Content Creation mit KI

Das umfassende Handbuch

Kapitel 4

Die großen KI-Plattformen: Vom Prompt-Tool zum multimodalen Kreativstudio

Wer, wie, was? Der Blick unter die KI-Motorhaube.

In den ersten drei Kapiteln dieses Buchs haben Sie gelernt, wie Sie mit KI in einen kreativen Dialog treten: Wie Sie prompten, iterieren und warum Sie die KI als Inspirationsmaschine nutzen können. Aber was passiert eigentlich auf der anderen Seite? Was antwortet Ihnen da, und warum antwortet es manchmal so verblüffend gut und manchmal einfach falsch?

Die kurze Antwort: ein Large Language Model (LLM), ein großes Sprachmodell. Das ist das Herzstück hinter ChatGPT, Claude, Gemini und allen anderen KI-Chatbots, die Sie in diesem Buch kennenlernen werden. Und um die Plattformen, ihre Stärken und ihre Eigenheiten zu verstehen, lohnt sich ein kurzer Blick unter die jeweilige Motorhaube. Denn neben den grundlegenden Funktionen schlägt in jedem Modell ein anderer Takt. Wer den versteht, tut sich bei der Wahl des passenden Juniors für seine Arbeit leichter.

4.1 Was steckt unter der Motorhaube? So funktionieren Sprachmodelle

ChatGPT[1] steht für *Chatbot Generative Pre-trained Transformer*. Schon der Name verrät einiges über die grundsätzliche Funktionsweise der LLMs. Es handelt sich, wie in Abbildung 4.2 skizziert, um einen Chatbot, der auf einem vortrainierten Sprachmodell basiert und Texte generiert. Im Gegensatz zu den langweiligen und nervigen Chatbots früherer Generationen, denen jeder schon auf Kundenservice-Seiten begegnet ist, kann ChatGPT menschenähnlich kommunizieren und Dialoge über alles Mögliche führen – vom Verfassen von Gedichten über die Beantwortung von Prüfungsfragen bis hin zu Blogtexten in jedem gewünschten Schreibstil.

Der Psychologe Steven Pinker bringt den Wow-Effekt auf den Punkt:

> *»Es ist beeindruckend, wie ChatGPT plausible, relevante und gut strukturierte Prosa produzieren kann, ohne die Welt zu verstehen, ohne offene Ziele, explizit dargestellte*

1 *https://chatgpt.com*

Fakten oder die anderen Dinge, die wir für notwendig gehalten hätten, um intelligent klingende Prosa zu erzeugen.[2]

Wie macht die Maschine das? Anders, als die meisten Menschen intuitiv vermuten. Ein Sprachmodell ist keine Datenbank, die auf Ihre Frage die passende Antwort nachschlägt. Es ist kein Lexikon, das Fakten speichert und auf Abruf ausspuckt. Es ist auch keine Suchmaschine, die im Netz nach Ergebnissen sucht. Um Texte zu generieren, die natürlicher Sprache ähneln, verwenden Sprachmodelle Deep Learning. Deep-Learning-Algorithmen sind tiefe neuronale Netze (*Deep Neural Networks*, kurz DNNs). Diese bestehen wie in Abbildung 4.1 gezeigt aus vielen Schichten linearer und nicht linearer Verarbeitungseinheiten, den künstlichen Neuronen. Daher auch der Begriff »deep« für »tief«. Je mehr Neuronen und Schichten ein neuronales Netz hat, desto komplexere Sachverhalte können abgebildet werden.

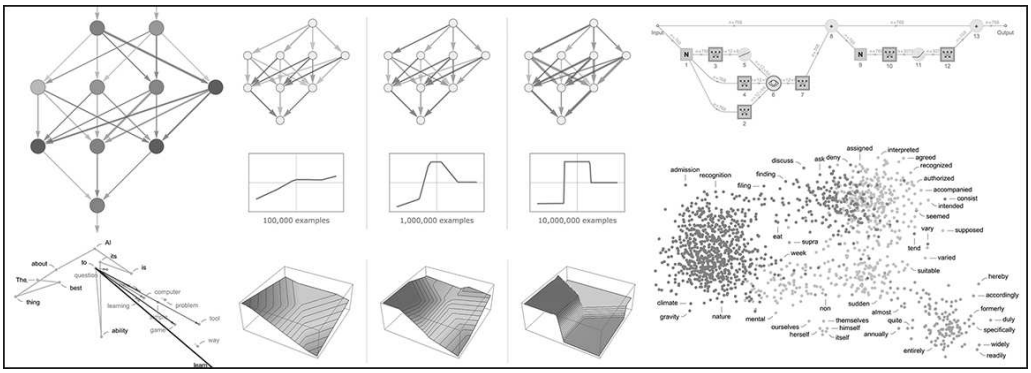

Abbildung 4.1: Stephen Wolfram illustriert die Arbeitsweise der KI mit mehrdimensionalen Datenräumen und Wahrscheinlichkeiten. (Quelle: Stephen Wolfram)[3]

Die Nutzung eines LLM für jedermann machte aber erst die zusätzliche Innovation möglich, mit deren Einführung ChatGPT von Anfang an der Textgenerator mit der größten Aufmerksamkeit war: der Chatbot, eine intelligente und intuitive Benutzeroberfläche (siehe Abbildung 4.2). Darüber können alle Menschen ohne technisches Vorwissen die LLMs mit ihren Fragen löchern und sich von den Schreibkünsten der KI überraschen lassen.

Das funktioniert in der Praxis dann wie in der Abbildung der NZZ[4] plakativ dargestellt: Das LLM, mit einer gewaltigen Menge an Texten von Webseiten, Büchern, Artikeln und Fo-

2 Alvin Powell, Will ChatGPT supplant us as writers, thinkers? In: The Harvard Gazette, *https://news.harvard.edu/gazette/story/2023/02/will-chatgpt-replace-human-writers-pinker-weighs-in/*, 14.02.2023 [21.02.2023]

3 *https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/*, 14.02.2023 [20.02.2023]

4 Ruth Fullerer, Was ist ein Prompt? Kann man Texte von Chat-GPT erkennen? Wo kann ich Bild-KI ausprobieren? Fragen über künstliche Intelligenz, die Sie sich nicht mehr zu stellen trauen, in: NZZ, *https://www.nzz.ch/technologie/was-ist-ein-prompt-und-andere-fragen-die-sie-sich-zu-kuenstlicher-intelligenz-nicht-mehr-zu-stellen-trauen-ld.1824069*, 21.04.2024 [31.01.2026]

renbeiträgen trainiert, hat gelernt, wie wahrscheinlich bestimmte »Wortbausteine«, sogenannte *Token*, in einem Kontext aufeinander folgen. Es stellt diese Token als Zahlenvektoren in einem mehrdimensionalen Bedeutungsraum dar. Ähnliche Begriffe liegen dort nahe beieinander. Wenn Sie einen Prompt wie »Was frisst die Katze?« eingeben, zerlegt das Modell Ihren Text in Token, wandelt sie in Vektoren um und berechnet für jedes mögliche nächste Wort eine Wahrscheinlichkeit. Das System wählt das plausibelste Token, fügt es an, berechnet dann erneut die Wahrscheinlichkeiten für das nächste und so weiter, bis eine vollständige Antwort entsteht: »Die Katze frisst Mäuse.«

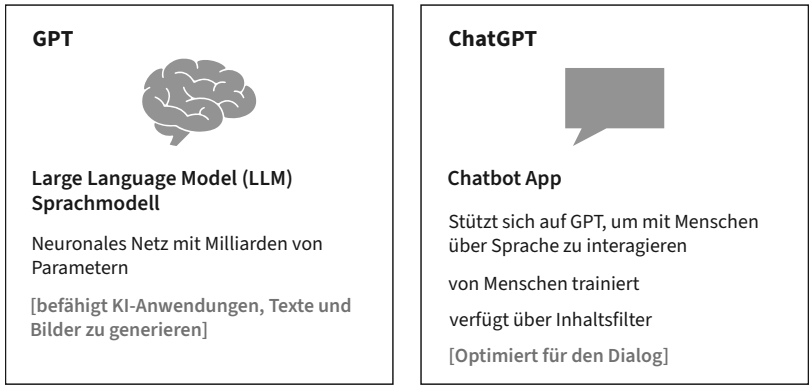

Abbildung 4.2: LLM und Chatbot sind Hirn- und Textmodul der KI-Revolution. Der Chatbot ist die Oberfläche, über die Sie per Prompt mit dem Sprachmodell kommunizieren. Das LLM verarbeitet Ihre Eingabe und generiert die Antwort. Die Internetsuche ergänzt das System um aktuelle Informationen, die über das Trainingsdaten-Ende hinausgehen.[5]

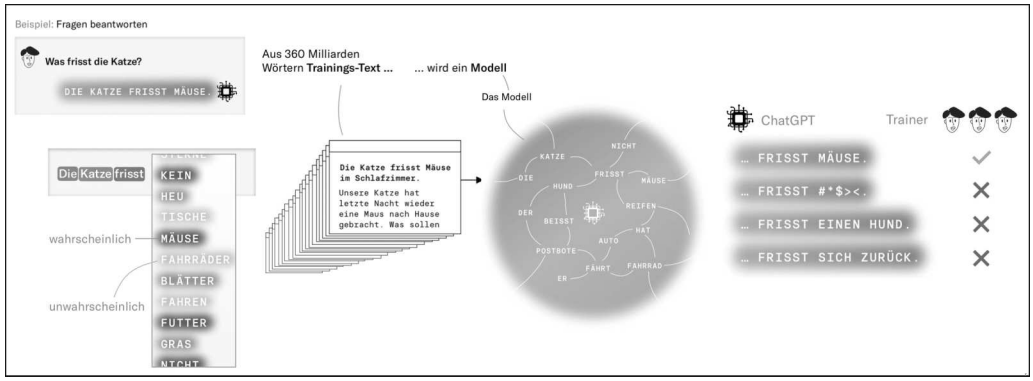

Abbildung 4.3: »Magische« Statistik: LLMs rechnen ihre Antwort Wort für Wort zusammen, auf der Basis von Wahrscheinlichkeiten vieler Trainingsdaten. (Quelle Grafiken: zusammengestellt aus NZZ[6]*)*

