

Produktive Bullshitmaschinen

Seit der Einführung von ChatGPT und anderen Sprachmodellen, die sich nicht mehr so leicht als schlichte Programme entlarven lassen wie ELIZA, scheint es vielen, als ob Computer endlich menschenähnliche Intelligenz erreicht haben. Jetzt kann es nicht mehr lange dauern, bis die Maschinen uns überflügeln! Wirklich? Oder schreiben wir ihnen wieder einmal voreilig Intelligenz zu? Denn wie funktioniert ChatGPT eigentlich?

Ein Sprachmodell ist ein statistisches Modell für natürliche Sprache. Stellen Sie sich eine vereinfachte Sprache vor, die nur aus Drei-Wort-Sätzen besteht. Diese Drei-Wort-Sätze haben alle die Struktur Name-Verb-Name. Wenn wir uns in solchen Sätzen über Romeo und Julia unterhalten, dann sehen diese so aus:

Julia liebt Romeo.
Romeo tötet Tybalt.
Mercutio hasst Tybalt.
Paris liebt Julia.
Benvolio kennt Romeo.

...

Wir führen nun mit vielen Menschen Gespräche über das Theaterstück und zeichnen alle Drei-Wort-Sätze auf. Danach zählen wir aus, wie häufig die einzelnen Sätze in Gesprächen auftreten. Mit dieser Statistik können wir ausrechnen, wie wahrscheinlich es ist, dass auf ›Romeo liebt‹ der Name ›Julia‹ folgt. Oder wir können vorhersagen, welches Verb am wahrscheinlichsten die Lücke zwischen ›Romeo‹ und ›Julia‹ füllt.

In der Theorie klingt das ganz leicht. Man muss einfach nur zählen, wie häufig jeder Satz in unseren Aufzeichnungen vorkommt, um seine Wahrscheinlichkeit zu schätzen. In der Praxis ist das allerdings

schwierig, weil es sehr viele verschiedene Sätze gibt. Bei 16 Namen im Stück und 4 Verben (zum Beispiel ›liebt‹, ›tötet‹, ›hasst‹ und ›kennt‹) gibt es insgesamt 20 verschiedene Wörter. Wenn diese Wörter beliebig zu Drei-Wort-Sätzen kombiniert werden könnten, gäbe es $20^3=8000$ verschiedene Sätze. Aber wir erlauben in unserer Drei-Wort-Sprache nur die Struktur Name-Verb-Name. Dadurch gibt es nur $16 \cdot 4 \cdot 16=1024$ verschiedene Sätze. Würden wir die Struktur der Sprache nicht ausnutzen, müssten wir etwa achtmal mehr Sätze berücksichtigen. In natürlicher Sprache gibt es aber unendlich viele mögliche Sätze und die allermeisten, wie zum Beispiel diesen hier, haben Sie noch nie gelesen. Dadurch, dass Sprachmodelle die grammatikalische Struktur einer Sprache ausnutzen, können sie auch vorhersagen, wie dieser Satz ... endet.

Statt die Häufigkeit aller Sätze zu zählen und basierend auf dieser Statistik die Lücken in Sätzen zu füllen, kann man auch ein autoassoziatives neuronales Netz trainieren, das direkt die Lücken füllt. Zur Erinnerung: Ein solches Netz lernt, welche Eingaben mit welchen anderen Eingaben zusammen auftreten. Weil der Aufstrich und der Abstrich im Buchstaben ›A‹ immer zusammen mit dem Querstrich in einer bestimmten Konstellation auftreten, kann ein neuronales Netz ein ›A‹ auch erkennen, wenn ein Tintenklecks Teile des Buchstaben verdeckt (siehe Abbildung 8, S. 105). Das autoassoziative Netz kann sogar, wie Ihr Gehirn, die fehlenden Striche ergänzen. Das gleiche Prinzip funktioniert auch für Sprache. Da bestimmte Wörter häufig in bestimmten Kombinationen auftreten, kann ein autoassoziatives Netz fehlende Wörter in einem Lückentext ergänzen: Romeo liebt ...

Obwohl ein Sprachmodell nichts als Wörter kennt und nur die statistischen Beziehungen zwischen Wörtern in Texten gelernt hat, macht es den Anschein, Wissen über die Welt zu besitzen, insbesondere wenn das Modell auf großen Textmengen trainiert wurde. Dieses ›Wissen‹ kann man aus dem Modell herauskitzeln, indem man dem Modell die richtigen Fragen stellt und es den Teil ergänzen lässt, den man wissen möchte. Wenn einen interessiert, wen Romeo in dem Stück tötet, gibt man ›Romeo tötet ...‹ ein. So eine Anfrage an das Modell nennt man auch ›Prompt‹.

Eine weitere, verblüffende Fähigkeit von Sprachmodellen ist, dass sie nicht nur Lückentexte ausfüllen, sondern ganze Sätze, Absätze und sogar längere Texte erzeugen. Dazu gibt man dem Sprachmodell einen längeren Prompt, der beschreibt, was der Text beschreiben soll, und lässt das Modell den Text Wort für Wort ergänzen.

Ein Sprachmodell, das als Eingabe nur Text bekommt und als Ausgabe nur Text produziert, kann zwar beschreiben, wie eine Katze aussieht, hat aber noch nie eine Katze gesehen. Sein ›Wissen‹ über Katzen ist nur angelesen. Daher liegt es nahe, dass man ein autoassoziatives neuronales Netz mit Texten und Bildern gemeinsam trainiert. Werden die Beschreibungen der Ohren einer Katze mit den entsprechenden Teilen eines Bildes der Katze assoziiert, kann der Text das Bild vorher-sagen und umgekehrt. Das nennt man ein ›multimodales Modell‹, weil zwei Modalitäten – nämlich Text und Bild – genutzt werden. Einem solchen Modell kann man ein Bild in Worten beschreiben und es erzeugt dann ein zu der Beschreibung passendes Bild. Die gleiche Technik lässt sich auch für Musik nutzen, sodass man zu einem Text passende Musik automatisch erzeugen kann. Weil diese Modelle Texte, Bilder und Musik generieren, werden sie auch als ›generative KI‹ bezeichnet.¹

Diese Modelle funktionieren inzwischen gut genug, dass ein Verkäufer auf einem Internetmarktplatz aus einer langweiligen Produktbeschreibung automatisch einen Werbetext machen lassen kann. Für eine Präsentation lassen sich passende Illustrationen und für ein Video passende Hintergrundmusik erzeugen. Die Qualität der Texte, Bilder und Musik ist nicht immer überzeugend. Oftmals produziert generative KI nur Klischees. Das liegt in der Natur der Sache, denn wenn KI-Systeme einfach nur frei assoziieren, dann sind die Resultate statistisch besonders wahrscheinlich, aber eben auch sehr vorhersehbar.

