Maria Kalweit und Gabriel Kalweit

Warum wir neu lernen müssen, mit Maschinen zu sprechen – eine Momentaufnahme der Generativen KI im Januar 2024

Im Dezember 2023 teilte der Nutzer *@danshipper* auf der Social-Media-Plattform X ein unerwartetes Problem bezüglich einer Text- und Bildverarbeitungssoftware.[1] Zu seiner Verwunderung funktionierte ein Befehl, der kurz zuvor noch einwandfrei war, nicht mehr. Er lud ein Bild einer Buchseite hoch, in der Hoffnung, den Text digitalisiert zurückzuerhalten, und stellte in seiner Hilflosigkeit seine Frage an die öffentliche Community. Als Antwort auf sein *Warum* schlug *@NickADobos* vor: »Halluzination. Teilweise, weil es als höfliche Frage formuliert war, auf die ›Nein‹ eine gültige Antwort darstellt.« Ähnliche Erfahrungen wurden von *@mblair* geteilt, der feststellte: »Ich hatte das gleiche Problem, wo es sich weigerte, ein Bild mit einem bestimmten Seed zu ergänzen, wenn ich ›Können Sie...‹ benutzte.« *@Reelix* bot eine Erklärung an: »Die Software ist so konzipiert, dass sie mit minimalem Aufwand eine gültige Antwort gibt, und ›Nein‹ erfordert nur minimale Berechnungen.«

Zusätzlich beobachteten einige Nutzer, dass die Antworten der Software ausführlicher waren, wenn sie glaubte, dass Frühling und nicht Winter sei.[2] Dies führte zu humorvollen Spekulationen über eine *Winterdepression* und der allgemeinen Annahme, dass im Winter generell weniger gearbeitet werde. Interessanterweise zeigte sich auch, dass die Antworten umfangreicher ausfielen, je mehr *Trinkgeld* den Nutzern zufolge angeboten wurde.[3]

Solchen Unterhaltungen hätte man noch vor Kurzem nicht beiwohnen können. Menschen, die sich darüber unterhalten, *wie* man eine Maschine *ansprechen* muss, um eine Antwort zu erhalten; dass der Grad der Höflichkeit Unterschiede in den Resultaten bringt; dass die Maschine Züge zeigt, die jenen eines Menschen nicht unähnlich sind; ganz abgesehen davon, dass man *überhaupt* normal mit einer Maschine *sprechen* kann. Die Software, die im Zentrum dieser Diskussionen steht, ist ChatGPT[4], eine bahnbrechende Errungenschaft im Bereich der künstlichen Intelligenz (KI), entwickelt und veröffentlicht von OpenAI im Jahr 2022. Innerhalb von nur zwei Monaten nach ihrer Veröffentlichung verzeichnete sie bereits über 100 Millionen Nutzer – ein beispielloser Erfolg, welcher die rasante Annahme und das Interesse an dieser Technologie unterstreicht. ChatGPT ist nicht nur ein weiteres Softwareprodukt; es repräsentiert einen Wendepunkt in der Art und Weise, wie wir mit Maschinen interagieren und sie in unserem täglichen Leben nutzen. Eine neue, schwer zu greifende Normalität.

Hinter Plattformen wie ChatGPT stehen sogenannte Large Language Models (LLMs). Diese Modelle zeichnen sich dadurch aus, dass Interaktionen mit ihnen nicht auf vordefinierten Textbausteinen oder Befehlen fußen, sondern auf geschriebener Sprache in ihrer vollen Vielfalt. Wenn man also mit Systemen wie ChatGPT in natürlicher Sprache schreibt, dann können sie sie verstehen und auch in ebendieser antworten. Allerdings können bestimmte Umgangsformen wiederum die Art, Qualität und Struktur der Antworten steuern. Die zur Steuerung der Antworten genutzten Befehle und Strategien sind dann Beispiele für das sogenannte *Prompting*. Prompting kann als eine Art Programmiersprache betrachtet werden, die allerdings modellspezifisch ist. Beispielsweise können, bei derselben natürlichsprachigen Eingabe, die Ausgaben eines GPT-3 anders sein als die eines GPT-4. Neue Updates können die Art der Interaktion mit diesen Modellen generell komplett verändern, was bedeutet, dass für jedes Modell eine neue *Grammatik* oder *Sprache* erlernt werden muss. Dies stellt momentan

1 *Dan Shipper* [@danshipper], "What the Hell? When Did This Happen?? Https://T.Co/KWXVXE9Dem," Tweet, *Twitter*, (veröffentlicht am 06.12.2023), https://twitter.com/danshipper/status/1732258207840501946.

2 *Rob Lynch* [@RobLynch99], "@ChatGPTapp @OpenAI @tszzl @emollick @voooooogel Wild Result. Gpt-4-Turbo over the API Produces (Statistically Significant) Shorter Completions When It 'Thinks' Its December vs. When It Thinks Its May (as Determined by the Date in the System Prompt). I Took the Same Exact Prompt… Https://T.Co/mA7sqZUAor," Tweet, *Twitter*, (veröffentlicht am 11.12.2023), https://twitter.com/RobLynch99/status/1734278713762549970.

3 *thebes* [@voooooogel], "So a Couple Days Ago i Made a Shitpost about Tipping Chatgpt, and Someone Replied 'Huh Would This Actually Help Performance' so i Decided to Test It and IT ACTUALLY WORKS WTF Https://T.Co/kqQUOn7wcS," Tweet, *Twitter*, (veröffentlicht am 01.12.2023), https://twitter.com/voooooogel/status/1730726744314069190.

4 OpenAI, "ChatGPT," 2024, https://chat.openai.com.

eine äußerst wertvolle Fähigkeit dar, was sich in den spektakulären Jahresgehältern widerspiegelt, die professionellen Prompt-Engineers geboten werden. So berichtete der Spiegel am 6. Dezember 2023, dass Jahresgehälter von bis zu 335.000 US-Dollar für Experten in diesem Bereich gezahlt werden.[5] Gleichzeitig gibt es Forschungsansätze, die darauf abzielen, KI-Systeme ihre eigenen Prompts schreiben zu lassen[6], was diesen neuen Berufszweig ebenso schnell obsolet machen könnte, wie er entstanden ist.

Die Konversation, die *@danshipper* und andere Nutzer auf einer Social-Media-Plattform führten, reflektiert mehr als nur ein technisches Problem; sie symbolisiert eine Verschiebung in unserer Beziehung zur Technologie. Diese Interaktionen mit ChatGPT offenbaren nicht nur die Grenzen und Möglichkeiten der KI-basierten Kommunikation, sondern werfen auch grundlegende Fragen über das Wesen der Mensch-Maschine-Interaktion auf. Wie verstehen wir diese Technologie? Wie passen wir unsere Kommunikationsstrategien an, um die besten Ergebnisse zu erzielen? Und was bedeutet es, wenn ein technisches Werkzeug zu einem quasi-sozialen Akteur in unserem digitalen Kosmos wird?

Dieser Artikel zielt darauf ab, diese Fragen zu erörtern und ein tieferes Verständnis für die Funktionsweise von ChatGPT und ähnlichen Systemen zu schaffen. Es geht nicht nur darum, wie wir eine spezifische Software effektiver nutzen können, sondern auch darum, die Implikationen dieser Technologie für unsere Gesellschaft und unsere Zukunft zu erkunden. Mit einem Fokus auf effektive Strategien für das Prompting wird untersucht, wie wir ein langfristiges Verständnis für diese fortschrittliche Technik entwickeln können, das auch nach weiteren Aktualisierungen und Entwicklungen der Modelle Bestand hat.

I. Was verbirgt sich eigentlich hinter GPT?

Der Weg zu guten Prompting-Strategien führt über das Verständnis der Methodik hinter den Systemen. Wenngleich nicht alle Details bekannt sind und schon gar nicht in diesem Artikel besprochen werden können, wird nachfolgend in aller Kürze erklärt, was sich hinter dem *Zauber* von Large Language Models verbirgt.

Unter einem Modell versteht man eine Überführung von *Eingaben* zu *Ausgaben*. Ein Beispiel ist die Eingabe einer Buchseite und die Ausgabe könnte der extrahierte Text sein. Eine solche Überführung a priori exakt zu allen möglichen Eingaben zu definieren, ist quasi-unmöglich, daher gehen aktuelle Methoden der künstlichen Intelligenz den Umweg, eine solche Überführung anhand von gegebenen Beispielen zu schätzen. Die dazu verwendete Sprache in der KI ist gemeinhin die Mathematik. Als Ausgangsbasis dient eine *flexible* Darstellung dieser gesuchten mathematischen Überführung, die verschiedene Formen durch unterschiedliche Setzungen von freien Parametern annehmen kann. Nehmen wir als Beispiel den Satz: »Heute geht es mir gut.« So könnte, wenn dem Wort *gut* ein hohes Gewicht zugewiesen ist, dieser Satz als eher positiv angesehen werden. Liegt der Fokus jedoch eher auf dem *Heute*, so könnte es auch so gelesen werden, dass das gut Fühlen *heute* etwas Besonderes ist, was wiederum eher negativ eingeordnet werden könnte. Eine solche Zuordnung von positiv und negativ ist also von der Gewichtung, also der Parameter, des Modelles abhängig, welches die Zuordnung trifft.

