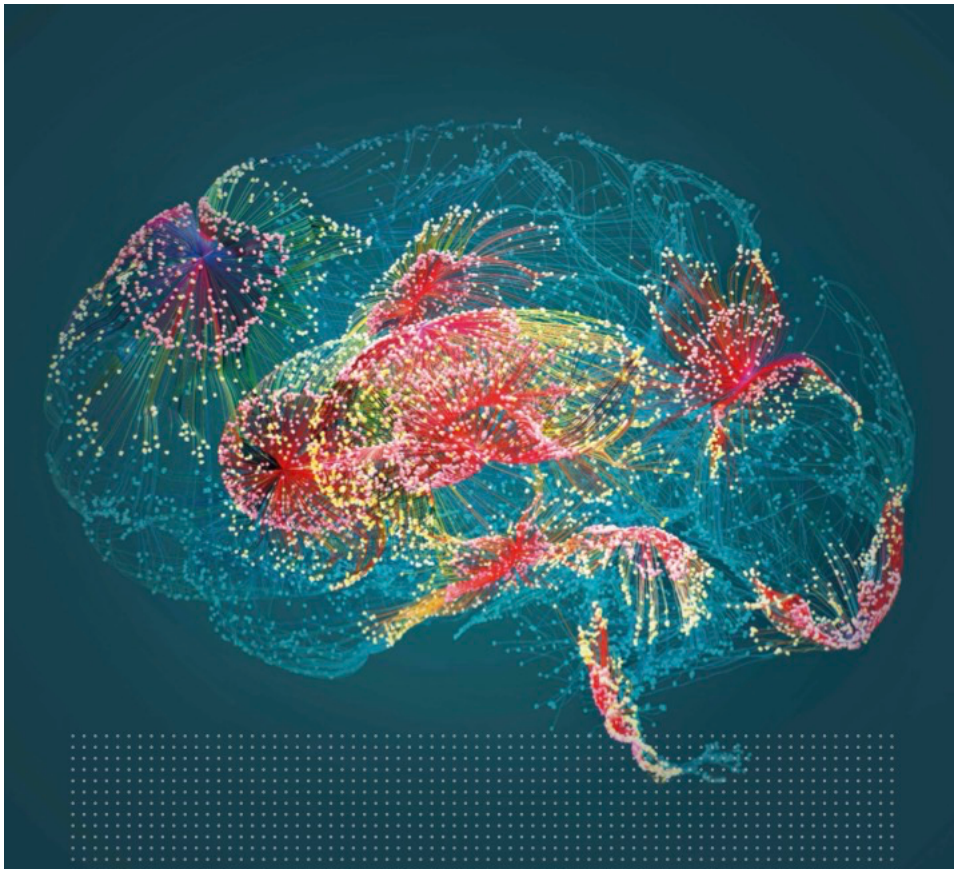
KI – genauso wie wir, nur anders

von Timm Richter

Es ist bemerkenswert, wie stark die generative KI – insbesondere die neuen Sprachassistenten wie ChatGPT, Claude oder Gemini – emotional polarisiert. Offenbar geht es dabei ans Eingemachte: unser Selbstverständnis als Menschen und Gesellschaft.

Das eine Lager appelliert – manchmal fast pathetisch – an die Einzigartigkeit des Menschen: Nur Menschen seien kreativ, besäßen Urteilsvermögen und könnten "richtig" verstehen. Wobei unklar bleibt, was mit diesem "richtig" gemeint ist. Im gleichen Atemzug werden der KI-Bias, Halluzination und Undurchschaubarkeit vorgeworfen – alles Eigenschaften, die auch Menschen nicht unbekannt sind. Wie interessant!

Das andere Lager hebt KI in den Himmel: Die allgemeine Superintelligenz – was auch immer das heißt, ein ähnlich unpräzise gebrauchter Begriff wie "Verstehen" – stehe kurz vor der Tür. Die KI sei oder werde uns in allen Belangen überlegen. Fast hat man den Eindruck, diese Protagonisten hofften auf die Überwindung jeglicher menschlichen Unzulänglichkeiten. Man kann sich fragen, ob das überhaupt geht – dass durch die KI alles klar und richtig wird. Spoiler: nein.



Wie in allen solchen Diskussionen wird sich wahrscheinlich ein mittlerer Weg zeigen – mit überraschenden Wendungen. Für eine fundierte Diskussion erscheint es mir nützlich, zwei Aspekte in den Blick zu nehmen:

Zum einen hilft es, ein - wenn auch grobes - Verständnis davon zu gewinnen, wie die neuen Sprachassistenten technisch funktionieren. Manches Argument, das Sprachassistenten mit dem mentalen Modell eines Computers oder Taschenrechners betrachtet, läuft dann ins Leere.

