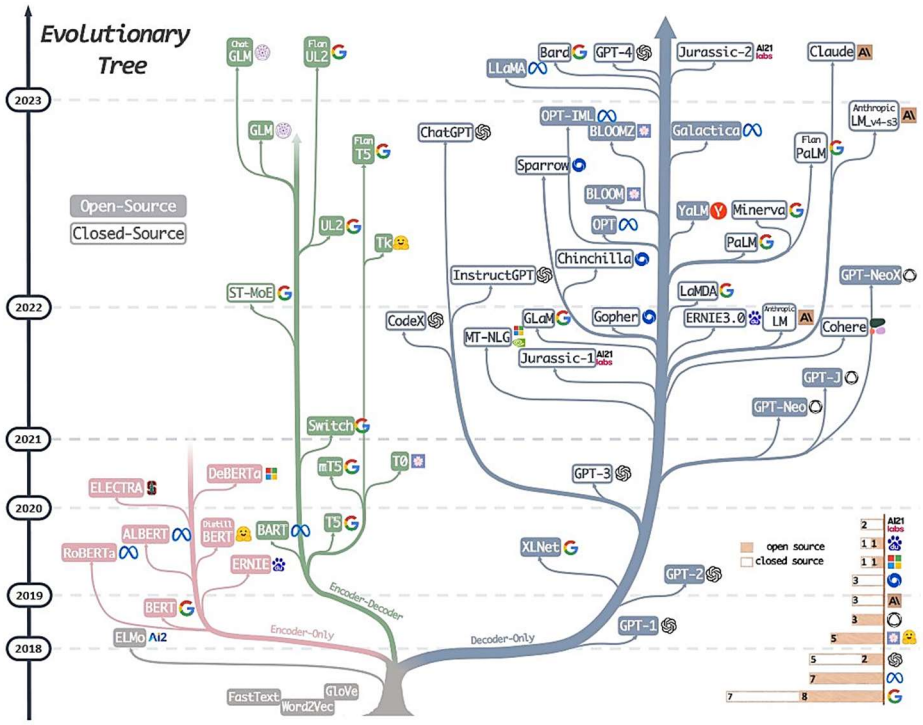



Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

*Abbildung 8: https://samim.io/p/2023-04-30-evolutionary-tree-of-llms/.*

Die Wurzeln reichen zurück zu früheren Modellen wie GPT-1 und GPT-2 von OpenAI. Diese bildeten den Grundstein für die Entwicklung leistungsstarker Sprachmodelle.

GPT-3, die dritte Iteration der Generative Pre-trained Transformer-Modelle, wurde im Jahr 2020 von OpenAI vorgestellt. Mit 175 Milliarden Parametern ist GPT-3 das bisher größte und leistungsfähigste Sprachmodell. Seine Fähigkeit, komplexe Zusammenhänge zu verstehen und qualitativ hochwertige Texte zu generieren, bildete die Grundlage für die Entwicklung von ChatGPT.

ChatGPT entstand durch Fine-Tuning von GPT-3 für die spezielle Anwendung im Dialog. Die Idee war, ein Sprachmodell zu schaffen, das natürliche und kohärente Gespräche mit Benutzern führen kann.

Die führenden Marktteilnehmer im Bereich Large Language-Models sind neben OpenAI Microsoft (das an OpenAI beteiligt ist), Meta und Google.

Eine Übersicht über die bedeutendsten Large Language-Models findet sich auf Wikipedia. Die Entwicklung ist dynamisch, sodass monatlich neue Versionen oder Modelle veröffentlicht werden.

