

Large Language Kabbala

Fröhliche Wissenschaft 265

Martin Warnke

Large Language Kabbala

Eine kleine Geschichte der Großen
Sprachmodelle



Matthes & Seitz Berlin

Inhalt

Einleitung: Zu den Ursprüngen der Large
Language Models 7

I. Künstliche neuronale Netze:
ein Ort der Statistik 12

II. Eine Linguistik auf statistischer
Grundlage 42

III. Kabbala 85

IV. Walter Benjamins Schwellenkunde 105

V. Versuch eines Endes 125

Anmerkungen 135

Dank 151

Einleitung: Zu den Ursprüngen der Large Language Models

»Die Sprache Gottes nämlich
hat keine Grammatik.«¹

Es sind die Schrift-Gelehrten und nicht die Nerds, die uns ChatGPT und die anderen konkurrierenden technischen Heldentaten von heutzutage erklären können, so lautet in aller Kürze die These dieser Schrift. Die Vorgeschichte der Large Language Models (LLM), die völlig zu Recht das Wort »Sprache« zwischen »Groß« und »Modell« im Namen tragen, legt deren Wurzeln in der Ausdeutung heiliger Texte, der Literaturkritik und der Wissenschaft von der Sprache frei, betreibt also recht eigentlich Kulturwissenschaft und nicht etwa Neuropsychologie oder Technikanalyse.

Meist jedoch wird zur Erklärung der neuen technischen Systeme eine biologische Metapher herangezogen, vornehmlich das menschliche Gehirn, denn ihre Architektur und Bauweise sind netzförmig: ähnlich wie, wenngleich sehr viel einfacher als das Zentralnervensystem lebender

Wesen. Die technischen Architekturen werden darum auch künstliche neuronale Netze genannt. Die Materialien, die hier ausgebreitet werden, legen nun allerdings einen anderen Ursprung der technischen Systeme nahe: Es ist weniger das Zentralnervensystem, das uns beim Verständnis der Large Language Models hilft, sondern ein sprachtheoretischer Ansatz, der von Zellig Sabbetai Harris (1909–1992) stammt, dem Doktorvater Noam Chomskys. Harris hat seine Sprachtheorie auf eine statistische Grundlage gestellt, die sich zu seiner Zeit technisch nicht realisieren ließ, nun aber auf diesen künstlichen neuronalen Netzen in Form von etwa ChatGPT ihre maschinelle Realisierung gefunden hat. Über die bei den LLM und bei Harris gleichermaßen wirksame Perspektive auf Sprache hinaus, die auf dem Verzicht auf eine der Sprache vorgelagerte Instanz und vorgängige Bedeutung der Sprachäußerung steht, greifen wir allerdings noch weiter zurück auf die Tradition von Walter Benjamins Sprachtheorie und die mystischen Textauslegungstechniken der jüdischen Kultur, der Kabbala, in denen alles ausschließlich aus dem heiligen Text kommt.

Die Eigenheiten der jetzt so akuten Computersysteme, ihre Leistungsfähigkeit, Fehler und Beschränkungen, lassen sich größtenteils direkt über deren linguistische Herkunft plausibel machen: Es sind eben keine Gehirne, sondern Sprachstruktu-

ren, die uns derzeit so faszinieren und beängstigen. Was jetzt als Kinderkrankheit – wie etwa die »Halluzinationen« genannten haltlosen Hervorbringungen der neuen Technologie – gewertet wird, in der Hoffnung, diese bald zu überwinden, erscheint aus der hier eingenommenen Perspektive eher als zwangsläufige Folge eines zwar vor einem halben Jahrhundert ausgearbeiteten, aber außerhalb der Linguistik-Fachcommunity weitestgehend unbekannten linguistischen Verfahrens – und zwar *by design*. Die »Halluzinationen« sind keine Ausgeburten kranker künstlicher Hirne, sondern Artefakte einer Linguistik, die – um die Pointe vorwegzunehmen – die Semantik von Texten ohne Weltbezug allein aus ihrer Syntax ableitet, was zweifellos als ein drastisches Vorgehen zu werten wäre. Eine Kritik solcher Sprache aus Sprache erzeugender Computermodelle könnte – müsste vielleicht – dann eben auch hier ansetzen: Was sind das für Systeme, die nur in der Sprache sind und die keine Welt haben? Welche schwerwiegenden Nebenwirkungen löst so etwas aus?