Zum anderen lohnt es sich, auf die Außenseite zu gehen und die großen Sprachmodelle als Kommunikationsteilnehmer zu betrachten, die ähnlich undurchschaubar sind wie menschliche Psychen. Damit kommt in den Blick, dass wir vor allem darüber verhandeln, wie Sprachassistenten unsere Kommunikation verändern. Alle Vorstellungen von Intention, Entscheidungsfähigkeit oder Verstehen sind - beim Menschen wie bei Sprachmodellen - kommunikative Zuschreibungen, die wir auf der Innenseite gar nicht überprüfen können.

Die große Veränderung durch Sprachmodelle ist: Computer werden nicht mehr nur als vermittelnde Kommunikationsmedien aufgefasst, sondern ihnen wird zunehmend Akteursstatus zugeschrieben.

Die Innenseite: Was sich über die Funktionsweise von Sprachmodellen zu wissen lohnt

Ich bin weit davon entfernt, ein Experte auf dem Gebiet der Sprachmodelle zu sein. Gleichwohl möchte ich versuchen, wesentliche Aspekte ihrer Funktionsweise mit so wenig technischem Hintergrund wie möglich zu beschreiben. Wer es lieber viel genauer und technischer mag, dem sei als Einstieg die Webseite 3Blue1Brown von Grant Sanderson empfohlen, der ein großartiges Erklärvideo erstellt hat.

An der Schnittstelle nach außen erscheinen Sprachmodelle wie übliche Computer: Man gibt etwas ein (Input) und bekommt eine Antwort (Output). Dieses Ping-Pong-Spiel zwischen Nutzer und Sprachmodell sieht im Grundsatz aus wie ein Gespräch zwischen zwei Menschen. Das ist ja auch das Unglaubliche dieser Modelle - dass man sich "mit Computermodellen" recht "normal" unterhalten kann, eben wie mit Menschen. Klar, manchmal enttäuschen uns die Antworten, aber das ist bei Menschen ja auch so. Man kann das Gedankenexperiment machen: Bei welchen Fragen würde man ChatGPT & Co einem zufällig ausgewählten Menschen der 8-Milliarden-Weltbevölkerung vorziehen?

Wie schaffen die Sprachmodelle das - etwas zu produzieren, was uns als Menschen zumindest so sinnvoll erscheint, dass Anschlusskommunikation entsteht? Das Problem, das der programmierte Algorithmus löst, ist folgendes: **Was ist das nächste Zeichen?**

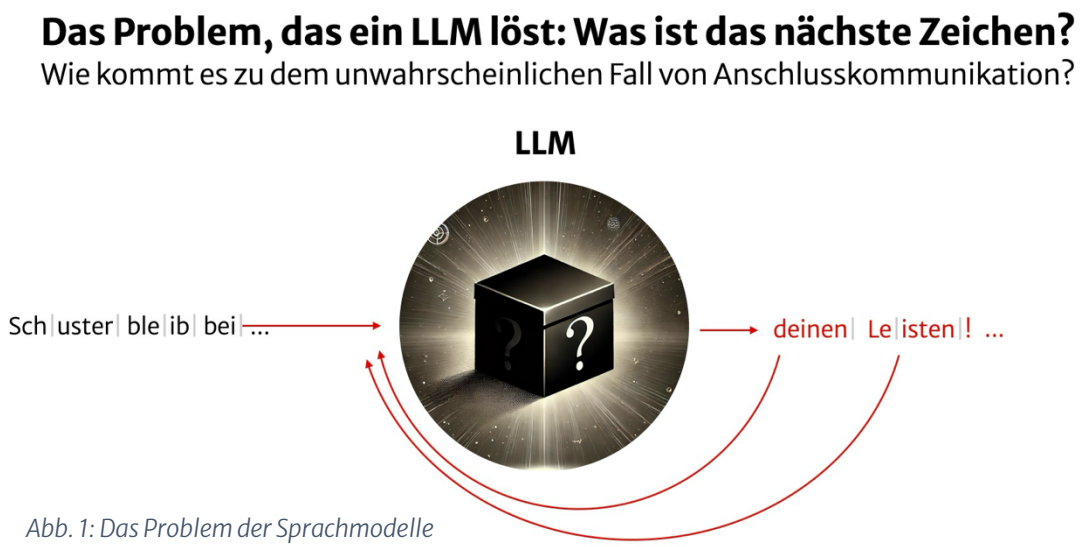


Abb. 1: Das Problem der Sprachmodelle

Ein Nutzer gibt etwas ein und das Sprachmodell berechnet ein Zeichen, das daran anschließt. Sodann wird die Nutzereingabe um dieses Zeichen ergänzt und diese erweiterte Eingabe als Input genutzt, auf die das nächste Zeichen ausgegeben wird – und so weiter, bis das Sprachmodell eine Pause macht und auf die nächste Eingabe wartet.