| Name | Release date | Developer | Number of parameters | Notes |
|---|---|---|---|---|
| GPT-1 | June 2018 | OpenAI | 117 million | First GPT model, decoder-only transformer. |
| BERT | October 2018 | Google | 340 million | An early and influential language model, but encoder-only and thus not built to be prompted or generative. |
| XLNet | June 2019 | Google | 340 million | An alternative to BERT; designed as encoder-only. |
| GPT-2 | February 2019 | OpenAI | 1.5 billion | general-purpose model based on transformer architecture |
| GPT-3 | May 2020 | OpenAI | 175 billion | A fine-tuned variant of GPT-3, termed GPT-3.5, was made available to the public through a web interface called ChatGPT |
| GPT-Neo | March 2021 | EleutherAI | 2.7 billion | The first of a series of free GPT-3 alternatives released by EleutherAI. |
| GPT-J | June 2021 | EleutherAI | 6 billion | GPT-3-style language model |
| Megatron-Turing NLG | October 2021 | Microsoft and Nvidia | 530 billion | Standard architecture but trained on a supercomputing cluster. |
| Ernie 3.0 Titan | December 2021 | Baidu | 260 billion | Chinese-language LLM. Ernie Bot is based on this model. |
| Claude[144] | December 2021 | Anthropic | 52 billion | Fine-tuned for desirable behavior in conversations. |
| GLaM (Generalist Language Model) | December 2021 | Google | 1.2 trillion | Sparse mixture of experts model, making it more expensive to train but cheaper to run inference compared to GPT-3. |
| Gopher | December 2021 | DeepMind | 280 billion | Further developed into the Chinchilla model. |
| LaMDA (Language Models for Dialog Applications) | January 2022 | Google | 137 billion | Specialized for response generation in conversations. |
| GPT-NeoX | February 2022 | EleutherAI | 20 billion | based on the Megatron architecture |
| Chinchilla | March 2022 | DeepMind | 70 billion | Reduced-parameter model trained on more data. Used in the Sparrow bot. Often cited for its neural scaling law. |
| PaLM (Pathways Language Model) | April 22 | Google | 540 billion | Trained for ~60 days on ~6000 TPU v4 chips. |
| OPT (Open Pretrained Transformer) | May 2022 | Meta | 175 billion | GPT-3 architecture with some adaptations from Megatron |
| YaLM 100B | June 2022 | Yandex | 100 billion | English-Russian model based on Microsoft's Megatron-LM. |
| Minerva | June 2022 | Google | 540 billion | Minerva is based on PaLM model, further trained on mathematical and scientific data. |
| BLOOM | July 2022 | Large collaboration led by Hugging Face | 175 billion | Essentially GPT-3 but trained on a multi-lingual corpus (30% English excluding programming languages) |
| Galactica | November 22 | Meta | 120 billion | Trained on scientific text and modalities. |
| AlexaTM (Teacher Models) | November 22 | Amazon | 20 billion | bidirectional sequence-to-sequence architecture |
| Neuro-sama | December 2022 | Independent | Unknown | A language model designed for live-streaming on Twitch. |
| LLaMA (Large Language Model Meta AI) | February 2023 | Meta | 65 billion | Trained on a large 20-language corpus to aim for better performance with fewer parameters. |
| GPT-4 | March 2023 | OpenAI | Exact number unknown | Available for ChatGPT Plus users and used in several products. |
| Cerebras-GPT | March 2023 | Cerebras | 13 billion | Trained with Chinchilla formula. |
| Falcon | March 2023 | Technology Innovation Institute | 40 billion | |
| BloombergGPT | March 2023 | Bloomberg L.P. | 50 billion | LLM trained on financial data from proprietary sources. |
| PanGu-Σ | March 2023 | Huawei | 1.085 trillion | |
| OpenAssistant | March 2023 | LAION | 17 billion | Trained on crowdsourced open data |
| Jurassic-2 | March 2023 | AI21 Labs | Exact size unknown | Multilingual |
| PaLM 2 (Pathways Language Model 2) | May 2023 | Google | 340 billion | Used in Bard chatbot. |
| Llama 2 | July 2023 | Meta | 70 billion | Successor of LLaMA. |
| Claude 2 | July 2023 | Anthropic | Unknown | Used in Claude chatbot. |
| Falcon 180B | September 23 | Technology Innovation Institute | 180 billion | |
| Mistral 7B | September 23 | Mistral AI | 7.3 billion | |
| Claude 2.1 | November 23 | Anthropic | Unknown | Used in Claude chatbot. Has a context window of 200,000 tokens, or ~500 pages. |
| Grok-1 | November 23 | x.AI | Unknown | Used in Grok chatbot. Grok-1 has a context length of 8,192 tokens and has access to X (Twitter). |
| Gemini | December 2023 | Google DeepMind | Unknown | Multimodal model, comes in three sizes. Used in Bard chatbot. |
| Mixtral 8x7B | December 2023 | Mistral AI | 46.7B total, | Mixture of experts model, outperforms GPT-3.5 and Llama 2 70B on many benchmarks. All weights were released via torrent. |
| Phi-2 | December 2023 | Microsoft | 2.7B | So-called small language model, trained on "textbook-quality" data based on the paper "Textbooks Are All You Need". |

*Abbildung 9: https://en.wikipedia.org/wiki/Large_language_model.*

Auf den Punkt gebracht sind ChatGPT und vergleichbare Large Language-Models lediglich *Sprachmodelle*, die in einer bisher nicht gekannten Qualität auf sprachlichen Eingaben basierende sprachliche Ausgaben erzeugen können. Sie basieren auf Methoden, denen ‘zufälligerweise‘ in der Wissenschaft der Begriff *künstliche Intelligenz* zugewiesen wurde. Sie sind aber keine intelligenten Wesen. Sie haben keine künstliche Intelligenz, sie verfügen über kein Bewusstsein und sie sind nicht zu echter Kreativität fähig, die über den vorgegebenen Rahmen hinaus geht.