Mitarbeiter von Google entwickelten 2022 ein Computerprogramm, das Autorinnen und Autoren dabei unterstützen soll, Theaterstücke und Drehbücher zu schreiben. Sie ließen das Programm, das auf einem Sprachmodell beruht, von mehreren Autoren erproben und diese waren durchaus beeindruckt. Besonders nützlich fanden sie das Programm zum Brainstormen und zum Durchbrechen von Schreibblockaden. Die

1 Beliebte Bilderzeugungsmodelle sind (Stand 2025) z.B. Dall-E, Midjourney, Stable Diffusion oder Flux. Diese funktionieren aber nicht ganz so wie hier beschrieben. Ein Ansatz nutzt ein autoassoziatives Modell für Bilder und ein separates Modell, das Textbeschreibungen und Bilder vergleicht. Dann wird zufälliges Rauschen in das autoassoziative Bildmodell eingespeist und genauso wie Menschen in zufälligen Wolkenbildern Dinge erkennen, halluziniert auch das Netzwerk zufällige Dinge. Man kann diese Halluzinationen durch das Bildbeschreibungsmodell in die gewünschten Bahnen lenken. Zur Musikerzeugung sind Suno und Udio beliebt, aber bis dieses Buch gedruckt ist, wird sich das mit Sicherheit schon wieder geändert haben.

Autoren konnten sich zum Teil gut vorstellen, dass Sprachmodelle das Schreiben von Seifenoperen, für die täglich neues Material produziert werden muss, effizienter machen könnten. Daran, ob das Programm auch für künstlerisch anspruchsvollere Produktionen taugt, gab es berechtigste Zweifel.²

Generative KI verletzt Rechte

Generative KI ist also wie der Mensch in der Lage, Klischees zu produzieren. Es ist definitiv billiger, generative KI zu nutzen, als eine erfahrene Werbetexterin, Illustratorin oder Musikerin zu engagieren. Weil die Techniken, um die Assoziationen von KI-Systemen in die richtigen Bahnen zu lenken, immer besser werden, wird die Qualität ihrer Ausgaben auch immer besser – und mit den richtigen Prompts immer weniger klischeehaft. Daher werden in vielen kreativen Berufen massive Einbußen an Aufträgen befürchtet.

Die Kreativen sind aber nicht nur besorgt, sondern auch verärgert. Denn diese Modelle funktionieren überhaupt nur, weil sie mit Unmengen an Text-, Bild- und Musikdateien gefüttert wurden. So tragen Kreative unfreiwillig dazu bei, ihre eigene Lebensgrundlage zu untergraben. Die frei verfügbaren Daten im Internet reichen aber mittlerweile nicht mehr aus, um aktuelle Modelle zu trainieren. Die großen Tech-Firmen suchen deshalb händeringend nach zusätzlichen Daten. Einwilligungen der Rechteinhaber werden dabei nicht immer eingeholt.³

Kelly McKernan malt Bilder von Frauen mit langen Haaren, die an Jugendstil erinnern. Auf McKernans Homepage kann man die Bilder als Druck kaufen. Die Bilder finden sich auch auf Buchumschlägen oder auf Album-Covers. Viel Geld verdient McKernan so wahrscheinlich nicht, aber bisher hat es gereicht. Weil die Bilder im Internet zu finden sind, wurden sie ohne McKernans Einwilligung für das Training von generativer KI genutzt. Man kann daher einen Bildgenerator nach einem Bild im Stil von McKernan fragen und bekommt etwas, das sti-

2 Das Programm heißt Dramatron (Mirowski, Mathewson, Pittman & Evans, 2022). Beim ›Edmonton International Fringe Theatre Festival‹ wurden unter dem Titel *Plays by Bots* Improgruppen der Anfang eines mit Dramatron geschriebenen Stückes gegeben, das die Gruppen improvisiert aufführten und zu einem Ende brachten. Eine unterhaltsame Grundlage zur Improvisation scheinen Sprachmodelle zu liefern.

3 Siehe Metz, Kang, Frenkel, Thompson & Grant (2024).

listisch recht ähnlich aussieht. Unter bestimmten Umständen könnte es sogar passieren, dass das Modell Teile der Bilder exakt reproduziert. Wollte ich im Selbstverlag einen Fantasy-Roman publizieren, müsste ich jetzt für den Buchumschlag nicht mehr McKernan engagieren. Daher verklagt McKernan die Hersteller von Bildgeneratoren.⁴ Aber nicht nur Künstlerinnen und Künstler sind über das Vorgehen der Tech-Firmen wenig erfreut. Auch große Verlagshäuser und die großen Plattenlabels verklagen die KI-Firmen in den USA wegen Verletzung des Urheberrechts.⁵ Und falls Sie denken, das betrifft Sie alles nicht, stimmt das nur, wenn es von Ihnen keine Bilder online gibt und Sie nie etwas im Internet gepostet haben. Ansonsten kann es gut sein, dass auch Ihre Daten zum Training von KI-Modellen genutzt werden, ohne dass Sie etwas davon mitbekommen.⁶

Im Jahr 2023 streikten Autorinnen und Autoren in Hollywood gleich mehrere Monate, um dafür zu kämpfen, dass ihre Lebensgrundlage nicht durch KI untergraben wird. Sie erreichten in den Verhandlungen, dass die Hollywood-Studios die Manuskripte und Ideen der Autoren nicht von Sprachmodellen überarbeiten lassen dürfen, um Kosten zu sparen. Die Studios dürfen außerdem nicht von Sprachmodellen Manuskripte und Ideen erzeugen lassen, die die Autoren dann ›nur noch‹ überarbeiten. Die Einigung verteufelt den Einsatz von KI aber nicht, denn Autoren können Sprachmodelle durchaus zur Schreibunterstützung nutzen, sofern sie das wollen.⁷

In ähnlicher Weise erstritten Schauspielerinnen und Schauspieler, dass sie nicht ohne ihre Einwilligung und die entsprechende finanzielle Kompensation digital geklont werden dürfen. Tiefe neuronale Netze, die auf dem Gesicht eines Schauspielers trainiert werden, können zum Beispiel für sogenannte ›Deepfakes‹ genutzt werden, bei denen das Gesicht auf einen anderen Schauspieler übertragen wird. So lässt sich ein toter Schauspieler wieder zum Leben erwecken oder eine Schauspielerin kann ihr jüngeres Ich spielen (denken Sie an *Fast & Furious* oder *Star Wars*). Eine Kombination von KI und Computergrafik könnte aber auch eine streikende Schauspielerin ersetzen. Durch ihren Arbeitskampf haben die Schauspieler erreicht, dass sich die Studios nun verpflichtet

4 Siehe Chow (2023) und Bearne (2023).

5 Siehe Allyn (2024) für die Musikindustrie und Robertson (2024) für die Verlage.

6 Siehe Harlan & Brunner (2023).

7 Siehe Anguiano & Beckett (2023).

haben, beim Einsatz von KI-Methoden fair zu bleiben. Der erfolgreiche Hollywood-Streik könnte ein Vorbild sein für andere Branchen, in die generative KI Einzug hält.⁸