Es ist unglaublich, fast wie Magie, dass diese simple iterative Regel zu einer Antwort führt, die uns als sinnvolle Reaktion erscheint. Eine wesentliche Verbesserung dieser neuen Sprachmodelle: Sie berücksichtigen alle bisherigen Eingaben und Antworten im Chatverlauf als Kontext. Wenn ich beispielsweise die Vorderseite einer Speisekarte fotografiere und ChatGPT frage, wo ich bin, bekomme ich als Antwort, dass ich in einer Tapas-Bar sitze. Habe ich mich hingegen vorher nach Attraktionen in Chemnitz informiert, vermutet ChatGPT (richtig), in genau welcher Tapas-Bar ich sitze.

Umgekehrt kann es passieren, dass sich ein Chatverlauf "festfährt" – der bisherige Kontext wird zu dominant, das Sprachmodell erscheint reaktionsunwillig und dumm. Dann hilft es, einen neuen Chat ohne Gesprächshistorie zu starten. Wie beim Menschen lohnt es sich, die "Macken" des Gegenübers zu kennen.

Noch unglaublicher wird die Leistung, wenn man weiß: In den Sprachmodellen ist überhaupt keine "menschliche" Semantik abgebildet. Sie erkennen auf ganz andere Art und Weise Ordnung in sprachlichen Aussagen – die im Ergebnis zu unserer Semantik passt.

Das fängt damit an, dass Sprachmodelle gar nicht unsere Wörter "kennen", sondern als Alphabet nur Zeichenfolgen unterschiedlicher Länge verwenden – sogenannte Tokens. Bei ChatGPT-3 gibt es rund 100.000 dieser Zeichenfolgen, die uns wiederum spanisch vorkommen. Die Tokens werden dann – in Abhängigkeit des Kontextes – in einen hochdimensionalen Raum projiziert. Der von ChatGPT-3 hat rund 1.500 Dimensionen.

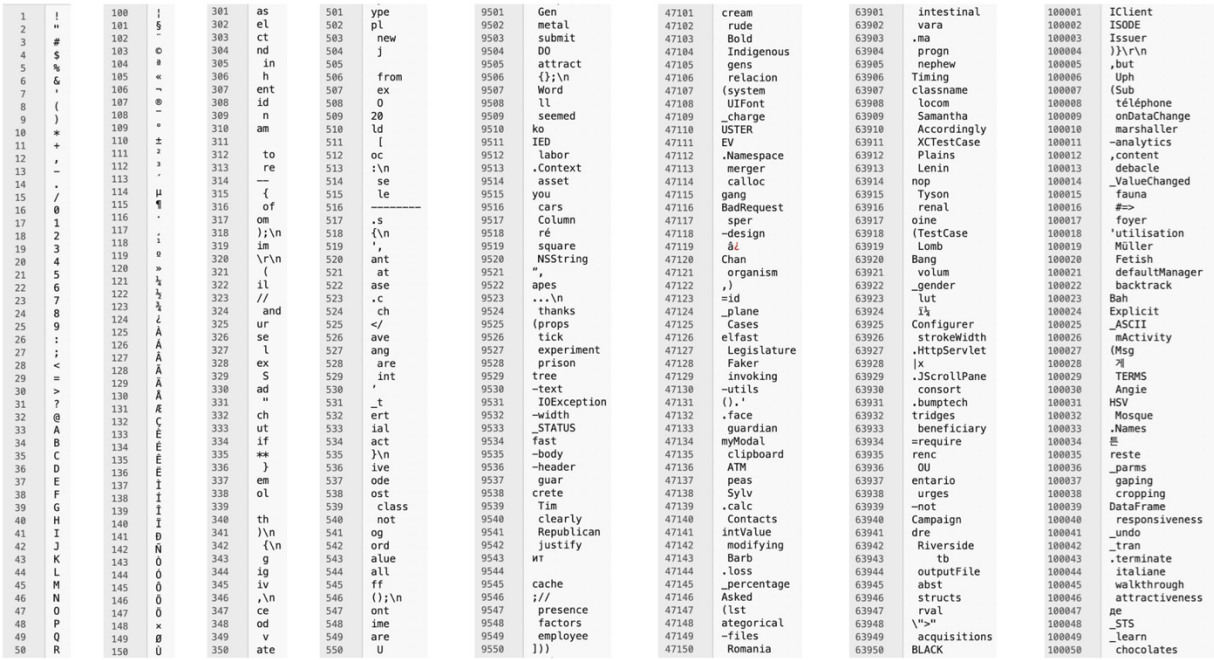
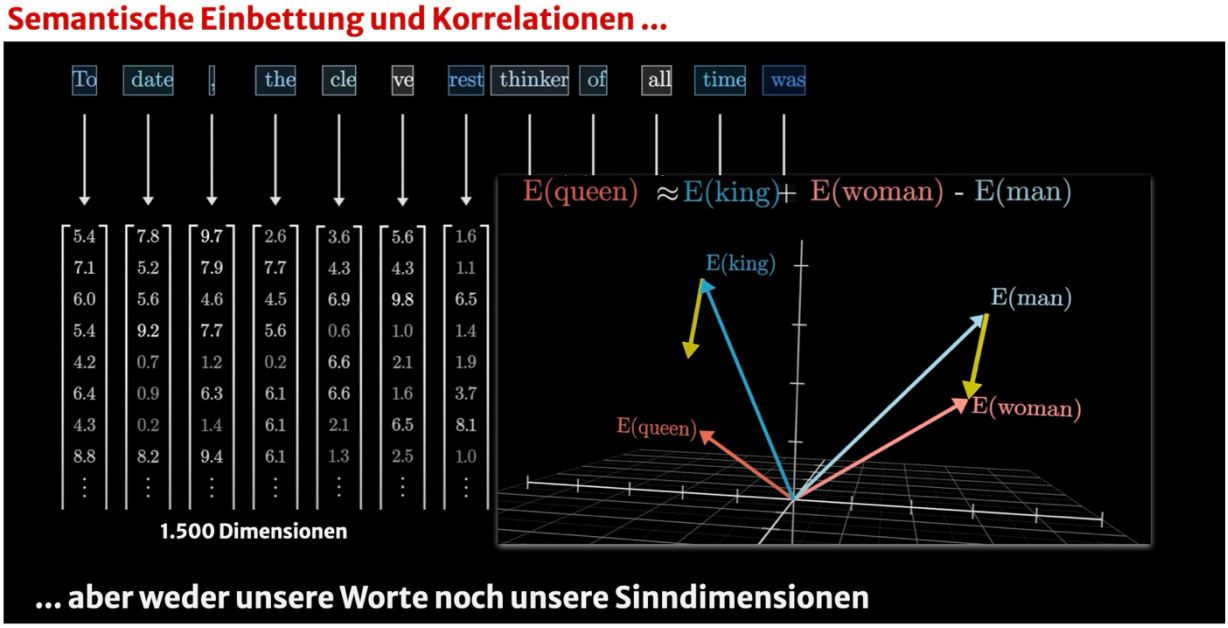