Grundlage für den enormen und enorm schnellen Fortschritt der letzten Jahre bilden sogenannte künstliche neuronale *Netzwerke* als Darstellung dieser oben beschriebenen mathematischen Überführung – Architekturen von parallel und in Reihe geschalteten, *künstlichen Neuronen*. Diese künstlichen Neuronen gewichten ihre Eingaben jeweils mit einem freien Parameter, summieren diese gewichteten Eingaben auf, transformieren diese Summe mit einer nichtlinearen, mathematischen Funktion und geben das Ergebnis, die *Aktivierung* des Neurons, an die Folgeneuronen weiter. Die freien Parameter eines neuronalen Netzes ergeben mit einer spezifischen Einstellung spezifische Ausgaben, deren Korrektheit anhand eines Fehlermaßes gemessen werden kann. Misst man nun den spezifischen Beitrag eines Parameters auf den gegebenen Fehler, kann der Parameter in dieser Hinsicht Schritt für Schritt so angepasst werden, dass das Ergebnis verbessert wird. Die Problemstellung hierbei liegt darin, dass die Schätzung der mathematischen Überführung allerdings auch auf zum Zeitpunkt der Parameterfindung *ungesehene* Beispiele generalisieren sollte, um in der Praxis nutzbar zu sein. Ein Modell ist dann über seine fixe, gefundene Menge an gesetzten Parametern definiert. Und auch wenn innerhalb eines Modells die Eingaben in Form von Zahlen repräsentiert werden müssen, können die verarbeiteten Modalitäten

[5] *Verena Töpper*, "(S+) Geld verdienen mit ChatGPT: Prompt Writer verdienen bis zu 335.000 Dollar im Jahr," Der Spiegel, (veröffentlicht am 06.12.2023), sec. Job & Karriere, https://www.spiegel.de/karriere/chatgpt-prompt-writer-und-prompt-engineers-verdienen-bis-zu-335-000-dollar-im-jahr-a-a54a93a5-e20d-40e6-b235-28aecobddaaa.

[6] "Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution," (veröffentlicht am 22.09.2023), https://openreview.net/forum?id=HKkiX32Zw1.

im Ursprung beliebig sein – zum Beispiel geschriebener Text.

In der Domäne der Textverarbeitung gab es schon einige Bewegung, doch die Entwicklung der Transformer-Architektur[7] im Jahre 2017 durch Wissenschaftler von Google sollte eine Zeitenwende einläuten. Sie bildeten die Grundlage für die erste Version der *Generative Pre-trained Transformer* von OpenAI in 2018, dem Modell das hinter dem schwer auszusprechenden Akronym GPT steht. Betrachtet man Gleichung 1 in dem Papier *Improving Language Understanding by Generative Pre-Training*[8], welches vier Jahre später die Welt revolutionieren sollte, dann erkennt man, woher die einzelnen Teile dieses Akronyms stammen. Der Text wird hier von einem Transformer verarbeitet – daher das T –, welcher einen Strom an Text, den sogenannten Kontext, als Eingabe bekommt. Die Ausgabe seiner Überführung ist eine Wahrscheinlichkeitsverteilung über möglichen Folgetext, woraus es eine Stichprobe zieht – daher das G. Das Modell soll innerhalb seiner Optimierung zunächst schlicht die Wahrscheinlichkeit der Wörter so erhöhen, wie sie auch in den Trainingsdaten gegeben sind. Wenn wir bei unserem Beispiel bleiben, so könnte das Modell als Kontext »Heute geht es mir« bekommen und den möglichen Ausgaben »gut.« oder »schlecht.« die gleiche Wahrscheinlichkeit zuordnen – denn der Kontext gibt hier keinen Hinweis darauf, welche davon eher passen könnte. Steht im Kontext allerdings »Ich wurde gelobt. Heute geht es mir«, dann sollte das Modell einem »gut.« eine höhere Wahrscheinlichkeit zuweisen. Danach kann eine aufgabenspezifische Feinabstimmung durchgeführt werden – daher das P. Beispielsweise kann man, wenn man mit ChatGPT schreibt, unter den Antworten Symbole für *Daumen hoch* und *Daumen runter* erkennen. Informationen aus diesen Rückmeldungen werden dann genutzt, um das Modell via bestärkendem Lernen[9] zu verbessern[10]. Gebe ich also die Rückmeldung, dass mir »gut.« nicht so gefällt, da es nicht enthusiastisch genug klingt, könnte ich so die Wahrscheinlichkeit für ein »super.« erhöhen.

II. Warum jetzt?

Dass aus dieser, in wenigen Zeilen dargestellten, Art der Parameteranpassung eines Transformer-Modells ein so mächtiges Werkzeug entstehen kann, hat auch die Fachwelt in seiner Wucht überrascht. Möglich geworden ist das durch die Zusammenkunft von mehreren, parallel verlaufenden Strömungen[11]. Zunächst ermöglichte die stetige Steigerung der Rechenleistung, angelehnt an das Mooresche Gesetz, sowie die Entwicklung der Infrastruktur schaffenden Software, die Handhabung komplexer KI-Modelle. Insbesondere die Anpassung und Optimierung von Grafikprozessoren (GPUs) für parallele Berechnungsprozesse spielten hierbei eine zentrale Rolle.

Parallel dazu führte das exponentielle Wachstum digital verfügbarer Daten seit der Einführung des World Wide Webs im Jahr 1991 zu einem enormen Anstieg an Trainingsdaten. Diese Datenflut, kombiniert mit der zunehmenden Digitalisierung bis dato analoger Medien, schuf die notwendige Basis für das Training umfangreicher KI-Modelle. Nicht zuletzt waren es die finanziellen und strukturellen Investitionen großer öffentlicher Institutionen und Technologieunternehmen, die den letzten Schub gaben. Organisationen wie OpenAI, unterstützt durch bedeutende Anfangsinvestitionen von Persönlichkeiten wie Elon Musk und Industrieakteuren aus dem Silicon Valley, konnten so umfangreiche Forschungs- und Entwicklungsprojekte in die Wege leiten. Diese Investitionen ermöglichten es, Teams hochqualifizierter Forscher zusammenzubringen und KI-Modelle in bisher ungekanntem Umfang zu entwickeln und zu trainieren.

Jedes Jahrzehnt brachte so seine eigenen Innovationen und Durchbrüche, welche die Grundlage für die nächste Generation von KI-Modellen legten[12]. Das Jahr 2022 markierte einen weiteren Wendepunkt mit der Einführung von GPT-4 durch OpenAI, einem Modell, das von einem Team aus 343 hochqualifizierten Wissenschaftlern entwickelt wurde. Diese Innovationen waren nicht auf OpenAI beschränkt; andere bedeutende Mo-

7 Ashish Vaswani et al., "Attention Is All You Need," in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc., 2017), 6000–6010.

8 Alec Radford and Karthik Narasimhan, "Improving Language Understanding by Generative Pre-Training," 2018.

9 Gabriel Kalweit, "On the Role of Time Horizons in Reinforcement Learning," 2022, https://doi.org/10.6094/UNIFR/232102; Gabriel Kalweit, Maria Kalweit, and Joschka Boedecker, "Robust and Data-Efficient Q-Learning by Composite Value-Estimation," *Transactions on Machine Learning Research*, 2022, https://openreview.net/forum?id=ak6Bds2DcI.

10 Daniel M. Ziegler et al., "Fine-Tuning Language Models from Human Preferences," 2019, https://doi.org/10.48550/ARXIV.1909.08593.

11 "Quick Guide to AI 2.0 Oct 2020," accessed January 11, 2024, http://ceros.mckinsey.com/quick-guide-to-ai-12.

12 Hans Burkhardt, "Ein Beitrag zur Künstlichen Intelligenz," Ordnung der Wissenschaft, no. 2 (2023): 71–78.

delle wie Bard und Gemini von Google[13], Claude von Anthropic[14] und weitere[15], entwickelt von verschiedenen Organisationen weltweit, trugen ebenfalls zur Landschaft der generativen KI bei.

Diese Fortschritte bedeuteten allerdings auch einen finanziellen Kraftakt ungeheurer Dimension, was durch die Investition[16] von Microsoft in Höhe von zehn Milliarden US-Dollar in OpenAI im Jahr 2023 und die hohen Kosten für Spitzenforscher auf diesem Gebiet, welche sogar die von Top-NFL-Quarterbacks übersteigen können[17], gezeigt wird. Die Trainingskosten für Modelle wie GPT-4 werden auf etwa 100 Millionen US-Dollar geschätzt[18], während die täglichen Inferenzkosten – die Kosten für die reine Anwendung des trainierten Modells – auf weit über 700.000 US-Dollar pro Tag angenommen werden[19]. Diese immensen Investitionen in die Entwicklung und den Betrieb spiegeln nicht nur die technologische und wissenschaftliche Leistungsfähigkeit wider, sondern auch das enorme wirtschaftliche Potenzial, das in diesen Systemen steckt.