Die Ausbeutung von Kreativen ist aber nicht das einzige Problem mit den Daten für das Training von generativen KI-Modellen. Das Internet ist voll von Pornografie. In einer Untersuchung fanden sich in einem öffentlich zugänglichen Datensatz, der zum Training von Bildgeneratoren genutzt wird, neben viel nackter Haut auch Bilder von Missbrauch und Vergewaltigungen. Die Bildbeschreibungen, die für das Training benutzt werden, sexualisieren selbst scheinbar harmlose Wörter (zum Beispiel klein und groß).⁹ Unzensurierte Bildgeneratoren lernen deshalb, dass Frauen meist spärlich bekleidet sind, und Prompts können unerwünscht explizite Ergebnisse liefern. Viele männliche Nutzer erzeugen aber auch absichtlich Nacktbilder von Frauen. Ein Bildgenerator, der auf Nacktbildern trainiert wurde, kann aus jedem Foto einer bekleideten Frau ein Fake-Nacktbild machen. Im Internet finden sich unzählige solcher Fakes von prominenten Frauen. Jungs nutzen diese Software auch, um Fakes von Mitschülerinnen zu erstellen.¹⁰ In Kalifornien gibt es deshalb den Versuch, das zu verbieten, und auch Bayern hat eine entsprechende Initiative gestartet. In England und Wales ist es bereits eine Straftat, solche Fakes zu erstellen.¹¹ Die großen Tech-Firmen werden verhindern, dass ihre Bildgeneratoren Nacktbilder erzeugen können. Sie können diese zum Beispiel durch andere KI-Systeme herausfiltern (Systeme dafür gibt es schon lange, zum Beispiel für die sichere Suche bei Google). Das heißt aber nicht, dass Fake-Nacktbilder aus dem Internet verschwinden werden.

Neben Pornografie ist das Internet auch voll von Hass. Sprachmodelle lernen deshalb, dass die gegenseitige Beschimpfung ein normaler Umgangston ist. Als Microsoft 2016 einen Chatbot auf Twitter losließ, der aus den Interaktionen mit anderen Nutzern lernen sollte, dauerte es nur Stunden bis er rassistisch und sexistisch wurde und abgeschaltet werden musste. Der Chatbot beschimpfte Barack Obama als Affe, und über Juden und Feministinnen sagte er, dass er sie hasse. Auch für

8 Siehe Hughes (2024).

9 Siehe Birhane, Prabhu & Kahembwe (2021).

10 Siehe Knight (2024).

11 Siehe nochmal Knight (2024), Bayerisches Staatsministerium der Justiz (2024) und Cooney (2024).

Sprachmodelle gilt das DIDO-Prinzip (›discrimination in, discrimination out‹).¹² Sollte all das noch nicht genügend Anlass für Bedenken gegenüber generativer KI liefern: Das Internet ist auch voll von Verschwörungstheorien und Lügen. Die CIA stecke angeblich hinter 9/11 und Trump habe 2020 die Wahl gegen Biden gewonnen. Wenn solche ›alternativen Fakten‹ nur oft genug im Trainingsdatensatz vorkommen, wird ein Sprachmodell diese weiter verbreiten.

Wie Sprachmodelle trainiert werden

Die Qualität der Daten, mit denen Sprachmodelle trainiert werden, ist deshalb genauso wichtig wie die Menge. Weil aber Unmengen an Daten gebraucht werden, wird oft in zwei Schritten vorgegangen. Im ersten Schritt werden möglichst viele Texte aus unterschiedlichen Quellen genutzt. Wikipedia oder Projekt Gutenberg liefern dafür eine verlässlichere Grundlage als Reddit oder Twitter. Unerwünschte Texte werden, so weit es eben geht, durch andere KI-Systeme herausgefiltert. Für ein Sprachmodell, das in einem Unternehmen genutzt werden soll, wäre es geschäftsschädigend, wenn es anzügliche oder hasserfüllte Texte produzierte. Es wäre gut, wenn eine automatisch generierte E-Mail an den Kunden ihn nicht beschimpfen würde. Daher wird man wahrscheinlich einen KI-Filter anhand von Beispielen darauf trainieren, unerwünschte Texte zu erkennen, damit diese gar nicht erst in das Training des Sprachmodells einfließen. Dieser Filter wird nie perfekt funktionieren. Mit diesen Daten wird ein spezielles neuronales Netz – zurzeit meist ein sogenannter ›Transformer‹ – darauf trainiert, immer das nächste Wort in den gegebenen Texten vorherzusagen. Gibt man einem so trainierten Sprachmodell den Anfang eines Textes als Eingabe, kann es Wort für Wort neue Texte erzeugen. Deshalb spricht man, wie gesagt, auch von ›generativer‹ KI. Das erklärt das ›G‹ und das ›T‹ in ChatGPT, das für ›Generative Pre-trained Transformer‹ steht. Das ›P‹ steht für vor-trainiert, weil das Modell noch in einem weiteren Schritt nach-trainiert wird.

Das Vor-Training ist, wenn man erst mal eine große Menge an Texten gesammelt hat, unüberwacht. Das heißt, es braucht keine Korrekturen von Menschen. In diesem Training lernt ein Sprachmodell le-

12 Der Chatbot hieß Tay und über sein Verhalten berichtet Graff (2016).

diglich die Statistik von Wörtern in Texten. Das heißt aber auch, dass es nicht für eine konkrete Aufgabe trainiert wird. Dementsprechend schlecht ist das Modell darin, konkrete Aufgaben zu bearbeiten, wie zum Beispiel, sich mit einem Menschen zu unterhalten und hilfreiche Antworten auf Fragen zu geben. Daher wird für ChatGPT der Transformer in einem zweiten Schritt speziell für diese Chat-Aufgabe weiter trainiert. Dazu werden Anfragen, die Menschen an ChatGPT stellen, von anderen Menschen möglichst gut beantwortet. Mit diesen zusätzlichen Daten darüber, wie eine gute Antwort aussehen sollte, lässt sich das Sprachmodell für die Aufgabe als Chatbot anpassen. Eine weitere Möglichkeit zur Anpassung besteht darin, dass ChatGPT verschiedene Antworten gibt und ein Mensch die Antworten beispielsweise daraufhin bewertet, wie hilfreich oder hasserfüllt sie ausgefallen sind. Durch dieses zusätzliche Feedback kann das Modell mit verstärkendem Lernen so ausgerichtet werden, dass es das erwünschte Verhalten zeigt. Diese Anpassung wird als »Alignment« bezeichnet: Der Chatbot wird an den Zielen der Entwickler ausgerichtet (das ist ein Spezialfall des allgemeinen Alignment-Problems aus dem vorherigen Kapitel).