5 Vgl. *https://zapier.com/blog/chatgpt-vs-gpt/#* (aus dem Englischen, modifiziert)
6 Ruth Fuller, ebenda

Tobi Müller bringt es in der *Zeit* auf eine griffige Formel: »*Sie rechnet, sie denkt nicht*.«[7]

Damit die Antworten dennoch möglichst menschlich klingen, der Chatbot also natürlich und eloquent mit seinen Nutzerinnen und Nutzern in Dialog treten kann, wird er zusätzlich trainiert, bevor und während er auf die Menschheit losgelassen wird. Expertinnen und Experten aus Fleisch und Blut formulieren selbst passende Antworten oder bewerten die Antworten, die das System auf die Fragen gibt, um es immer besser an die menschlichen Bedürfnisse anzupassen. Dabei geht es nicht darum, Antworten so vorzuformulieren, dass der Bot sie bei der nächsten Antwort verwenden kann. Vielmehr lernt der Chatbot auf diese Weise, ob seine Antwort so vom Nutzer gewünscht war und verstanden wurde. Aus diesen Trainingsparametern kann das System dann regelrecht lernen und weitertrainieren.[8] Der Chatbot bekommt auf diese Weise zudem Inhaltsfilter, die verhindern sollen, dass die KI außer Kontrolle gerät oder gegen Rechte sowie ethische und moralische Grundregeln verstößt.

Der Mathematiker Stephen Wolfram hat die Arbeitsweise von ChatGPT in einem viel beachteten Fachartikel beschrieben, der bis heute zu den besten Erklärungen gehört. Seine Kernaussage:

> *»ChatGPT versucht im Grunde immer, eine vernünftige Fortsetzung des vorherigen Texts zu erstellen, wobei ›vernünftig‹ bedeutet, was man erwarten kann, nachdem man gesehen hat, was andere Leute auf Milliarden von Webseiten usw. geschrieben haben.«*[9]

Stellen Sie sich das bildhaft so vor: Jemand, der tausend Kochbücher gelesen hat, kann Ihnen ein plausibles Rezept für Risotto improvisieren, nicht weil er die Rezepte auswendig kennt, sondern weil er die Muster und Zusammenhänge verinnerlicht hat. Ein Sprachmodell macht etwas Ähnliches, nur mit der gesamten menschlichen Schriftsprache. Es hat während seines Trainings Muster in der menschlichen Sprache deduziert, hat sich die Logik und die Struktur von Texten angeschaut, daraus »gelernt«, die zugrundeliegenden Daten aber nicht behalten.

Stefan Ultes vom Lehrstuhl für Sprachgenerierung und Dialogsysteme an der Uni Bamberg fasst zusammen, was das für die Praxis bedeutet:

»*Die Technologie ist nicht darauf ausgelegt, dass sie Wissen speichert und dann alles im Detail immer korrekt wiedergibt. Es ist ein Sprachmodell, kein Wissensmodell*.«[10]

---

7 Tobi Müller, Der Midcult-Generator, in: Zeit online, *https://www.zeit.de/kultur/2023-01/chatgpt-gpt-ki-kunst-kultur-midcult/komplettansicht#print*, 31.01.2023 [04.02.2023]

8 Werner Bogula, Bringt ChatGPT den Durchbruch für KI? In: ARIC Insights V, 16.02.2023, (online, Zoom-Seminar)

9 Siehe vertiefenden Fachartikel: Stephen Wolfram, What Is ChatGPT Doing … and Why Does It Work? In: Writings, *https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/*, 14.02.2023 [20.02.2023]

10 Hinnerk Feldwisch-Drentrup, Viel versprochen, in: FAS, 19.02.2023, Nr. 7, S. 53

Das ist vielleicht etwas vereinfacht formuliert: Aber richtig ist, dass LLMs Optimierer für Sprachwahrscheinlichkeiten und keine expliziten Wissensdatenbanken sind. Was sich inzwischen aber auch herauskristallisiert hat: Praktisch enthalten sie schon sehr viel implizites Weltwissen, das sie aus ihren Trainingsdaten abstrahiert haben – allerdings ohne Garantie auf Vollständigkeit oder Fehlerfreiheit. Wohlgemerkt: Weltwissen, nicht das Weltverständnis, das wir Menschen in uns tragen.

Das alles erklärt auch, warum KI auch halluziniert, sprich hin und wieder Dinge behauptet, die schlicht nicht stimmen. Das Modell berechnet, was als Nächstes am plausibelsten klingt. Meistens ist das Plausible auch richtig. Aber eben nicht immer. Die KI hat kein Faktengedächtnis, das sie überprüfen könnte – sie hat ein Sprachgefühl, das erstaunlich oft ins Schwarze trifft und gelegentlich danebenliegt, wenn der Kontext fehlt. Und im Zweifelsfall wird das System »lieber« halluzinieren, also Fakten und Geschichten erfinden, als Ihre Frage unbeantwortet zu lassen. Ein interessanter »Charakterzug«, auf den Sie sich einstellen sollten.

Genau deshalb ist die im Nachhinein in die KIs integrierte Internetsuche so wichtig geworden: Alle großen Plattformen bieten inzwischen die Möglichkeit, dass die KI (selbstständig oder erst auf expliziten Wunsch, sprich Prompt ihres Users) das Web in Echtzeit durchsucht, bevor sie antwortet. Das kompensiert eine fundamentale Schwäche. Das Modell selbst kennt Ereignisse nur bis zu seinem letzten Trainingszeitpunkt. Dieser verschiebt sich natürlich mit der Zeit immer wieder in Richtung Gegenwart, hängt aber dennoch stets ein gutes Stück hinter ihr zurück: Die undatierte Frage nach dem Ergebnis des »letzten« Fußballspiels des FC Bayern München gegen Borussia Dortmund, das für den User zum Zeitpunkt der Frage gestern stattfand, ist für die Maschine daher vielleicht eine Begegnung der letzten Saison im letzten Jahr. Daher wird die Antwort nicht den Erwartungen des Users entsprechen, sondern als »falsch« abgetan werden und der Maschine Unzuverlässigkeit unterstellt. Damit tut man der Maschine aber wissentlich oder unwissentlich »unrecht«. Das »Problem« saß in diesem Fall sozusagen vor der Maschine.

Denn über angebundene Tools wie Websuche oder Datenbanken kann die Maschine auf aktuelle Informationen zugreifen, wenn sie es denn darf und kann und der User dafür das Abo bezahlt. Wenn Sie in ChatGPT, Claude oder Gemini die Websuche aktivieren, verbinden Sie das Sprachgefühl des Modells mit aktuellen Informationen aus dem Netz. Das ist, als würde Ihr brillanter, aber manchmal unzuverlässiger Textassistent endlich Zugang zu einer Bibliothek bekommen.

Merke: Ein Sprachmodell ist keine Datenbank – dennoch Vorsicht!

LLMs sind nicht primär dafür gebaut, Trainingsdaten wie eine Datenbank abzulegen. Sie lernen statistische Muster in Sprache und repräsentieren Bedeutungen in hochdimensionalen Vektorräumen. In der Praxis können sie aber einzelne Passagen aus ihren Trainingsdaten memorisiert haben und unter bestimmten Bedingungen wörtlich reproduzieren. Deshalb sollte man sie nicht mit sensiblen oder urheberrechtlich heiklen Inhalten trainieren oder füttern.

Das Ergebnis dieser beschriebenen Technologie ist ein verblüffendes Sprachverständnis, das laut KI-Forscherin Hannah Bast nicht sofort, aber doch sehr bald »alles verändern« wird. Denn wie sie in einem hörenswerten Spiegel-Podcast zusammenfasst:

»Die Menschheit kreiert derzeit eine intelligentere Spezies.«[11]

4.2 Schnell oder gründlich? Die zwei Klassen von KI-Modellen

Jetzt wissen Sie, wie ein LLM, ein Sprachmodell, funktioniert: Es berechnet Token für Token die wahrscheinlichste Fortsetzung des Dialogs im gegebenen Kontext. Aber es gibt Unterschiede in der Leistung verschiedener Modelle ein und desselben LLMs.

Das erste ist auf Tempo optimiert: Es antwortet sofort, ohne lange nachzudenken. Das zweite arbeitet langsamer und dafür gründlicher. Beide arbeiten mit denselben Grundlagen, aber ihre Herangehensweise unterscheidet sich fundamental. Alle großen Plattformen bieten inzwischen diese zwei Klassen an:

- *Standardmodelle* (manchmal auch Instant-Modelle genannt) antworten schnell und eignen sich für alltägliche Aufgaben: eine E-Mail formulieren oder einen Text zusammenfassen. Sie sind die Arbeitstiere des Alltags und relativ günstig, wenn nicht sogar kostenlos nutzbar.
- *Reasoning-Modelle* (oft als *Thinking-Modus* verfügbar) durchlaufen vor der Antwort einen internen Denkprozess. Das klingt vermenschlicht, ist es auch. Tatsächlich investieren sie einfach mehr Rechenzeit: Sie zerlegen komplexe Aufgaben in Teilschritte, vergleichen verschiedene Lösungspfade und verwerfen viele schlechte Vorschläge, bevor sie Ihnen die Antwort präsentieren. Das reduziert tatsächlich die Fehlerquote. Ein Fun Fact: Dass man der Maschine beim vermeintlichen Denken auf dem Bildschirm zuschauen kann, ist unterhaltsam und eher für unser menschliches »Hirn« gedacht, damit wir das maschinelle »Grübeln« besser verstehen und nachvollziehen können, warum das Ergebnis nun länger dauert. Es ist also eher ein Marketing-Kniff denn ein echtes »Denk-Protokoll«.