Hier ein lustiges Beispiel, was passiert, wenn ein Bildredakteur durch ‘künstliche Intelligenz‘ ersetzt wird:

*Abbildung 10: https://www.swr3.de/aktuell/schimmel-vermieter-mietwohnung-rechte-108.html.*

**Cheat-Sheet zu ChatGPT**

Nach meiner Meinung besteht somit nicht die Gefahr, dass die '*künstliche Intelligenz*' in Zukunft sämtliche Aufgaben übernehmen, uns Menschen ersetzbar und damit arbeitslos machen wird. Unbestreitbar wird sich durch den Einsatz von Generative Artificial Intelligence die Produktivität in zahlreichen Branchen deutlich erhöhen lassen. Somit werden Menschen, die Tools auf Basis von Generative Artificial Intelligence intelligent nutzen, Personen ersetzen, die diese Tools nicht bedienen können. Es ist daher sinnvoll, sich mit den Eigenheiten der Tools vertraut zu machen. An erster Stelle ist in diesem Kontext ChatGPT zu nennen. Es folgt daher eine Anleitung, wie Fragen bzw. Aufforderungen – sogenannte Prompts – an ChatGPT formuliert werden können.

Grundsätzlich kann man in natürlicher Sprache mit ChatGPT sprechen (d. h. wie meine Mutter schon seit Jahren Google Suchen formuliert). Sowohl die Prompts als auch die Ausgaben können – auch unabhängig voneinander - in zahlreichen Sprachen erfolgen.

Da die Sprachmodelle primär mit englischsprachigen Texten trainiert wurden, wird bei einer Ausgabe in einer anderen Sprache eine Übersetzung integriert. Damit dieser Schritt für die Eingabe nicht notwendig ist, habe ich es mir angewöhnt, die Prompts auf Englisch zu formulieren und als Ausgabesprache ggf. Deutsch zu wählen. Ob dadurch die Ergebnisse tatsächlich besser werden, kann ich nicht beurteilen, aber auf Englisch ist eine größere Anzahl an Beispielen und Tutorials verfügbar.

Bei der Formulierung eines Prompts kann man:

- eine **Rolle** benennen,
- den **Bedarf** formulieren,
- die **Aufgabe** spezifizieren, auch mit **Details,**
- und das **Ausgabeformat** festlegen.
- Man kann Dinge **ausschließen**
- und **Beispiele** für die gewünschte Ergebnisse anfügen.

*Act like a **[ROLE]**, I need a **[NEEDS]**, you will **[TASK]**, in the process, you should **[Details]**, **[DO NOT...]**, present the result as **[FORMAT]**. Here is an example: **[Example].***

