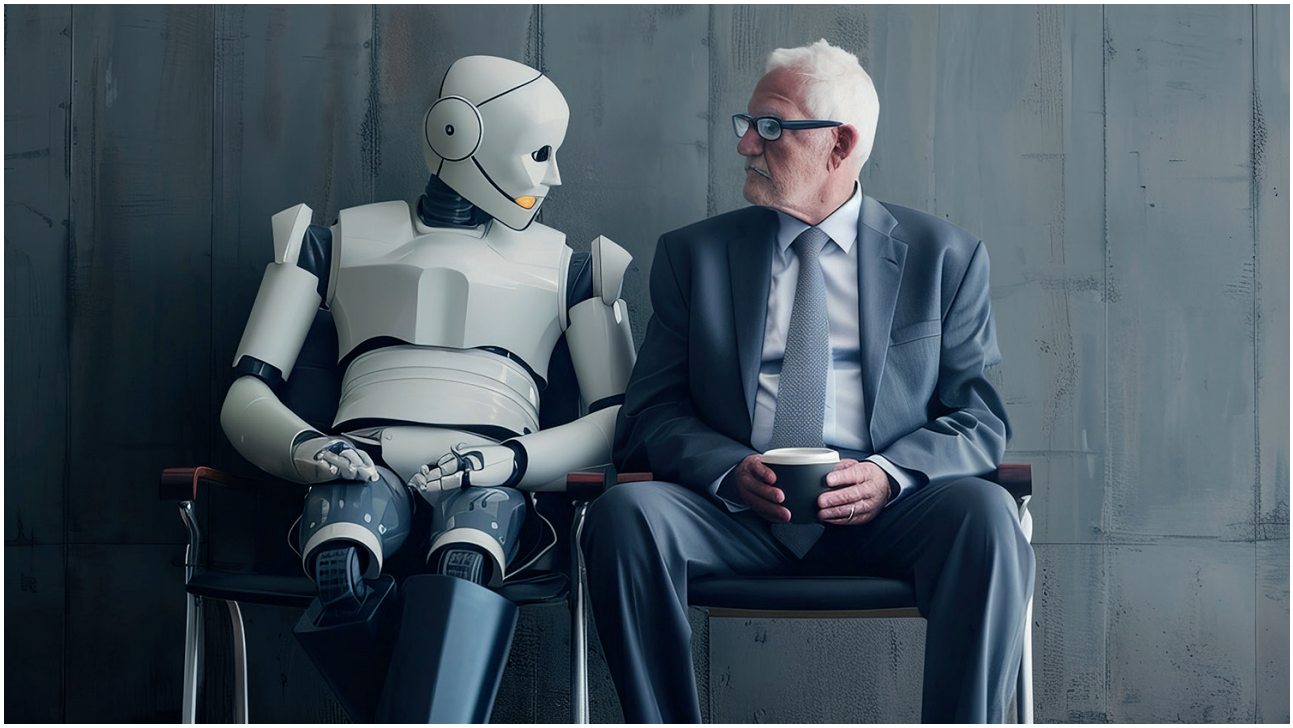
Versteht KI die Sprache doch? – Antworten auf Ihre Einwände



Ein Roboter sitzt auf einem Stuhl neben einem Menschen – versteht die KI Sprache wie ein Mensch?
Bild: KI-generiert – MD Media - stock.adobe.com

Kann eine KI Sprache je wirklich verstehen? Was ist, wenn man simuliertes Sprachverständnis nicht mehr von richtigem Verständnis unterscheiden kann? Steckt in unseren Köpfen nicht auch irgendwie ein Large Language Model? Letzte Woche habe ich mit den Gedanken des Sprachphilosophen Gottlob Frege gezeigt, warum eine KI zwar verblüffend gute Texte produzieren, Sprache aber nicht verstehen kann. Darauf habe ich sehr viele Rückmeldungen erhalten. Vielen Dank dafür. Zustimmung und Ablehnung haben sich dabei etwa die Waage gehalten. Ich habe die fünf wichtigsten Einwände herausgesucht und versuche hier, darauf einzugehen. Es geht um den Unterschied zwischen korrekten Ergebnissen und Verstehen, Beispiele für das Nichtverstehen der KI und die ganz grundsätzliche Frage, was unser Gehirn von einem solchen Sprachmodell unterscheidet – abgesehen davon natürlich, dass wir mehr Rechtschreibfehler machen und deutlich langsamer sind.