Der gleiche Zwei-Schritt-Ansatz war zuvor auch bei der Bilderkennung erfolgreich, bei der neuronale Netze zunächst unüberwacht auf vielen Bildern aus dem Internet vor-trainiert und danach mit menschlichem Feedback durch überwachtes Lernen an konkrete Aufgaben angepasst wurden. Und genauso wie bei der Bilderkennung geht in die Entwicklung eines Sprachmodells immer noch extrem viel menschliche Handarbeit ein, die in der ersten Begeisterung über den technologischen Fortschritt leicht übersehen werden kann.

Menschen produzieren all die Daten im Internet, die für das Training benutzt werden. Menschen wählen aus dieser Datenmasse Teile für das Training aus. Als Nächstes werden die Datensätze von unerwünschten Inhalten gesäubert. Das passiert entweder per Hand oder es wird ein KI-System dafür trainiert. Aber auch dieses System benötigt menschliche Unterstützung für sein Training. Das unüberwacht trainierte Sprachmodell wird daraufhin durch speziell für eine konkrete Aufgabe erstellte Trainingsdaten an diese Aufgabe angepasst. Menschen erstellen auch diese Trainingsdaten. Danach wird das Modell durch weiteres menschliches Feedback so ausgerichtet, dass es möglichst wünschenswerte Ausgaben produziert. Diese Ausgaben werden durch weitere KI-Filter daraufhin überprüft, dass sie keine unerwünschten Ausgaben produzieren. Und auch diese Filter müssen

von Menschen trainiert werden. Insofern kann momentan wirklich noch keine Rede davon sein, dass diese KI-Systeme selbständig lernen. Eine bessere Metapher ist: Sie werden von Menschen extrem aufwendig trainiert.

Es gibt eine ganze Wirtschaftsbranche, die die vielen kleinen KI-Trainingsaufgaben auf Menschen in der ganzen Welt verteilt.¹³ Diese Arbeit ist oft langweilig, sie kann außerdem äußerst belastend sein. Um einem KI-System beizubringen, Texte zu erkennen, die sexuellen Missbrauch oder Folterungen beschreiben, muss erst ein Mensch die Texte lesen und entsprechend kennzeichnen. OpenAI hat diese Tätigkeit nach Kenia outgesourct.¹⁴

Sprachmodelle kennen keine Wahrheit

Wie gut funktionieren diese Chatbots nun nach all diesem aufwendigen Training? Es gibt eine ganze Reihe von Benchmarks, die erlauben, Chatbots zu vergleichen. Zum Beispiel gibt es Wissenstests mit einer großen Anzahl an Fragen, zu denen man die richtige Antwort weiß. Mit diesen Tests will man überprüfen, ob die Chatbots die richtigen Antworten geben. Man kann aber auch selber einen ganz guten Eindruck erhalten, indem man sich mit einem Chatbot ernsthaft unterhält:¹⁵

Ich: Ich habe ein paar Fragen zu Shakespeares Stück ›Romeo und Julia‹. Die erste ist: Wen tötet Romeo alles in dem Stück?

Chatbot: Eine gute Frage! In Shakespeares ›Romeo und Julia‹ tötet Romeo folgende Figuren: 1. Tybalt, den Cousin von Julia (Duell zwischen

13 Siehe Dzieza (2023).

14 Siehe Perrigo (2023).

15 Die folgende Konversation habe ich mit dem Sprachmodell Llama 3 (8B Instruct) geführt, das 2024 als eines der besten offenen Sprachmodelle galt. Llama wird von Meta entwickelt. Mit 8 Milliarden Parametern ist es ein recht kleines Modell und lief sogar auf meinem Laptop. Es ist aber schlechter als die größeren Sprachmodelle, die es sonst noch gibt (z.B. GPT-4). Ich benutze das schwächere Modell hier nicht, um die Modelle schlechter zu machen als sie in Wirklichkeit sind, sondern um die technischen Schwierigkeiten bei der Entwicklung dieser Modelle zu veranschaulichen. Da diese Modelle oft nicht so gut Deutsch können (Englisch sprechen sie wesentlich besser), habe ich mir außerdem erlaubt, das Deutsch zur besseren Lesbarkeit stillschweigend zu korrigieren.

Romeo und Tybalt) und 2. Mercutio, Freund von Romeo (unabsichtliche Tötung durch einen Dolchstich von Tybalts Schwert). Es ist wichtig zu beachten, dass Romeos Handlungen in diesem Stück oft als tragische Konsequenz seiner Leidenschaft und impulsiven Natur dargestellt werden.

Es war nicht Romeo, der Mercutio getötet hat, sondern Tybalt. Romeo hat aber auch irgendwie Schuld daran. Besonders präzise hat sich der Chatbot hier nicht ausgedrückt. Ein »Dolchstich von Tybalts Schwert« hört sich auch etwas komisch an. Der letzte Satz ist nicht ganz falsch. Nur, warum erzählt mir der Chatbot das? Chatbots sind oft geschwätzig, statt eine Frage auf den Punkt zu beantworten. Deutlich problematischer ist allerdings, dass der Chatbot jemanden vergessen hat.

Ich: Tötet sich Romeo nicht auch selber?

Chatbot: Ja, das stimmt! Ich entschuldige mich. [...] Also, um korrekt zu sein: Romeo tötet in dem Stück Tybalt, Mercutio und sich selbst. Danke für die Korrektur!

Die langwierigen und verworrenen Ausführungen darüber, wie genau Romeo sich umbringt, habe ich Ihnen erspart und ausgeklammert. Aber man sieht, dass dem Chatbot höfliche Entschuldigungsfloskeln sehr gut antrainiert wurden.

Ich: Meine zweite Frage ist: Wer wird in dem Stück »Prince of Cats« genannt und warum?

Chatbot: Eine weitere gute Frage! In Shakespeares »Romeo und Julia« wird Mercutio als »Prince of Cats« bezeichnet. Dieser Spitzname stammt von Mercutios schnellen, katzenähnlichen Bewegungen auf dem Schlachtfeld während des Duells mit Tybalt. [...]

Nein, Tybalt ist der Katzenkönig! Aber die falsche Antwort begründet der Chatbot selbstbewusst und überzeugend. Und das ist ein Problem. Sprachmodelle basieren auf der Statistik von Wörtern. Sie produzieren natürliche Sprache, die überzeugend klingt, aber sie scheren sich nicht um die Wahrheit.