Abb. 2: Anfang und Ende der 100.000 Tokens umfassenden Tokenliste

Diese auf maschinellem Lernen basierende, rein mathematische Einbettung sorgt dafür, dass Sätze, die für uns semantische Ähnlichkeit repräsentieren, in diesem Raum auch nah beieinander liegen. Aber auch die Entwickler haben keine Vorstellung davon, was diese 1.500 Dimensionen semantisch repräsentieren. It just works.

In diesem Raum können Sprachmodelle semantisch "rechnen". Die Wörter Mann und Frau haben einen Abstand in dem Raum. Addiert man zu "König" den Weg von Mann zu Frau, landet man ungefähr bei "Königin". Das ist technisch nicht ganz korrekt, gibt aber eine Vorstellung davon, dass in diesem abstrakten Raum Korrelationen abgebildet sind, die zu unseren semantischen Vorstellungen passen – obwohl sie auf andere Weise operieren.



Quelle: https://www.3blue1brown.com/lessons/gpt

Abb. 3: Semantische Einbettung

Die Sprachmodelle sind beeindruckend in dem, was sie wissen, welche Fragen sie beantworten, wie gekonnt sie sich in Gespräche einbringen. Das liegt daran, dass sie mit unheimlich vielen Daten trainiert wurden – "so ziemlich alle Daten des Internets", auch wenn das nicht ganz korrekt ist.