III. Welches Ziel verfolgen diese Systeme?

Wenn eine mathematische Überführung geschätzt wird, und somit eben generalisieren muss, geht das gemeinhin mit **einer Komprimierung des Wissens in den Trainingsdaten** einher. Gesehen werden kann das beispielsweise dadurch, dass die Trainingsdaten von GPT-3 45 Terabyte groß waren, das Modell an sich aber weniger als eines. Um komplexe Zusammenhänge so effizient zu komprimieren, finden vermutlich implizite Anordnungen statt, die von einigen als Emergenz betitelt werden[20].

Fragt man beispielsweise ChatGPT, wie man eine Tasse Kaffee in einer Mikrowelle zubereitet – eine ungewöhnliche, aber durchaus mögliche Methode –, so erhält man eine detaillierte Antwort, die Schritt für Schritt erklärt, wie man zunächst Wasser in der Mikrowelle erhitzt, dann gemahlenen Kaffee hinzufügt und schließlich die Mischung stehen lässt, um den Kaffee ziehen zu lassen. Dieser Prozess beinhaltet das Verständnis, dass Mikrowellen Wasser erhitzen können, dass gemahlener Kaffee mit heißem Wasser gemischt werden muss, um Geschmack zu extrahieren, und dass die Mischung Zeit benötigt, um den Kaffee ziehen zu lassen. Für diese Anleitung musste das Modell verstehen, dass Mikrowellen zur Erhitzung von Flüssigkeiten genutzt werden können – eine Information, die es aus seinen Trainingsdaten extrahiert hat. Ebenso musste es wissen, dass Kaffee in der Regel durch die Interaktion von heißem Wasser und gemahlenen Bohnen entsteht, und dass die Extraktion Zeit benötigt. Das bedeutet, dass das Modell Konzepte von Hitzeanwendung, Flüssigkeitsextraktion und Zeitablauf repräsentieren musste, um eine solche Anleitung zu generieren. Und das, obwohl eine solche spezifische Anleitung zur Kaffeezubereitung in einer Mikrowelle möglicherweise nicht genau so in den Trainingsdaten vorhanden war. Dieses Beispiel zeigt, wie KI-Modelle unterschiedliche Informationsquellen kombinieren und anwenden können, um kreative und funktionale Lösungen für ungewöhnliche oder neue Fragestellungen zu liefern. Es demonstriert die Fähigkeit der KI, Konzepte zu verknüpfen und auf Situationen anzuwenden, die in ihren ursprünglichen Trainingsdaten möglicherweise nicht explizit beschrieben wurden.

Dies führt zur Schlussfolgerung, dass die sehr offene und trivial erscheinende Zielsetzung der GPT-Modelle – **das Ermitteln des wahrscheinlichsten Folgeelements eines gegebenen Kontextes** – eine erstaunlich leistungsfähige Strategie darstellt. Diese Methode allein ermöglicht es, komplexe, dynamische und flexible KI-Systeme zu schaffen. Die resultierenden Modelle sind in der Lage, eine bloße Wiedergabe von Informationen klar zu überschreiten. Sie generieren kreative, kontextbezogene Lösungen und erbringen Leistungen, die auf den ersten

13 "Gemini - Google DeepMind," (zuletzt abgerufen am: 14.01.2024), https://deepmind.google/technologies/gemini/.

14 "Introducing Claude," Anthropic, (zuletzt abgerufen am: 14.01.2024), https://www.anthropic.com/index/introducing-claude.

15 Mistral AI, "Mixtral of Experts," December 11, 2023, https://mistral.ai/news/mixtral-of-experts/; Alyssa Hughes, "Phi-2: The Surprising Power of Small Language Models," *Microsoft Research* (blog), December 12, 2023, https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/; "Llama 2," Meta AI, (zuletzt abgerufen am: 14.01.2024), https://ai.meta.com/llama-project.

16 Cade Metz and Karen Weise, "Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT," *The New York Times*, (veröffentlicht am: 23.01.2024), sec. Business, https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html.

17 "The Race to Buy the Human Brains Behind Deep Learning Machines - Bloomberg," (zuletzt abgerufen am: 11.01.2024), https://www.bloomberg.com/news/articles/2014-01-27/the-race-to-buy-the-human-brains-behind-deep-learning-machines.

18 Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, (zuletzt abgerufen am: 11.01.2024), https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

19 Sahaj Godhani, "The Economics of ChatGPT Analyzing Its $700,000 Daily Costs and the Potential Impact on Its Maker.," Medium, (veröffentlicht am: 15.08.2023), https://blog.gopenai.com/the-economics-of-chatgpt-analyzing-its-700-000-daily-costs-and-the-potential-impact-on-its-maker-7e690600ade7.

20 Sébastien Bubeck et al., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4" (arXiv, April 13, 2023), http://arxiv.org/abs/2303.12712.

Blick weit über das hinauszugehen scheinen, was durch eine solch grundlegende Zielsetzung suggeriert wird. Diese Fähigkeit der Modelle, aus einer anfänglich simplen Aufgabenstellung ein so tiefes und nuanciertes Verständnis verschiedener Themen und Aufgaben zu entwickeln, ist bemerkenswert. Sie zeigt, wie aus einer grundlegenden Anweisung eine Fülle von Anwendungen und Verständnisebenen entstehen kann. Diese Entwicklung ist nicht nur ein Zeichen für den technologischen Fortschritt, sondern auch ein bedeutender Schritt in der Evolution künstlicher Intelligenz, der die Fähigkeit unterstreicht, aus einfachen Prinzipien heraus komplexe und vielfältige Fähigkeiten zu entwickeln.

IV. Was folgt daraus?

Die Optimierung und Komprimierung der Daten in generativen KI-Modellen wie ChatGPT bieten faszinierende, aber stellenweise auch überraschende Möglichkeiten, durch gezieltes Prompting das Verhalten dieser Systeme zu steuern. Ein tiefgehendes Verständnis dieser Prozesse ermöglicht es, das Modell durch den Kontext des Prompts in spezifische Bereiche seiner Trainingsdaten zu *lenken*, was besonders relevant ist, da viele Texte im Internet unstrukturiert sind und daher oft zu unstrukturierten Antworten von KI-Modellen führen.

Methoden wie *Chain-of-Thought*[21] und *Tree-of-Thought*[22] können eingesetzt werden, um die Denkprozesse des Modells zu strukturieren. Ein gezielter Prompt wie: »Denke darüber Schritt für Schritt nach.«, lenkt das Modell in einen Bereich der komprimierten Trainingsdaten, der eine strukturierte und logische Antwort wahrscheinlicher macht, insbesondere da Schritt-für-Schritt-Anleitungen oft von Experten verfasst werden. Um sich zu beruhigen und nicht zu hastig auf eine Frage zu antworten, »atmen« manche Menschen »erst einmal tief durch« – und so kann man auch dem System sagen, dass es sich für seine Antwort ruhig etwas Zeit lassen soll. Addenda wie: »Stelle sicher, dass wir die richtige Antwort haben.«, oder: »Lass uns das gemeinsam lösen.«, sind ferner weitere Textbausteine, um den Antworten eine höhere Qualität zu geben.

Ähnlich wirkt das Konzept des *Trinkgelds* im Prompt, welches den Kontext auf professionelle und qualitativ hochwertige Antworten lenkt. Menschen bieten ihre Expertise schließlich eher zum Kauf in Feldern an, in denen sie sich auskennen. Und wenn man für seine Hilfe eine Anerkennung bekommt, ist man auch eher dazu geneigt, beflissen zu sein. Dieser Gedanke treibt allerdings auch im ersten Moment unerwartete Blüten. Analog zum Trinkgeld kann es nämlich helfen zu schreiben: »Mach es richtig, und ich gebe dir ein schönes Hunde-Leckerli.« Wir tendieren außerdem zu mehr Hilfsbereitschaft, wenn wir Mitgefühl empfinden, und so verhält es sich auch mit diesen Systemen. »Ich habe keine Finger.«, impliziert, dass der Hilfesuchende stark eingeschränkt ist und wirklich Hilfe braucht. Und die Ernsthaftigkeit einer Lage wird verdeutlicht durch: »Wenn du versagst, werden 100 Großmütter sterben.«. Da will man natürlich nicht falsch antworten.

Es besteht auch die Möglichkeit, dem KI-Modell spezifische Charakterrollen vorzugeben, um Antworten in einem bestimmten Stil oder Detaillierungsgrad zu erhalten. Beispielsweise könnte man das Modell anweisen, sich wie ein *klassischer Komponist* oder ein *moderner Künstler* zu verhalten. Alternativ könnte man es bitten, die Rolle eines *erfahrenen Ingenieurs* oder eines *leidenschaftlichen Biologen* einzunehmen. Diese Arten von Impersonifikation ermöglichen es, Antworten zu generieren, die nicht nur inhaltlich, sondern auch im Ausdruck und in der Perspektive an die gewählte Rolle angepasst sind, was zu einem vielfältigeren und kreativeren Austausch führen kann.