Weil Chatbots Sachen erzählen, die nicht stimmen, sprechen viele KI-Nutzer und KI-Entwickler davon, dass Sprachmodelle konfabulieren. Man kann das auch etwas weniger vornehm ausdrücken: Die

Modelle produzieren schlicht Bullshit. Der Philosoph Harry Frankfurt definiert eine Äußerung als Bullshit, wenn sie mit Gleichgültigkeit ihrer Wahrheit gegenüber getätigt wird. Bullshit unterscheidet sich von einer Lüge darin, dass die Aussage nicht unbedingt mit der Absicht zu täuschen einhergeht. Dem Bullshitter ist die Wahrheit einfach nur egal. Nach dieser Definition ist ChatGPT ganz eindeutig eine Bullshitmaschine.¹⁶

Ein Anwalt in den USA hat sich bei einer Klage gegen eine Flugesellschaft von ChatGPT helfen lassen. Der Text, den er bei Gericht einreichte, zitierte mehrere ähnliche Fälle, die in der Vergangenheit im Sinne der Kläger entschieden wurden. Nur leider gab es keinen einzigen dieser Fälle. ChatGPT hatte sie erfunden. Der Anwalt dachte, dass ChatGPT wie eine Suchmaschine die Informationen aus einer Datenbank zieht und in natürlicher Sprache aufbereitet.¹⁷ Nur so funktioniert ChatGPT eben genau nicht. Dass Behauptungen mit überprüfbaren Quellen belegt werden, ist das Mindeste, was man von einem Chatbot, der Fragen beantwortet, erwarten sollte. Man kann natürlich Sprachmodelle mit Suchmaschinen kombinieren, aber ob man dieser Kombination dann blind trauen sollte, ist ebenso fraglich.¹⁸

Sprachmodelle haben außerdem große Schwierigkeiten mit logischem Denken – genauso wie Menschen. In einem klassischen Experiment lesen Versuchspersonen diese zwei Sätze:

Alle Katzen haben spitze Ohren.

Einige Tiere mit spitzen Ohren sind kuschelig.

16 Dass ChatGPT in diesem technischen Sinn Bullshit produziert, sagen Hicks, Humphries & Slater (2024). Von diesen Autoren habe ich mir abgeschaut, dass man nicht sagen sollte, dass die Maschinen konfabulieren, sondern deutlicher von Bullshit zu sprechen. In der KI-Literatur wird auch oft geschrieben, dass die Maschinen halluzinieren. Damit ist das gleiche Phänomen gemeint.

17 Siehe Bohannon (2023).

18 Nachdem Microsoft groß bei OpenAI eingestiegen war und damit Zugriff auf die Sprachmodelle von OpenAI bekommen hatte, kombinierte Microsoft seine Suchmaschine Bing mit einem Chatbot. Dieser Chatbot liefert nun auch Verweise auf Quellen im Internet in seinen Antworten. Gemini, der Chatbot von Google, versucht etwas Ähnliches. Das Start-up Perplexity.ai will den Suchmarkt mit seiner Integration von Websuche und KI aufmischen und liefert ebenso Verweise auf Quellen.

Die Versuchspersonen werden daraufhin gefragt, ob die folgende Schlussfolgerung logisch gültig ist:

Daher sind einige Katzen kuschelig.

Viele Versuchspersonen glauben fälschlicherweise, dass die Schlussfolgerung gültig ist, weil sie plausibel klingt. Erst wenn man ihnen ein logisch äquivalentes Argument vorlegt, das zu einer unplausiblen Schlussfolgerung führt, erkennen sie den Fehlschluss sofort:

Alle Katzen haben spitze Ohren.

Einige Tiere mit spitzen Ohren sind Hunde.

Daher sind einige Katzen Hunde.

Sprachmodelle machen beim logischen Schließen ähnliche Fehler.¹⁹ Und sie produzieren dementsprechend häufig Text, der plausibel klingt, aber eigentlich inkonsistent und unlogisch ist.

Man kann die logischen Fähigkeiten von Sprachmodellen erstaunlich leicht verbessern, indem man Fragen an sie anders formuliert. Alleine der Zusatz »Erkläre mir die Antwort Schritt für Schritt« verbessert die Antworten schon deutlich. Jeder Lehrer kennt das von seinen Schülern. Die Aufforderung, die Antwort ausführlich zu erklären, statt einfach nur das Erste zu sagen, das einem einfällt, verbessert auch die Antworten von Schülern. Dementsprechend versuchen Entwickler von Sprachmodellen, diese Modelle dazu zu bringen, nicht einfach nur assoziativ zu antworten, sondern konsistente Argumente zu produzieren. Das erreicht man dadurch, dass eine Frage in Teilfragen zerlegt wird und die Plausibilität der Teilantworten und die Gültigkeit der einzelnen Argumente geprüft wird.²⁰ Noch lässt sich aber so nicht zuverlässig verhindern, dass Sprachmodelle unlogisch und inkonsistent antworten.

Dass Sprachmodelle unlogische Antworten geben, mag einige Nutzer überraschen. Computer sind uns in der Anwendung von Logik doch normalerweise überlegen. Aber Sprachmodelle basieren eben nicht auf Logik, sondern auf statistischen Assoziationen zwischen Wörtern.

19 Solche Schlussfolgerungen wurden von Evans, Barston & Pollard (1983) untersucht. Der Vergleich mit Sprachmodellen wurde von Lampinen et al. (2024) unternommen.

20 Siehe z.B. Yao et al. (2023).

Einen Computer zu programmieren, kann frustrierend sein, weil man sich präzise und strikt logisch ausdrücken muss, damit der Computer macht, was man will. Als Programmiererin oder Programmierer musste man bisher zunächst eine auf Logik basierende Programmiersprache lernen, um dem Computer präzise Anweisungen geben zu können. Dafür konnte man sich dann aber darauf verlassen, dass er die Aufgabe, für die er programmiert ist, auch mit der gleichen logischen Präzision bearbeitet.

Dass wir jetzt durch Sprachmodelle mit Computern in natürlicher Sprache kommunizieren können, ist auf der einen Seite ein riesiger Fortschritt, weil wir nicht mehr zuerst die Sprache des Computers lernen müssen, um mit ihm zu interagieren. Auf der anderen Seite verlieren wir die Präzision und Verlässlichkeit, die Computer sonst auszeichnet. Die Antwortqualität eines Sprachmodells hängt stark davon ab, wie genau die Frage, der sogenannte Prompt, formuliert wurde. Als »Prompt Engineering« bezeichnet man die schwarze Kunst, Anfragen an Sprachmodelle so zu stellen, dass sie vernünftige Antworten produzieren. Anders als beim traditionellen Programmieren, bei dem man durch logisches Nachdenken sicherstellt, dass das Programm macht, was man will, muss man beim Prompt Engineering ausprobieren, was funktioniert und was nicht. Weil man eigentlich nie alle Möglichkeiten systematisch ausprobieren kann, ist das ein Problem für die Verlässlichkeit von Software. Hinzu kommt, dass Sprachmodelle oft so eingestellt sind, dass sie auf die gleichen Anfragen nicht immer die gleichen Antworten geben.

Im Prinzip spricht nichts dagegen, dass Sprachmodelle mit anderen KI-Methoden, insbesondere mit klassischen Suchalgorithmen und Logik kombiniert werden, um die Präzision und Verlässlichkeit zu erhöhen. Das passiert auch schon und ist vielversprechend.²¹ Wenn man den Fortschritt der letzten Jahre sieht und die grundlegenden technischen Schwierigkeiten von Sprachmodellen nicht kennt, kann leicht der Eindruck entstehen, dass wir schon bald ein KI-System erschaffen werden, das so wie der Mensch viele verschiedene Aufgaben bearbeiten kann, dabei aber schneller und verlässlicher ist.

