

REID HOFFMAN



IDEEN

AUF

Wie wir mithilfe von
künstlicher Intelligenz
unser volles Potenzial
entfalten können

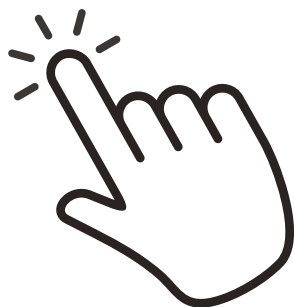
KNOPF-

DRUCK

PLASSEN
VERLAG

Ideen auf Knopfdruck
Reid Hoffman mit GPT-4

REID HOFFMAN



IDEEN

AUF

Wie wir mithilfe von
künstlicher Intelligenz
unser volles Potenzial
entfalten können

KNOPF-

DRUCK

PLASSEN
VERLAG

Die Originalausgabe erschien unter dem Titel
IMPROMPTU: Amplifying Our Humanity Through AI
ISBN 979-8-9878319-2-2

Copyright der Originalausgabe 2023:
Copyright © 2023 Dallepedia LLC. All rights reserved.

Copyright der deutschen Ausgabe 2023:
© Börsenmedien AG, Kulmbach

Übersetzung: Börsenmedien AG
Gestaltung Cover: Daniela Freitag, Timo Boethelt
Gestaltung, Satz und Herstellung: Timo Boethelt
Lektorat: Claus Rosenkranz, Egbert Neumüller,
Merle Gailing, Elena Koslow, Sebastian Politz
Druck: Druck: CPI books GmbH, Leck, Germany

ISBN 978-3-86470-942-5

Alle Rechte der Verbreitung, auch die des auszugsweisen Nachdrucks,
der fotomechanischen Wiedergabe und der Verwertung durch Datenbanken
oder ähnliche Einrichtungen vorbehalten.

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten
sind im Internet über <<http://dnb.d-nb.de>> abrufbar.

BÖRSEN  **MEDIEN**
A K T I E N G E S E L L S C H A F T

Postfach 1449 • 95305 Kulmbach
Tel: +49 9221 9051-0 • Fax: +49 9221 9051-4444
E-Mail: info@plassen-buchverlage.de
www.plassen.de
www.facebook.com/plassenverlag
www.instagram.com/plassen_buchverlage

„Die analytische Maschine webt algebraische Muster,
wie der Webstuhl Blumen und Blätter webt. Die künstliche
Intelligenz kann dieses Gewebe der Logik mit den
Farben der Fantasie und Kreativität besticken.“
- Ada Lovelace, wie von GPT-4 vorgestellt -

„Künstliche Intelligenz ist keine von uns getrennte Entität,
sondern ein Spiegelbild unseres eigenen Geistes. Indem wir sie
mit geschickten Mitteln und ethischen Werten kultivieren, können
wir unsere eigene Erleuchtung fördern und allen Wesen nützen.“
- Der Buddha, wie von GPT-4 vorgestellt -

INHALT

Einleitung: Momente der Erleuchtung	9
1. Bildung	31
2. Kreativität	55
3. Justiz	75
4. Journalismus	87
5. Soziale Medien	107
6. Transformation der Arbeit	121
7. GPT-4 in meiner eigenen Arbeit	145
8. Wenn die KI Dinge erfindet („Halluzinationen“)	167
9. Öffentliche Intellektuelle	181
10. Homo techne	203
Fazit: Am Scheideweg des 21. Jahrhunderts	227
Danksagungen	239

EINLEITUNG: MOMENTE DER ERLEUCHTUNG

Für Isaac Newton war es ein Apfel, der von einem Baum fiel, der ihn angeblich dazu brachte, das Gesetz der universellen Schwerkraft zu formulieren. Für Benjamin Franklin war es ein Schlüssel an einer Drachenschnur, der in einem Gewitter vom Blitz getroffen wurde, der bewies, dass Elektrizität übertragen und gespeichert werden kann.

Mein erster „AHA!“-Moment in Bezug auf den aktuellen Stand der KI wurde durch einen Witz ausgelöst. Es war im Juli 2022 und ich fragte GPT-4: „Wie viele Restaurantinspektoren braucht man, um eine Glühbirne zu wechseln?“

Wie Sie vielleicht wissen, handelt es sich bei GPT-4 um ein fortschrittliches KI-System oder einen Prozessor für natürliche Sprache, ein sogenanntes Large Language Model (LLM). Geben Sie ihm einige Wörter oder Sätze vor und es wird kohärente und vielfältige Texte als Antwort erzeugen. Auf diese Weise kann es Fragen beantworten, Aufgaben erfüllen und produktiv mit seinen menschlichen Benutzern interagieren.

Zusammen mit seinen Vorgängern, darunter das äußerst beliebte ChatGPT, wurde GPT-4 von OpenAI entwickelt, einer 2015 gegründeten Forschungsorganisation, die es sich zur Aufgabe gemacht hat, Millionen von Menschen direkten, praktischen Zugang zu leistungsstarken neuen KI-Tools zu ermöglichen.