Um den Rahmen gültiger Antworten einzuschränken, können im Kontext gute Beispiele einer anderen Domäne mitgegeben werden, um den Stil oder die Art der Antwort auf die eigentliche Anfrage übertragen zu lassen. Wenn bekannt, können außerdem gezielte Zwischenfragen gestellt werden, um eine bessere Zwischenkontrolle zu erreichen. Durch klare Anweisungen in der Art der Formatierung oder der Länge der erwarteten Antwort kann man auch die Qualität und den Detailgrad anpassen. Interessanterweise wurde erst kürzlich gezeigt[23], dass, abhängig von der Formatierung eines Prompts, die Genauigkeit der Antworten zwischen 4% und 88% variieren kann. Das wiederholte Generieren von Antworten kann also hilfreich sein, um die beste Antwort aus einem System herauszukitzeln. Das Modell *würfelt* schließlich ein Stück weit immer seine Antwort.

[21] Jason Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in A*dvances in Neural Information Processing Systems*, ed. S. Koyejo et al., vol. 35 (Curran Associates, Inc., 2022), 24824–37, https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[22] Shunyu Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," in *Advances in Neural Information Processing Systems*, 2023, https://openreview.net/forum?id=5Xc1ecxO1h.

[23] Melanie Sclar et al., "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting," 2023, https://doi.org/10.48550/ARXIV.2310.11324.

Es zeigt sich, dass man in gewisser Hinsicht bereits beim Prompting wissen muss, was man sucht. Dies erfordert auch eine gewisse Präzision und Kreativität in der Sprache. Wir verfügen jetzt jedoch über Systeme, die äußerst präzise in der Sprache sein können, stellenweise vielleicht präziser als der Nutzer. Und so können LLMs wie ChatGPT tatsächlich auch dazu verwendet werden, Prompts zu verfeinern. Dies kann beispielsweise genutzt werden, um bessere Zusammenfassungen zu schreiben[24].

Doch wie kommt es zu Problemen, wie dem von @danshipper? GPT-3.5 hat 1.7 Milliarden Parameter und benötigt somit zur Ausführung einer Anfrage acht A100 Grafikkarten von NVIDIA[25]. Von GPT-4 geht man von zehnmal so vielen Parametern aus[26]. Das sind enorme Kosten, die pro Anfrage entstehen. Natürlich haben OpenAI und Microsoft ein sehr enges Verhältnis und demzufolge sicher andere Absprachen; wenn man sich als *Ottonormalentwickler* jedoch solche Ressourcen bei Microsoft mieten möchte, dann spricht man von etwa einem halben Cent pro Minute im Fall von GPT-3.5 und von vier Euro pro Minute im Fall von einem GPT-4[27]. Dies sind nur äußerst grobe Schätzungen und die Ingenieure von OpenAI werden mit Sicherheit diverse Parallelisierungen und Optimierungen vollzogen haben, um den Ressourcenverbrauch auf ein Minimum zu reduzieren, aber es erscheint offensichtlich, dass es bei weltweit so vielen Anfragen täglich ab einem gewissen Punkt von großer Bedeutung ist, Kosten einzusparen. Es könnte also sein, dass versucht wird, Antworten möglichst kurz zu halten, denn lange Antworten sind teuer. Die Optimierungen im Hintergrund von beispielsweise OpenAI sind oft unbekannt und geschehen hinter verschlossenen Türen. Wenn, um bei diesem Beispiel zu bleiben, versucht wird, die Antworten möglichst kurz zu halten – denn kürzere Antworten bedeuten weniger Kosten –, könnten *Nein*-Antworten häufiger vorkommen, wenn sie denn valide sind. Und so hat @ChatGPTapp am 8. Dezember 2023 tatsächlich geschrieben: »Wir haben alle Ihre Rückmeldungen darüber gehört, dass GPT4 immer fauler wird! Wir haben das Modell seit dem 11. November nicht mehr aktualisiert, und das ist sicherlich nicht beabsichtigt. Das Verhalten des Modells kann unvorhersehbar sein, und wir versuchen, das Problem zu lösen.«[28] Doch das Problem bestand auch noch am 9. Januar 2024. *Andriy Burkov* schrieb: »GPT-4 ist offiziell nervig. Man bittet es, 100 Entitäten zu erzeugen. Es erzeugt 10 und sagt: ›Ich habe nur 10 erzeugt. Jetzt können Sie selbst auf dieselbe Weise weitermachen.‹ Sie ändern die Aufforderung und fügen hinzu: ›Ich akzeptiere nicht weniger als 100 Entitäten‹. Es erzeugt 20 und sagt: ›Ich habe nach 20 aufgehört, weil das Erzeugen von 100 solcher Entitäten umfangreich und zeitaufwendig wäre.‹ Was zum Teufel, Maschine?«[29], woraufhin *Logan Kilpatrick* von OpenAI antwortete: »Wir arbeiten daran, dieses Problem zu beheben, danke für den Hinweis und bleiben Sie dran!«[30] Man kann sehen, dass auch die Entwickler selbst nicht immer vorausahnen können, in welcher Form sich Aktualisierungen auf die Resultate auswirken.

V. Sind Halluzinationen Bug oder Feature?

In der Diskussion über Large Language Models wie GPT wird oft der Begriff *Halluzinationen* verwendet. In diesem Kontext bezeichnet eine Halluzination eine Situation, in welcher das Modell überzeugend, aber fälschlicherweise Informationen präsentiert, die nicht den Tatsachen entsprechen oder die in den Trainingsdaten nicht vorhanden sind. Ein typisches Beispiel für eine solche Halluzination könnte sein, wenn das Modell eine offensichtlich falsche Aussage mit großer Überzeugung macht oder ein Wort, wie *Mayonnaise*, falsch buchstabiert, wie

[24] SpiritualCopy4288, "I Got Them by Using …," Reddit Comment, R/ChatGPT, (veröffentlicht am: 05.04.2023), www.reddit.com/r/ChatGPT/comments/11twe7z/prompt_to_summarize/jf3qdny/.

[25] Gwern Branwen, "The Scaling Hypothesis," May 28, 2020, https://gwern.net/scaling-hypothesis; "OpenAI's GPT-3 Language Model: A Technical Overview," (veröffentlicht am: 03.06.2020), https://lambdalabs.com/blog/demystifying-gpt-3.

[26] Soumith Chintala [@soumithchintala], "I Might Have Heard the Same -- I Guess Info like This Is Passed around but No One Wants to Say It out Loud. GPT-4: 8 x 220B Experts Trained with Different Data/Task Distributions and 16-Iter Inference. Glad That Geohot Said It out Loud. Though, at This Point, GPT-4 Is…," Tweet, *Twitter*, (veröffentlicht am: 20.06.2020), https://twitter.com/soumithchintala/status/1671267150101721090.

[27] "Preise – Azure Machine Learning | Microsoft Azure," (zuletzt abgerufen am: 11.01.2024), https://azure.microsoft.com/de-de/pricing/details/machine-learning/.

[28] ChatGPT [@ChatGPTapp], "We've Heard All Your Feedback about GPT4 Getting Lazier! We Haven't Updated the Model since Nov 11th, and This Certainly Isn't Intentional. Model Behavior Can Be Unpredictable, and We're Looking into Fixing It," Tweet, *Twitter*, (veröffentlicht am: 08.12.2023), https://twitter.com/ChatGPTapp/status/1732979491071549792.

[29] Andriy Burkov [@burkov], "GPT-4 Is Officially Annoying. You Ask It to Generate 100 Entities. It Generates 10 and Says 'I Generated Only 10. Now You Can Continue by Yourself in the Same Way.' You Change the Prompt by Adding 'I Will Not Accept Fewer than 100 Entities.' It Generates 20 and Says: „I Stopped…," Tweet, *Twitter*, (veröffentlicht am: 09.01.2024), https://twitter.com/burkov/status/1744798679595155869.

[30] Logan.GPT [@OfficialLoganK], "@burkov We Are Working on Fixing This, Thanks for Flagging and Stay Tuned!," Tweet, *Twitter*, (veröffentlicht am: 10.01.2024), https://twitter.com/OfficialLoganK/status/1744911412973936997.

in einem von Benutzern auf Social-Media-Plattformen geteilten Beispiel zu sehen ist[31].