Die Sprache ist ein Wunder. Mit zwei Dutzend Buchstaben und ein paar Tausend Wörtern lässt sich alles ausdrücken: von der Liebeserklärung bis zur Anleitung zum Bau einer Atombombe, von «Hamlet» bis zum Kochrezept. Möglich ist das, weil Wörter Symbole sind: Es sind Zeichen, die für etwas anderes stehen. Das Wort «Baum» steht für ein Konzept in unseren Köpfen, das wir uns erarbeitet haben, indem wir auf Bäume geklettert sind, Herzen in deren Rinde geschnitzt haben und uns im lauen Schatten einer Baumkrone küssten.

Der Sprachphilosoph Gottlob Frege sagt: Ein Wort hat nicht nur einen aussersprachlichen Sinn, dieses Konzept in unseren Köpfen, sondern auch eine Bedeutung: der Gegenstand selbst. Wer die Welt nicht erlebt, hat keinen Zugriff auf die Bedeutung der Wörter. Das gilt auch für die KI: Sie kann unendlich viele Zeichen verarbeiten, Wörter zählen und daraus Wahrscheinlichkeiten kalkulieren. Weil sie aber nie auf einen Baum klettern kann, nie

eine Rinde berühren und nie eine andere KI unter einem Baum küssen kann, versteht sie letztlich das Wort «Baum» nicht.

Das sind, in etwa, die grundsätzlichen Einwände, die Gottlob Frege gegen die Behauptung ins Feld führen würde, dass die KI unsere Sprache versteht. Es geht dabei also nicht darum, wie korrekt die Sätze sind, die eine KI formuliert und ob sie zutreffen. Es sind grundsätzliche Überlegungen zu Sprache und Verständnis. Das stösst nicht nur auf Gegenliebe.

Der Taschenrechner

Michael Pfiffner ist damit nicht einverstanden. Er schreibt: *Der Taschenrechner versteht die Wirtshausrechnung auch nicht und liefert trotzdem wahre Ergebnisse.*

Das ist ein schönes Bild. Es leuchtet sofort ein: Der Taschenrechner rechnet perfekt, ohne die Welt verstehen zu müssen. Also sollte die KI doch auch Wahrheit produzieren können. Stellen wir uns die Szene kurz vor: Ein Wirtshaus, die Bedienung legt die Rechnung auf den Tisch. Was passiert jetzt? Ein Gast nimmt den Taschenrechner in die Hand, tippt die Preise ein und hat, schwupps, das richtige Resultat auf dem Display. Wo liegt der Haken?

Wir haben den Gast vergessen. Der Taschenrechner tut nur, was der Gast eintippt. Der Taschenrechner kann dabei keine Fehler machen, er rechnet immer richtig. Nicht so der Gast: Er kann sich vertippen oder eine Zahl falsch ablesen. Und genau das ist der Punkt: Innerhalb seiner symbolischen Welt ist der Taschenrechner perfekt. Aber ein Mensch muss ihm die Welt in Zahlen übersetzen und ihn bedienen. Das ist bei der KI genauso. Ich glaube, wir überschätzen die Fähigkeit der KI, weil wir die Bedeutung des Menschen unterschätzen.