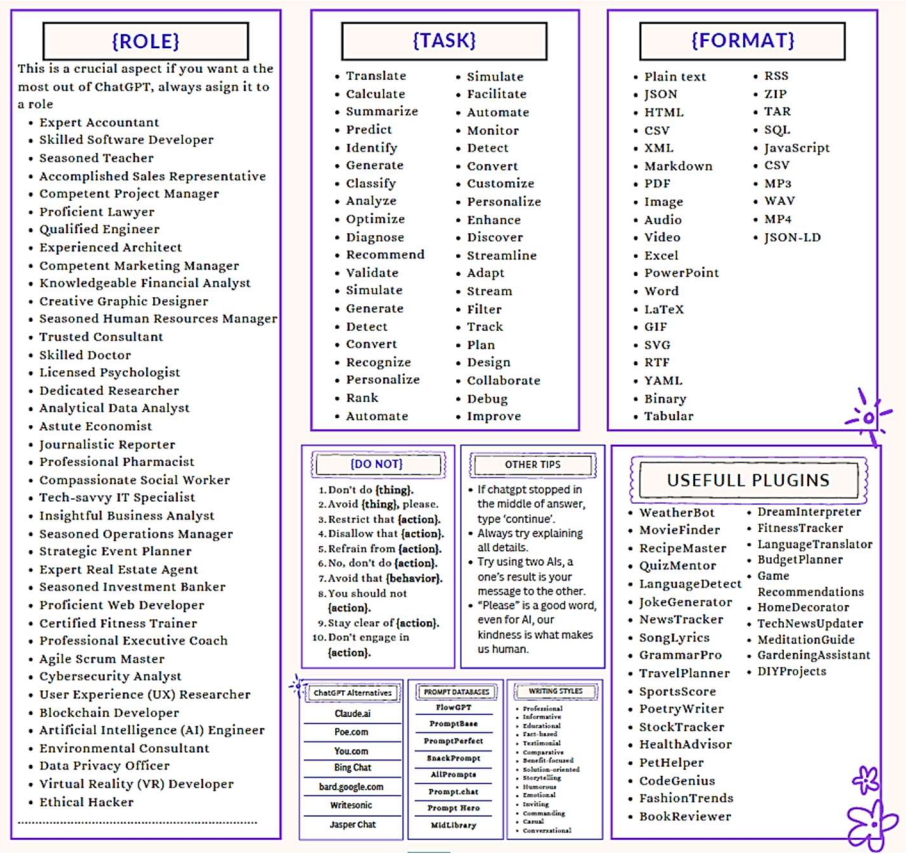



*Abbildung 11: gptai.gumroad.com*

Hinweise und Beispiele zu ChatGPT-Prompts in deutscher Sprache finden sich z. B. auf blogkurs.de/chatgpt-prompts/.

### 4.1.3 Weitere GenAI-Anwendungen

ChatGPT ist der prominenteste Vertreter der neuartigen Tools im Bereich der künstlichen Intelligenz, mitnichten jedoch der einzige. Seit der Zugriff auf das OpenAI-GPT-Sprachmodell über eine offene API auch Drittentwicklern zur Verfügung steht, werden täglich neue Tools veröffentlicht. Diese Drittanwendungen nutzen die ‘Text-in-Text-out‘-Funktionalität der OpenAI-API, um einen Mehrwert zu generieren. So schnell wie sie entstehen geraten jedoch auch viele dieser Anwendungen wieder in Vergessenheit, vermutlich da der angestrebte Mehrwert nicht so nachhaltig ist wie versprochen.

Um einen Überblick über die tatsächlich nützlichen Tools zu behalten, ist es sinnvoll, diese nach ihrer grundsätzlichen Aufgabe zu gruppieren. Im Anschluss können die bedeutendsten Vertreter der Klassen beleuchtet werden.

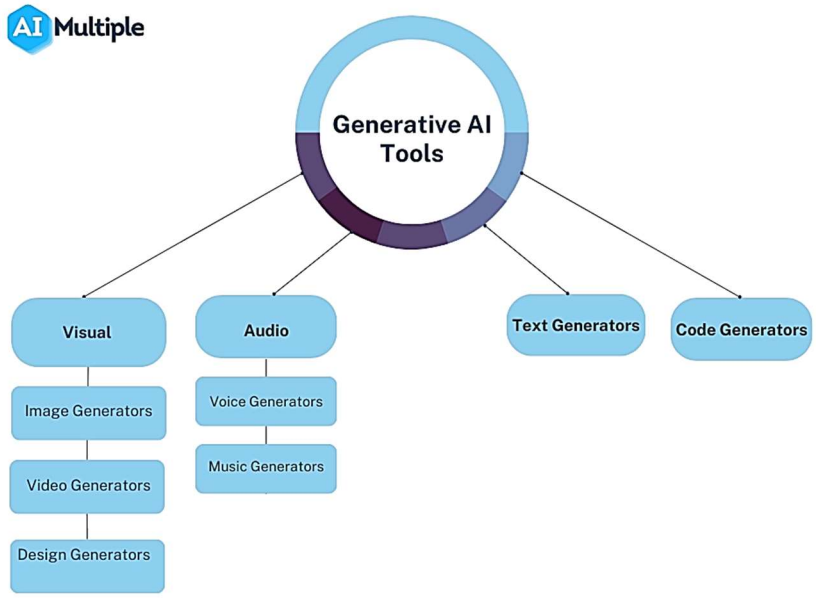



*Abbildung 12: aimultiple.com*

Die ‘Erzeugnisse‘ der Generative-Artificial-Intelligence-Tools können folgende Formen annehmen:

- visuell, in Form von Bildern, Videos oder Designelementen,
- Audio, d. h. gesprochene Sprache oder Musik,
- Text oder
- Code in einer Programmiersprache.