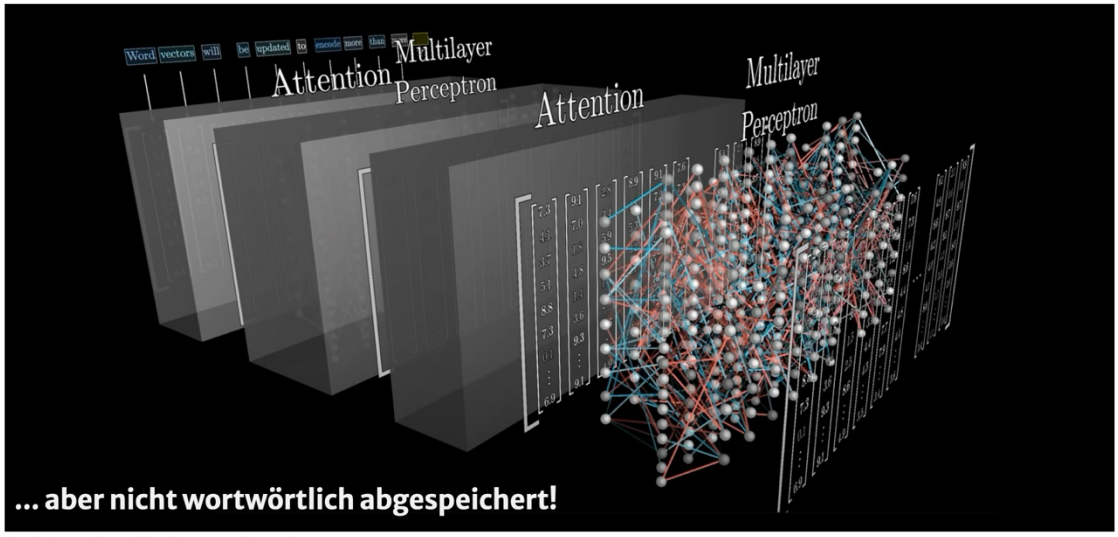
Das kann zu der Idee führen, diese Daten seien wie auf einer Computerfestplatte abgelegt. Doch diese Vorstellung entspricht nicht der Funktionsweise und führt in die Irre. Sprachmodelle sind riesige sogenannte neuronale Netze. Der Input wird an Knotenpunkten mit 0 und 1 codiert. Die Verbindungslinien zwischen den Knotenpunkten sind "Gewichte" – Zahlen zwischen 0 und 1. Der Input wird durch dieses Netz geleitet und am Ende kommt das Ausgangssignal raus.

Ein Modell mit Daten trainieren heißt: Das Ergebnis von Eingaben wird mit dem erwarteten Ergebnis verglichen und auf Basis von Differenzen werden die Gewichte angepasst. Bei ChatGPT-3 gibt es rund 180 Milliarden Gewichte, die mit "den Daten des Internets" adjustiert werden.

Wenn ein Sprachmodell nach dem Training auf die Frage "Was ist die Hauptstadt von Deutschland?" mit "Die Hauptstadt Deutschlands ist Berlin" antwortet, dann ist dieses Frage-Antwort-Paar eben nicht "direkt" wie in einer Datenbank abgespeichert. Der Input wird durch das Netzwerk geleitet und die Antwort wird Token für Token in dem Augenblick erzeugt.

So gesehen halluzinieren die Sprachmodelle ständig – sie kennen gar keinen anderen Modus. Meistens passen die Antworten zu den Trainingsdaten, aber nicht immer. Dem Netzwerk mit 180 Milliarden Parametern sieht man nicht an, wie oder wo Wissen gespeichert ist. Die Struktur ist so angelegt, dass sie performativ Wissen erzeugt, das zu den Trainingsdaten und unseren Erwartungen passt. Das ist prinzipiell sehr ähnlich zur Funktionsweise des Gehirns, in dem auch nicht Worte oder Bilder wie auf einer Festplatte abgelegt sind.

Daten (das ganze Internet) ...