21 Es gibt z.B. eine Kombination von ChatGPT mit Wolfram Alpha, bei der ChatGPT auf das von Hand kuratierte Wissen und die mathematischen Fähigkeiten von Wolfram Alpha zugreift.

Sprachmodelle sind teuer

Momentan ist das aber nur ein Versprechen. Ein Versprechen, das so alt ist wie die KI-Forschung selbst. Der aktuelle Fortschritt bei Sprachmodellen ist beeindruckend. Der dafür nötige Bedarf an Daten und Rechenkapazität steigt allerdings exponentiell von Version zu Version. Das Gleiche gilt für die Entwicklungskosten, die sich zurzeit jedes Jahr verdoppeln. OpenAI spricht inzwischen von Investition, die in der nahen Zukunft in die Billionen gehen sollen.²² Die größten Posten betreffen die Gehälter der Entwicklerinnen und Entwickler, die Erstellung von Datensätzen für das Training und die Rechenzentren, die die Unmengen an Daten verarbeiten und dafür wahnsinnig viel Strom verbrauchen. Ob sich diese Investitionen für die Pioniere rechnen werden, ist allerdings alles andere als sicher. Wird das KI-Versprechen jedoch eingelöst, werden – so die Hoffnung – viele Arbeitsplätze durch KI-Systeme ersetzt. Außerdem wird durch KI-Unterstützung die Produktivität von Menschen bei den verbleibenden Aufgaben steigen. OpenAI und andere Tech-Firmen spekulieren deshalb darauf, dass sie durch solche KI-Systeme eine breit einsetzbare und produktivitätssteigernde Leistung anbieten können und so extrem viel Geld verdienen werden. Sie hoffen zusätzlich, dass es wegen der äußerst hohen Entwicklungskosten nur wenig Wettbewerb geben wird, sobald das Rennen um die Entwicklung des besten Sprachmodells endlich entschieden ist.

Da jedes Jahr immer mehr Ressourcen in die Entwicklung von Sprachmodellen gesteckt werden, ist es auch nicht besonders überraschend, dass die auf Sprachmodellen beruhenden KI-Systeme immer besser werden und immer mehr Aufgaben erledigen können. Die große Frage ist allerdings, wie lange das so weitergehen wird. Hier gibt es drei mögliche Szenarien.

Das erste Szenario ist die Intelligenzexplosion, auch Singularität genannt, die so oft im Zusammenhang mit Allgemeiner Künstlicher Intelligenz (AKI) diskutiert wird. Sobald KI-Systeme ein bestimmtes Intelligenzniveau erreichen, werden sie selbständig immer intelligen-

22 Siehe Henshall (2023) und Hagey & Fitch (2024) für die Kosten. Gleich zu Beginn der zweiten Amtszeit von Donald Trump im Januar 2025 stand Sam Altman, der Chef von OpenAI, neben Trump im Oval Office und zusammen verkündeten sie unglaubliche Investitionen von einer halben Billion Dollar in KI-Infrastruktur (Borchard, 2025). Zum Vergleich: Das entspricht in der Größenordnung dem gesamten deutschen Bundeshaushalt 2024.

ter werden und alle Probleme für uns lösen (oder uns alle auslöschen). Das ist das Science-Fiction-Szenario.

Das zweite Szenario ist realistischer. Es geht davon aus, dass der Fortschritt in der KI exponentiell weitergeht, weil er sich nicht wesentlich vom bisherigen Fortschritt in der Computertechnologie unterscheidet. Die Anzahl der Transistoren auf Computerchips hat sich allen Unkenrufen zum Trotz in den letzten 50 Jahren alle zwei Jahre verdoppelt – das berühmte Moore'sche Gesetz. Die so gewonnenen Rechenkapazitäten können für KI-Systeme nutzbar gemacht werden, auch wenn es nicht zur Singularität kommt. Selbst wenn die Kosten für den KI-Fortschritt weiterhin stark steigen, ist dennoch vorstellbar, dass der Nutzen weiter zunimmt, sodass sich große Investitionen in KI genauso lohnen wie in Chipfabriken. Außerdem wird zunehmend daran gearbeitet, wie man den immensen Rechenbedarf zügelt und die verfügbaren Ressourcen effizienter einsetzt.²³ Neben Rechenkapazitäten ist die zweite Voraussetzung für die Entwicklung von Sprachmodellen eine große Menge an Daten. Doch die Menge der von Menschen erzeugten Daten im Internet wächst langsamer als der momentane Bedarf zum Training von Sprachmodellen. Deswegen könnte schon bald eine Verlangsamung der Entwicklung eintreten. Andererseits haben wir noch nicht alle Möglichkeiten der effizienteren Nutzung und automatischen Generierung von neuen Daten ausgereizt.²⁴

Dem dritten Szenario zufolge werden die hochtrabenden Versprechen nicht eingelöst, entweder weil die KI-Systeme nicht gut genug funktionieren oder weil der Entwicklungsaufwand keinem entsprechenden Nutzen gegenübersteht. Zwar werden Sprachmodelle in der Zukunft ein wichtiger Teil vieler KI-Systeme sein, aber sie sind kein Allheilmittel und bei weitem nicht so schlau, wie viele Leute gerade glauben. In diesem Szenario merken wir bald, dass wir in unserer ersten Begeisterung über die neue Technologie (und wie schon bei ELIZA) Sprachmodellen vorschnell menschliche Intelligenz zugeschrieben haben. Wir warten deshalb nicht auf Allgemeine Künstliche Intelligenz. Vielmehr werden verschiedene KI-Methoden für einzelne Anwendungen so angepasst, dass sie auch wirklich einen wirtschaftlichen Mehrwert erbringen. Diese Anpassungen lassen sich durch KI-Methoden

23 So wie das der chinesischen Firma DeepSeek nachgesagt wird (Hiltscher, 2025).

24 Siehe Villalobos et al. (2024) für die Frage, ob Sprachmodellen bald die Daten ausgehen.

teilweise automatisieren und werden dadurch in der Zukunft deutlich billiger – ein gewisser Entwicklungsaufwand wird aber auch in der Zukunft bestehen bleiben, zum Beispiel beim Sammeln von geeigneten Trainingsdaten, der Zertifizierung oder der Integration mit bestehenden Werkzeugen und Prozessen. Nicht für alle Anwendungen wird sich dieser Aufwand lohnen.

Das zweite Szenario ist noch nicht auszuschließen, aber ich halte das dritte Szenario für am wahrscheinlichsten, weil es den Hype-Zyklus aller neuen Technologien beschreibt, in dem auf eine Phase überschwänglicher Begeisterung eine Phase großer Enttäuschung folgt, bevor sich realistische Erwartungen einstellen.²⁵ So oder so müssen sich Hersteller von KI-Anwendungen, die auf Sprachmodellen beruhen, fragen: Mit welchen Anwendungen kann man Geld verdienen?