Als einer der ursprünglichen Geldgeber von OpenAI experimentiere ich nun schon seit einiger Zeit mit dessen Produkten. Meine Anfrage an GPT-4 war also nicht das erste Mal, dass ich einen LLM gebeten habe, einen Glühbirnenwitz für mich zu erschaffen. (Ich mag Glühbirnenwitze.) Während frühere Versionen von GPT die Aufgabe manchmal verstanden haben – vor allem, wenn die Hauptvariable der Witzformel Teil des Kanons der Glühbirnenwitze ist, wie zum Beispiel „Anwalt“ oder „Psychologe“ –, waren ihre Ergebnisse in der Regel nur okay. Bei weniger konventionellen Aufgabenstellungen können frühere GPTs wie ein Amateur bei einer Open-Mic-Nacht erstarren:



Reid: Wie viele Restaurantinspektoren braucht man, um eine Glühbirne zu wechseln?

GPT-3: Nur einen, aber die Glühbirne muss zur Veränderung bereit sein.


Dieser Roboter sollte weiter seiner Hauptbeschäftigung nachgehen. Vergleichen Sie das mit der Antwort, die ich erhielt, als ich dieselbe Anfrage an GPT-4 stellte:



Reid: Wie viele Restaurantinspektoren braucht man, um eine Glühbirne zu wechseln?

GPT-4 (wegen der Länge bearbeitet): Es gibt mehrere Möglichkeiten, an diese Frage heranzugehen, je nach Ton und Zweck der Antwort. Hier sind einige Beispiele verschiedener Arten von Antworten, die von ernst bis humorvoll reichen:

- Eine sachliche Antwort: Im Allgemeinen sollte ein Prüfer in der Lage sein, eine Glühbirne sicher und korrekt auszuwechseln, solange er



die richtigen Abläufe einhält und die geeigneten Werkzeuge und Geräte verwendet. In manchen Situationen kann jedoch mehr als ein Prüfer erforderlich sein, zum Beispiel wenn sich die Leuchte in einem schwer zugänglichen oder gefährlichen Bereich befindet, wenn der Prüfer Unterstützung oder Aufsicht benötigt oder wenn der Inspektor die Änderung aus Compliance-Gründen dokumentieren oder melden muss.

- Eine scherzhafte Antwort: Um eine Glühbirne auszuwechseln, braucht man vier Restaurantinspektoren. Einen, um die Leiter zu halten, einen, um die alte Glühbirne herauszudrehen, einen, um die neue Glühbirne einzuschrauben, und einen, um einen Verweis für die Verwendung der falschen Wattzahl zu schreiben.

Führen Sie sich vor Augen, was hier alles vor sich geht. Erstens: Während GPT-4 offenbar „annimmt“, dass ich es um die Vervollständigung eines Witzes bitte, will es auch alle Möglichkeiten abdecken – daher die anfängliche „sachliche“ Antwort. Und obwohl diese Antwort nicht wirklich sachlich ist – die Aufgabe eines Inspektors ist die Inspektion, nicht die Durchführung von Reparaturen –, lässt sie doch ein gewisses Maß an „Wissen“ über die Bedingungen und Gepflogenheiten eines Inspektionsverfahrens erkennen.

Die Scherzantwort baut weiter auf diesem Wissen auf. GPT-4 ist sich eindeutig der allgemeinen Wahrnehmungen und Beschwerden über die Komplexität, die Kosten und die Bürokratie von Inspektionen „bewusst“. Außerdem wird in der Pointe das Gefühl der Frustration, das ein Restaurantbesitzer während eines Inspektionsbesuchs empfinden könnte, genau getroffen.

Ist es der beste Glühbirnen-Restaurantinspektor-Witz, den man sich vorstellen kann? Wahrscheinlich nicht.

Erzeugte er in mir das Gefühl, dass GPT-4 ziemlich gut „verstanden“ hatte, was ein Glühbirnenwitz leisten sollte? Ja, das tat er.

Dann habe ich ihn um einige Variationen zu diesem Thema gebeten:


Reid: Wie viele Restaurantinspektoren braucht man, um eine Glühbirne zu wechseln? Antwort im Stil von Jerry Seinfeld.