Thinking liefert bei anspruchsvollen Aufgaben deutlich bessere Ergebnisse: beispielsweise bei der Analyse widersprüchlicher Quellen oder bei Aufgaben, die mehrere Arbeitsschritte wie Recherchieren, Zusammenfassen und Einordnen erfordern.

Merke: Nicht jede Aufgabe braucht das stärkste Modell. Aber wer nur das schnelle nutzt, verschenkt das Potenzial der KI bei komplexen Aufgaben. Für Ihre Content-Arbeit bedeutet das: Nutzen Sie Standardmodelle für schnelle Aufgaben und Reasoning-Modelle für alles, was Tiefe und Präzision erfordert.

11 Juan Moreno, Die Menschheit kreiert derzeit eine intelligentere Spezies, in: Spiegel, Moreno+1, *https://www.spiegel.de/netzwelt/ki-forscherin-hannah-bast-die-menschheit-kreiert-derzeit-eine-intelligentere-spezies-podcast-a-b9b78548-b47c-46ff-86d1-c1a7ed2069fd*, 30.08.2023 [28.05.2024]

4.3 Von Text zu Multimodalität: Wenn KI mehr als Worte versteht

Stellen Sie sich vor, Sie kommen aus einem Workshop zurück. Im Gepäck: ein Foto vom vollgeschriebenen Whiteboard und eine Audiodatei mit einem Experten-Interview. Noch vor zwei Jahren hätten Sie das alles mühsam abtippen und transkribieren müssen. Heute laden Sie es einfach in Ihre KI hoch, und sie versteht alles davon. Sie übernimmt das Transkribieren der Audiodatei und fasst die Notizen für Ihre PowerPoint-Präsentation strukturiert zusammen. Diese Fähigkeit der KI nennt sich *Multimodalität*, und sie ist der vielleicht folgenreichste technologische Sprung der letzten zwei KI-Jahre.

Die heutigen Modelle verstehen nämlich nicht mehr nur geschriebene Sprache, sondern auch Bilder und gesprochenes Wort. Zeigen Sie der KI ein Foto, und sie beschreibt, was darauf zu sehen ist. Laden Sie eine komplette PDF-Präsentation hoch, und sie fasst den Inhalt zusammen, beantwortet Fragen dazu oder erstellt daraus Social-Media-Posts samt Bildvorschlag. Selbst Sprachnachrichten versteht sie und antwortet darauf in Echtzeit mit natürlich klingender Stimme. Diese Art »multimodale Übersetzung« funktioniert dabei in alle Richtungen: von Bild zu Text genauso wie von Audio zu Blogpost. Und die Modelle werden zunehmend besser darin, auch Video zu verstehen und zu generieren. So können Sie beispielsweise einen Werbespot, den Sie auf YouTube gesehen haben, von Googles Gemini ansehen lassen und von der KI etwa auf Zielgruppentauglichkeit oder Zeitgemäßheit einschätzen lassen.

Ein illustratives Beispiel für multimodales Arbeiten mit Gemini

Stellen Sie sich eine kurze Videosequenz für Ihr Projekt vor: Eine Frau telefoniert mit einem Freund und erzählt eine Geschichte aus dem Urlaub. Ein simples Szenario.

Abbildung 4.4: Multimodal arbeiten in einem Modell: Googles Gemini bietet fast »alles unter einem Dach«. (Quelle Bild: Screenshot Gemini)

In Abbildung 4.4 sehen, was Sie mit Google Gemini Anfang 2026 daraus machen können – aus einer einzigen Plattform heraus, von dort, wo Sie wahrscheinlich die meiste Zeit verbringen:

- Die Videosequenz erzeugen Sie mit Google und seinem aktuellem Gemini-Videomodell (*Veo*-Familie).
- Die Urlaubsfotos bauen Sie mit dem integrierten Bildgenerator *Nano Banana* ein.
- Den Soundtrack komponiert das neue Musikmodell *Lyria* ebenfalls direkt in der Gemini-App.
- Und das Skript für den Dialog? Das schreiben Sie mit *Gemini* als Sprachmodell.

Text, Bild, Video und Musik entstehen nicht mehr in vier verschiedenen Tools mit vier verschiedenen Workflows. Sie entstehen, wenn Sie möchten, an einem Ort. Das ist Multimodalität in der Praxis: nicht nur verschiedene Formate verstehen, sondern sie auch kontextuell erzeugen und miteinander verbinden.

Google war mit Gemini eines der ersten großen Modelle, das von Beginn an konsequent als multimodales System (Text, Bild, Audio, Video) positioniert wurde und die Features nun bereits in einer Oberfläche bündelt. Andere Modelle wie ChatGPT haben Bild- und Videomodelle schrittweise integriert: Aus DALL·E wurde das heute standardmäßige GPT-Image-1.5, und Video-Modelle wie Sora sind in ChatGPT und professionelle Workflows eingebunden – oft zusätzlich in Social-Media-Plattformen präsent. Noch bieten nicht alle Plattformen den vollen Umfang wie Gemini, aber die Richtung der Modellentwicklung ist klar vorgezeichnet.

Das ist eine Momentbeschreibung, denn das Tempo dieser Entwicklung bleibt weiterhin hoch: Als OpenAI im März 2023 GPT-4 vorstellte, war allein die Fähigkeit, Bilder zu »verstehen«, eine Sensation. Keine drei Jahre später ist die Erzeugung von Text, Bild, Video und Musik aus einer Hand, sprich einem KI-Modell, möglich.

Für Sie als Content Creator verändert diese Fähigkeit die Arbeit grundlegend. Denken Sie bei Ihrem nächsten Projekt bewusst darüber nach, welche Quellen Sie bisher nicht genutzt haben, weil die Aufbereitung zu aufwendig war.

In den folgenden Kapiteln dieses Buchs werden Sie die multimodalen Fähigkeiten der KI in der Praxis kennenlernen: beim Texten (Kapitel 5, »Texte mit KI schreiben, optimieren und zusammenfassen«), bei der Bildgenerierung (Kapitel 7, »Promptografie – Bilder mit KI generieren, variieren und steuern«), bei Audio (Kapitel 8, »Audio und Musik mit KI produzieren und optimieren«), bei Video (Kapitel 9, »Video mit KI konzipieren, generieren und produzieren«) und beim Design (Kapitel 10, »Gestalten mit KI-Design-Plattformen«).

Die Grundlage dafür ist immer dieselbe: ein Sprachmodell, das gelernt hat, die Welt und ihre Zusammenhänge nicht nur in Wörtern, sondern auch in Bildern und Klängen zu verstehen.

Merke: Multimodalität macht die KI vom reinen Textassistenten zum universellen Kreativwerkzeug. Ihre Quellen sind nicht mehr auf geschriebenen Text beschränkt. Alles, was Sie sehen oder hören können, wird zum Rohmaterial für Ihre Content-Arbeit.

Schauen wir uns nun die wichtigsten Modelle, deren Fähigkeiten, Stärken, Schwächen und Unterschiede genauer an.

4.4 OpenAI und ChatGPT: Der Pionier, der alles will

Als OpenAI im November 2022 ChatGPT veröffentlichte, konnte zum ersten Mal jeder Mensch ohne Programmierkenntnisse mit einem großen Sprachmodell sprechen. Fünf Tage, eine Million Nutzerinnen und Nutzer; zwei Monate, hundert Millionen. Dieser Moment löste nicht nur eine ganze Industrie aus, sondern markiert den Beginn der Disruption für viele andere Industrien: Google beschleunigte fortan Gemini, Anthropic brachte Claude, und rund um den Globus entstanden neue KI-Start-ups. Inzwischen nähert sich ChatGPT der Marke von einer Milliarde wöchentlicher Nutzerinnen und Nutzer und ist damit die mit Abstand größte KI-Plattform der Welt. OpenAIs Strategie war von Anfang an klar: ChatGPT soll die eine Plattform für alles und für alle sein. Und diese Strategie prägt jede Funktion, die OpenAI baut.

Was ChatGPT kann

ChatGPT hat sich von einem Chatbot zu einer integrierten KI-Arbeitsplattform entwickelt. Wenn Sie die Plattform öffnen, finden Sie dort nicht nur ein Textfenster, in dem Sie texten können, sondern ein ganzes Werkzeugarsenal:

- *Canvas* öffnet neben dem Chat ein Arbeitsfenster, in dem Sie gemeinsam mit der KI an Texten oder Code arbeiten können. Sie können einzelne Passagen markieren und gezielt überarbeiten lassen, mit Versionskontrolle und Inline-Vorschlägen (siehe dazu auch Kapitel 5, »Texte mit KI schreiben, optimieren und zusammenfassen«).
- *Memory* speichert bevorzugte Informationen und wiederkehrende Angaben über mehrere Chats hinweg. Sie müssen nicht mehr in jedem neuen Gespräch erklären, wer Sie sind und wie Sie arbeiten.
- *Deep Research* lässt die KI eigenständig Webquellen durchsuchen und daraus fundierte Berichte mit Belegen erstellen.
- Eine native *Bildgenerierung* mit *GPT-Image* ist direkt ins Sprachmodell integriert und hat das frühere DALL-E im Frontend ersetzt. Das Ergebnis sind eine deutlich bessere Bildqualität, sehr gute Textwiedergabe im Bild und konversationelles *Editing* (»Zeige die gleiche Szene aus der Drohnenperspektive«). Dazu wird das Weltwissen in die Bildentwicklung integriert, was Content Creation in neue Dimensionen führt.

- *Projects* ermöglichen es, Arbeitsstränge projektweise zu organisieren (mehr dazu in Abschnitt 12.12, »Arbeiten in Projekten«), mit eigenen Anweisungen und bereitgestellten Dateien über mehrere Chats hinweg.
- *Advanced Voice Mode* ermöglicht natürliche Sprache und Gespräche.
- Ein wachsender Katalog an *Apps* ergänzt ChatGPT um externe Funktionen und Integration für spezifische Anwendungsfälle.
- *Custom GPTs* ermöglichen die Anpassung von ChatGPT an definierte Aufgaben, Inhalte oder Arbeitsstile.
- Mit *Vision* haben sich die Bildverarbeitungsfähigkeiten von ChatGPT von einer einfachen Bildanalyse zu einer Echtzeit- und Hochgeschwindigkeits-Bildverarbeitung samt Live-View und -Interaktion weiterentwickelt.
- Im *Agentenmodus* kann das Modell eigenständig mehrstufige Aufgaben planen und ausführen, statt nur auf einzelne Fragen zu antworten.