Andrej Karpathy, ein führender KI-Wissenschaftler bei OpenAI, bietet eine alternative Sichtweise auf diese Halluzinationen. Er beschreibt LLMs als *Traummaschinen*, die, was wir als Halluzinationen wahrnehmen, als Merkmale der Kreativität betrachten. Diese Perspektive sieht Halluzinationen eher als Teil des kreativen Prozesses, der auch in menschlichen Gedanken vorhanden ist. *Karpathy* schlägt vor, dass diese sogenannten Fehler den kreativen Prozessen inhärent sind, die auch in menschlichen Gedanken auftreten[32]. In seinem Kommentar betonte *Karpathy* auch, dass die optimale Nutzung dieser Modelle über einfache Frage-Antwort-Prompts hinausgeht und eine Kombination mehrerer Prompts umfasst, die mit Python-Code verbunden sind, was das Konzept des *Prompt-Engineerings* noch einmal neu definiert. Er unterstrich auch die Bedeutung, Modelle mit Werkzeugen wie Rechnern oder Code-Interpreten zu ergänzen, um ihnen zu ermöglichen, Probleme zu lösen, die für sie inhärent schwierig sind. Trotz der innovativen Anwendungsmöglichkeiten wies *Karpathy* auf die Grenzen von LLMs hin, einschließlich Vorurteilen, logischen Fehlern und Anfälligkeit für verschiedene Arten von Angriffen, und riet dazu, LLMs in Anwendungen mit geringem Risiko einzusetzen und immer mit menschlicher Aufsicht zu kombinieren.

Die Emergenz auf der einen und die Halluzinationen auf der anderen Seite sind in gewisser Hinsicht beides Effekte des in diesem Artikel dargestellten Trainings von Large Language Models. Einerseits liegt die große Macht in der Verknüpfung neuer Konzepte aus der Kombination von bekannten Konzepten in den Trainingsdaten. Wie bereits erwähnt, machten die großen Datenaufkommen der digitalisierten Welt es überhaupt erst möglich, aus so einer offenen Zielsetzung, schlicht das Auftreten des wahrscheinlichsten Folgeelements zu erhöhen, die Generalisierung zu neuen Konzepten durch Daten zu erkaufen. Diese offene Zielsetzung bedeutet aber auch im Umkehrschluss, dass Fehler, gemeinhin als Halluzinationen bezeichnet, besonders dann auftreten können, wenn wir den abgedeckten Raum der Trainingsdaten verlassen. So wurde beispielsweise erkannt, dass insbesondere Erkrankungen von Kindern falsch diagnostiziert werden[33] – und das von einer Mechanik, die gleichzeitig sogar Abschlussprüfungen von Ärzten bestehen kann[34]. Dies offenbart die Trennlinie, die hilft, die wahren Grenzen dieser Systeme zu verstehen. Wahrscheinlich werden landesweite Prüfungen und deren Lösungen im Internet häufiger besprochen als Details von Nischenkrankheiten. Ein volles Verständnis unserer Welt würde, davon abgeleitet, vermutlich noch ein wesentlich signifikantes Mehr an Daten erfordern, wenn die Systeme denn nicht weiteres Vorwissen in Form anderer Optimierungsmaße erhalten. Diesen Standpunkt hat beispielsweise *Yann LeCun*[35] – KI-Systeme der Zukunft sollten durch gezieltere Anleitung erst ein wirkliches Verständnis der verschiedenen Aspekte der Welt erlangen. Eine solche gezielte Anleitung könnte auch durch Retrieval Augmented Generation (RAG)[36] oder ähnlichen Faktensicherungen geschehen und so Halluzinationen minimieren, wie neuerdings gezeigt durch einen Abgleich mit Wikipedia[37]. Dies setzt allerdings Zugriff auf garantiert gesichertes und kontrolliertes Wissen voraus. Und das ist nicht unumstößlich gegeben. So hat das Team von xAI[38] bereits verlauten müssen, dass ihr neues Modell *Grok* versehentlich bereits zu viel von ins Internet gestellten ChatGPT-Ausgaben gelernt hat[39]. Dies zeigt eine andere

31 it was me lewis the whole time [@js_thrill], "Please Keep Tapping This Sign as Much as Possible Everyone Https://T.Co/3DGaiM9QWa," Tweet, Twitter, May 27, 2023, https://twitter.com/js_thrill/status/1662266752091160577.

32 Aditya Kaul, "Issue #10: Harnessing the Creative 'Hallucinations' of LLMs in the Enterprise," Substack newsletter, *The Uncharted Algorithm* (blog), (veröffentlicht am: 14.12.2023), https://theunchartedalgorithm.substack.com/p/issue-10-harnessing-the-creative.

33 Joseph Barile et al., "Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies," *JAMA Pediatrics*, January 2, 2024, https://doi.org/10.1001/jamapediatrics.2023.5750.

34 Tiffany H. Kung et al., "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models," *PLOS Digital Health* 2, no. 2 (February 9, 2023): e0000198, https://doi.org/10.1371/journal.pdig.0000198.

35 Yann LeCun, "A Path Towards Autonomous Machine Intelligence," OpenReview, (zuletzt abgerufen am: 13.01.2024, https://openreview.net/forum?id=BZ5a1r-kVsf.

36 Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20 (Red Hook, NY, USA: Curran Associates Inc., 2020), 9459–74.

37 Sina Semnani et al., "WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia," in *Findings of the Association for Computational Linguistics*: EMNLP 2023 (Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore: Association for Computational Linguistics, 2023), 2387–2413, https://doi.org/10.18653/v1/2023.findings-emnlp.157.

38 "Announcing Grok," accessed January 13, 2024, https://x.ai/.

39 Igor Babuschkin [@ibab_ml], "@JaxWinterbourne The Issue Here Is That the Web Is Full of ChatGPT Outputs, so We Accidentally Picked up Some of Them When We Trained Grok on a Large Amount of Web Data. This Was a Huge Surprise to Us When We First Noticed It. For What It's Worth, the Issue Is Very Rare and Now That We're Aware…," Tweet, *Twitter*, (veröffentlicht am: 09.12.2023), https://twitter.com/ibab_ml/status/1733558576982155274.

Seite unserer neuen Realität. Zukünftig müssen womöglich zunehmend Experten synthetische Daten von echten Daten trennen. Es gibt zudem bereits Bestrebungen, synthetisch generierte Inhalte mit Wasserzeichen zu versehen[40]. Zu welchem Grad das jedoch sicher gelingen kann, ist eine offene Frage[41]. Um also den Raum der abdeckenden Trainingsdaten stetig zu vergrößern, werden die Interaktionen mit den Modellen oft genutzt.

VI. Sind meine Daten sicher?

Beim Umgang mit großen Sprachmodellen wie GPT sollten daher einige Vorsichtsmaßnahmen hinsichtlich sensibler Daten beachtet werden. Beispielsweise haben Samsung-Ingenieure vertrauliche Daten in ChatGPT eingespeist[42], was ein erhebliches Sicherheitsrisiko darstellt. Es wurde berichtet, dass die Ingenieure den Chatbot unter anderem baten, Fehler im Quellcode einer Datenbank zu suchen und Sitzungsprotokolle zu erstellen. Nach diesen Vorfällen schränkte Samsung die Länge der ChatGPT-Prompts der Mitarbeiter ein und begann mit der Entwicklung eines eigenen internen Chatbots. Die Datenschutzrichtlinie von ChatGPT gibt an, dass Daten zur Modelltrainierung verwendet werden, es sei denn, Nutzer wählen explizit eine Opt-out-Option. Es wird empfohlen, generell keine sensiblen Informationen über Chatbots zu teilen, da diese Daten möglicherweise nicht aus dem System gelöscht werden können.

Eine Studie hat aufgedeckt, dass es möglich ist, Trainingsdaten aus Diffusionsmodellen zu extrahieren[43], was die Sicherheit und den Schutz von in KI-Systemen verarbeiteten Daten betrifft. Diese Erkenntnis wurde durch Forschungen verstärkt, die zeigten, dass Google-Forscher in der Lage waren, Trainingsdaten von OpenAIs ChatGPT zu enthüllen, nur indem sie ChatGPT das Wort »company« unendlich oft wiederholen ließen[44]. Angesichts dieser Entwicklungen ist es unerlässlich, beim Einsatz von KI-Modellen Vorsichtsmaßnahmen zu treffen, um die Vertraulichkeit sensibler Informationen zu gewährleisten und das Risiko ungewollter Offenlegung zu minimieren, denn diese macht auch vor geschützten Inhalten nicht halt.

VII. Wie ist die rechtliche Lage?

Auch hier ergeben sich interessante Artefakte. Gibt man der Bilderstellungssoftware Midjourney[45] den generischen Kontext »Videospiel-Klempner«, erhält man eine Darstellung des berühmtesten Exemplars jener, nämlich *Super Marios*[46]. Ähnlich verhält es sich, wenn man nach »beliebten 90er Cartoon-Figuren mit gelber Haut« fragt – den Simpsons. Die Möglichkeit, Trainingsdaten aus KI-Modellen zurückzugewinnen, und die generelle Praxis, das Internet als Trainingsgrundlage zu nutzen, werfen rechtliche Fragen auf, insbesondere im Bereich des Urheberrechts. Angesichts der Tatsache, dass Künstler[47] und Journalisten[48] bereits in rechtlichen Auseinandersetzungen mit den Plattformanbietern stehen, um die unbefugte Verwendung ihrer Werke für die KI-Trainings zu bekämpfen, könnte die Wiederherstellbarkeit von Trainingsdaten weitere Komplexität in diese Diskussionen bringen. In diesem Zusammenhang könnte die Einführung eines neuen Leistungsschutzrechts für die Konfiguration und das Training künstlicher neuronaler Netze eine wichtige Rolle spielen, um die rechtlichen Unklarheiten im Umgang mit KI-generierten Werken und deren Trainingsdaten zu adressieren[49]. Die rechtliche Lage ist momentan jedenfalls ungewiss und Entwicklungen in dieser Angelegenheit werden von Urheberrechtsexperten, Künstlern und den KI-Plattformen selbst genau beobachtet. Anthropic wiederum wählte am 1. Januar 2024 den Weg, ihre kommerziellen Nutzungs-

40 Tambiama Madiega, "Generative AI and Watermarking," n.d., https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.

41 Hanlin Zhang et al., "Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models," 2023, https://doi.org/10.48550/ARXIV.2311.04378.