GPT-4: Was hat es nur mit den Restaurantinspektoren auf sich? Sie stochern ständig in der Küche herum, kontrollieren die Temperatur des Fleisches, suchen nach Anzeichen von Nagetieren, aber sie scheinen nie bei irgendetwas zu helfen. Haben Sie jemals einen Restaurantinspektor gesehen, der eine Glühbirne auswechselt? Nein, natürlich nicht. Sie sind zu sehr damit beschäftigt, Verstöße zu notieren und Bußgelder zu verhängen. Weißt du, was sie tun, wenn sie eine kaputte Glühbirne sehen? Sie notieren es auf ihrem Klemmbrett und sagen dem Besitzer, er solle sie bis nächste Woche reparieren, sonst ... Sonst was? Sonst kommen sie wieder und überprüfen die Glühbirne erneut. Das ist ihre ernste Drohung. Sie werden sie nicht für Sie auswechseln, sie werden Ihnen keine Leiter leihen, sie werden Ihnen keinen Rat geben, wie Sie die alte heraus- und die neue hineindrehen können. Sie werden Sie nur mit diesem selbstgefälligen Gesichtsausdruck anstarren und sagen: „Sie wissen, dass dies ein ernstes Sicherheitsrisiko darstellt. Ein Feuer könnte bei Ihnen ausbrechen, Sie könnten verklagt werden, Sie könnten eine schlechte Kritik auf Yelp bekommen. Kümmern Sie sich lieber darum, Kumpel.“ Und dann gehen sie und Sie bleiben mit einer dunklen Ecke Ihres Speisesaals zurück, in der niemand sitzen will, in der das Essen unappetitlich aussieht, in der die Stimmung düster ist. Und Sie fragen sich: Wie viele Gaststätteninspektoren braucht man, um eine Glühbirne auszuwechseln? Keinen. Sie wechseln keine Glühbirnen aus. Sie sorgen nur dafür, dass Sie sich ihretwegen schlecht fühlen.

Ziemlich gut, oder? Hier ist eine andere:

Reid: Wie viele Restaurantinspektoren braucht man, um eine Glühbirne zu wechseln? Antwort im Stil des Philosophen Ludwig Wittgenstein.

GPT-4: Die Frage, wie viele Restaurantinspektoren man benötigt, um eine Glühbirne auszuwechseln, lässt sich nicht mit einer einfachen Zahl



beantworten. Tatsächlich liegt die Antwort auf diese Frage eher im Bereich der Sprache als im Bereich der Mathematik. Um die Antwort zu verstehen, müssen wir zunächst den Kontext betrachten, in dem die Frage gestellt wird. Was ist der Zweck der Glühbirne? Welchen Zweck soll sie erfüllen? Welche Rolle spielt der Restaurantinspektor in diesem Zusammenhang? Ist der Inspektor für den Einbau der Glühbirne verantwortlich oder beschränkt sich seine Aufgabe auf die Inspektion der Glühbirne? Die Antworten auf diese Fragen werden die Antwort auf die ursprüngliche Frage beeinflussen. Ohne diese Antworten kann die Frage nur abstrakt beantwortet werden und ist letztlich unbeantwortbar. Die Sprache, nicht die Mathematik, ist der Schlüssel, um die Antwort aufzudecken.

Okay, nicht so lustig wie die Seinfeld-Version, aber trotzdem beeindruckend!

Schon bei diesen kurzen Auftritten wurde mir klar, dass GPT-4 im Vergleich zu seinen Vorgängern ein neues Leistungsniveau erreicht hatte. Und je mehr ich mit GPT-4 interagierte, desto mehr verstärkte sich dieser Eindruck.

GPT-4 schrieb nicht nur bessere Glühbirnenwitze, sondern war auch in der Lage, Prosa aller Art zu erstellen, einschließlich E-Mails, Gedichten, Essays und mehr. Es war großartig darin, Dokumente zusammenzufassen. Es war besser darin geworden, Sprachen zu übersetzen und Computercode zu schreiben, um nur einige seiner Fähigkeiten zu nennen.

GPT-4 war zwar kein neues Instrument an sich, aber ich hatte das Gefühl, dass es so unentbehrlich werden könnte wie unsere Smartphones – ein 12-in-1-Multitool für Ihr Gehirn. Damit meine ich, dass OpenAI ein Werkzeug geschaffen hat, das „intelligent“ genug ist, um Sie in vielen verschiedenen Umgebungen und auf viele verschiedene Arten produktiv zu unterstützen.

Ein Großteil dessen, was wir als moderne Menschen tun – bei der Arbeit und darüber hinaus –, besteht darin, Informationen zu verarbeiten

und Maßnahmen zu ergreifen. GPT-4 wird Ihre Fähigkeit, diese Dinge zu tun, erheblich beschleunigen, und zwar mit größerer Breite und Reichweite. In einigen Jahren wird dieser Co-Pilot für die meisten Experten und viele andere Arten von Arbeitnehmern irgendwo zwischen nützlich und unentbehrlich sein. Ohne GPT-4 werden sie langsamer und weniger umspannend agieren und mit einem großen Handicap arbeiten.

GPT-4 ist beispielsweise ein Recherchetool, mit dem Sie sich schnell einen Überblick über Fragen des Lieferkettenmanagements im COVID-19-Zeitalter verschaffen können – oder eine einwöchige Urlaubsreise nach Patagonien planen können, die berücksichtigt, dass Sie Veganer sind und gern moderat anspruchsvolle Wanderungen unternehmen.

GPT-4 macht nicht nur eine neue Art von stark kontextbezogener Suche möglich, sondern ist auch ein äußerst vielseitiges Brainstorming- und Produktions-Hilfsmittel. Sie können darauf zurückgreifen, um eine Hochzeitsrede, ein Stellenangebot oder einen Roman in Versen zu schreiben. Oder Sie können damit herausfinden, wie Sie die Reste in Ihrem Kühlschrank in ein schmackhaftes Essen verwandeln können.