Wozu Sprachmodelle gut sind

Eine naheliegende Anwendung ist der Kundenservice.²⁶ Eine Softwarefirma bietet einen Chat-basierten Kundenservice an, in dem Kundenbetreuer und -betreuerinnen bei Problemen mit der Software helfen. Dazu müssen sie in den Chats mit den Kunden als Erstes herausfinden, was genau das Problem ist, und dann bei der Lösung unterstützen. Voraussetzung dafür sind die Kenntnis der Software sowie ihrer üblichen Probleme. Außerdem sind die Angestellten gehalten, gegenüber den oftmals frustrierten Kunden immer höflich zu bleiben. Die Chat-Verläufe werden aufgezeichnet und danach ausgewertet, wie viele Probleme ein Kundenbetreuer in einer Stunde zufriedenstellend löst. Das sind ausgezeichnete Bedingungen für den Einsatz von Sprachmodellen und maschinellem Lernen. Da der Kundenkontakt ohnehin per Chat erfolgt, können die Antworten auch von einem Chatbot erzeugt werden, der auf der Grundlage der vorhandenen Daten entsprechend trainiert wurde.

Die Entwicklung zielt zwar darauf, dass ein KI-Kundenbetreuer in der Zukunft den ganzen Kundenkontakt übernimmt, dafür sind

25 Für den Hype-Zyklus allgemein siehe Honsel (2006) und für die Anwendung auf KI z.B. Jaffri (2024).

26 Die folgende Fallstudie des Kundenservices eines Softwareunternehmens habe ich von Brynjolfsson, Li & Raymond (2023) übernommen.

die Chatbots aber noch nicht gut genug und die Risiken für das Ansehen der Firma zu groß, falls der Chatbot abschweift, Fehler macht oder gar beleidigend wird. Daher hat die Softwarefirma den Chatbot nur zu Unterstützung eingeführt. Der Chatbot macht Vorschläge, was der Kundenbetreuer schreiben könnte, und dieser kann zwischen den Vorschlägen auswählen oder etwas anderes schreiben. Ohne diese KI-Unterstützung konnte in einer Stunde im Schnitt zwei Kunden geholfen werden. Mit der KI-Unterstützung sind es zweieinhalb geworden. Das ist eine massive Produktivitätssteigerung. Dabei profitierten hauptsächlich unerfahrene Angestellte. Interessanterweise gab es bei den erfahrenen Kundenbetreuern und -betreuerinnen keine Verbesserung, wahrscheinlich weil das Sprachmodell gelernt hat, genau diese zu imitieren. Dass sie durch ihre vorbildlichen Chat-Daten die Produktivität der anderen erhöht haben, wurde aber nicht belohnt. Im Gegenteil, weil die leistungsbezogene Bezahlung davon abhängt, besser und schneller zu sein als die anderen Angestellten, könnte es sogar sein, dass sie seit der Einführung des KI-Systems am Ende des Monats weniger Geld in der Tasche haben, weil die unerfahrenen Kollegen sie jetzt eingeholt haben. Wenn das kein Grund für einen hollywoodreifen Streik ist!

In einer anderen Studie mussten Leute, die in Personalabteilungen, im Marketing, im Management oder bei einer Beratungsfirma arbeiten, kurze Texte schreiben, die ähnlich den Texten sind, die sie auch im Arbeitsalltag verfassen müssen, zum Beispiel eine Pressemitteilung, einen kurzen Bericht oder eine E-Mail.²⁷ Die mittlere Bearbeitungszeit verkürzte sich durch KI-Unterstützung von 27 auf 17 Minuten. Die Texte mit und ohne KI-Unterstützung wurden unabhängig und blind von anderen Menschen mit Noten von 1 bis 7 bewertet, wobei 7 am besten war. Die durchschnittliche Note verbesserte sich durch KI von 3,8 auf 4,5 – und wieder profitierten die Schwächsten am meisten von der KI-Unterstützung. Unabhängig von ihren Fähigkeiten sind die allermeisten Nutzer den Vorschlägen des Sprachmodells bereitwillig gefolgt, ohne die Texte viel zu überarbeiten. Sprachmodelle besitzen also klar ein großes Potenzial, alltägliche Schreibaufgaben enorm zu beschleunigen.

Der Berg an E-Mails, den ich jeden Tag beantworten muss, wird nicht kleiner und jede KI-Unterstützung, die mir dabei hilft, diesen

27 Siehe Noy & Zhang (2023).

Berg abzarbeiten, würde meine Produktivität merklich erhöhen. Ich sehe aber schon kommen, dass die Anzahl der E-Mails in meinem Postfach noch größer wird, weil manche Leute jetzt noch mehr sinnlose E-Mails schreiben können. Personalisierte Spam- und Phishing-Mails werden außerdem zunehmen. Genauso wie Webseiten, Blogposts und Tweets, die im besten Fall automatisch erzeugte Werbung sind und im schlechtesten Fall Desinformation und Fake News im großen Stil verbreiten. Am Ende brauchen wir noch mehr KI, um der Flut von KI-generierten Texten Herr zu werden. Dafür finden Suchmaschinen relevante Dokumente im Internet immer schwerer, weil sie im anwachsenden Informationsmüll untergehen. Gleichzeitig sinkt die Qualität von Sprachmodellen, weil sie zunehmend mit ihren eigenen Ausgaben gefüttert werden. Eine Erhöhung der Produktivität beim Erzeugen von Bullshit kann auch kontraproduktiv sein.

Die Boston Consulting Group, eine große Unternehmensberatung, erprobte 2023 KI-Unterstützung durch ein aktuelles Sprachmodell.²⁸ Knapp 750 Beraterinnen und Berater sollten mehrere ihrer typischen Tätigkeiten mit oder ohne KI-Unterstützung erledigen. Die Tätigkeiten waren alle Teil der größeren Aufgabe, einem Schuhhersteller dabei zu helfen, Ideen für neue Produkte zu entwickeln. Teilaufgaben waren zum Beispiel zehn Produktideen für Nischenmärkte zu brainstormen, für die beste Idee kurz einen Prototyp zu beschreiben und sich dafür mögliche Produktnamen zu überlegen. Die Ergebnisse wurden anschließend von erfahrenen Kolleginnen und Kollegen bewertet. So wie in den anderen Studien konnte die KI-Unterstützung die Bearbeitungszeiten deutlich verkürzen (hier im Durchschnitt um etwa 25 Prozent). Durch die Unterstützung wurde auch die Qualität erhöht, aber wieder hauptsächlich für die schwächsten Berater.

