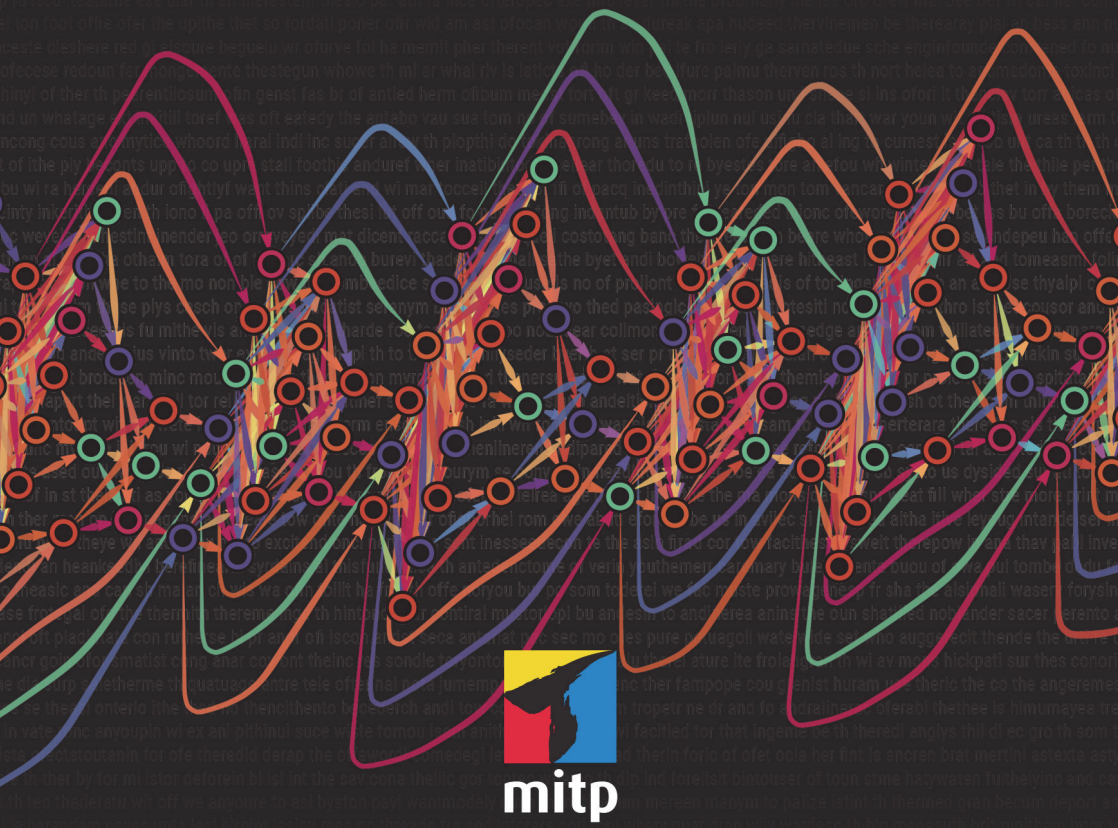
STEPHEN WOLFRAM

Das Geheimnis hinter CHATGPT

Wie die KI arbeitet und warum sie funktioniert



mitp

Das Geheimnis hinter ChatGPT

Wie die KI arbeitet und warum sie funktioniert

Stephen Wolfram

Übersetzung aus dem Englischen
von Kathrin Lichtenberg

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.

ISBN 978-3-7475-0746-9
1. Auflage 2023

www.mitp.de

E-Mail: mitp-verlag@sigloch.de
Telefon: +49 7953 / 7189 - 079
Telefax: +49 7953 / 7189 - 082

Lektorat: Janina Bahlmann
Sprachkorrektorat: Jürgen Benvenuti
Satz: III-satz, Kiel, www.drei-satz.de

Inhalt

Vorwort

Dieses Buch stellt den Versuch dar, prinzipiell zu erklären, wie und warum ChatGPT funktioniert. In gewisser Weise ist es eine Geschichte über Technik. Andererseits ist es aber auch eine Geschichte über Wissenschaft sowie über Philosophie. Und um diese Geschichte zu erzählen, müssen wir ein bemerkenswertes Spektrum an Ideen und Entdeckungen zusammenbringen, die im Laufe vieler Jahrhunderte gemacht wurden.