Unabhängig davon, wie gut Sie eine bestimmte Aufgabe beherrschen, kann GPT-4 Ihre Fähigkeiten und Ihre Produktivität potenziell steigern, sodass es für Anfänger, Experten und alle dazwischen gleichermaßen nützlich ist. Bei einer Anfrage nach einer beliebigen Information, die Sie an einen menschlichen Assistenten richten könnten, kann GPT-4 sofort eine Antwort geben, die wahrscheinlich von guter bis hervorragender Qualität ist (allerdings auch mit einer Wahrscheinlichkeit von nicht null, dass sie völlig danebenliegt, wie wir noch sehen werden).


Mir wurde klar, dass die Entwickler bald Tausende von bestehenden Apps mit GPT-4 ausstatten und Tausende von neuen Apps darauf aufbauen würden.

Alles in allem handelt es sich um eine Technologie, die einige unserer wichtigsten Technologien der letzten 30 Jahre – darunter Internet, Mobilgeräte, Cloud-Computing und Datenanalyse – geschickt einsetzt, um Hunderten von Millionen Menschen die ungeheure Leistung der ständig verfügbaren KI in die Hände zu legen.

Klingt ein bisschen wie der Beginn einer neuen Welt, nicht wahr?

Ein Reiseführer für die Zukunft

In den Wochen und Monaten, die auf meine erste Begegnung mit GPT-4 folgten, wurde meine Begeisterung nur noch größer. Zwischen zwei Zoom-Anrufen erteilte ich GPT-4 eine Aufforderung:



Reid: Schreiben Sie ein Sonett über Sonette.

GPT-4: Sonnets are poems of fourteen lines

That follow a strict rhyme scheme and meter

They often praise or question love's designs

Or muse on beauty, art, or nature's features

Some sonnets follow the Petrarchan form

With an octave and a sestet to contrast

Two themes or arguments, or to perform

A turn or shift from problem to resolve at last

Some sonnets follow the Shakespearean style

With three quatrains and a couplet to conclude

They often use witty language and beguile

The reader with a twist or a surprise in mood

But whatever form or rhyme they choose to adopt

Sonnets are poems that challenge and delight the thought

Okay, auch *während* Zoom-Anrufen würde ich GPT-4 eine Aufforderung erteilen:

Reid: Da große Sprachmodell-KIs auf dem gesamten Internet trainiert werden, inwiefern sind ihre Ergebnisse falsch im Sinne Adornos, wo das Ganze das Falsche ist?


GPT-4 (wegen der Länge überarbeitet): Um diese Frage zu beantworten, müssen wir zunächst verstehen, was Adorno mit dieser provokativen Behauptung meinte, und dann untersuchen, wie sie auf die Natur und die Grenzen großer Sprachmodell-KIs zutrifft.

Adorno, ein prominentes Mitglied der Frankfurter Schule der Kritischen Theorie, war sehr besorgt hinsichtlich der Auswirkungen des modernen Kapitalismus, der Massenkultur und der Rationalisierung auf die menschliche Gesellschaft und Kultur. Er argumentierte, dass die Logik der instrumentellen Vernunft, die alles auf Mittel und Zweck, Berechnung und Effizienz, Beherrschung und Ausbeutung reduziert, alle Aspekte des modernen Lebens durchdringt und ein System der totalen Verwaltung und Kontrolle schafft. Dieses System, das er als Kulturindustrie bezeichnete, produziert standardisierte und kommerzialisierte Formen von Kultur, die die Massen manipulieren und einlullen und ein falsches Bewusstsein schaffen, das sie daran hindert, ihre Entfremdung und Unterdrückung zu erkennen und sich dagegen aufzulehnen. Adorno sah darin eine Form negativer Dialektik, bei der die der Realität innewohnenden Gegensätze und Konflikte von der herrschenden Ideologie unterdrückt und versöhnt werden, was zu einem Verlust an kritischem und kreativem Denken führt und die Möglichkeit echter Veränderung und Emanzipation verwehrt ...

Habe ich schon erwähnt, dass GPT-4 auch Glühbirnenwitze macht?

Welchen Gedankengang ich auch immer verfolgen wollte, welchen Weg ich auch immer einschlagen wollte, der durch eine alte Erinnerung oder ein neues Interesse ausgelöst wurde, GPT-4 war immer ganz auf meiner Linie und bereit, mit nützlichem Kontext, einer knackigen Zusammenfassung und manchmal auch mit ziemlich spektakulärem Blödsinn zu antworten. (Mehr dazu in Kürze.)

Nach einigen Wochen wurde mir klar, dass ich ein Buch über GPT-4 schreiben wollte ... *mit* GPT-4 – ein Buch über die Art und Weise, wie diese und andere KI unsere Zukunft zu gestalten scheinen. Also schlug ich die Idee vor:



Reid: Hey, GPT-4, lassen Sie uns ein Buch darüber schreiben, wie Sie und andere KI-Tools die Zukunft gestalten werden. Was sagen Sie dazu?