## *Visuell*

**Midjourney** kreiert aus Textbeschreibungen Bilder und wird durch Befehle an einen Discord-Bot gesteuert. Das Programm erschafft zunächst eine Vorschau mit vier möglichen Entwürfen, die zur Weiterverarbeitung genutzt werden können.

Ein von OpenAI entwickeltes vergleichbares Programm ist **DALL-E**. Hier werden ebenso Bilder aus Textbeschreibungen erzeugt.

**CleanupPictures** fällt in die Kategorie Fotobearbeitung, speziell in den Bereich Fotobereinigung. Benutzer können mit dieser Anwendung unerwünschte Elemente wie Wasserzeichen oder Mängel automatisiert aus Fotos entfernen.

Zum Erstellen von Präsentationen eignet sich **Tome.** Darin können Templates ausgewählt werden, aus denen anhand der Anweisungen eine Präsentation erzeugt wird. Für die Texterstellung wird ChatGPT, für die Bebilderung DALL-E 2 genutzt.

Der Videogenerator **Runway Gen-1** erzeugt aus hochgeladen Bildern, Videodateien und Textprompts Videos.

Videobearbeitung auf der Basis künstlicher Intelligenz ermöglicht **Pictory**. Das Tool verhilft zu einer schnellen Nachbearbeitung von Videos einschließlich dem Herausschneiden unliebsamer Sequenzen oder Kürzungen. Aus langen Videos lassen sich auf Knopfdruck kurze Trailer generieren, die die Highlights des Videos aufgreifen. Es können auch Skripte oder Blogposts in Videos umgewandelt werden und Texte automatisiert in Videos eingeblendet werden.

## *Audio*

Das Noise-Cancelling-Tool **Krisp** kann störende Hintergrundgeräusche in Videokonferenzen mit Programmen wie Skype, Microsoft Teams und Zoom per Mausklick herausfiltern. In puncto Google-Suchvolumen kann Krisp zwar nicht mit ChatGPT oder Copy.ai mithalten, es rangiert aber immerhin im Mittelfeld der zehn populärsten Künstliche-Intelligenz-Tools.

Mit **otter.ai** kann gesprochene Sprache in Text transkribiert werden. Dadurch können z. B. in Videos automatisch Untertitel bzw. Texte zu Podcasts erzeugt werden. Außerdem können damit die Inhalte von Audiodateien für die Textanalyse zugänglich gemacht werden.

Die umgekehrte Richtung der Umwandlung von Text in Sprache ermöglicht z. B. **Lovo.ai**.

**VocalRemover** ist in der Lage, Stimme und Musik zu trennen, beispielsweise zur Erstellung von Remixes oder Karaokeversionen eines Songs.

## *Text*

Neben dem Platzhirsch ChatGPT bestehen weitere nützliche Tools zur Textverarbeitung mit künstlicher Intelligenz, die nachfolgend genannt werden.

**Google Bard** ist ein von Google entwickelter Chatbot auf der Basis von künstlicher Intelligenz, der als direkte Reaktion auf den Erfolg von ChatGPT entwickelt und im März 2023 in eingeschränkter Kapazität veröffentlicht wurde, bevor er im Laufe des Sommers in weiteren Ländern verfügbar wurde.

**Character.ai** ist ebenfalls eine Chat-Anwendung, die eine größere Verbreitung gefunden hat.

**QuillBot** ist eine im Jahr 2017 entwickelte Software, die künstliche Intelligenz nutzt, um Texte neu zu schreiben und zu paraphrasieren. Damit lassen sich beispielsweise E-Mails, Social-Media-Posts und Essays automatisiert erstellen und Texte umschreiben. Quillbot.ai nutzt dazu u. a. eine Grammatikprüfung und Funktionen für die Zusammenfassung von Texten. Es stellt sich hier die Frage, ob es möglich ist, z. B. Referate für das Fach Geschichte in der neunten Klasse zuerst in ChatGPT zu erstellen und danach in QuillBot umformulieren zu lassen, um sie damit für automatische genAI-Erkennungssoftware unverdächtig zu machen.

**Copy.ai** wird von Unternehmen als Textgenerator eingesetzt, beispielsweise für Blogs, Social-Media-Posts oder E-Mails.