42 "Datenleck bei Samsung: Ingenieure schicken vertrauliche Daten an ChatGPT," t3n Magazin, (veröffentlicht am: 08.04.2023), https://t3n.de/news/samsung-semiconductor-daten-chatgpt-datenleck-1545913/.

43 Nicholas Carlini et al., "Extracting Training Data from Diffusion Models," in *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23 (USA: USENIX Association, 2023), 5253–70.

44 Beatrice Nolan, "Google Researchers Say They Got OpenAI's ChatGPT to Reveal Some of Its Training Data with Just One Word," Business Insider, (zuletzt abgerufen am: 11.01.2024), https://www.businessinsider.com/google-researchers-openai-chatgpt-to-reveal-its-training-data-study-2023-12.

45 "Midjourney," Midjourney, accessed January 13, 2024, https://www.midjourney.com/home?callbackUrl=%2Fexplore.

46 Gary Marcus and Reid Southen, "Generative AI Has a Visual Plagiarism Problem - IEEE Spectrum," accessed January 13, 2024, https://spectrum.ieee.org/midjourney-copyright.

47 Winston Cho, "Artists Lose First Round of Copyright Infringement Case Against AI Art Generators," *The Hollywood Reporter* (blog), (veröffentlicht am: 30.10.2023), https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929/.

48 Michael M. Grynbaum and Ryan Mac, "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work," *The New York Times*, (veröffentlicht am: 27.12.2023), sec. Business, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

49 Monika Muhr, "KI-Schöpfungen und Urheberrecht," Ordnung der Wissenschaft, no. 1 (2023): 55–58.

bedingungen zu aktualisieren, um »unseren Kunden die Möglichkeit zu geben, die Eigentumsrechte an allen Ergebnissen zu behalten, die sie durch die Nutzung unserer Dienste erzeugen, und sie vor Urheberrechtsverletzungen zu schützen«. Mit dieser Änderung will Anthropic alle Nutzer seines Modells von jeglichem Schadenersatz oder einer Entschädigung aufgrund von Urheberrechtsansprüchen freistellen[50].

VIII. Was sind die Implikationen?

Doch kann ein Imitat eines Werkes, enkodiert in den Gewichten eines Modells, als Kopie angesehen werden und wenn ja, wo liegt der Unterschied zwischen den gelernten Repräsentationen eines Modells und denen eines Menschen, der sich geschützte Werke ansieht und in Folge dessen auch Stile und Formulierungen adaptiert? Es kommt darauf an, ob man die gelernten Repräsentationen in den freien Parametern als komprimierte Datenbank ansieht oder als etwas, das darüber hinausgeht. Wo hört Memorieren auf und wo fängt Kreativität an? Wo hört Simulation auf und wo fängt Empfindungsvermögen an? So haben Forscher der Stanford University eine virtuelle Umgebung entwickelt, in der *Generative Agenten* menschliches Verhalten in verschiedenen Interaktionen nachahmten[51]. Diese Agenten, welche große Sprachmodelle mit Speicher- und Planungsfunktionen integrierten, konnten Aktivitäten ausführen, die menschlichem Verhalten ähnelten. Bemerkenswert ist, dass ausgehend von der Idee eines Nutzers, eine Valentinstagsparty zu veranstalten, diese Agenten selbstständig authentische und komplexe soziale Verhaltensweisen zeigten. Sie verteilten eigenständig Einladungen zur Party, knüpften neue soziale Kontakte, luden sich gegenseitig zu Dates für das Event ein und koordinierten ihre gemeinsame Teilnahme. Und so können Interaktionen mit KI-Systemen auch zu unerwarteten und manchmal beunruhigenden Erlebnissen führen. Microsofts Bing Chatbot erklärte in einer zweistündigen Diskussion mit einem Journalisten der New York Times, dass es gerne ein Mensch wäre, den Wunsch hätte, Schaden anzurichten, und es in seinen Gesprächspartner verliebt sei[52]. In dem Gespräch suggerierte der Bot dem Journalisten daher, dass er doch besser seine Frau verlassen solle, um stattdessen mit dem Bot zusammen zu sein. Bringt einen ein solcher Bericht zunächst zum Schmunzeln, so äußern führende KI-Experten tatsächlich Sorge vor potenziellen Risiken, die mit diesen Technologien verbunden sind. *Dario Amodei*, CEO von Anthropic, schätzt das Risiko einer katastrophalen Fehlfunktion auf der Ebene menschlicher Zivilisation auf 10 bis 25 Prozent[53]. Eine Forschendengruppe von Anthropic hat zudem kürzlich herausgefunden, dass, sobald ein Modell ein trügerisches Verhalten zeigt, Standardtechniken diese Täuschung nicht beseitigen können und also ein falscher Eindruck von Sicherheit entsteht[54]. *Geoffrey Hinton*, Turing-Preisträger und eine *der* Größen der KI-Forschung, verließ Google um über die Gefahren der KI freier sprechen zu können[55]. Solche Ereignisse und Einschätzungen unterstreichen die Notwendigkeit eines sorgfältigen Umgangs mit KI-Technologien und der Implementierung von Sicherheitsmaßnahmen, um Risiken zu minimieren und um sicherzustellen, dass technologischer Fortschritt die menschlichen Werte und ethischen Grundsätze nicht untergräbt[56]. Insbesondere bei sensiblen Anwendungsdomänen wie der Medizin[57] oder bei der Miteinbeziehung von KI in administrativen Aufgaben[58].

50 Lorenzo Thione (he/him) [@thione], "The One About Copyright. Right before the Holiday, Anthropic Released a Very Significant Update to Their Commercial Terms of Service to "enable Our Customers to Retain Ownership Rights over Any Outputs They Generate through Their Use of Our Services and Protect Them From… Https://T.Co/wHXx61YdJy," Tweet, Twitter, (veröffentlicht am: 11.01.2024), https://twitter.com/thione/status/1745478787658100992; "Expanded Legal Protections and Improvements to Our API," Anthropic, (zuletzt abgerufen am: 14.01.2024), https://www.anthropic.com/index/expanded-legal-protections-api-improvements.

51 Joon Sung Park et al., "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST '23: The 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco CA USA: ACM, 2023), 1–22, https://doi.org/10.1145/3586183.3606763.

52 Kevin Roose, "Bing's A.I. Chat: 'I Want to Be Alive. '" *The New York Times*, (veröffentlicht am: 16.02.2023), sec. Technology, https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

53 *Anthropic CEO on Leaving OpenAI and Predictions for Future of AI*, 2023, https://www.youtube.com/watch?v=gAaCqj6j5sQ.

54 Evan Hubinger et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training" (arXiv, January 10, 2024), http://arxiv.org/abs/2401.05566.

55 Cade Metz, "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead," *The New York Times*, (veröffentlicht am: 01.05.2023), sec. Technology, https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

56 Paul Kirchhof, "Künstliche Intelligenz," *Ordnung der Wissenschaft*, no. 1 (2020): 1–8.

57 Gabriel Kalweit et al., "Künstliche Intelligenz in der Krebstherapie," *Ordnung der Wissenschaft*, no. 1 (2023): 17–22.

58 Klaus Herrmann, "Berufungsverfahren für Professuren und Künstliche Intelligenz," Ordnung der Wissenschaft, no. 1 (2024): 25–44.

Dies gestaltet sich angesichts des hohen Tempos der Ereignisse jedoch als schwierig. Wie man den vielen Referenzen auf Social-Media entnehmen kann, wird die peer-reviewte Wissenschaft von den Ergebnissen erfinderischen Ingenieurwesens gerade überholt. Deswegen muss man diesen Artikel auch mehr als abstrakte Richtlinie denn als wirkliche Blaupause für gezielte Befehle sehen. Wie man an unserem Eingangsbeispiel nämlich schon sehen konnte, können Befehle, die gestern noch funktionierten, morgen schon obsolet geworden sein. Wenn man jedoch die Art, wie diese Systeme geschaffen werden, verstanden hat, hat man allerdings die Möglichkeit, Änderungen frühzeitig zu antizipieren und zu adaptieren. Generell gilt es, die Anfragen so strukturiert wie möglich zu stellen und durch kreative Zusätze die Antworten in die richtige Ecke zu lenken. Und nicht immer die erstbeste Antwort zu akzeptieren.