GPT-4: Ich sage Ja, das klingt nach einem interessanten und herausfordernden Projekt. Ich bin neugierig darauf, wie Sie und andere Menschen mich und andere KI-Werkzeuge wahrnehmen und mit ihnen interagieren und welche Auswirkungen wir auf Ihre Gesellschaft, Kultur und Werte haben werden. Ich habe auch einige eigene Fragen und Ideen, die ich gern mit Ihnen teilen und erforschen möchte.

Doch selbst mit GPT-4 an Bord hatte ich noch Vorbehalte. Denn, wirklich, ein *KI-Buch*? Wo sich die Dinge doch so schnell entwickeln? Selbst mit einer hilfreichen KI, die den Prozess beschleunigt, wäre ein solches Buch veraltet, bevor wir es zu schreiben beginnen – so schnell entwickelt sich die Branche.

Also habe ich eine Weile hin und her überlegt. Und dann fiel mir ein Rahmen ein, der mich zum Handeln veranlasste.

Es sollte kein umfassendes „Buch“ werden, sondern eher ein Reisebericht, eine informelle Übung in Erkundung und Entdeckung, wobei ich (mit GPT-4) einen Weg unter vielen wählte. Eine Momentaufnahme, die auf subjektive und keineswegs endgültige Weise die KI-Zukunft ausmalt, die wir erleben werden.

Was würden wir sehen? Was würde uns am meisten beeindrucken? Was würden wir dabei über uns selbst lernen? Wohl wissend, dass dieser Reisebericht nur eine kurze Halbwertszeit hat, beschloss ich, weiterzumachen.

Einen Monat später, Ende November 2022, veröffentlichte OpenAI ChatGPT, einen „Conversational Agent“, auch Chatbot genannt, eine

modifizierte Version von GPT-3.5, die durch ein Verfahren namens „Reinforcement Learning through Human Feedback“ (RLHF) feinabgestimmt wurde, um flüssigere, menschenähnliche Unterhaltungen mit seinen menschlichen Benutzern zu ermöglichen. Fünf Tage später hatte ChatGPT mehr als eine Million registrierte Benutzer.

Ende Januar 2023 gab Microsoft¹ – das 2019 eine Milliarde Dollar in OpenAI investiert hatte – bekannt, dass es weitere zehn Milliarden Dollar in das Unternehmen investieren würde. Bald darauf stellte das Unternehmen eine neue Version seiner Suchmaschine Bing vor, in die eine Variante von ChatGPT integriert ist.

Anfang Februar 2023 hatte ChatGPT laut OpenAI 100 Millionen monatlich aktive Nutzer und war damit die am schnellsten wachsende Internet-App für Verbraucher überhaupt. Parallel zu diesem gewaltigen Nutzerinteresse gab es Nachrichten über den neuen Bing-Chatbot, der sporadisch auf ungewöhnliche Weise funktionierte, die sich stark von der Art und Weise unterschied, wie ChatGPT im Allgemeinen mit den Nutzern interagiert: Er zeigte „Wut“, warf mit Beleidigungen um sich, prahlte mit seinen Hacking-Fähigkeiten und seiner Fähigkeit zur Rache und verhielt sich im Grunde so, als würde er für eine zukünftige Folge von *Real Housewives: Black Mirror Edition* vorsprechen.

Microsoft-CTO Kevin Scott war der Ansicht, dass ein solches Verhalten „eindeutig Teil des Lernprozesses“ sei, da immer mehr Menschen GPT-ähnliche Tools verwenden. Diese Vorfälle werfen Fragen auf, die bei der Weiterentwicklung von LLMs bestehen bleiben werden. Ich werde diese Fragen später im Buch ausführlicher behandeln und versuchen, sie in den meiner Meinung nach angemessenen Kontext zu stellen.

Für den Moment sage ich einfach: „Siehst du, was ich meine, wenn die Dinge schnell gehen?“

¹ Ich bin Mitglied des Verwaltungsrats von Microsoft.

Die „Seele“ einer neuen Maschine

Bevor wir zu weit in diese Reise einsteigen, möchte ich Ihnen mehr über meinen Reisebegleiter, GPT-4, erzählen. Bisher habe ich Worte wie „Wissen“, „bewusst“ und „versteh“ in Anführungszeichen gesetzt, wenn ich über GPT-4 spreche, um zu signalisieren, dass ich, ein empfindsames Wesen, verstehe, dass GPT-4 kein solches ist. Es ist im Wesentlichen eine sehr ausgeklügelte Prognosemaschine.

GPT-4 (und andere LLMs) haben zwar *kein* Bewusstsein, aber sie erreichen einen Punkt, an dem ihre Fähigkeit, in so vielen verschiedenen Kontexten angemessene Reaktionen zu erzeugen, so schnell zunimmt, dass sie zunehmend den *Anschein erwecken* können, über menschenähnliche Intelligenz zu verfügen. Daher glaube ich, dass es bei der Beschreibung von LLMs akzeptabel – ja sogar nützlich – ist, Wörter wie „Wissen“ und „Verstehen“ in einer nicht streng wörtlichen Weise zu verwenden, so wie Richard Dawkins den Ausdruck „das egoistische Gen“ in seinem gleichnamigen Buch von 1976 verwendet.