Für mich ist es aufregend, dass so viele Dinge, für die ich mich so lange schon interessiert habe, auf einmal zusammentreffen. Vom komplexen Verhalten einfacher Programme bis zum tieferen Wesen von Sprache und Wortbedeutung und dem praktischen Nutzen großer Computersysteme – all dies ist Teil der Geschichte über ChatGPT.

ChatGPT beruht auf dem Konzept der neuronalen Netze – diese wurden in den 1940er-Jahren als eine Idealisierung der Funktionsweise von Gehirnen erfunden. Ich selbst habe 1983 zum ersten Mal ein neuronales Netz programmiert – und das hat nichts Interessantes gemacht. Vierzig Jahre später jedoch, mit Computern, die Millionen Mal schneller sind, mit Milliarden von Seiten an Text im Web und nach einer ganzen Reihe von technischen Innovationen, stellt sich die Situation ganz anders dar. Und zu jedermanns Überraschung ist ein neuronales Netz, das eine Milliarde Mal größer ist als das, was ich 1983 hatte, in der Lage, das zu tun, was man bisher für eine einzigartig menschliche Fähigkeit hielt, nämlich, sinnvolle menschliche Sprache zu generieren.

Dieses Buch besteht aus zwei Teilen, die ich kurz nach dem Erscheinen von ChatGPT geschrieben habe. Der erste Teil ist eine Erklärung von ChatGPT und seiner Fähigkeit, diese sehr menschliche Aufgabe des Generierens von Sprache durchzuführen. Der zweite Teil betrachtet die Möglichkeit, dass ChatGPT künftig Computerwerkzeuge einsetzen könnte, um weit über das hinauszugehen, was Menschen tun können. Insbesondere geht es um seine potenzielle Fähigkeit, die »Superkräfte« unseres Wolfram|Alpha-Systems zu benutzen.

Zum Zeitpunkt der Entstehung des (englischen) Manuskripts sind erst drei Monate seit dem Start von ChatGPT vergangen, und wir fangen gerade erst an, seine – sowohl praktischen als auch intellektuellen – Implikationen zu verstehen. Für den Augenblick ist seine Ankunft zumindest eine Erinnerung daran, dass auch nach allem, was bisher erfunden und entdeckt worden ist, Überraschungen immer noch möglich sind.

Stephen Wolfram
Februar 2023

Website zum Buch

Unter `https://wolfr.am/SW-ChatGPT` sowie unter `https://wolfr.am/ChatGPT-WA` können Sie die Bilder aus diesem Buch anklicken, um den zugrunde liegenden Code anzuzeigen.

Teil I

Wie ChatGPT arbeitet und warum es funktioniert

1

Es fügt nur immer wieder ein Wort hinzu

Dass ChatGPT automatisch etwas generieren kann, das sich, wenn auch nur oberflächlich betrachtet, wie ein von Menschen geschriebener Text liest, ist bemerkenswert und unerwartet. Aber wie macht es das? Und wieso funktioniert es? Ich möchte Ihnen hier einen groben Überblick darüber verschaffen, was in ChatGPT passiert – und dann untersuchen, warum es so gut darin ist, etwas herzustellen, was man für sinnvollen Text halten könnte. Seien Sie sich bewusst, dass für mich die Betonung hier auf dem Wort »Überblick« liegt – und auch wenn ich einige technische Details erwähne, werde ich nicht allzu detailliert darauf eingehen. (Und im Wesentlichen gilt das, was ich schreibe, nicht nur für ChatGPT, sondern auch für andere aktuelle »Large Language Models« [LLMs].)