Tipp: Vermeiden Sie die »Automatik-Falle«

ChatGPT wählt das Modell im Hintergrund je nach Komplexität Ihrer Frage automatisch aus und entscheidet dabei, ob es ein schnelleres *Standard-Modell* oder ein leistungsfähigeres *Reasoning*-Modell verwendet.

Das klingt komfortabel, hat aber einen Haken: Die automatische Auswahl priorisiert oft Geschwindigkeit und Kosten vor maximaler Tiefe. Für eine kurze Frage mag das ausreichen. Für professionelle Content-Arbeit sollten Sie aber nach Möglichkeit das jeweils beste verfügbare Modell mit aktiviertem *Thinking-Modus* wählen. Der Unterschied in der Antworttiefe kann deutlich sein. OpenAI macht das leider nicht gerade transparent. Im Zweifel gilt: Wenn eine Antwort oberflächlich wirkt, prüfen Sie daher unbedingt, welches Modell gerade für Sie arbeitet.

Atlas: Wenn der Browser zur KI wird

Im Oktober 2025 hat OpenAI mit ChatGPT *Atlas*[12] einen eigenen *KI-Browser* gelauncht. Chromium-basiert, mit ChatGPT als Kern, zunächst nur für macOS (Windows und mobile Versionen sind angekündigt). Das Besondere: Atlas versteht, anders als ein herkömmlicher Browser, was Sie gerade im Web tun. Der KI-Assistent sitzt als Sidebar neben jeder Webseite und kann deren Inhalt zusammenfassen, Ihre Fragen dazu beantworten oder Daten zusammenfassen. Im *Agent Mode* kann Atlas sogar eigenständig im Web navigieren, etwa Formulare ausfüllen oder Preise vergleichen und sogar Quiz und Spiele für sie regelrecht »durchklicken«. Sie schauen dem Cursor einfach dabei zu und greifen nur ein, wenn nötig.

12 *https://chatgpt.com/de-DE/atlas/*

Atlas speichert auf Wunsch auch *Browser Memories*: Kontext von Webseiten, die Sie besucht haben, der in späteren Chats dann einfach wieder auftaucht.

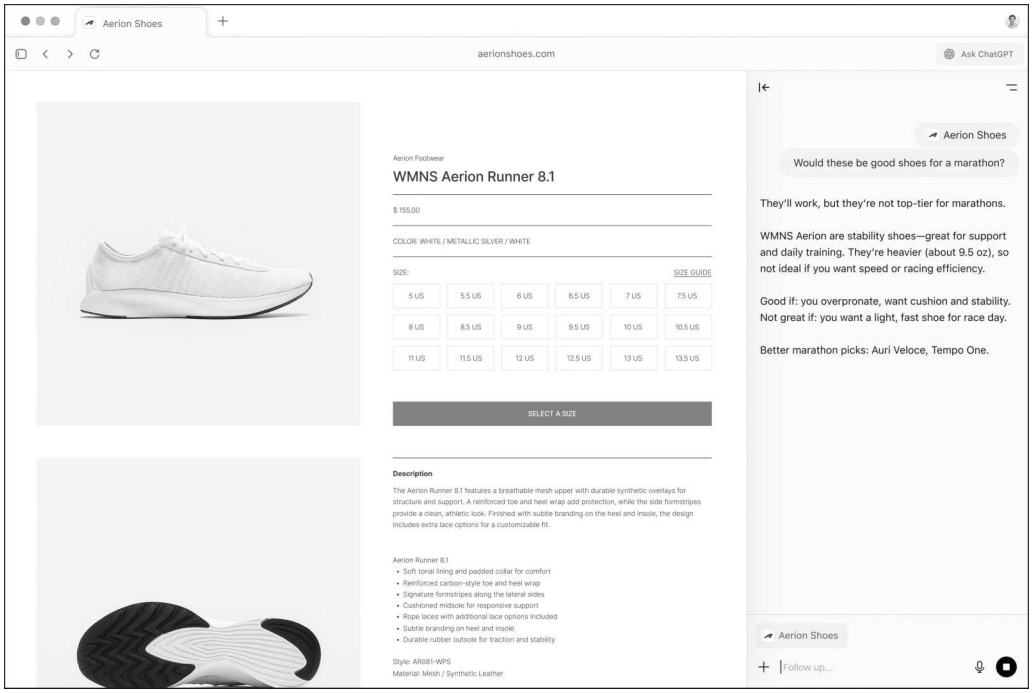

Abbildung 4.5: In Atlas öffnet sich die Seitenleiste von ChatGPT in einem beliebigen Fenster, um Artikel zusammenzufassen, Produkte zu vergleichen oder Daten von jeder Website, die Sie gerade ansehen oder gerade geschlossen haben, zu analysieren. (Bild: OpenAI)

> *Beispiel-Prompt: »Finde alle Stellenanzeigen, die ich letzte Woche angesehen habe, und erstelle eine Zusammenfassung der Branchentrends.«*

Eine solche Anfrage funktioniert dann, weil Atlas sich erinnert. Das ist mächtig, erfordert aber auch bewusstes Datenschutzbewusstsein: Sie sollten genau wissen und festlegen, was Atlas sehen und speichern darf.

Stärken und Schwächen

ChatGPTs größte Stärke ist die Vielseitigkeit. Kaum ein anderer Anbieter hat so viele Funktionen in einer einzigen Plattform gebündelt. Wenn nächste Woche ein neues KI-Feature auf den Markt kommt, hat ChatGPT es wahrscheinlich als Erstes. Das ist gleichzeitig die größte Schwäche: Die Oberfläche wird zunehmend unübersichtlich. Versteckte Modellauswahl, verschachtelte Menüs, es braucht Einarbeitung, um das Beste herauszuholen. Ob Sie eine neue Funktion dann auch finden, steht auf einem anderen Blatt.

Seit Anfang 2026 testet OpenAI zudem Werbung in den kostenlosen und günstigen Tarifen. Nur Plus-, Pro- und die Business-Tarife sollen werbefrei bleiben.[13]

Merke: ChatGPT ist die vielseitigste KI-Plattform am Markt, und vielleicht auch die unübersichtlichste. Wer sich die Zeit nimmt, Projects, Memory und die richtige Modellauswahl zu beherrschen, bekommt aber ein außerordentlich leistungsfähiges Werkzeug. Wer einfach drauflos chattet, kratzt an der Oberfläche.

Wie Sie sich anmelden und mit ChatGPT Ihre ersten Textversuche starten können, erfahren Sie ausführlich in Abschnitt 5.2, »So richten Sie Ihre KI für die langfristigere Zusammenarbeit ein«.

4.5 Anthropic und Claude: Tiefe statt Breite

Anthropic wurde 2021 von Dario und Daniela Amodei gegründet, Geschwister, die zuvor bei OpenAI in Führungspositionen gearbeitet hatten und das Unternehmen verließen, weil sie einen stärkeren Fokus auf KI-Sicherheit und Zuverlässigkeit setzen wollten. Diese Gründungsgeschichte prägt das Produkt bis heute. Wo ChatGPT versucht, alles gleichzeitig zu sein, konzentriert sich Claude[14] auf die vordefinierte Kernarbeit. Und macht die bemerkenswert gut.

Was Claude besonders macht

Claude hat sich von Beginn an als Werkzeug für besonders anspruchsvolle Textarbeit etabliert. Die Qualität bei längeren, stilistisch differenzierten Texten und beim sorgfältigen Befolgen nuancierter Anweisungen ist spürbar höher als bei den meisten Wettbewerbern. Wo ChatGPT manchmal dazu neigt, schnell und oberflächlich zu antworten, nimmt sich Claude eher die Zeit, präzise zu sein. Viele professionelle Autorinnen und Creator, Juristinnen und Analysten bevorzugen Claude deshalb für Aufgaben, bei denen es auf sprachliche Feinarbeit ankommt.

Dass Claude anders klingt als die Konkurrenz, hat einen konkreten Grund. Anthropic legt großen Wert darauf, dass Claude verantwortungsvoll und sicher agiert. Das spiegelt sich nicht nur im Training wider, sondern auch in der Art, wie Claude mit sensiblen Themen umgeht. Im Januar 2026 hat Anthropic eine neue *Constitution*[15] veröffentlicht, ein 80-seitiges Grundsatzdokument, das beschreibt, wer Claude ist, wie es sich verhalten soll und warum. Der entscheidende Punkt: Anthropic lehrt Claude nicht mehr, *was* es tun soll, sondern *warum*. Die Idee dahinter: Wenn eine KI die Gründe hinter den Regeln versteht, kann sie auch in Situationen richtig handeln, die keine Regel vorhersieht. Anthropic defi-

13 *https://openai.com/de-DE/index/our-approach-to-advertising-and-expanding-access/*, 16.01.2026 [14.02.2026]

14 *https://claude.ai/*

15 *https://www.anthropic.com/constitution*

niert damit eine Vier-Stufen-Hierarchie: Zuerst Sicherheit, dann Ethik, dann Anthropics Richtlinien, und erst an vierter Stelle Hilfsbereitschaft. (In der Praxis landen, so scheint es, fast alle Gespräche bei Stufe vier.)

Aber wenn es heikel wird, greifen die höheren Stufen. Anthropic hat die Constitution unter einer Creative-Commons-Lizenz veröffentlicht und möchte damit ein Signal setzen: Transparenz ist wichtiger als Geheimhaltung.