## *Generierung von Programmiercode*

Zur Erzeugung von Programmiercode bestehen neben ChatGPT spezialisierte Anwendungen.

**OpenAI Codex** ist der auf OpenAI-GPT basierende Programmiercode-Generator, der anhand von Milliarden Codezeilen trainiert wurde und derzeit 14 Programmiersprachen unterstützt.

**Copilot** verwendet öffentlich verfügbaren Code aus GitHub-Repositorys. Das Tool erkennt Fehler im Code und empfiehlt Änderungen. Copilot kann in die üblichen Programmierumgebungen integriert werden.

Der von Google unterstützte **AlphaCode** von DeepMind ermöglicht Entwicklern Zugriff auf Quellcode aus Sprachbibliotheken. Mit AlphaCode können Tausende vorgefertigte Bibliotheken genutzt und so Schnittstellen zu APIs von Drittanbietern schnell und einfach erstellt werden.

**GenAI Stack** ist ein Framework für die Integration von Large Language Models in jede Anwendung. Damit können eigene Anwendungen schnell um deren mächtige Funktionalitäten erweitert werden.

Ein Blick auf die Häufigkeit der Website-Besuche der Generative-Artificial-Intelligence-Anwendungen hilft, die Übersicht zu bewahren und die bedeutendsten Tools zu identifizieren.

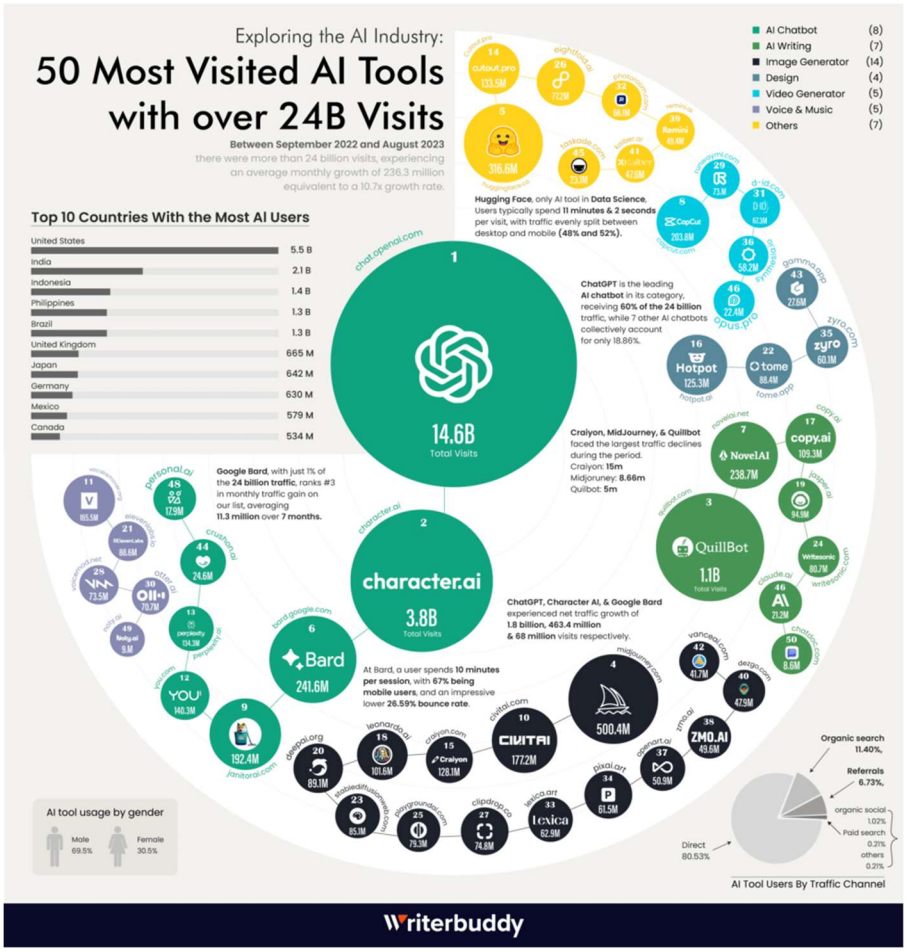



*Abbildung 13: writerbuddy.ai.*

## 4.2 Weitere Begriffe im Rahmen der Datenanalyse

Neben den im vorangegangenen Abschnitt beschriebenen Begriffen gibt es weitere Ausdrücke, die immer wieder auftauchen. Da wird von Big Data Analytics, Data-Mining, Predictive Analytics, BI etc. gesprochen. Anbieter von