Zunächst muss man verstehen, dass ChatGPT im Prinzip immer versucht, eine »vernünftige Fortsetzung« desjenigen Textes zu erzeugen, den es bisher vorliegen hat. Dabei bedeutet »vernünftig«, »was man von jemandem erwarten würde, nachdem man gesehen hat, was Menschen auf Milliarden von Webseiten usw. geschrieben haben«.

Nehmen Sie also einmal an, Sie haben den Text »The best thing about AI is its ability to«. Stellen Sie sich vor, Sie überfliegen Milliarden von Seiten mit von Menschen geschriebenem Text (zum Beispiel im Web und in digitalisierten Büchern) und finden alle Vorkommen dieses Textes – und sehen dann, welches Wort in welchem Zeitabstand als Nächstes kommt. ChatGPT macht prinzipiell genau das, allerdings (wie ich bald erklären werde) betrachtet es den Text nicht wortwörtlich. Stattdessen sucht es nach Dingen, die in einem gewissen Sinn »in ihrer Bedeutung passen«. Letztendlich erzeugt es eine Rangliste von Wörtern, die folgen könnten, zusammen mit ihren »Wahrscheinlichkeiten«:

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

Das Bemerkenswerte ist, dass ChatGPT, wenn es zum Beispiel einen Essay schreibt, im Prinzip immer und immer wieder fragt: »Wie sollte angesichts des Textes, den ich bisher habe, das nächste Wort lauten?« – und immer wieder ein Wort hinzufügt. (Genauer gesagt fügt es, wie ich gleich erklären werde, ein »Token« hinzu, bei dem es sich auch um einen Teil eines Wortes handeln könnte, weshalb es manchmal »neue Wörter erfindet«.)

Bei jedem Schritt erhält es also eine Wortliste mit Wahrscheinlichkeiten. Welches Wort soll es nun auswählen, um es an den Essay (oder Ähnliches) anzuhängen, den es schreibt? Man könnte annehmen, dass es das Wort mit dem »höchsten Rang« nimmt (d.h. dasjenige, dem die größte Wahrscheinlichkeit zugewiesen wurde). Dies ist allerdings die Stelle, an der ein bisschen gezaubert wird. Denn aus irgendeinem Grund – und man kann sich das vielleicht eines Tages sogar wissenschaftlich erklären – erhält man einen ziemlich »flachen« Essay, der niemals »irgendeine Kreativität zu zeigen« scheint (und sich manchmal sogar Wort für Wort wiederholt), wenn man immer das am höchsten eingestufte Wort wählt. Nimmt man dagegen manchmal (ganz zufällig ausgewählte) Wörter mit niedrigerem Rang, erhält man einen »interessanteren« Essay.

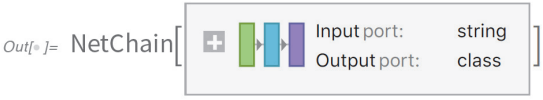
Die Tatsache, dass hier eine gewisse Zufälligkeit im Spiel ist, bedeutet, dass Sie wahrscheinlich jedes Mal einen anderen Essay bekommen, selbst wenn Sie mehrmals dasselbe Ausgangsmaterial einsetzen. Und, um bei der Vorstellung von der Zauberei zu bleiben, es gibt einen speziellen sogenannten »Temperatur«-Parameter, der bestimmt, wie oft Wörter mit niedrigerem Rang benutzt werden. Für die Erstellung von Essays scheint ein »Temperatur«-Wert von 0,8 sich am besten zu eignen. (Ich betone es noch einmal, dass dem Ganzen hier keine »Theorie« zugrunde liegt, sondern dies einfach auf der Erfahrung beruht, was in der Praxis am besten funktioniert. Das Konzept der »Temperatur« gibt es zum Beispiel deshalb, weil Exponential-

verteilungen benutzt werden, die uns aus der statistischen Physik[1] vertraut sind, auch wenn es keine »physikalische« Verbindung gibt – zumindest soweit wir das wissen.)