Der 20-nach-20-Uhr-Fehler

Herr Berg schreibt: *In diesem Zusammenhang würde mich ein konkretes Beispiel interessieren, welche Frage an ein Sprachmodell zu einer Antwort führt, die ein Nicht-Verstehen deutlich zeigen kann. In der praktischen Erfahrung erleben wir, dass heutige Sprachmodelle weit über das Neu-Gruppieren von Sprache nach Regeln der Wahrscheinlichkeit hinausgehen. So ein Modell «sieht» die Welt auf eine andere Weise als wir.*

Es gibt ein einfaches Beispiel, mit dem sich zeigen lässt, dass Sprachmodelle auf Statistik beruhen und nicht auf dem Verstehen der Welt. Fordern Sie ChatGPT oder ein anderes LLM auf, das Bild eines analogen Zifferblatts zu generieren, das 20.20 Uhr zeigt. Das Resultat wird das Bild einer Uhr sein, die 10.10 Uhr zeigt. Warum?

Weil Uhrenhersteller ihre Uhren gerne mit dieser Uhrzeit abbilden. Es ist das Smiley-Face des Zifferblatts. Offensichtlich hat die KI gelernt, dass Uhren so aussehen. Es ist gut möglich, dass dieser spezifische Fehler irgendwann behoben wird, das ändert aber nichts an der Tatsache, dass KI-Systeme nach dem Prinzip der Wahrscheinlichkeit funktionieren und keinen Begriff von der Wahrheit haben. Es ist wie bei einer Wetterprognose: Ob die Prognose zutrifft oder nicht, ist egal. Eine Prognose ist immer eine Aussage über die Wahrscheinlichkeit und nie über die Wahrheit.

Auch zutreffende Prognose ist nicht wahr

Pinke Helga schreibt: *Haben Sie noch nie mit verschiedenen Modellen gearbeitet? KI ist viel mehr als nur ChatGPT. Meine Erfahrungen zeigen, daß einige Modelle sehr gut komplexe Texte, die sich sogar selbst referenzieren, verstehen können.*

Das glaube ich gern. Wir landen aber immer wieder am gleichen Ort: Auch wenn die Wetterprognose zutrifft, ist es keine Aussage über das Wetter, sondern eine Vorhersage. Es ist ein kategorialer Unterschied, ob ich aus dem Fenster blicke und feststelle, dass es regnet, oder ob ich eine Wetterprognose konsultiere, die mir sagt, dass die Wahrscheinlichkeit sehr hoch ist, dass es jetzt gerade regnet.

Der Clever-Hans-Effekt

Das gilt auch für den Einwand von @emojized. Er oder sie schreibt: *Das Ganze erinnert mich an die Simulationstheorie: Wenn die Wahrnehmung und das Ergebnis für uns identisch sind, spielt es letztlich keine Rolle, ob es sich um eine «echte» Realität handelt oder nicht. Genauso wenig wie es einen Unterschied macht, ob wir mit einem einfachen LLM oder einem komplexen neuronalen Netz interagieren – solange die Auswirkungen und Erfahrungen für uns gleich sind.*

Vielleicht kennen Sie die Geschichte vom «klugen Hans». Wilhelm von Osten, ein Lehrer aus Preussen, trat um 1900 in Berlin mit seinem Pferd «Hans» auf: Das Pferd konnte rechnen, lesen, musizieren und den Kalender benutzen. Wenn man Hans fragte «Wie viel ist drei plus zwei?», so klopfte das Pferd fünfmal mit dem Huf. Das Berliner Publikum war begeistert, Zeitungen berichteten, und das Pferd wurde zu einer Sensation.

Aber konnte das Pferd wirklich rechnen? Viele Experten zweifelten daran. Also wurde eine wissenschaftliche Kommission damit beauftragt, das rechnende Pferd zu untersuchen. In dieser «Hans-Kommission» arbeitete der

auch der Psychologe Oskar Pfungst. Er testete das Pferd Hans ausführlich und stellte etwas Interessantes fest:

- Wenn der Fragesteller die Lösung selbst kannte, antwortete Hans korrekt.
- Wenn der Fragesteller die Lösung nicht kannte, scheiterte Hans.
- Wenn Hans die Fragesteller nicht sehen konnte, klappte es ebenfalls nicht.