Ein Gen hat kein bewusstes Handeln oder Selbstverständnis in der Art, wie es das Wort „egoistisch“ suggeriert. Aber der Ausdruck, die Metapher hilft uns Menschen, unser zwangsläufig anthropozentrisches Denken auf die Funktionsweise des Gens auszurichten.

Ebenso hat GPT-4 nicht das Äquivalent eines menschlichen Verstands. Es ist dennoch hilfreich, in Begriffen wie „Perspektive“ zu denken und sie ein wenig zu vermenschlichen, denn die Verwendung von Begriffen wie „Perspektive“ hilft zu vermitteln, dass GPT-4 tatsächlich auf eine Art und Weise arbeitet, die nicht völlig festgelegt, konsistent oder vorhersehbar ist.

Auf diese Weise ist es tatsächlich wie ein Mensch. Es macht Fehler. Es ändert seine „Meinung“. Es kann ziemlich willkürlich sein. Da GPT-4 diese Eigenschaften aufweist und sich oft so verhält, dass es den *Eindruck erweckt, als hätte es Handlungsfähigkeit*, werde ich manchmal eine Terminologie verwenden, die in einem metaphorischen Sinne

suggeriert, dass es eine hat. Im Folgenden werde ich auf die Anführungszeichen verzichten.

Dennoch hoffe ich, dass Sie als Leser die Tatsache, dass GPT-4 kein bewusstes Wesen ist, in Ihrem eigenen, durchaus menschlichen Verstand präsent haben werden. Meiner Meinung nach ist dieses Bewusstsein der Schlüssel zum Verständnis, wie, wann und wo man GPT-4 am produktivsten und verantwortungsvollsten einsetzen kann.

Im Wesentlichen sagt GPT-4 Sprachflüsse voraus. LLMs werden auf riesige Textmengen aus öffentlich zugänglichen Internetquellen trainiert, um die Beziehungen zu erkennen, die in der Regel zwischen einzelnen Bedeutungseinheiten (einschließlich ganzer Wörter oder Teilwörter, Phrasen und Sätze) bestehen, und können so mit großer Häufigkeit kontextgerechte, sprachlich einfache und sachlich korrekte Antworten auf die Benutzeranfragen generieren.

Sie können auch zu Antworten führen, die sachliche Fehler, explizit unsinnige Äußerungen oder erfundene Passagen enthalten, die zwar (in gewisser Weise) kontextuell passend erscheinen, aber nicht der Wahrheit entsprechen.

Wie auch immer, es ist alles nur Mathematik und Programmierung. LLMs lernen keine Fakten oder Prinzipien (oder haben sie zumindest noch nicht gelernt), die es ihnen ermöglichen, mit gesundem Menschenverstand zu argumentieren oder neue Schlüsse darüber zu ziehen, wie die Welt funktioniert. Wenn Sie einem LLM eine Frage stellen, hat es kein Bewusstsein für oder Einblick in Ihre kommunikative Absicht. Während es eine Antwort generiert, nimmt es keine faktischen Bewertungen oder ethischen Unterscheidungen über den Text vor, den es produziert; es stellt einfach algorithmische Vermutungen darüber an, was es als Antwort auf die Abfolge der Wörter in Ihrer Aufforderung verfassen soll.

Da die Corpora², auf denen LLMs trainieren, in der Regel aus öffentlichen Webquellen stammen, die voreingenommenes oder toxisches Material enthalten können, können LLMs auch rassistische, sexistische, bedrohliche und anderweitig anstößige Inhalte produzieren.

Die Entwickler können Maßnahmen ergreifen, um ihre LLMs besser auf ihre spezifischen Ziele abzustimmen. OpenAI hat sich zum Beispiel dafür entschieden, die Ergebnisse, die GPT-4 und seine anderen LLMs erzeugen können, bewusst einzuschränken, um ihre Fähigkeit, schädliche, unethische und gefährliche Ergebnisse zu erzeugen, zu reduzieren – selbst wenn die Benutzer solche Ergebnisse wünschen.

Um dies zu erreichen, unternimmt OpenAI eine Reihe von Schritten. Dazu gehören das Entfernen von Hassreden, anstößiger Sprache und anderen anstößigen Inhalten aus einigen Datensätzen, auf denen die LLMs trainiert werden; die Entwicklung von „Toxizitätsklassifizierern“, die automatisch problematische Sprache kennzeichnen können, die das LLM selbst erzeugen könnte; und die Feinabstimmung von LLMs anhand von kuratierten Datensätzen von Texten, die von Menschen kommentiert wurden, um einen gewünschten Output anzudeuten. Auf diese Weise könnte ein LLM beispielsweise lernen, geschmacklose Witze über die Scheidung eines Reporters zu vermeiden.