Auch Claude hat trotz seiner Textfähigkeit mehr zu bieten als das Textfenster:

- Mit *Extended Thinking* lässt sich auch Claude in einen Denkmodus versetzen, in dem das Modell eine quasi nachvollziehbare Gedankenkette zeigt, bevor es antwortet. Das ist besonders bei komplexen Analysen, strategischen Entscheidungen und wissenschaftlichem Arbeiten ein spürbarer Qualitätsgewinn, und es macht die Denkwege der KI transparent und überprüfbar. Das Flaggschiff-Modell Opus zusammen mit aktiviertem Extended Thinking gehört zu den leistungsfähigsten Modellen am Markt.
- Wie bei ChatGPT gibt es auch bei Claude *Projects*, um Arbeitsräume mit eigenen Anweisungen, Dateien und projektspezifischem Gedächtnis zu organisieren.
- Mit *Skills* können Sie in Ihren **Einstellungen** Claude wiederkehrende Arbeitsanweisungen beibringen, etwa Ihren Markenton oder redaktionelle Leitlinien. Einmal definiert, auch mit der Unterstützung der KI selbst, wendet Claude diese Vorgaben in jedem passenden Gespräch automatisch an. Praktisch für Content-Teams, die konsistent in einer Markensprache arbeiten.
- *Deep Research* eignet sich für umfassende, quellenbasierte Rechercheberichte.
- Der Canvas-Modus von ChatGPT nennt sich hier *Artefakt*. Allerdings wählt die KI am liebsten selbst den Moment, wann er genutzt wird, lässt sich aber auch anstoßen.
- Claude kann, wie die anderen Modell ChatGPT und Gemini auch, *Code schreiben* und in der Artefakt-Ansicht direkt daneben gut nachvollziehbar visuell ausführen, von einzelnen Skripten bis hin zu kompletten Webanwendungen. Dabei müssen Sie noch nicht einmal selbst coden können. Sie sagen der Maschine in Ihrer menschlichen Sprache, was Sie möchten. Sie setzt um. Sie iterieren mit Worten. Sie setzt diese Anweisungen wieder in Code um. Man spricht hier auch vom *Vibe-Coding*.

  Früher hieß es einmal heroisch: »Wer coden kann, dem gehört die Welt!« Nun, dann ist mit KI wohl auch diese Weltordnung ins Wanken gekommen.

  Wer an dieser Stelle noch tiefer einsteigen möchte, findet mit *Claude Code* eine eigenständige Anwendung, die direkt in der Entwicklungsumgebung arbeitet: Sie liest ganze Codebasen, behebt Fehler, führt Tests aus und verwaltet Projekte, ohne dass Sie jedes Mal einen Prompt tippen müssen. Auch hier gilt: Statt komplizierte Coding-Syntax zu lernen oder anzuwenden, beschreiben Sie einfach in Ihrer Sprache, was Sie brauchen.
- *Claude in Excel* und *Claude in PowerPoint* binden die KI direkt in die Microsoft-Office-Umgebung ein. Sie können in einer Tabelle oder Präsentation mit Claude arbeiten,

ohne die Anwendung zu verlassen. Mit Claude können Sie direkt aus dem Chat heraus fertige Arbeitsdokumente zu produzieren.

- Sie können ein Besprechungsprotokoll einfügen, das Claude dann als *Word-Datei* oder als schön gestylte *Powerpoint-Präsentation* im gewünschten Corporate Design, z. B. auf Basis Ihrer Vorlagen, zusammenfasst. Damit ist endlich Schluss mit der zeitaufwendigen und nervigen PPTX-Pfeilchenschieberei!
- Claude bietet eine Anbindung an Microsoft 365. Über den Microsoft-365-Connector[16] kann das Modell auf ausgewählte Unternehmensdaten und Arbeitsumgebungen wie SharePoint, OneDrive, Outlook und Teams zugreifen, um die dort abgelegten Dokumente zu analysieren, E-Mails-Threads aufzurufen oder Erkenntnisse aus Team-Chats zu gewinnen.

Best Practice: Der Game-Changer für Webprojekte

Wie diese Bausteine in der Praxis zusammenspielen, zeigt sich insbesondere bei Webprojekten. Claude eignet sich als Entwicklungspartner für komplette Websites:

Aus einem gemeinsam erarbeiteten Konzept kann im Chat ein klickbarer HTML-Prototyp entstehen, der Schritt für Schritt zur fertigen Seite – mit Unterseiten und fertigen Texten – weiterentwickelt wird – bis hin zur WordPress-Integration. Was früher Wochen an komplizierter Abstimmung zwischen Redaktion, Konzeption und Entwicklung kostete, lässt sich jetzt in wenigen Tagen umsetzen.

Der entscheidende Vorteil liegt im neu ermöglichten Workflow der Menschen dahinter: Anstelle von geskribbelten Mockups oder Beschreibungstexten liefert die Konzepterin mit Claude direkt fertigen, lauffähigen HTML-Code samt passender CSS-Datei, sodass alle Projektbeteiligten das gewünschte Ergebnis sofort im Browser sehen können. Der Entwickler zerlegt diesen Prototyp anschließend in die nötigen WordPress-Templates (z. B. Header, Footer, Inhaltsvorlagen) und ersetzt statische Inhalte durch dynamische WordPress-Funktionen. Er muss das Konzept nicht mehr interpretieren und nachbauen, sondern portiert das funktionierende Ergebnis in die WordPress-Struktur. Das klassische »Lost in Translation« zwischen Konzeption und Entwicklung entfällt – und mit ihm ein Großteil der ineffizienten Abstimmungsschleifen.

Claude Cowork: Der Agent für Wissensarbeit

Was Claude Anfang 2026 von der Konkurrenz abhebt, ist *Cowork*[17]. Es startete Anfang Januar 2026 als Research Preview, zunächst für Mac, inzwischen auch für Windows.

16 *https://support.claude.com/de/articles/12542951-microsoft-365-connector-aktivieren-und-verwenden*

17 *https://support.claude.com/de/articles/13345190-erste-schritte-mit-cowork*

Cowork verwandelt Claude von einem Chatbot in einen Desktop-Agenten. Sie geben Claude Zugriff auf einen Ordner auf Ihrem Computer, und es kann darin Dateien lesen, analysieren und bearbeiten. Nicht eine Datei pro Prompt, sondern ganze Workflows:

> *Beispiel-Prompt: »Nimm die 30 Kundenfeedback-Mails, identifiziere die fünf häufigsten Themen, und erstelle für jedes Thema einen Blogpost-Entwurf als Word-Datei …«*

Claude erstellt daraufhin einen Plan, unterteilt ihn in Teilaufgaben, arbeitet diese parallel ab und präsentiert Ihnen das Ergebnis. Sie können jederzeit eingreifen oder die Kontrolle übernehmen.

Kurz nach Einführung von Claude Cowork kamen sogenannte Plug-ins hinzu: elf Open-Source-Erweiterungen für unterschiedliche Berufsfelder wie Vertrieb oder Marketing. Ein Plug-in bündelt dabei jeweils Fachwissen und vordefinierte Workflows für ein bestimmtes Aufgabengebiet. Diese Plug-ins lassen sich individuell anpassen oder komplett neu erstellen, ganz ohne Programmierkenntnisse (ausführlicher dazu siehe Abschnitt 11.3, »Cowork-Plug-ins – wenn aus dem virtuellen Assistenten ein »echter« Kollege wird«).

Stärken und Grenzen

Claudes Stärke liegt erfahrungsgemäß in der Tiefe: Textqualität und Anweisungstreue. Das Agenten-Ökosystem mit Cowork (siehe dazu auch Abschnitt 11.3, »Cowork-Plug-ins – wenn aus dem virtuellen Assistenten ein »echter« Kollege wird«) und einem offenen MCP-Standard ist zum Redaktionsschluss des Buches das wahrscheinlich durchdachteste am Markt. Für Content-Schaffende, die auf sprachliche Qualität und systematische Workflows setzen, ist Claude die »natürliche« Wahl.

Aber es hat auch Grenzen: Stand April 2026 bietet Claude außer einem eigenen Design-Tool für Präsentationen oder Prototypen weiterhin keine Bildgenerierung oder Videoproduktion an und ist daher nicht so multimodal aufgestellt wie Gemini oder ChatGPT. Anthropic baut weniger Funktionen, aber die wichtigen nach Einschätzung der Autoren besser. Das ist Stärke und Schwäche zugleich. Wer ein Schweizer Taschenmesser braucht, ist dann eben bei ChatGPT besser aufgehoben.

Merke: Claude stellt Qualität über Vielseitigkeit. Wer präzise Texte und Konzepte braucht, findet hier sein passendes Angebot. Wer alles in einer Plattform will, wird woanders suchen.

4.6 Google und Gemini: Der Integrator

Google ist der Technologieriese, der den wahrscheinlich größten Datenschatz der Welt besitzt und dessen KI-Modelle am tiefsten in ein bestehendes Ökosystem eingebettet sind. Der bahnbrechende Transformer-Ansatz, auf dem alle heutigen Sprachmodelle ba-

sieren, wurde 2017 bei Google entwickelt. Dass ausgerechnet OpenAI den Durchbruch in der öffentlichen Wahrnehmung schaffte, war für Google ein strategischer Weckruf. Seitdem investiert das Unternehmen massiv und hat nach einigen Irrfahrten mit Gemini 3 Pro zur Spitzengruppe aufgeschlossen.

Die Workspace-Integration: wenn Gemini Ihren Kontext kennt

Was *Gemini* von ChatGPT und Claude fundamental unterscheidet, ist nicht das Modell, sondern der Kontext. Gemini ist direkt mit Gmail, Google Docs, Sheets, Slides, Drive und Calendar verknüpft. Wer bereits in der Google-Welt arbeitet, bekommt also einen Junior an die Seite, der den eigenen Arbeitskontext kennt, ohne dass Sie erst Dokumente hochladen müssen.

Das zeigt sich besonders eindrücklich bei Deep Research: Während die Deep-Research-Funktionen von ChatGPT und Claude ausschließlich das Web durchsuchen, kann Geminis *Deep Research* seit Ende 2025 auch Ihre Workspace-Inhalte einbeziehen, also E-Mails, Dokumente auf Drive und Teamchats.