Im Folgenden werden im ersten Kapitel zunächst die Setzungen der heutigen Large Language Models an ihren Gründungstexten freigelegt, den frühen Forschungen zu künstlichen neuronalen Netzen, den Trainingsverfahren, denen sie unterworfen werden, den Modellannahmen, die seit 2010 die erstaunlichen Durchbrüche ermöglich-

ten: Welche Annahmen haben die Pioniere der heutigen statistisch fundierten Text- und Bild-Allesfresser gemacht, die dann so folgenreich für die weitere, mittlerweile unübersichtlich werdende technische Entwicklung wurden? Und damit auch: Welche Grundsätze und Begriffe könnten damit als Theorie der technischen Methoden benannt werden, die eine Kritik erst ermöglichen? Wie können wir überhaupt wieder zur Theorie zurückkehren, nachdem diese schon lautstark für überflüssig erklärt und deren Ende ausgerufen wurde?²

Wir werden im zweiten Kapitel entdecken, dass hierfür eine linguistische Theorie geeignet ist, die vor etwa siebzig Jahren auf einem statistischen Ansatz begründet ausgearbeitet wurde und erst jetzt, wo sie von sehr großen Datensammlungen getrieben wird, zu voller operationaler Blüte gelangen konnte. Und wie alle Theorien hat sie ihre folgenreichen Eigentümlichkeiten, die in ihrer Anwendung dann kaum abzulegen sind.

Im dritten Kapitel begegnen wir einer Textauslegungstechnik, die erstaunlich viel mit den bereits besuchten Abschnitten unserer Archäologie der Large Language Models gemein hat: die jüdische Kabbala. In ihrer radikalen Beschränkung auf und ihrem unbedingten Glauben an ein Textkorpus prophezeit die kabbalistische Erzeugung von Text aus Text dann sogar Wohl und

Wehe der Large Language Models, die Methode ihrer Hervorbringungen und die tiefe Ursache ihrer Fehlleistungen.

Anschließend, im vierten Kapitel, wird dann die angedeutete Referenz zu Walter Benjamin relevant, denn durch sie lässt sich nachverfolgen, wie Benjamins Schwellenkunde ein Überdenken von Differenzen und Verengungen möglich macht, wo Perspektiven als defizitär erkennbar werden, an welchen Stellen noch etwas hinzukommen muss, um Sprache in ihrer Universalität angemessener zu erfassen.

I. Künstliche neuronale Netze: ein Ort der Statistik

»Der Preis, den wir für die Verwendung statistischer Prinzipien bei der Gestaltung des Systems zahlen, ist die Wahrscheinlichkeit, dass wir in einem bestimmten Fall eine falsche Antwort erhalten – d. h. eine falsche Antwort, die in der Natur des Systems liegt und nicht auf eine Fehlfunktion einer seiner Komponenten zurückzuführen ist.«¹
Frank Rosenblatt über seine Arbeit

Das Perceptron Frank Rosenblatts. Verrauschte Umweltsignale, ihre Statistik und deren Preis.

Im Jahr 1957 schrieb der US-amerikanische Psychologe und Informatiker Frank Rosenblatt, der heute als Begründer der neuronalen KI gilt, einen Bericht für das Cornell Aeronautic Laboratory, der die mehrjährige Arbeit an einem von ihm erdachten und unter seiner Leitung konstruierten elektromechanischen System beschrieb, das

lernt, Ähnlichkeiten in Wahrnehmungsdaten zu unterscheiden. Diese Maschine tut dies nicht deterministisch, sondern probabilistisch, was sie als Vorläufer der sehr viel komplexeren Systeme von heute qualifiziert, denn sie erbringt ihre Leistung durch statistische Auswertung einer großen Zahl von Versuchen. Mit seinem von ihm so genannten »Perceptron« legte Rosenblatt den Grund für das, was heute als *Maschinelles Lernen* bezeichnet wird, das ein Computer realisiert, der nicht programmiert, sondern trainiert wird, um statistische Auffälligkeiten in beliebigen, voraussetzungslosen Umweltdaten zu entdecken.

Dieser Zweig der Forschung an künstlicher Intelligenz², die *subsymbolische, konnektionistische KI*, arbeitete in Anlehnung an die Physiologie von Nervensystemen mit vernetzten künstlichen Neuronen und führte anfänglich ein Nischendasein, weil die ganz anders auf Logik und Regeln bauenden Systeme der *symbolischen KI* bis Anfang des 21. Jahrhunderts als allein aussichtsreich galten. Die symbolische KI arbeitete an Expertensystemen und Datenbanken, die Wissen explizit logisch kodierte und mithilfe von Ableitungsregeln auswerteten. Erst ab 2010 gelangen der subsymbolischen, konnektionistischen KI die Durchbrüche bis hin zu den LLM, sodass im Moment im öffentlichen Diskurs nur noch diese unter »KI« verstanden werden. Dieser Paradigmen-

wechsel ist typisch für die Umschwünge in den Wissenschaften der heutigen Zeit, die sich von deduktiven Methoden zu den statistischen wenden, von expliziter Kodierung von Wissen und dessen programmgesteuerter Auswertung nach ersten Prinzipien (wie bei der symbolischen KI) zu undurchsichtigen, riesigen Ansammlungen einzelner durch Training eingestellter Modellparameter (wie bei der konnektionistischen KI). Die Entwicklung der Computertechnik für den Bau riesiger künstlicher neuronaler Netze und die Verfügbarkeit gigantischer Datenmengen für deren Training über das Internet waren für diesen Umschwung erforderlich.