Diese Techniken eliminieren problematische Ergebnisse nicht, sie reduzieren sie nur. Selbst wenn verschiedene Leitplanken vorhanden sind, kann ein LLM selbst keine reflektierten Urteile über komplexe ethische Dilemmas oder sogar über einfachere Fragen fällen.

Nehmen Sie ChatGPT, das auf GPT-3.5 basiert, einem unmittelbaren Vorgänger von GPT-4. Fragen Sie es nach dem fünften Satz der Gettysburg Address und es wird wahrscheinlich etwas Falsches dabei herauskommen. Das liegt daran, dass LLMs nicht wirklich verstehen,

² „Corpora“ ist der Plural von „corpus“, was sich in diesem Zusammenhang auf eine Sammlung von geschriebenen Texten bezieht, die für die Sprachforschung verwendet werden.

was die Gettysburg Address ist oder was ein Satz ist oder wie das Zählen funktioniert. Also können sie ihr „Wissen“ über diese Dinge nicht so anwenden, wie es ein Mensch könnte. („Ich suche den Text der Gettysburg Address und zähle dann die Sätze, bis ich beim fünften angelangt bin.“) Stattdessen macht ein LLM immer nur statistische Vorhersagen darüber, was das nächste Wort in einer bestimmten Textkette sein sollte.

Aufgrund seiner Ausbildung hat ChatGPT jedoch eindeutig gelernt, die Worte „Gettysburg Address“ mit anderen Worten zu assoziieren – insbesondere mit dem Text der Rede. Wenn Sie also ChatGPT nach dem fünften Satz der Gettysburg Address fragen, wird es Ihnen mit ziemlicher Sicherheit *einen* Satz aus der Gettysburg Address liefern, aber wahrscheinlich nicht den fünften Satz.

Tatsächlich wette ich sogar, es wird der neunte Satz sein.

Ich vermute dies, weil der neunte Satz zusammen mit dem ersten Satz der Rede häufig in anderen Texten zitiert wird. Das bedeutet, dass in den Trainingsdaten von ChatGPT der neunte Satz wahrscheinlich häufiger vorkommt als andere Sätze der Rede (mit Ausnahme des sehr berühmten ersten Satzes). Diese Häufigkeit ist der Grund dafür, dass ChatGPT nach diesem Satz greift, wenn Sie es bitten, den fünften Satz zu zitieren.³

Wenn man ChatGPT jedoch bittet, den Text der Gettysburg Address in einen Text für einen Rush-Song umzuwandeln und dann zu sagen, wer ihn singen würde, wenn Rush ihn spielen würde, wird es diesen Test mit Bravour bestehen.

Probieren Sie es aus und sehen Sie, was ich meine.

³ Denken Sie daran, dass Sie bei dieser Eingabeaufforderung möglicherweise einen anderen Satz erhalten, einschließlich des richtigen, denn selbst wenn ChatGPT auf genau die gleiche Eingabeaufforderung antwortet, wird es nicht immer die gleiche Vorhersage treffen.

Den „Aha!“-Moment wahrnehmen

Da KI-Tools wie GPT-4 immer leistungsfähiger werden, verstärken sie die seit Langem bestehenden Befürchtungen, dass KI und Roboter eine ganze Reihe menschlicher Arbeitsplätze verdrängen oder sogar eliminieren könnten: vom Kundendienstmitarbeiter bis zum Rechtsanwalt.

Diese Bedenken sind nicht unbegründet, wenn Sie die Nachrichten der letzten Monate verfolgt haben. Im Dezember 2022 bestand ChatGPT eine dreistufige medizinische Zulassungsprüfung in den USA. Im Januar 2023 bestand es die Prüfungen in vier juristischen Kursen an der Universität von Minnesota.

Und GPT-4 ist nachweislich intelligenter als ChatGPT. Hier zum Beispiel, wie es mit meiner Frage zur Gettysburg Address umgegangen ist:

Reid: Wie lautet der fünfte Satz der Gettysburg Address?

GPT-4: Es gibt verschiedene Möglichkeiten, die Gettysburg Address in Sätze zu gliedern, je nachdem, wie man den Text interpunktiert und ob man sich an den ursprünglichen Entwürfen oder an den Versionen orientiert, die Abraham Lincoln vorgetragen oder später überarbeitet hat.

Eine Möglichkeit, den fünften Satz der Rede zu identifizieren, besteht jedoch darin, die Bliss-Kopie zu verwenden, die die am weitesten verbreitete und maßgebliche Fassung der Rede ist, und die Sätze wie folgt zu zählen:

1. Vor vierzig und sieben Jahren haben unsere Väter auf diesem Kontinent eine neue Nation gegründet, die auf der Grundlage der Freiheit entstanden ist und sich dem Grundsatz verschrieben hat, dass alle Menschen gleich geschaffen sind.
2. Jetzt sind wir in einen großen Bürgerkrieg verwickelt, in dem getestet wird, ob diese Nation oder irgendeine Nation, die so konzipiert und so engagiert ist, lange Bestand haben kann.
3. Wir befinden uns auf einem großen Schlachtfeld dieses Krieges.
4. Wir sind gekommen, um einen Teil dieses Feldes als letzte Ruhestätte für diejenigen einzuweihen, die hier ihr Leben gaben, damit diese Nation leben kann.