Das brave Pferd konnte also nicht rechnen, das Tier reagierte auf mikromotorische Signale. Sobald Hans die richtige Anzahl von Hufschlägen erreicht hatte, veränderte der Fragesteller minimal seinen Gesichtsausdruck und seine Körperhaltung. Diese Signale konnte Hans interpretieren – und hörte auf, mit den Hufen zu klopfen.

Die Sache mit den Wölfen

Psychologe Oskar Pfungst hatte den «Clever-Hans-Effekt» entdeckt: Versuchstiere, aber auch Menschen reagieren oft auf unbewusste Hinweise der Forschenden. Diese Entdeckung führte unter anderem dazu, dass für Versuche strengere Bedingungen eingeführt wurden, zum Beispiel die doppelblinde Versuchsanordnung. Dabei wissen weder die Probanden noch die Forscher, ob ein Placebo oder das Medikament verabreicht wird.

Von einem solchen «Clever-Hans-Effekt» spricht man auch im Umgang mit der KI: Er tritt dann auf, wenn die KI die richtigen Resultate ausspuckt, aber aus den falschen Gründen. Ein bekanntes Beispiel in der Forschungsliteratur ist die Klassierung von Wölfen auf Fotos: In einem Experiment wurde die KI mit Fotos von Wölfen trainiert und erkannte die Tiere danach auf anderen Bildern perfekt.

Bis die Forscher die Bilder variierten: Sie konnten zeigen, dass die KI sich nicht den Wolf gemerkt hatte, sondern den Schnee im Bild. Die KI hatte, wenigstens zunächst, korrekte Resultate geliefert, aber aus den falschen Gründen – ein Clever-Hans-Effekt. Das Problem dabei: Anders als bei einem herkömmlichen Computerprogramm gibt es keinen Code, den man nachschlagen könnte. Die Lernvorgänge in einem neuronalen Netz sind eine Blackbox.

Ist das Gehirn ein Sprachmodell?

Kommen wir zum letzten Einwand, er stammt von @ ErhabeneWahrheit: *Der menschliche Verstand ist ein Sprachmodell, ähnlich einer Konversations-KI wie GPT. Es scheint, dass dies schwer zu akzeptieren ist. Aber: der menschliche Verstand hat auch keinen direkten Bezug zur Wahrheit oder zur Erfahrung. Es gibt nur Erinnerungen, die als Informationen vorliegen. Diese Informationen hat die KI im Prinzip ebenso.*

Das ist ein spannender Einwurf. Ist das Gehirn nichts anderes als eine Art Computer voller Informationen? Ist das Gehirn eine Art Informationen verarbeitende Maschine, die in unserem Schädel eingesperrt ist wie Prozessor und Speicherchip in einem Computer? Wenn man davon ausgeht, dann ist unser Sprachvermögen tatsächlich eine Art LLM, ein Sprachmodell, das durch Augen und Ohren Input erhält und daraus einen Output generiert.

Waldhorn und Skifahren

Ich spiele Horn. Waldhorn. Ein schwieriges Instrument. Ich erinnere mich gut, wie mein ungarischer Hornlehrer mich manchmal ruppig auf Fehler hinwies. «Matthias! Das habe ich Dir doch schon tausendmal gesagt, dass Du hier nicht die 3 drücken sollst sondern die 1 und 2». Er meinte die Ventile. Ich wusste es ja. Meine Finger machten trotzdem etwas anderes. Das ist der Unterschied zwischen dem deklarativen Gedächtnis, zu dem das Sprachgedächtnis gehört, und dem prozeduralen Gedächtnis. Anders gesagt: In unserem Gehirn hat es viel mehr als Sprache.

Typische Beispiele für das prozedurale Gedächtnis sind Tanzen und Fahrradfahren, Klavierspielen (oder Horn), Töpfern und Nähen, Schwimmen und Skifahren, mit zehn Fingern schreiben oder Essen mit Stäbchen, Schuhe binden und Zähne putzen. Sie «wissen», wie man sich die Schuhe bindet – aber versuchen Sie einmal, das präzise zu beschreiben.