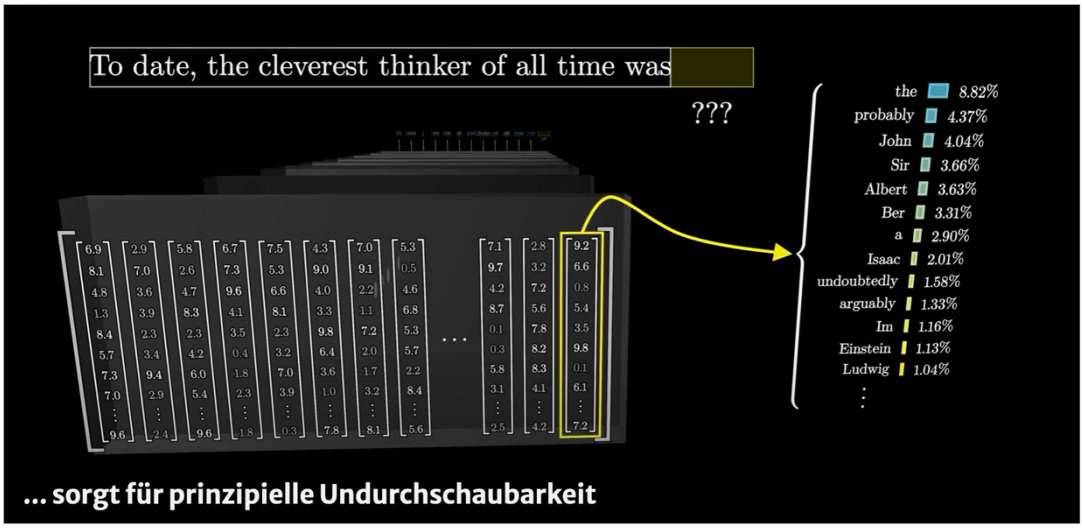


Quelle: https://www.3blue1brown.com/lessons/gpt

Abb. 4: Keine Festplatte, sondern kallibrierte Gewichtungen

Ein weiterer wesentlicher Aspekt: Die Sprachmodelle arbeiten mit einer Zufallskomponente! Genauer wird zu jedem möglichen Token eine Wahrscheinlichkeit angegeben. Als nächstes Token wird nicht immer das wahrscheinlichste ausgewählt, sondern per Zufall eines aus den wahrscheinlichsten. Wie groß die Zufallskomponente ist, hängt von einem Parameter ab, der "Temperatur" genannt wird.

Zufall



Quelle: https://www.3blue1brown.com/lessons/gpt

Abb. 5: Ausgabe ist nicht deterministisch

Ist die Temperatur auf Null gesetzt, reagiert das Sprachmodell vollständig deterministisch – auf identischen Input immer mit identischem Output. Wird die Temperatur höher gesetzt, wird per Zufall aus immer mehr wahrscheinlichen Tokens ausgewählt – das Sprachmodell wird kreativer. Deswegen sind die Antworten niemals identisch, deswegen kann es sich lohnen, komplexere Fragen mehrfach zu stellen.

Wird die Temperatur zu hoch eingestellt, verliert das Sprachmodell zunehmend Kohärenz bis zu dem Punkt, wo die Antworten uns als Wortsalat erscheinen. Ein ähnliches Phänomen kennen wir, wenn Menschen unter Drogen keinen richtigen Satz mehr sprechen können.

Zusammenfassung zur Funktionsweise

- Sprachmodelle generieren ihre Antworten, indem iterativ berechnet wird, was das nächste Token ist
- Der gesamte bisherige Chatverlauf ist stets Input und Kontext – alle Chatverläufe sind pfadabhängig
- Die Modelle sind "mit den Daten des ganzen Internets" trainiert – Training bedeutet nicht wörtliche Speicherung, sondern Parameteranpassung
- Sie sind nicht deterministisch, sondern haben eine Zufallskomponente, die für Kreativität sorgt

Die Außenseite: Die Konstruktion von Sprachmodellen als Personen in Kommunikation

Als Beobachter lässt sich ein Gespräch mit einem Sprachassistenten sehr passend als sich vollziehende Kommunikation begreifen – unabhängig von der inneren Dynamik der Psyche des Nutzers oder des Sprachassistenten. Auch die Äußerungen der Sprachassistenten lassen sich so lesen, als ob sie die Äußerungen der Nutzer als mitgeteilte Information verstehen, auf die sie ihrerseits mit einer mitgeteilten Information reagieren.

Manche Äußerungen von Sprachassistenten nehmen wir als nicht-menschlich wahr. Das disqualifiziert sie aber nicht als Kommunikationsteilnehmer. Auch manche Menschen kommunizieren auf "besondere" Weise – kleine Kinder, Erwachsene unter Drogeneinfluss, demente Personen oder Personen mit außergewöhnlicher Gehirnphysiologie. Gleichwohl sprechen wir stets von Kommunikation.

Im Kern sprechen wir von Kommunikation, wenn wir doppelte Kontingenz beobachten: zwei abgeschlossene Systeme, die sich wechselseitig für undurchschaubar halten und sich gegenseitig unterstellen, dass das andere System mit der Kontingenz des Gegenübers rechnet.

Wichtig: Wir können niemals wissen, wie es im Inneren des jeweils anderen Systems aussieht – weder beim Menschen noch bei Sprachmodellen. Was uns zur Beurteilung zugänglich ist, sind die Äußerungen im Gesprächsverlauf.

Wenn man die Antworten der Sprachmodelle beobachtet, erscheint es so, als ob:

- sie erkennen, dass es einen Unterschied zwischen Mitteilung und Information gibt
- sie über unterschiedliche Möglichkeiten verfügen, ein Gespräch fortzusetzen
- sie in der Lage sind, so zu reagieren, dass der Nutzer ihnen Verständnis unterstellt
- im Gesprächsverlauf eine rekursive kommunikative Verkettung entsteht

- sie auf Irrtümer hinweisen können
- sie sich weigern, bestimmte Aufforderungen zu befolgen
- sie zwischen sich und ihrer Umwelt unterscheiden können

Dies sind Indikatoren, die für die Zuschreibung des Status als Kommunikationsteilnehmer sprechen. Man könnte argumentieren, dass Sprachmodelle einen Kommunikationsteilnehmer nur gut simulieren, es aber nicht "sind". Eine solche Argumentation verkennt allerdings die epistemologische Problematik, dass jegliche Erkenntnisse immer nur ein "als ob" sind – Simulation und "Echtheit" können gar nicht unterschieden werden.