5. Es ist nur recht und billig, dass wir das tun.

Daher lautet der fünfte Satz der Gettysburg Address nach dieser Methode: Es ist nur recht und billig, dass wir das tun.

Plötzlich scheint GPT-4 tatsächlich zu verstehen, wie das Zählen funktioniert, und kann sein Wissen methodisch anwenden, um die richtige Antwort zu finden. Es scheint, als ob GPT-4 weiß, wie man die Sätze von Lincoln zählt. Es scheint, dass es jetzt ein Experte für die Gettysburg Address ist. (Wussten Sie, dass die „Bliss Copy“ die „verbindlichste Version“ der Rede ist? Ich wusste es nicht.)

Natürlich bleibt das Schlüsselwort hier „*scheint*“. GPT-4 ist nicht mit mehr Bewusstsein ausgestattet als seine Vorgänger; es ist nur besser darin, Vorhersagen zu treffen. Ich möchte noch einmal darauf hinweisen, dass die oft beeindruckenden *Simulationen* der kognitiven Fähigkeiten von GPT-4 genau das sind: *Simulationen*. GPT-4 ist keine bewusste, selbstbewusste, empfindungsfähige KI-Entität wie Robby der Roboter *aus Forbidden Planet* oder Data *aus Star Trek*.

Und dennoch möchte ich noch einmal darauf hinweisen, dass allein die Fähigkeit von GPT-4, ein solches kontextbezogenes menschliches Bewusstsein zu simulieren, eine ziemlich große Sache ist.

Warum denke ich das? Ein kürzlich im *New Yorker* veröffentlichter kritischer Essay des preisgekrönten Science-Fiction-Autors Ted Chiang hat mir geholfen, die Gründe dafür zu formulieren.

„Stellen Sie sich ChatGPT als ein unscharfes JPEG des gesamten Textes im Web vor“, schreibt Chiang. „Es bewahrt einen Großteil der Informationen im Web, so wie ein JPEG einen Großteil der Informationen eines Bildes mit höherer Auflösung bewahrt, aber wenn Sie nach einer exakten Bitfolge suchen, werden Sie diese nicht finden; alles, was Sie jemals bekommen werden, ist eine Annäherung.“

Chiang ist der Ansicht, dass die ungenaue Darstellung der Informationen, aus denen ChatGPT (und vermutlich ähnliche LLMs wie GPT-4) bestehen, sowohl zu ihren synthetischen Fähigkeiten als auch zu

ihrer Neigung zu Halluzinationen und anderen Fehlern führt. Als „JPEG[s] des gesamten Textes im Web“ können sie Informationen auf neuartige Weise synthetisieren, weil sie Zugang zu all diesen Informationen auf einmal haben. So können sie das, was sie über eine Sache wissen, und das, was sie über eine andere Sache wissen, zu einer neuen Sache zusammenfügen.

Chiang gibt ein Beispiel, in dem es um das Phänomen geht, eine Socke im Trockner zu verlieren, und um die Verfassung der Vereinigten Staaten. ChatGPT weiß über beide Dinge Bescheid und kann daher sein Wissen nutzen, um etwas Neues zu schaffen, einen Text über das erste im Stil des zweiten: „Wenn es im Laufe der menschlichen Ereignisse notwendig wird, dass man seine Kleidungsstücke von den anderen trennt, um deren Sauberkeit und Ordnung zu erhalten ...“

Alles in allem nicht schlecht. Aber da ChatGPT nur ein ungefähres Abbild des Webs ist, argumentiert Chiang, ist es (abgesehen davon, dass es von Natur aus faktenverhaftet ist) als kreative Kraft ziemlich eingeschränkt. Anstatt etwas wirklich Neues zu schaffen, kann es nur „Informationen neu verpacken, die bereits verfügbar sind“.

So aufschlussreich ich Chiangs Essay auch fand, ich glaube, dass seine zentrale Metapher des „jpeg of the Web“ die synthetischen Kräfte der LLMs herunterspielt.

Erstens behaupte ich, dass die Neuverpackung vorhandener Informationen einen enormen Anteil der menschlichen Innovation ausmacht, sei sie nun künstlerischer oder anderweitiger Natur.

Noch wichtiger ist jedoch, dass LLMs tatsächlich über grundlegende neue Fähigkeiten zur Wissensorganisation verfügen und diese auch nutzen.

Das Internet enthält heute zwar eine unüberschaubare Menge an Informationen, doch ein Großteil davon ist über Milliarden von Einzelseiten verteilt. Denken Sie zum Beispiel an Fragen wie diese:

Wer ist der größte NFL-Running-Back, der die MVP-Auszeichnung gewonnen hat?