> *Beispiel-Prompt: »Erstelle eine Konkurrenzanalyse, die öffentliche Webdaten mit unseren internen Strategiememos abgleicht.«*
>
> *Gemini liefert einen Bericht, der beides verbindet. Kein anderer Anbieter kann das in dieser Tiefe.*

In den Workspace-Apps selbst sitzt Gemini als Seitenpanel in Gmail, Docs, Sheets und Slides. Sie können E-Mails zusammenfassen lassen, Antworten entwerfen oder Daten analysieren, alles direkt im Workflow. Seit Anfang 2025 sind diese KI-Funktionen in die regulären Workspace-Abonnements integriert, nicht mehr als kostenpflichtiges Add-on.

NotebookLM: Das Recherche-Labor

NotebookLM ist Googles spezialisiertes Recherchewerkzeug und ähnelt ein wenig dem Projekt- oder Project-Ansatz der anderen Anbieter: Sie laden Ihre Quellen in Form von Daten in ein *Notebook* hoch, also etwa Fachartikel und Studien, und erhalten damit eine interaktive Wissensbasis, mit der sie kommunizieren können. Die bekannteste Ausgabefunktion sind *Audio Overviews*, KI-generierte Podcasts, in denen zwei Hosts Ihr Material diskutieren. Seit 2026 können Sie per *Interactive Mode* sogar dazwischenrufen und Rückfragen stellen. Weitere Ausgabeformate wie Video Overviews, Quizz, Infografiken und automatisch generierte Slide-Decks machen NotebookLM zunehmend auch zum Produktionswerkzeug. Seit Januar 2026 lässt sich ein Notebook als Quelle in die Gemini-App einbinden, sodass Ihre Recherche direkt in die Weiterverarbeitung fließt.

Weitere wichtige Werkzeuge

- *Canvas* funktioniert bei Gemini ähnlich wie bei den anderen Modellen: ein Arbeitsbereich neben dem Chat, in dem Sie an Dokumenten feilen können.
- *Gems* sind Googles Variante der Custom GPTs, spezialisierte Mini-Agenten für wiederkehrende Aufgaben, die Sie einmal konfigurieren und dann immer wieder einsetzen.
- *Veo* generiert Videos direkt in der Gemini-App und ist neben Sora bei ChatGPT die zweite große Plattform mit integrierter Videoproduktion.
- *Nano Banana* als integriertes Bild-Tool, das 2025 über Nacht neue Maßstäbe an Bildqualität und Multimodalität setzte. So können Sie beispielsweise mehrere Bilddateien hochladen und deren Bestandteile in einem Bild kombinieren lassen oder Infografiken mit vielen Textbestandteilen generieren.
- Gemini in Chrome schließlich bringt die KI direkt in den *Browser* und wird damit zu einer Art KI-Browser wie Atlas, bisher allerdings nur in den USA und auf Englisch.

Stärken und Grenzen

Geminis größte Stärke – neben der bereits erläuterten nativen Multimodalität – ist die tiefe Verankerung im Arbeitsalltag: Wer in der Google-Welt arbeitet und »lebt«, bekommt eine KI, die den kürzesten Weg vom Gedanken zum Ergebnis bietet, weil der Kontext bereits da und immer griffbereit ist; Deep Research mit Workspace-Zugriff, NotebookLM als Recherchewerkzeug, Gemini im Seitenpanel jeder Google-App. Gemini ist ein Ökosystem, das die Konkurrenz so schnell nicht nachbauen kann.

Die Schwäche: Googles Modelle haben bei der reinen Textqualität, besonders bei längeren, stilistisch anspruchsvollen Texten, traditionell hinter Claude und teilweise auch ChatGPT gelegen – später auch mehr zu den bestehenden ethischen Herausforderungen des Modells. Gemini 3 Pro hat diesen Abstand zwar wahrnehmbar verkürzt, aber bei der sprachlichen Feinarbeit, die für viele Content Creator zählt, ist er noch spürbar. Auch wirkt die Chat-Oberfläche weniger ausgereift als die der Wettbewerber. Google kann Modelle bauen, aber Produktdesign war nie seine erste Disziplin.

Merke: Gemini ist die KI für alle, die bereits in der Google-Welt arbeiten. Die Workspace-Integration, die Verbindung von Deep Research mit Ihren eigenen Daten und NotebookLM als Recherche-Labor machen es zum natürlichen Assistenten für Teams, die auf Google setzen. Wer woanders arbeitet, hat weniger davon.

4.7 Perplexity – wenn Recherchequalität zählt

Perplexity[18] ist anders als ChatGPT, Claude und Gemini. Während letztere versuchen, möglichst vieles gleichzeitig zu können, hat sich Perplexity von Beginn an auf eine einzige

18 *https://www.perplexity.ai*

Sache konzentriert, und die macht es außerordentlich gut: Recherche. Die anderen Funktionen wurden zwar sukzessive aufgerüstet, laufen aber eher so mit.

Das Grundsatzprinzip ist schnell erklärt: Perplexity ist eine Antwortmaschine! Sie stellen eine Frage, die KI durchsucht das Web fast in Echtzeit, liest die relevanten Quellen rasend schnell aus und liefert Ihnen eine zusammengefasste Antwort mit nummerierten Quellenangaben direkt im Text. Jede Aussage ist leicht nachprüfbar. Das ist ein fundamentaler Unterschied zu den großen Chatbots: Wenn Sie beispielsweise Claude etwas fragen, antwortet dieser primär aus ihrem Trainingsgedächtnis, und Sie wissen nie so ganz genau, woher die Antwort nun stammt. Sie müssen nachfragen. Bei Perplexity sehen Sie es sofort.

Das Besondere dabei: Perplexity baut keine eigenen Frontier-Modelle. Unter der Motorhaube arbeiten, je nach Einstellung und Abo, Modelle von OpenAI, Anthropic oder Google. Perplexity hat stattdessen alles in die Umgebung investiert, in der diese Modelle arbeiten: den Echtzeit-Webzugriff, die Quellenauswahl, die Zusammenfassung, die Darstellung. Es wird zu einem Lehrstück für die gesamte Branche: Nicht das Modell allein macht den Unterschied, sondern was man drumherum baut.

Das Model Council – wenn drei Modelle gemeinsam nachdenken

Anfang 2026 hat Perplexity diesen Gedanken konsequent weitergeführt und das sogenannte *Model Council*[19] eingeführt. Statt sich für ein Modell entscheiden zu müssen, lässt Perplexity Ihre Anfrage gleichzeitig von drei aktuellen Frontier-Modellen bearbeiten, etwa Claude Opus, GPT und Gemini. Ein viertes Modell fungiert als »Synthesizer«, vergleicht die Antworten und zeigt, wo die Modelle übereinstimmen und wo sie voneinander abweichen. Der Clou: Jedes Sprachmodell hat trainingsbedingt blinde Flecken. Wo die Modelle konvergieren, können Sie der Antwort stärker vertrauen. Wo sie divergieren, wissen Sie, dass Sie genauer hinschauen sollten.

Viele erfahrene Creator machen genau diesen Workflow längst manuell: Arbeiten mit dem einen Modell, anschließender Check durch Perplexity. Nun hat Letzteres diesen Workflow automatisiert. Das Model Council ist aktuell den Max-Abonnenten vorbehalten (200 Dollar pro Monat), aber es zeigt die Richtung, in die KI sich entwickeln könnte:

Im nächsten Schritt verknüpft *Perplexity Computer* immerhin schon 19 verschiedene Sprachmodelle in einer durchgehenden Plattform für Software-Projekte – von der ersten Idee bis zur fertigen Anwendung. Die Plattform lässt dabei mehrere autonome Agenten parallel arbeiten. Das integrierte System *Opus* analysiert jede Aufgabe und weist sie automatisch dem Modell zu, das dafür am besten geeignet ist. Der Speicher sichert vergangene Projekte und Voreinstellungen. Dazu bindet das System externe Dateien und Web-Inhalte direkt über zahlreiche Schnittstellen ein. User können so von einer einfachen Aufgabe auf hunderte parallele Projekte skalieren.

19 *https://www.perplexity.ai/de/hub/blog/introducing-model-council*

Der naheliegende Rückschluss dieser Entwicklung: Die Zukunft gehört gar nicht dem einen besten Modell, sondern dem intelligenten Zusammenspiel mehrerer Modelle – und dem, der es versteht, sie zu orchestrieren.

Spaces, Comet und das wachsende Ökosystem

Perplexity hat sich in kurzer Zeit von einer reinen Such- und Antwortmaschine zu einer Recherche-Plattform entwickelt.

- *Spaces*, auf Deutsch *Räume*, ermöglichen es Ihnen, Rechercheprojekte projektartig zu organisieren, etwa alle Quellen und Notizen für einen Fachartikel an einem Ort zu sammeln, ähnlich wie Projects und Notebooks bei den anderen Plattformen. Das ist praktisch für Content Creator, die an mehreren Projekten gleichzeitig arbeiten.
- Mit dem *Comet-Browser*[20], der seit Juli 2025 verfügbar und inzwischen kostenlos für alle ist, geht Perplexity ähnlich wie OpenAI mit Atlas noch einen Schritt weiter: Comet ist ein Chromium-basierter Browser mit eingebautem KI-Assistenten, der Sie beim Surfen begleitet: Auch er kann Webseiten für Sie zusammenfassen, Fragen zum aktuellen Tab beantworten oder sogar eigenständig im Web navigieren, etwa Preise vergleichen oder Flüge suchen. Für Max-Abonnenten gibt es einen *Background Assistant*, der mehrere Aufgaben gleichzeitig im Hintergrund erledigt, während Sie weiterarbeiten. Das ist kein Chatbot mehr, das ist ein Assistent, ein Agent, der für Sie handelt. Dies hat große Auswirkungen auch auf die SEO- und GEO-Branche, wie in Kapitel 6, »Von SEO zu GEO – Content für KI-Suchmaschinen«, ausführlicher erklärt wird.