Umgekehrt nützt es nichts, wenn Ihnen jemand genauestens erklärt, wie Sie auf der Skipiste den nächsten Schwung nehmen sollen – Ihr Körper muss es ins Gefühl kriegen. Der Philosoph Gilbert Ryle unterscheidet deshalb zwischen knowing that und knowing how.

Wir Menschen verfügen nicht nur über sprachliches Wissen. Wir lernen auch mit dem Körper, durch Bewegung und sensorische Rückkopplung. Diese Art des Wissens fehlt einem LLM. In ihrem Buch «The extended mind – The power of Thinking outside the Brain» argumentiert Annie Murphy Paul, dass unser übliches Verständnis von Denken viel zu stark «hirnzentriert» sei.

Denken, Lernen, Problemlösen passiere keineswegs nur im Kopf. Sie zeigt, dass sich viele kognitive Prozesse über den ganzen Körper erstrecken, ja über den Körper hinaus in die physische Umgebung und über soziale Beziehungen. Annie Murphy Paul sagt, diese «extra-neuralen» Ressourcen – Körper, Raum, Mitmenschen – seien wesentlich dafür verantwortlich, wie wir denken, verstehen, uns erinnern und wie wir kreativ sein können.

Vielleicht ist unser Sprachzentrum ein kleines LLM – aber darum herum haben wir Menschen noch viel mehr im Kopf. Und nicht nur um Kopf, sondern auch im Herzen und in der Hand.

Basel 12. September 2025, Matthias Zehnder
mz@matthiaszehnder.ch

Anmmerkungen

Zu Frege: Frege versteht Sinn nicht als «Konzept in unseren Köpfen», sondern als die «Art und Weise der Gegebenheit». Für unsere Zwecke spielt das keine Rolle, es sei aber angemerkt.

Zum Taschenrechner: Natürlich können theoretisch auch Taschenrechner Fehler haben, etwa bei einem Hardwaredefekt. Grundsätzlich implementiert ein Taschenrechner aber deterministische Algorithmen und die werden immer gleich abgearbeitet.

Zu den Wetterprognosen: Ganz präzise formuliert müsste ich sagen, dass Prognosen probabilistische Aussagen sind und solche Aussagen können nicht wahr oder falsch sein, sondern nur zutreffend oder nicht zutreffend.

Zum LLM als Sprachgedächtnis: Ich habe versucht, möglichst stark auf den Einwand einzugehen. Ich möchte aber noch einmal unterstreichen: Ein LLM ist ein Modell für statistische Regularitäten, kein Modell für das menschliche Sprachgedächtnis.

Quellen

Frege, Gottlob: Über Sinn und Bedeutung (1892). In: ders., Funktion, Begriff, Bedeutung. Fünf logische Studien. Hrsg. von Günther Patzig. Göttingen: Vandenhoeck & Ruprecht, 1962.

Frege, Gottlob: Die Grundlagen der Arithmetik. Hamburg: Meiner, 1986 (Erstausgabe 1884).

Frege, Gottlob: Grundgesetze der Arithmetik. 2 Bde. Stuttgart-Bad Cannstatt: Frommann-Holzboog, 1998ff (Erstausgabe Jena 1893/1903).

Krishnapuram, Balaji (2016): Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY 2016 ACM Conferences.

Paul, Annie Murphy (2021): The extended mind: the power of thinking outside the brain, New York 2021.

Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos (2016): „Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA 2016, S. 1135–1144, https://dl.acm.org/doi/10.1145/2939672.2939778 [12.09.2025].

Ryle, Gilbert (2015): Der Begriff des Geistes, übers. v. Kurt Baier, Günther Patzig, Ulrich Steinvorth, [Nachdruck] 2021, Ditzingen 2015 Reclams Universal-Bibliothek Nr. 19345.