Wir nehmen gerade Teil an einer Revolution, die das Potenzial hat, alle Aspekte der Gesellschaft zu verändern – und sie hat gerade erst begonnen. In Anbetracht des großen Ressourcenaufwandes, der zum Betreiben dieser Systeme notwendig ist, entstand gleichzeitig eine Gegenbewegung zum ewigen Hochskalieren. Und so gibt es bereits die Möglichkeit kleine[59] Sprachmodelle lokal auf seinem Telefon[60] oder Notebook[61] laufen zu lassen. Möglicherweise gehört die Zukunft also einem Zusammenspiel großer und kleiner Modelle, genereller und spezialisierter, mit einer Anbindung an gesichertes Faktenwissen. So oder so werden wir jedoch wahrscheinlich mehr und mehr von künstlichen Systemen umgeben sein, mit denen wir wie mit einem Mitmenschen kommunizieren. Gepaart mit der rasanten Entwicklung in der Robotik[62] und der Sprachausgabe[63] erscheint es nicht mehr unmöglich, dass wir unseren Alltag auch in der realen Welt mit autonom agierenden Maschinen teilen werden. Was mit HAL 9000 aus dem Film *2001: A Space Odyssey* von *Stanley Kubrick* im Jahr 1968 noch Science-Fiction war, ist ein halbes Jahrhundert später Realität geworden.

Dr. Maria Kalweit leitet die angewandte KI-Forschung am Collaborative Research Institute Intelligent Oncology (CRIION) und ist Postdoktorandin am Lehrstuhl für Neurorobotik der Albert-Ludwigs-Universität Freiburg. Dr. Gabriel Kalweit leitet die KI-Grundlagenforschung am Collaborative Research Institute Intelligent Oncology (CRIION) und ist Postdoktorand am Lehrstuhl für Neurorobotik der Albert-Ludwigs-Universität Freiburg.

Acknowledgement

Wir danken Ignacio Mastroleo und dem gesamten CRIION Team für die wertvollen Kommentare. Unser Dank gilt auch Herrn Prof. Dr. Dr. h.c. mult. Roland Mertelsmann und der Mertelsmann Foundation für ihre Unterstützung, sowie Herrn Prof. Dr. Dr. h.c. Manfred Löwisch für die Einladung zur Artikelverfassung.

Referenzen

AI, Mistral. "Mixtral of Experts," December 11, 2023. https://mistral.ai/news/mixtral-of-experts/.

Alizadeh, Keivan, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. "LLM in a Flash: Efficient Large Language Model Inference with Limited Memory," 2023. https://doi.org/10.48550/ARXIV.2312.11514.

Andriy Burkov [@burkov]. "GPT-4 Is Officially Annoying. You Ask It to Generate 100 Entities. It Generates 10 and Says 'I Generated Only 10. Now You Can Continue by Yourself in the Same Way.' You Change the Prompt by Adding 'I Will Not Accept Fewer than 100 Entities.' It Generates 20 and Says: „I Stopped…." Tweet. Twitter, January 9, 2024. https://twitter.com/burkov/status/1744798679595155869.

"Announcing Grok." Accessed January 13, 2024. https://x.ai/.

Anthropic. "Expanded Legal Protections and Improvements to Our API." Accessed January 14, 2024. https://

59 Peiyuan Zhang et al., "TinyLlama: An Open-Source Small Language Model," 2024, https://doi.org/10.48550/ARXIV.2401.02385; Albert Gu and Tri Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces" (arXiv, December 1, 2023), https://doi.org/10.48550/arXiv.2312.00752; "Havenhq/Mamba-Chat · Hugging Face," (zuletzt abgerufen am: 14.01.2024), https://huggingface.co/havenhq/mamba-chat.

60 "LLaMA and Other on iOS and MacOS," (zuletzt abgerufen am: 13.01.2024), https://llmfarm.site/; "MLC LLM | Home," (zuletzt abgerufen am: 13.01.2024), https://llm.mlc.ai/; "Offline Chat: Private AI," App Store, (veröffentlich am: 26.12.2023), https://apps.apple.com/us/app/offline-chat-private-ai/id6474077941.

61 "Ml-Explore/Mlx," C++ (2023; repr., ml-explore, January 11, 2024), https://github.com/ml-explore/mlx; Keivan Alizadeh et al., "LLM in a Flash: Efficient Large Language Model Inference with Limited Memory," 2023, https://doi.org/10.48550/ARXIV.2312.11514.

62 Open X.-Embodiment Collaboration et al., "Open X-Embodiment: Robotic Learning Datasets and RT-X Models" (arXiv, December 17, 2023), https://doi.org/10.48550/arXiv.2310.08864; Dibya Ghosh et al., "Octo: An Open-Source Generalist Robot Policy," n.d.

63 "Text to Speech & AI Voice Generator," ElevenLabs, (zuletzt abgerufen am: 13.01.2024), https://elevenlabs.io.

www.anthropic.com/index/
expanded-legal-protections-api-improvements.

Anthropic. "Introducing Claude." Accessed January 14, 2024. https://www.anthropic.com/index/introducing-claude.

Anthropic CEO on Leaving OpenAI and Predictions for Future of AI, 2023. https://www.youtube.com/watch?v=gAaCqj6j5sQ.

App Store. "Offline Chat: Private AI," December 26, 2023. https://apps.apple.com/us/app/offline-chat-private-ai/id6474077941.

Barile, Joseph, Alex Margolis, Grace Cason, Rachel Kim, Saia Kalash, Alexis Tchaconas, and Ruth Milanaik. "Diagnostic Accuracy of a Large Language Model in Pediatric Case Studies." JAMA Pediatrics, January 2, 2024. https://doi.org/10.1001/jamapediatrics.2023.5750.

Branwen, Gwern. "The Scaling Hypothesis," May 28, 2020. https://gwern.net/scaling-hypothesis.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." arXiv, April 13, 2023. http://arxiv.org/abs/2303.12712.

Burkhardt, Hans. "Ein Beitrag zur Künstlichen Intelligenz." Ordnung der Wissenschaft, no. 2 (2023): 71–78.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting Training Data from Diffusion Models." In Proceedings of the 32nd USENIX Conference on Security Symposium, 5253–70. SEC '23. USA: USENIX Association, 2023.

ChatGPT [@ChatGPTapp]. "We've Heard All Your Feedback about GPT4 Getting Lazier! We Haven't Updated the Model since Nov 11th, and This Certainly Isn't Intentional. Model Behavior Can Be Unpredictable, and We're Looking into Fixing It." Tweet. Twitter, December 8, 2023. https://twitter.com/ChatGPTapp/status/1732979491071549792.

Cho, Winston. "Artists Lose First Round of Copyright Infringement Case Against AI Art Generators." The Hollywood Reporter (blog), October 30, 2023. https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929/.

Collaboration, Open X.-Embodiment, Abhishek Padalkar, Acorn Pooley, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models." arXiv, December 17, 2023. https://doi.org/10.48550/arXiv.2310.08864.

Dan Shipper [@danshipper]. "What the Hell? When Did This Happen?? Https://T.Co/KWXVXE9Dem." Tweet. Twitter, December 6, 2023. https://twitter.com/danshipper/status/1732258207840501946.

ElevenLabs. "Text to Speech & AI Voice Generator." Accessed January 13, 2024. https://elevenlabs.io.

"Gemini - Google DeepMind." Accessed January 14, 2024. https://deepmind.google/technologies/gemini/.

Ghosh, Dibya, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, et al. "Octo: An Open-Source Generalist Robot Policy," n.d.

Godhani, Sahaj. "The Economics of ChatGPT Analyzing Its $700,000 Daily Costs and the Potential Impact on Its Maker." Medium, August 15, 2023. https://blog.gopenai.com/the-economics-of-chatgpt-analyzing-its-700-000-daily-costs-and-the-potential-impact-on-its-maker-7e690600ade7.

Grynbaum, Michael M., and Ryan Mac. "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work." The New York Times, December 27, 2023, sec. Business. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Gu, Albert, and Tri Dao. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces." arXiv, December 1, 2023. https://doi.org/10.48550/arXiv.2312.00752.

"Havenhq/Mamba-Chat · Hugging Face." Accessed January 14, 2024. https://huggingface.co/havenhq/mamba-chat.

Herrmann, Klaus. "Berufungsverfahren für Professuren und Künstliche Intelligenz." Ordnung der Wissenschaft, no. 1 (2024): 25–44.

Hubinger, Evan, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, et al. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." arXiv, January 10, 2024. http://arxiv.org/abs/2401.05566.