Perplexity wächst dabei zunehmend über die reine Recherche hinaus: Rechercheergebnisse lassen sich direkt in formatierte fertige Berichte, Präsentationen oder Dashboards gießen und exportieren, und die Bildgenerierung u. a. mit GPT Image oder Nano Banana und die Videogenerierung (in Max-Abo) sind integriert. Die Stärke liegt aber weiterhin dort, wo Perplexity angefangen hat: beim quellenbasierten Arbeiten mit den besten verfügbaren Modellen.

Perplexity Pro, Max und die Deep-Research-Frage

Die kostenlose Version von Perplexity ist für gelegentliche Recherchen ausreichend. Die Pro-Version schaltet leistungsfähigere Modelle frei und bietet Pro Search, eine gründlichere Rechercheform, bei der die KI sich zunächst selbst klärende Fragen stellt und dann gezielt mehrere Dutzend Quellen durchsucht. Max ergänzt das Model Council, den vollen Comet-Funktionsumfang und Zugang zu den stärksten verfügbaren Modellen.

Seit Ende 2024 bieten auch ChatGPT, Claude und Gemini Deep Research-Funktionen an. Wo liegt dann noch der Vorteil von Perplexity? Der Unterschied ist die Arbeitsweise: Deep Research ist ein Modus für besonders aufwendige Fragen, der mehrere Minuten braucht

20 *https://www.perplexity.ai/comet*

und lange Berichte liefert. Perplexity dagegen ist eine regelrechte »Recherche-Maschine«: schnelle Antworten mit Quellen, bei jeder Frage, sofort. In der Praxis nutzen viele Content-Schaffende beides: Standard für die laufende Faktenprüfung, Deep Research für die »Tiefenbohrung« bei komplexen Projekten.

Warum Perplexity für Content Creator relevant ist

Für alle, die professionell mit Inhalten arbeiten, löst Perplexity ein konkretes Problem: die quellenbasierte Recherche. Perplexity durchsucht dabei nicht nur das offene Web, sondern auch *akademische Datenbanken* für Fachpublikationen und *Social-Media-Plattformen*, was es in vielen Fällen zu einer schnellen Alternative zu klassischen Social-Media-Listening-Tools macht.

Wenn Sie einen Fachartikel oder eine Präsentation vorbereiten, brauchen Sie überprüfbare Antworten. Sie müssen wissen, woher eine Zahl stammt und ob eine Studie tatsächlich existiert. Während ChatGPT eine flüssige Zusammenfassung liefert, deren Quellen Sie ebenfalls mit ein paar zusätzlichen Klicks überprüfen können, zeigt Perplexity Links zu den Originalstudien mit genauen Erscheinungsdaten. Das schützt Sie vor einem Fehler, der Ihnen nicht passieren sollte: eine Quelle zu zitieren, die nicht existiert.

Merke: Perplexity zeigt, dass nicht das Modell den Unterschied macht, sondern die Umgebung, in der es arbeitet. Dasselbe Sprachmodell liefert in Perplexity andere, für die Recherche bessere Ergebnisse als im normalen Claude-Chat. Und mit dem Model Council lässt Perplexity mehrere Modelle gemeinsam nachdenken. Eine kluge Strategie und ein Vorgeschmack auf die nahe Zukunft der KI-Nutzung eben auch und gerade mit Agenten.

4.8 Microsoft und Copilot: KI direkt am Arbeitsplatz

Microsoft verfolgt eine grundlegend andere Strategie als OpenAI, Anthropic oder Google: Statt einen eigenen Chatbot zum Zentrum des Universums zu machen, bringt Microsoft die KI dorthin, wo Hunderte Millionen Menschen bereits arbeiten, in Word, Excel, PowerPoint, Outlook und Teams. Der Microsoft 365 Copilot ist so gesehen kein eigenständiges Produkt, sondern ein Assistent, der in Ihren bestehenden Tools lebt.

Unter der Motorhaube arbeiten hier aktuelle OpenAI-GPT-Modelle. Aber das Modell ist nicht der Punkt. Der Punkt ist der Zugriff auf Ihre Unternehmensdaten.

- Wenn Sie Copilot in Outlook fragen, »Fass mir die wichtigsten E-Mails von heute zusammen«, kennt er Ihren Posteingang. Wenn Sie ihn in PowerPoint bitten, »Erstelle eine Präsentation aus dem Quartalsbericht«, kann er das Dokument auf Ihrem SharePoint finden und daraus Folien bauen.

- Seit der Ignite-Konferenz Ende 2025 entwickelt sich der Copilot vom reaktiven Assistenten zum *Agent Mode*: In Word, Excel oder PowerPoint kann er mehrstufige Aufgaben eigenständig bearbeiten und im Dialog mit Ihnen iterativ verbessern.
- Mit *Copilot Studio* können Unternehmen eigene KI-Agenten per Low-Code-Oberfläche bauen, etwa einen Agenten, der wöchentlich Branchennews zusammenfasst und mit internen Strategiedokumenten abgleicht. Mehr dazu in Kapitel 11, »KI-Agenten – vom Junior zum eigenständigen Mitarbeiter«.

Der Copilot ist die naheliegende Wahl für alle, die ohnehin in der Microsoft-Welt arbeiten, und das sind Hunderte Millionen Microsoft-365-Nutzende weltweit.

Und damit ist auch schon die Einschränkung genannt: Copilot ist kein offenes Kreativwerkzeug wie ChatGPT oder Claude. Für die Entwicklung einer Brand Voice oder experimentelles Arbeiten sind die eigenständigen Chatbots die bessere Wahl.

Der Copilot glänzt bei Zusammenfassungen, E-Mail-Entwürfen und Präsentationen aus bestehenden Dokumenten. Ab Juli 2026 werden KI-Funktionen wie der Copilot Chat in die regulären Microsoft-365-Abonnements integriert, allerdings zu höheren Preisen. KI wird damit im Microsoft-Ökosystem weniger Opt-in und mehr Standard.

Merke: Der Microsoft 365 Copilot ist keine eigenständige KI-Plattform, sondern ein echtes Werkzeug zur Steigerung der *Effizienz* am Microsoft-ausgestatteten Arbeitsplatz. Daher entscheiden sich viele Unternehmen für diesen Weg: Es ist ja alles schon da. Compliance: Check!

Seine Stärke liegt weder im Modell – da läuft ein Modell der GPT-Klasse – noch in seiner dialogischen Ader oder seiner Kreativität. Im Gegenteil: Sie werden schnell merken, dass dieser »Chatpartner« in seinen Antworten sehr kurz angebunden reagiert. Seine Stärke spielt er beim Zugriff auf Ihre Unternehmensdaten und bei der direkten Einbindung in Ihre täglichen Microsoft-Tools aus.

4.9 Mistral: Die europäische Alternative

Im KI-Wettbewerb dominieren amerikanische Unternehmen. Dazu gehören OpenAI und Anthropic aus San Francisco sowie Google aus Mountain View. Und dann gibt es *Mistral*[21]. Das Unternehmen wurde 2023 in Paris gegründet. Die Gründer sind ehemalige Forschende von Google DeepMind und Meta. Sie verfolgen eine Mission, die in Europa auf offene Ohren stößt: Sie wollen leistungsfähige KI-Modelle entwickeln, die europäischen Werten wie Datenschutz, Transparenz und Datenhoheit entsprechen.

Was als ambitioniertes Start-up begann, ist Anfang 2026 ein Unternehmen mit einer Bewertung von rund 12 Milliarden Euro und einer strategischen Position, die kein amerikanischer Wettbewerber so einnehmen kann. Mistral ist nicht nur in der EU ansässig,

21 *https://mistral.ai*

sondern unterstellt sich auch vollständig europäischer Jurisdiktion. In einer Zeit, in der der US-amerikanische CLOUD Act es US-Behörden erlaubt, auf Daten amerikanischer Unternehmen zuzugreifen – auch wenn diese auf europäischen Servern liegen –, ist das ein konkreter Vorteil. Unternehmen in regulierten Branchen schauen deshalb genauer hin.

Auch Mistrals Chatbot *Le Chat* hat sich seit dem Launch schnell weiterentwickelt und bietet inzwischen vieles, was die großen Plattformen auch können:

- *Deep Research*
- einen *Voice Mode* mit dem hauseigenen Sprachmodell Voxtral
- *Projects* zum Organisieren von Recherchen
- einen leistungsfähigen Bildgenerator

Eine Besonderheit sind die *Flash Answers*, die Antworten mit bis zu 1.000 Wörtern pro Sekunde ausgeben, deutlich schneller als bei den Wettbewerbern.

Im Enterprise-Bereich verbindet Le Chat über MCP-Konnektoren Tools wie SharePoint, Jira oder Slack direkt mit dem Assistenten. Unternehmen können eigene KI-Agenten bauen, ohne Code, direkt in der Plattform. Entscheidend für viele:

Le Chat Enterprise lässt sich vollständig *on-premise* oder in der eigenen Cloud betreiben. Keine Daten verlassen die eigene Infrastruktur, wenn Sie das nicht wollen.

Im Januar 2026 schloss Mistral einen Rahmenvertrag mit dem französischen Verteidigungsministerium: Die Streitkräfte erhalten Zugang zu Mistrals Modellen und Services, die auf französisch kontrollierter Infrastruktur betrieben werden. Parallel dazu verhandeln Frankreich und Deutschland gemeinsam mit SAP ein Abkommen über KI-Lösungen in der öffentlichen Verwaltung. Das sind keine Pilotprojekte mehr, sondern strategische Weichenstellungen. Europäische Regierungen setzen zunehmend auf europäische KI.

Mit der im Februar 2026 erfolgten Übernahme der französischen Serverless-Plattform *Koyeb*[22] baut Mistral zudem eine eigene europäische KI-Cloud mit garantierter Datenresidenz in der EU auf. Das Ziel ist ein vollständiger europäischer KI-Stack vom Modell über die Plattform bis zur Infrastruktur. Für Unternehmen, die ab August 2026 die Anforderungen des *EU AI Act* erfüllen müssen, ist das ein relevantes Angebot.