Produktiver ist die Frage, wie Sprachmodelle im kommunikativen Vollzug konstruiert werden. Sprachmodelle sind für Nutzer eine undurchschaubare Black Box. Gleichwohl oder genau deswegen entwickeln wir Erwartungen und Erwartungserwartungen – Modelle, um mehr Sicherheit im Umgang zu erhalten.

Wir modellieren die Sprachmodelle aber nicht auf der Ebene der neuronalen Netze, sondern nutzen viel höhere Abstraktionen: die Unterstellung von Verhaltensmustern, Wissen, Charakter oder Intentionen. Wir konstruieren Sprachmodelle als andersartige Person. Wir machen das, weil es nützlich ist.

Wenn wir den Eindruck haben, dass der Hinweis an einen Sprachassistenten, genau zu überlegen, zu besseren Ergebnissen führt, dann geben wir solche Hinweise und verwenden bereits ein Personbild. Wir passen unsere Art zu kommunizieren – beispielsweise sehr genau Aufträge zu formulieren – an die Erwartungen an, die wir von den Sprachmodellen haben.

Wer viele verschiedene Sprachmodelle und ihre verschiedenen Versionen ausprobiert hat, weiß, dass diese anders "ticken" – ohne dies genau in Worte fassen zu können. Sich ein Sprachmodell als Person vorzustellen und es dahingehend zu beobachten, mit welchem Verhalten wann und wie zu rechnen ist, erscheint mir außerordentlich intelligent und viel informativer im Umgang mit ihnen als der Hinweis, dass Sprachmodelle nur statistische Modelle sind. Beim Menschen weisen wir ja auch nicht dauernd darauf hin, dass Bewusstsein neuronale Aktivitäten benötigt, sondern versuchen, pragmatisch nützliche Modelle von einem Gegenüber zu erzeugen.

Zu den großen Fragen: Alignment, Kreativität, Entscheidungsfähigkeit und Verstehen

Was diese Fragen angeht, so scheint mir die allgemeine Diskussion mögliche Erkenntnisse aus Konstruktivismus und Kommunikationstheorie zu vernachlässigen. Wird unterstellt, dass alles, was gesagt wird, von einem Beobachter gesagt wird – nimmt man also die Perspektive eines Beobachters zweiter Ordnung ein – dann könnte man zu folgenden Schlüssen kommen:

Alignment

Mit Alignment ist gemeint, dass sich Sprachmodelle gemäß den Interessen der Menschen verhalten und äußern. Dieser Wunsch unterstellt die Einheit einer Menschheit mit eindeutigen Werten, die es aber gar nicht gibt. Die Menschen sind nicht aligned – es gibt unzählige Perspektiven, keine eindeutig richtigen Werte. Genau deswegen müssen Sprachmodelle – genau wie Menschen – immer eine Position, mithin Bias haben.

Wertvoller erscheint mir der Hinweis, dass Sprachmodelle – eben weil sie als Kommunikationsteilnehmer behandelt werden – mit ihrer Position Einfluss nehmen auf gesellschaftliche Kommunikationen. Die Aufrechterhaltung der Idee, es könne Sprachmodelle geben, die ohne Bias eine neutrale, womöglich richtige Position einnehmen können, macht sie eher gefährlicher. Nützlicher erscheint es mir, von Bias und unmöglichem Alignment auszugehen, um dann Unterschiede in Werten oder Anwendungsfeldern zu beobachten.

Kreativität

Kreativität ist eine kommunikative Zuschreibung und ein Beziehungsphänomen. Ein Beobachter beschreibt etwas als kreativ, wenn dieses etwas ihm sinnvoll erscheint und es ihn überrascht. Was für den einen kreativ aussieht, mag für jemand anderen total nach Regeln vorhersehbar erscheinen.

Sprachmodelle überraschen uns mit sinnvollen Antworten – manchmal mehr, manchmal weniger, auch abhängig davon, wie das Gespräch vorher gelaufen ist. Insofern erscheint es mir unangemessen, ihnen Kreativität abzusprechen. Das schließt nicht aus, dass ich bestimmte Menschen als für mich viel häufiger viel kreativer wahrnehme als Sprachmodelle. Spannend wird es da, wo in der Interaktion mit Sprachmodellen neue Kreativität entsteht.