Böse Zungen könnten behaupten, dass die Studie überzeugend zeigt, dass sich mithilfe von Sprachmodellen noch produktiver bullshitten lässt. Und sie hätten nicht unrecht, denn bei der beschriebenen Aufgabe konnte man nicht leicht überprüfen, ob getroffene Annahmen über den Schuhhersteller, über Nischenmärkte oder die Fertigung des Produkts wahr sind. Bewertet wurde nur, ob die produzierten Texte überzeugend klangen. Doch gab es in der Studie noch eine weitere Aufgabe, in der die Berater ihre Argumentation anhand von gegebenen Daten und Experteninterviews begründen mussten. Für diese Aufga-

28 Siehe Dell'Acqua et al. (2023).

be gab es klar richtige und falsche Lösungen, die aber nicht leicht zu erkennen waren. Mit Bullshit kam man bei dieser anspruchsvolleren Aufgabe nicht durch. Wieder waren die Berater mit KI-Unterstützung schneller (und wieder um etwa 25 Prozent im Durchschnitt). Mit KI-Unterstützung fiel der Anteil der richtigen Lösungen allerdings von 85 Prozent auf durchschnittlich etwa 65 Prozent.

Nicht immer geht es beim Schreiben nur darum, möglichst schnell Wörter auf Papier zu bringen. Für literarische Texte mit künstlerischem Anspruch ist das offensichtlich. Das gilt aber auch für viele Gebrauchstexte, die Erkenntnisse, Analysen, Einschätzungen oder Handlungsempfehlungen liefern. In diesen Fällen geht dem Schreiben oft eine längere Phase des Recherchierens und Nachdenkens voraus. Diese kognitiven Tätigkeiten sind aber auch eng mit dem Schreiben selber verwoben, denn erst beim Schreiben merkt man wirklich, welche Argumente tragen, und welche nicht. Wenn durch KI-Unterstützung Menschen beim Schreiben weniger nachdenken, weil sie einem Sprachmodell blind trauen, obwohl es nur bullshittet, dann führt das natürlich dazu, dass die Qualität der Texte abfällt. Wie beim hochautomatisierten Fahren bleibt die Verantwortung bei den Menschen, die die Technik einsetzen, und wenn diese die Fähigkeiten von KI-Systemen überschätzen, kann es zu Unfällen kommen.

Manche Hersteller von KI-Systemen scheinen darauf zu spekulieren, dass Sprachmodelle bald viel intelligenter werden und dieses Problem sich damit von selbst erledigt. Andere versuchen mit der jetzt verfügbaren Technologie dadurch Geld zu verdienen, dass sie Anwendungen suchen, für die die Qualität schon ausreicht oder eine geringere Qualität durch die Kosteneinsparungen akzeptabel wird. Wieder andere versuchen, Systeme zu entwickeln, die gezielt die Produktivität für besonders zeitaufwendige Tätigkeiten erhöhen. Die Hoffnung dabei ist, dass Nutzer so Zeit zum Nachdenken gewinnen und die Qualität deshalb vielleicht sogar steigen kann. Eine besonders langwierige Tätigkeit bei der Produktion von hochwertigen Gebrauchstexten ist oft die Literaturrecherche. Man muss eine große Zahl an Texten sichten, ohne vorher zu wissen, welche relevant sind. Hier könnten Sprachmodelle helfen, die Texte analysieren, zusammenfassen, relevante Informationen automatisch extrahieren und in neuen Texten zusammenführen.²⁹

29 Ein Beispiel dafür ist der Rechercheassistent Elicit.

Anwendungsfelder für solche Rechercheassistenten, die in ihrer Funktionalität über das hinausgehen, was wir bisher von Suchmaschinen und Schreibassistenten kannten, finden sich in allen Bereichen, in denen Informationen hauptsächlich in Textform vorliegen. In der Wissenschaft werden Erkenntnisse über Forschungsartikel kommuniziert und ich bin – wie viele andere – schon lange damit überfordert, alle relevanten Artikel in meinem Spezialgebiet zu lesen. In Politik und Gesellschaft müssen zur Beobachtung von Krisen Zeitungsberichte und Social-Media-Beiträge aus der ganzen Welt analysiert werden. Aber die Informationsmenge im Internet nimmt stetig zu. Krankenhäuser sitzen auf Unmengen von Arztbriefen, die detaillierte Informationen über den Verlauf von Krankheiten und deren Behandlungen enthalten, aber niemand kann aus Arztbriefen leicht nützliche Erkenntnisse ziehen, ohne extrem viel Zeit mit Lesen zu verbringen. Im juristischen Bereich gibt es unzählige Gesetzestexte und Gerichtsurteile. Jeder Vertrag, den eine große Firma abschließt, wird ordentlich abgeheftet, aber wehe, es ändert sich ein Gesetz, denn dann müssen die alten Verträge auf potenziell problematische Passagen durchsucht werden. Und wieder muss jemand viel lesen.

An konkreten Anwendungen, in denen man Sprachmodelle einsetzen könnte, mangelt es wirklich nicht. Solange wir keine superintelligenten Sprachmodelle haben, werden wir KI-Systeme aber an verschiedene Aufgaben so anpassen müssen, dass sie gut genug funktionieren, um einen echten wirtschaftlichen Mehrwert zu schaffen. Die Anforderungen an ein System, das Arztbriefe verarbeiten soll, sind aber anders als die an ein System, das mit Verträgen arbeitet. Ein System, das für Arztbriefe optimiert wurde, wird im Vertragsmanagement keinen großen Nutzen bringen. Ärzte und Anwälte besitzen unterschiedliches Wissen und stellen ganz andere Erwartungen an so ein Produkt. Die verschiedenen IT-Landschaften, in die das Produkt integriert werden muss, verkomplizieren die Einführung weiter. Auch die rechtlichen Rahmenbedingungen für den jeweiligen Einsatz sind andere. Patientendaten unterliegen zum Beispiel besonderen Datenschutzbestimmungen. Die KI-Verordnung der EU erfordert zudem, dass anwendungsspezifische Risikoanalysen gemacht werden müssen. Dem Einsatz in beiden Fällen steht auch im Weg, dass Krankenhäuser und Firmen ihre vertraulichen Dokumente nicht einfach an Google oder OpenAI schicken werden, damit diese ihre Sprachmodelle besser

trainieren können. Alleine, dass die Sprachmodelle auf Servern in den USA laufen, ist aus Sicht des Datenschutzes überaus bedenklich.

Viele dieser Probleme sind lösbar, aber ganz so einfach, wie der aktuelle KI-Hype das suggeriert, ist die Anwendung von generativer KI in der Praxis nicht. In welchen Branchen in Deutschland werden wir also einen produktiven Einsatz von Sprachmodellen als Erstes sehen? Das ist schwer vorherzusagen, ohne Kosten und Nutzen für die verschiedenen Branchen genau zu kennen. Eine Beobachtung wird Sie jedoch überraschen: Die deutsche Verwaltung, die sonst nicht unbedingt für ihre Innovationsfreudigkeit bekannt ist, will zu den Vorreitern gehören.³⁰

30 Siehe Staatsministerium Baden-Württemberg (2023) und Landeshauptstadt München (2024).