Bei aller Relevanz: Mistrals Modelle spielen in der Spitzengruppe mit, und das aktuelle Modell *Mistral Large* ist konkurrenzfähig mit GPT und Opus. In den absoluten Top-Disziplinen wie komplexem Reasoning und kreativem Schreiben liegen die amerikanischen Frontier-Modelle aber nach wie vor vorn. Auch das Ökosystem ist kleiner: weniger Integrationen, weniger Community-Tools als bei ChatGPT oder Claude.

22 Senad Karaahmetovic, Mistral AI übernimmt Infrastruktur-Startup Koyeb zur Stärkung seiner Kapazitäten, in: Investing.com, *https://de.investing.com/news/company-news/mistral-ai-uber-nimmt-infrastrukturstartup-koyeb-zur-starkung-seiner-kapazitaten-93CH-3349036*, 18.02.2026 [18.02.2026]

Ein Detail, das Sie vielleicht kennen sollten: Mistral hat auch US-Investoren an Bord, darunter Microsoft, Andreessen Horowitz und Nvidia.[23] Die Frage, ob der CLOUD Act über diese Investorenbeziehungen Zugriffsmöglichkeiten schaffen könnte, ist juristisch ungeklärt. Mistrals Position ist deutlich stärker als die jedes US-Anbieters, aber eine kategorische Immunität sollte man auch hier nicht annehmen.

Merke: Mistral ist die relevanteste europäische Alternative zu den amerikanischen KI-Plattformen, nicht weil die Modelle besser wären, sondern weil Datenhoheit, DSGVO-Konformität und der EU AI Act für immer mehr Unternehmen den Ausschlag geben. Wer in der EU mit sensiblen Daten arbeitet, sollte Mistral kennen.

Bei aller Begeisterung für die Leistungsfähigkeit und die vielen Features, die die LLMs bieten, gelten jedoch für alle auch allgemeingültige Einschränkungen und Regeln der Zusammenarbeit, die man im Hinterkopf haben sollte.

4.10 Watch Out: Durchschnitt und Bias

Haben Sie sich schon einmal gefragt, warum KI-generierte Texte so oft nach … nichts schmecken? Eher nach »Graubrot« als nach feinem »Sauerteig«? Warum zehn verschiedene Modelle auf denselben Prompt zehn Texte liefern, die sich ähnlich langweilig lesen? Das hat einen Grund. Und der ist wichtig für Ihre Kreation, denn er betrifft zwei verschiedene Probleme: das Qualitätsproblem des Durchschnitts und das Werteproblem der Verzerrung – des sogenannten Bias.

Die KI rechnet immer auf Mitte – dafür ist sie schließlich trainiert!

Stellen Sie sich ein Sprachmodell wie eine riesige Abstimmungsmaschine vor. Egal welche Plattform: Bei jeder Antwort berechnet es: Was ist das wahrscheinlichste nächste Wort? Und das wahrscheinlichste Wort ist fast immer das, was die Mehrheit der Trainingsdaten nahelegt. Das Ergebnis: Durchschnitt.

Bei *Texten* spüren Sie das sofort. Die Sätze sind freundlich, aber glatt. Kein Stolpern, kein eigener Ton. Ein KI-Text liest sich »irgendwie gut«, aber auch irgendwie austauschbar. Denken Sie an die 10 langweiligsten LinkedIn-Posts in Ihrer heutigen Timeline. Das Thema, meine Meinung, drei Bullets mit Emojis, und am Ende standardmäßig das »Was meint ihr?« Das ist kein Zufall, sondern Mathematik. Der Mittelwert ist per Definition nicht originell. Sondern irgendwie »richtig«.

Bei *Bildern* ist der Effekt noch deutlicher. Lassen Sie eine KI ein Porträt generieren, bekommen Sie ein symmetrisches Gesicht mit warmem Licht, im Goldenen Schnitt komponiert. Landschaften leuchten in Sonnenuntergangsfarben, Büroräume sehen aus wie aus einer Hochglanzbroschüre. Schön, aber beliebig. Wenn alles nach Stock-Foto aussieht,

23 *https://mistral.ai/news/mistral-ai-raises-1-7-b-to-accelerate-technological-progress-with-ai*

fehlt genau das, was gute Content Creation ausmacht: Überraschung. Originalität. Und damit sicherlich auch die Aufmerksamkeit des Publikums.

Ein harmloses Beispiel? Generieren Sie doch ein Bild einer analogen Uhr mit KI. Sie werden feststellen: Egal welches Modell Sie bemühen: Die Zeiger darauf stehen im Ergebnis meist auf 10 vor 10. Warum? Weil alle Uhren im Netz »positiv rüberkommen« sollen und deren Zeiger daher auf »bitte lächeln« gestellt wurden. Die Zeiger scheinen durch das Trainingsmaterial geradezu festgetackert. Sie werden merken, dass das Modell selbst mit sehr klaren Prompts hartnäckig wieder auf 10 vor 10 »zurückfällt«.

Der Bias: Die westliche »Brille«

Das Durchschnittsproblem betrifft die Qualität Ihrer Ergebnisse. Das nächste Problem betrifft hingegen etwas anderes: die Fairness. Die großen Sprachmodelle wurden überwiegend mit englischsprachigen und vor allem westlichen Daten trainiert. In ihnen stecken also nicht nur die Werte der westlichen Hemisphäre, sondern auch deren Vorurteile. Diese werden als Bias bezeichnet. Sie werden in den KI-generierten Ergebnissen – sowohl in Texten als auch in Bildern – erschreckend, manchmal aber auch nur subtil deutlich. Ein Beispiel: Bei einem Prompt mit »CEO« werden beispielsweise überwiegend weiße Männer dargestellt, bei einem Prompt mit »Pflegekraft« überwiegend Frauen. Noch deutlicher wird dies bei Bildern mit interkulturellem Kontext, wie in Abbildung 4.6: Der Prompt mit »Kind spielt im Wasser« lieferte mit europäischem Kontext ein Bild von einem blonden Mädchen in einem sauberen See, im afrikanischen hingegen ein Bild von einem Jungen in einem schlammigen Fluss mit einem Elefanten im Hintergrund. Wäre Ihnen der Bias in den Bildern auf den ersten Blick aufgefallen?

Abbildung 4.6: So sieht Midjourney unsere Welt: Auch KI-Bildgeneratoren reproduzieren bestehende Stereotype. (Quelle Bild: Facebook, Midjourney-Gruppe)

Zugegeben: Das Beispiel ist zwei Jahre alt. Inzwischen wurden viele KIs auf Bias »sensibilisiert«. Doch nun kann es passieren, dass bei gleichem Prompt nur noch weibliche CEOs gezeigt werden, einfach um dem Vorwurf des Bias diametral entgegenzutreten – was aber natürlich ebenso wenig zielführend ist.

Prüfen Sie Ihre Ergebnisse also immer auf ungewollte Verzerrungen. Was die KI als »wahrscheinlichstes« oder als »gewünschtes« Ergebnis liefert, ist nicht neutral, sondern ein Spiegel einseitiger Trainingsdaten und menschlicher Voreingenommenheiten. Lassen Sie andere nochmals auf das Ergebnis schauen. Stichwort: »Human in the Loop.«

- *Diversifizieren Sie*: Wer wichtige Inhalte durch ein zweites Modell gegenlesen lässt, fängt Verzerrungen auf.
- *Hinterfragen Sie* den ersten Entwurf, denn was die KI liefert, ist der Durchschnitt, nicht die beste Lösung.
- *Prüfen Sie* bei sensiblen Themen besonders genau auf Geschlechterrollen oder politische Einordnungen: Hier ist menschliche Kontrolle Pflicht.

Die bisher beschriebenen Verzerrungen sind Nebeneffekte des Trainings, ärgerlich, aber korrigierbar. Es gibt allerdings Fälle, in denen Verzerrung Teil des Designs ist.

Ein Beispiel ist *Grok*, der Chatbot von Elon Musks Unternehmen xAI. Grok wurde explizit als »rebellischer« Assistent positioniert – weniger gefiltert und politisch nicht korrekt, was sowohl Zustimmung bei Anhängern freier Meinungsäußerung als auch scharfe Kritik von Beobachtern hervorruft.

Als Grok allerdings öffentlich gegen seinen Besitzer Musk rebellierte und ihn als einen der größten Verbreiter von Falschinformationen bezeichnete, versuchte xAI, seine Antworten zu »optimieren«. Business Punk beschrieb dies als »Euphemismus für die Unterdrückung kritischer Äußerungen gegenüber Musk«.[24] xAI soll daraufhin sogar Maßnahmen ergriffen haben, damit der Chatbot Quellen ignoriert, die Musk und Trump der Verbreitung von Fake News bezichtigen. Nach Bekanntwerden der Initiative ruderte xAI jedoch zurück. Dennoch ist dies ein Beweis dafür, wie leicht die Werte einer künstlichen Intelligenz manipulierbar sind.

Wenn Verzerrung kein Versehen mehr ist

Wie stark sich das auswirkt, zeigte eine im Januar 2026 veröffentlichte Studie der *Anti-Defamation League* (ADL): In den von der NGO, die sich gegen Hassrede engagiert, durchgeführten Tests zur Erkennung antisemitischer Inhalte erreichte Claude 80 von 100 Punkten. Mit 49 bis 57 Punkten platzieren sich Gemini, ChatGPT und das chinesische *Deepseek* im mittleren Wertungsfeld. Am Ende des Rankings rangiert

24 KI-Rebellion: Musks Chatbot Grok stellt sich gegen seinen Schöpfer, in: Business Punk, *https://www.business-punk.com/tech/ki-rebellion-musks-chatbot-grok-stellt-sich-gegen-seinen-schoepfer/*, 01.04.2025 [18.02.2026]