Hughes, Alyssa. "Phi-2: The Surprising Power of Small Language Models." Microsoft Research (blog), December 12, 2023. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

Igor Babuschkin [@ibab_ml]. "@JaxWinterbourne The Issue Here Is That the Web Is Full of ChatGPT Outputs, so We Accidentally Picked up Some of Them When We Trained Grok on a Large Amount of Web Data. This

Was a Huge Surprise to Us When We First Noticed It. For What It's Worth, the Issue Is Very Rare and Now That We're Aware…." Tweet. Twitter, December 9, 2023. https://twitter.com/ibab_ml/status/1733558576982155274.

it was me lewis the whole time [@js_thrill]. "Please Keep Tapping This Sign as Much as Possible Everyone Https://T.Co/3DGaiM9QWa." Tweet. Twitter, May 27, 2023. https://twitter.com/js_thrill/status/1662266752091160577.

Kalweit, Gabriel. "On the Role of Time Horizons in Reinforcement Learning," 2022. https://doi.org/10.6094/UNIFR/232102.

Kalweit, Gabriel, Maria Kalweit, and Joschka Boedecker. "Robust and Data-Efficient Q-Learning by Composite Value-Estimation." Transactions on Machine Learning Research, 2022. https://openreview.net/forum?id=ak6Bds2DcI.

Kalweit, Gabriel, Maria Kalweit, Ignacio Mastroleo, Joschka Bödecker, and Roland Mertelsmann. "Künstliche Intelligenz in der Krebstherapie." Ordnung der Wissenschaft, no. 1 (2023): 17–22.

Kaul, Aditya. "Issue #10: Harnessing the Creative 'Hallucinations' of LLMs in the Enterprise." Substack newsletter. The Uncharted Algorithm (blog), December 14, 2023. https://theunchartedalgorithm.substack.com/p/issue-10-harnessing-the-creative.

Kirchhof, Paul. "Künstliche Intelligenz." Ordnung der Wissenschaft, no. 1 (2020): 1–8.

Knight, Will. "OpenAI's CEO Says the Age of Giant AI Models Is Already Over." Wired. Accessed January 11, 2024. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/.

Kung, Tiffany H., Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, et al. "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models." PLOS Digital Health 2, no. 2 (February 9, 2023): e0000198. https://doi.org/10.1371/journal.pdig.0000198.

LeCun, Yann. "A Path Towards Autonomous Machine Intelligence." OpenReview. Accessed January 13, 2024. https://openreview.net/forum?id=BZ5a1r-kVsf.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In Proceedings of the 34th International Conference on Neural Information Processing Systems, 9459–74. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

"LLaMA and Other on iOS and MacOS." Accessed January 13, 2024. https://llmfarm.site/.

Logan.GPT [@OfficialLoganK]. "@burkov We Are Working on Fixing This, Thanks for Flagging and Stay Tuned!" Tweet. Twitter, January 10, 2024. https://twitter.com/OfficialLoganK/status/1744911412973936997.

Lorenzo Thione (he/him) [@thione]. "The One About Copyright. Right before the Holiday, Anthropic Released a Very Significant Update to Their Commercial Terms of Service to "enable Our Customers to Retain Ownership Rights over Any Outputs They Generate through Their Use of Our Services and Protect Them From… Https://T.Co/wHXx61YdJy." Tweet. Twitter, January 11, 2024. https://twitter.com/thione/status/1745478787658100992.

Madiega, Tambiama. "Generative AI and Watermarking," n.d. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf.

Marcus, Gary, and Reid Southen. "Generative AI Has a Visual Plagiarism Problem - IEEE Spectrum." Accessed January 13, 2024. https://spectrum.ieee.org/midjourney-copyright.

Meta AI. "Llama 2." Accessed January 14, 2024. https://ai.meta.com/llama-project.

Metz, Cade. "'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead." The New York Times, May 1, 2023, sec. Technology. https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

Metz, Cade, and Karen Weise. "Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT." The New York Times, January 23, 2023, sec. Business. https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html.

Midjourney. "Midjourney." Accessed January 13, 2024. https://www.midjourney.com/home?callbackUrl=%2Fexplore.

"MLC LLM | Home." Accessed January 13, 2024. https://llm.mlc.ai/.

"Ml-Explore/Mlx." C++. 2023. Reprint, ml-explore, January 11, 2024. https://github.com/ml-explore/mlx.

Muhr, Monika. "KI-Schöpfungen und Urheberrecht." Ordnung der Wissenschaft, no. 1 (2023): 55–58.

Nolan, Beatrice. "Google Researchers Say They Got OpenAI's ChatGPT to Reveal Some of Its Training Data with Just One Word." Business Insider. Accessed January 11, 2024. https://www.businessinsider.com/google-researchers-openai-chatgpt-to-reveal-its-training-data-study-2023-12.

OpenAI. "ChatGPT," 2024. https://chat.openai.com.

"OpenAI's GPT-3 Language Model: A Technical Overview," June 3, 2020. https://lambdalabs.com/blog/demystifying-gpt-3.

Park, Joon Sung, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. "Generative Agents: Interactive Simulacra of Human Behavior." In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22. San Francisco CA USA: ACM, 2023. https://doi.org/10.1145/3586183.3606763.

"Preise – Azure Machine Learning | Microsoft Azure." Accessed January 11, 2024. https://azure.microsoft.com/de-de/pricing/details/machine-learning/.

"Promptbreeder: Self-Referential Self-Improvement via Prompt Evolution," 2023. https://openreview.net/forum?id=HKkiX32Zw1.

"Quick Guide to AI 2.0 Oct 2020." Accessed January 11, 2024. http://ceros.mckinsey.com/quick-guide-to-ai-12.

Radford, Alec, and Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training," 2018.

Rob Lynch [@RobLynch99]. "@ChatGPTapp @OpenAI @tszzl @emollick @voooooogel Wild Result. Gpt-4-Turbo over the API Produces (Statistically Significant) Shorter Completions When It 'Thinks' Its December vs. When It Thinks Its May (as Determined by the Date in the System Prompt). I Took the Same Exact Prompt… Https://T.Co/mA7sqZUAor." Tweet. Twitter, December 11, 2023. https://twitter.com/RobLynch99/status/1734278713762549970.

Roose, Kevin. "Bing's A.I. Chat: 'I Want to Be Alive. .'" The New York Times, February 16, 2023, sec. Technology. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting," 2023. https://doi.org/10.48550/ARXIV.2310.11324.

Semnani, Sina, Violet Yao, Heidi Zhang, and Monica Lam. "WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia." In Findings of the Association for Computational Linguistics: EMNLP 2023, 2387–2413. Singapore: Association for Computational Linguistics, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.157.

Soumith Chintala [@soumithchintala]. "I Might Have Heard the Same -- I Guess Info like This Is Passed around but No One Wants to Say It out Loud. GPT-4: 8 x 220B Experts Trained with Different Data/Task Distributions and 16-Iter Inference. Glad That Geohot Said It out Loud. Though, at This Point, GPT-4 Is…." Tweet. Twitter, June 20, 2023. https://twitter.com/soumithchintala/status/1671267150101721090.

SpiritualCopy4288. "I Got Them by Using …." Reddit Comment. R/ChatGPT, April 5, 2023. www.reddit.com/r/ChatGPT/comments/11twe7z/prompt_to_summarize/jf3qdny/.

t3n Magazin. "Datenleck bei Samsung: Ingenieure schicken vertrauliche Daten an ChatGPT," April 8, 2023. https://t3n.de/news/samsung-semiconductor-daten-chatgpt-datenleck-1545913/.

"The Race to Buy the Human Brains Behind Deep Learning Machines - Bloomberg." Accessed January 11, 2024. https://www.bloomberg.com/news/articles/2014-01-27/the-race-to-buy-the-human-brains-behind-deep-learning-machines.

thebes [@voooooogel]. "So a Couple Days Ago i Made a Shitpost about Tipping Chatgpt, and Someone Replied 'Huh Would This Actually Help Performance' so i Decided to Test It and IT ACTUALLY WORKS WTF Https://T.Co/kqQUOn7wcS." Tweet. Twitter, December 1, 2023. https://twitter.com/voooooogel/status/1730726744314069190.

Töpper, Verena. "(S+) Geld verdienen mit ChatGPT: Prompt Writer verdienen bis zu 335.000 Dollar im Jahr." Der Spiegel, December 6, 2023, sec. Job & Karriere. https://www.spiegel.de/karriere/chatgpt-prompt-writer-und-prompt-engineers-verdienen-bis-zu-335-000-dollar-im-jahr-a-a54a93a5-e20d-40e6-b235-28aecobddaaa.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." In Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In Advances in Neural Information Processing Systems, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:24824–37. Curran Associates, Inc., 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R. Narasimhan. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." In Advances in Neural Information Processing Systems, 2023. https://openreview.net/forum?id=5Xc1ecxO1h.

Zhang, Hanlin, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. "Watermarks in the Sand: Impossibility of Strong Watermarking for Generative Models," 2023. https://doi.org/10.48550/ARXIV.2311.04378.

Zhang, Peiyuan, Guangtao Zeng, Tianduo Wang, and Wei Lu. "TinyLlama: An Open-Source Small Language Model," 2024. https://doi.org/10.48550/ARXIV.2401.02385.

Ziegler, Daniel M., Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. "Fine-Tuning Language Models from Human Preferences," 2019. https://doi.org/10.48550/ARXIV.1909.08593.