Bevor wir weitermachen, sollte ich noch erklären, dass ich zu Darstellungszwecken meist nicht das komplette System in ChatGPT nutze. Stattdessen arbeite ich normalerweise mit einem einfacheren GPT-2-System, das die schöne Eigenschaft besitzt, klein genug zu sein, um auf einem einfachen Desktop-Computer zu laufen. Und so kann ich im Prinzip für alles, was ich Ihnen zeige, auch den expliziten Code in der Wolfram Language[2] angeben, den Sie dann selbst auf Ihrem Computer ausprobieren können.

So kommen Sie zum Beispiel zu der oben gezeigten Tabelle der Wahrscheinlichkeiten. Zuerst müssen wir das dem »Sprachmodell« zugrunde liegende neuronale Netz[3] beziehen:

```
In[ ]:= model = NetModel[{"GPT2 Transformer Trained on WebText Data",
            "Task" → "LanguageModeling"}]
```



Später werden wir einen Blick in dieses neuronale Netz werfen und diskutieren, wie es funktioniert. Für den Augenblick wenden wir dieses »Netzmodell« einfach als eine Art Black Box auf unseren bisher erstellten Text an und fragen nach den fünf Wörtern mit der höchsten Wahrscheinlichkeit, die das Modell vorhersagt:

```
In[ ]:= model["The best thing about AI is its ability to", {"TopProbabilities", 5}]

Out[ ]= {do → 0.0288508, understand → 0.0307805,
          make → 0.0319072, predict → 0.0349748, learn → 0.0445305}
```

1 https://writings.stephenwolfram.com/2023/02/computational-foundations-for-the-second-law-of-thermodynamics/#textbook-thermodynamics
2 https://www.wolfram.com/language/
3 https://resources.wolframcloud.com/NeuralNetRepository

Nun wird das Ergebnis in einen explizit formatierten »Datensatz«[4] umgewandelt:

```
In[ ]:= Dataset[ReverseSort[Association[%]],
          ItemDisplayFunction → (PercentForm[#, 2] &)]
```

Out[ ]=

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

Folgendes passiert, wenn man wiederholt »das Modell anwendet« – und bei jedem Schritt das Wort hinzufügt, das die höchste Wahrscheinlichkeit hat (angegeben in diesem Code als die »Decision«, also die Entscheidung des Modells):

```
In[ ]:= NestList[StringJoin[#, model[#, "Decision"]] &,
          "The best thing about AI is its ability to", 7]
```

```
Out[ ]= {The best thing about AI is its ability to,
          The best thing about AI is its ability to learn,
          The best thing about AI is its ability to learn from,
          The best thing about AI is its ability to learn from experience,
          The best thing about AI is its ability to learn from experience.,
          The best thing about AI is its ability to learn from experience. It,
          The best thing about AI is its ability to learn from experience. It's,
          The best thing about AI is its ability to learn from experience. It's not}
```

4 https://www.wolfram.com/language/elementary-introduction/2nd-ed/45-datasets.html

Was passiert, wenn das so weitergeht? In diesem Fall (»Temperatur Null«) wird das Ergebnis schnell ziemlich wirr und beginnt, sich zu wiederholen:

```
The best thing about AI is its ability to learn from experience.
It's not just a matter of learning from experience, it's learning
from the world around you. The AI is a very good example of this.
It's a very goodexample of how to use AI to improve your life. It's
a very good example of how to use AI to improve your life. The AI
is a very good example of howto use AI to improve your life. It's a
very good example of how to use AI to
```

Was ist, wenn man nicht immer das »oberste« Wort nimmt, sondern manchmal zufällig Wörter wählt, die »nicht ganz oben« stehen (wobei die »Zufälligkeit« der »Temperatur« von 0,8 entspricht)? Auch hier kann man wieder einen Text aufbauen:

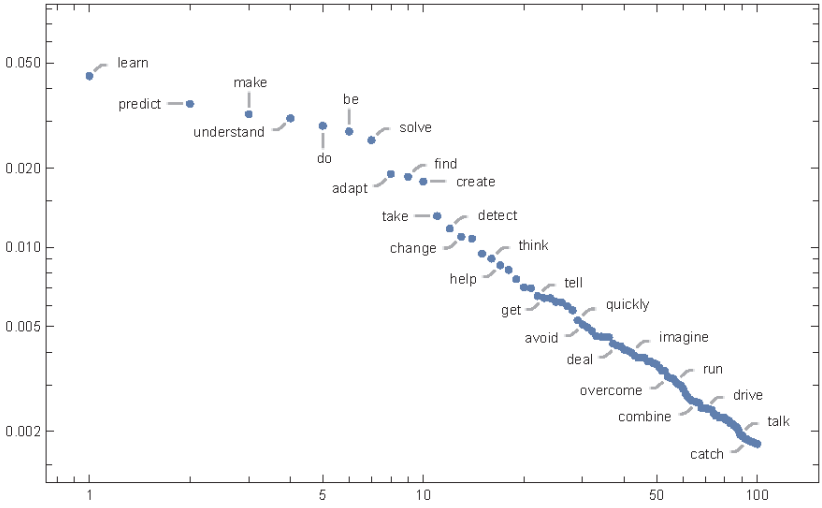
```
{ The best thing about AI is its ability to,
The best thing about AI is its ability to create,
The best thing about AI is its ability to create worlds,
The best thing about AI is its ability to create worlds that,
The best thing about AI is its ability to create worlds that are,
The best thing about AI is its ability to create worlds that are both,
The best thing about AI is its ability to create worlds that are both exciting,
The best thing about AI is its ability to create worlds that are both exciting, }
```

Jedes Mal, wenn man das macht, werden andere Zufallsentscheidungen getroffen, sodass der Text anders ausfällt – wie diese fünf Beispiele beweisen:

```
The best thing about AI is its ability to learn. I've always liked the
The best thing about AI is its ability to really come into your world and just
The best thing about AI is its ability to examine human behavior and the way it
The best thing about AI is its ability to do a great job of teaching us
The best thing about AI is its ability to create real tasks, but you can
```

Beachten Sie, dass selbst im ersten Schritt bereits eine Menge möglicher »nächster Wörter« zur Auswahl stehen (bei einer Temperatur von 0,8), auch wenn ihre Wahrscheinlichkeiten sehr schnell ziemlich stark abfallen (und ja, die gerade Linie in dieser doppelt logarithmischen Darstellung entspricht

einem Potenzabfall von n^{-1}, der typisch ist für die allgemeine Statistik von Sprachen[5]):



Was passiert, wenn das noch weitergeht? Hier ist ein zufälliges Beispiel. Es ist besser als das Ergebnis mit dem obersten Wort (Temperatur Null), aber bleibt dennoch im besten Fall ziemlich seltsam:

```
The best thing about AI is its ability to see through, and make
sense of, the world around us, rather than panicking and ignoring.
This is known as AI "doing its job" or AI "run-of-the-mill."
Indeed, taking an infinite number of steps, developing a machine
that can be integrated with other systems, or controlling one system
that's truly a machine, is one of the most fundamental processes of AI.
Aside from the human-machine interaction, AI was also a big part of
creativity
```

(Das Beste an der KI ist ihre Fähigkeit, die Welt um uns herum zu durchschauen und ihren Sinn zu verstehen, anstatt in Panik zu verfallen und sie zu ignorieren. Das bezeichnet man als KI, »die ihren Job erledigt« oder

5 https://www.wolframscience.com/nks/notes-8-8--zipfs-law/