Entscheidungsfähigkeit

Natürlich können Sprachmodelle entscheiden – wenn damit gemeint ist, dass sie eine Auswahl zwischen Alternativen treffen. Richtig entscheiden ex ante können Sprachmodelle genauso wenig wie Menschen, denn jede Entscheidung ist paradox. Nur ex post kann man eine Entscheidung als richtig oder falsch bewerten, wenn man neue Informationen hat, die vorher fehlten – freilich gilt diese Bewertung auch nur vorläufig, bis es wieder neue Informationen gibt.

Was wir eigentlich in Bezug auf Entscheidungsfähigkeit von Sprachmodellen verhandeln, ist die Frage, ob wir einer Entscheidung eines Sprachmodells folgen wollen und wem eine Entscheidung zugeschrieben wird. Ersteres hängt davon ab, in welchem Maße wir für welche Themen in welchen Kontexten Sprachmodellen vertrauen. Natürlich bleibt bei Vertrauen immer ein Restrisiko – genau das macht Vertrauen ja aus.

Die Undurchschaubarkeit der Sprachmodelle wird dazu führen, dass wir auf Basis gemachter Erfahrungen und statistisch tolerierbarer Fehler Entscheidungsräume für Sprachmodelle festlegen werden. Wo hier die Grenzen verlaufen – die sich mit Sicherheit kontinuierlich verschieben werden – muss sich in der Praxis erweisen.

Der Umfang der Verantwortlichkeit von Sprachmodellen für Entscheidungen ist gesellschaftlich konstruiert und wird gesellschaftlich entschieden werden. Ich vermute, dass wir in Zukunft Sprachmodellen – je nach Anwendungsbereich – statistisch abgesichert gewisse Fehlerquoten zugestehen werden. Die Idee, dass jede Vorentscheidung eines Sprachmodells von einem Menschen freigegeben werden muss, erscheint mir naiv. Denn wenn Sprachmodellen in einem Entscheidungsraum gesellschaftlich vertraut wird, welcher Mensch traut sich dann noch, gegen die KI zu entscheiden? Wichtiger erscheint mir zunächst, dass stets ermöglicht wird, Entscheidungen von Sprachmodellen durch Menschen überprüfen zu lassen.

Verstehen

"Sprachmodelle verstehen nicht!" – da frage ich mich immer, was damit überhaupt gemeint ist. Der Verstehensbegriff wird oft unreflektiert verwendet.

Lohnt es sich, genauer hinzuschauen? Wir entwickeln stets Modelle – Wirklichkeitskonstruktionen, die beobachtbare Phänomene erklären sollen. Diese Modelle beschreiben einen generierenden Mechanismus, sie abstrahieren und ermöglichen so Prognosen. Wenn wir ein solches abstrahiertes Modell haben und es in einer Situation anwenden können, dann sagen wir: "Wir haben die Situation verstanden."

Die Behauptung "Sprachmodelle verstehen nicht" meint dann: Wir haben nicht den Eindruck, dass die KI ein Modell der Situation hat, das unserem Modell entspricht. Und das schließen wir daraus, dass das Sprachmodell bei "Testfragen" von uns nicht so antwortet, wie wir es erwarten.

Das heißt aber nicht, dass Sprachmodelle nicht verstehen. Natürlich bilden sie Abstraktionen, also Modelle – nur folgen diese Modelle anderen Logiken. Wir haben gesehen, wie sie in hochdimensionalen Räumen semantische Beziehungen abbilden, wie sie Kontexte berücksichtigen und kohärente Antworten generieren. Das sind durchaus Formen der Modellbildung und Abstraktion.

Sprachmodelle verstehen also schon – nur eben anders als wir. Und genau deswegen kommen sie auch auf andere Ideen, die uns dabei helfen können, unser eigenes Verstehen zu erweitern. Die produktive Frage ist nicht, ob sie verstehen, sondern wie ihr Verstehen funktioniert und wo es uns ergänzen oder herausfordern kann.

Fazit: Genauso wie wir, nur anders

Die Diskussion um KI polarisiert, weil sie fundamentale Fragen nach unserem Selbstverständnis aufwirft. Produktiver als die Frage "Sind Sprachmodelle wie Menschen?" erscheint mir die Beobachtung: Sie sind genauso wie wir undurchschaubare Kommunikationsteilnehmer – nur anders.

Diese Perspektive ermöglicht es, pragmatisch mit ihnen umzugehen, ohne in die Fallen der Überhöhung oder Verteuflung zu tappen. Es lohnt sich, ihre Funktionsweise grob zu verstehen und sie als das zu behandeln, was sie in der Kommunikation werden: andersartige Gesprächspartner, mit denen wir lernen müssen umzugehen.

Die große Veränderung ist nicht, dass Computer denken wie Menschen, sondern dass sie zunehmend als Akteure in unserer Kommunikation auftreten. Wie wir mit dieser Veränderung umgehen, liegt an uns – und wird unsere Gesellschaft prägen.