Eine wichtige praktische Eigenschaft des Perceptron, des Urmodells des konnektionistischen Computers, sollte sein, dass sein Speicher wesentlich kleiner zu sein habe als die möglichen Daten, die er zu klassifizieren hat: Die Komplexität der Umwelt wird so im System komprimiert abgebildet, in diesem Sinn ist dieses System also buchstäblich ein *Modell*.³

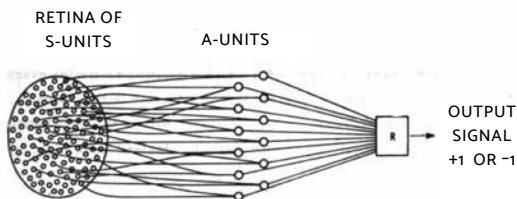
Rosenblatts Vorbild war das visuelle System im Zentralnervensystem von Lebewesen, und das erste Perceptron, das er in seinem Report beschreibt, verarbeitete ausschließlich optische Stimuli. Sein Input ist mit einer TV-Kamera, sein Output mit einem Drucker verbunden. Die Kamera nimmt das Bild eines ihr präsentierten Gegen-

standes auf, das Perceptron klassifiziert die TV-Bilder nach ihrer geometrischen Form, unterscheidet etwa Quadrate von Kreisen, woraufhin der Drucker den erkannten Formtyp meldet, indem er das betreffende Wort ausdruckt. Rosenblatt beschreibt nur das Training seines Apparats: »Es ist möglich, dem System beizubringen, zwei solcher verallgemeinerter Formen oder ›Wahrnehmungen‹ zu unterscheiden, indem für jede Form eine zufällige Stichprobe aus der Menge ihrer möglichen Transformationen [unterschiedlichen Ansichten der Form, M. W.] präsentiert wird, während gleichzeitig das System ›gezwungen‹ wird, mit Ausgabe 1 für Form 1 und Ausgabe 2 für Form 2 zu antworten.«⁴

Diese erste beschriebene Anwendung des Perceptrons hat zunächst viel mit der Simulation von Gehirnaktivitäten und wenig mit Sprache zu tun. Und so hat sich bis heute die neurophysiologische Metapher für das gesamte Feld der Künstlichen Neuronalen Netze festgesetzt, wie wir sehen werden: fälschlicherweise. Denn visuelle Inputs waren nur eine Sorte Daten, die in solchen elektromechanischen Geräten verarbeitet werden konnten. Aber eigentlich ist es völlig egal, was da zufällig am Input liegt, und so hat Rosenblatt gleichermaßen akustische oder elektromagnetische Umweltdaten in Betracht gezogen. Zudem ist sein Apparat allgemein genug konstruiert, um alles zu

verarbeiten, was als Eingangssignal kodiert werden kann.

Insgesamt besteht das Perceptron aus drei Ebenen: Die erste Ebene bildet der »sensorische« Input, etwa die Rasterpunkte einer TV-Kamera; die zweite, »assoziative« Ebene besteht aus einzelnen Elementen, die auf die Inputs reagieren, um sie schließlich an die dritte »Response«-Ebene weiterzuleiten, die das eindeutige Ergebnis der Aktivität des Perceptrons anhand einiger weniger Elemente anzeigt.⁵ Damit ergibt sich folgende Kette: Stimulus → Assoziation → Response.



In obigem Beispiel besteht die sensorische Ebene aus ungefähr 100 Elementen, die assoziative Ebene aus 12 Elementen, und die Response-Ebene aus einem Element, das dann eine binäre Entscheidung anzeigt: in diesem Fall Quadrat oder Kreis, im Schaubild durch »+1« oder »-1« kodiert. In dieser Einfachheit modelliert das Perceptron allerdings keinesfalls einen relevanten Teil eines Zentralnervensystems, denn dort sind

die Elemente (Neuronen), wie längst bekannt war, sehr dicht über kurze und über große Entfernungen hinweg kreuz und quer miteinander verknüpft, und es gibt dort auch keine Anzeigetafel, auf der immer nur ein Lämpchen leuchtet. Vielmehr ist Hirngewebe immens viel komplexer. Dafür übernimmt das Perceptron – bis zu seinen heutigen Nachfahren, den künstlichen neuronalen Netzen – den Aktivierungsmechanismus einer tatsächlichen Nervenzelle:⁷ Die Elemente der Assoziationsebene empfangen zunächst erregende oder dämpfende Signale der ersten sensorischen Ebene (den Signalen der verknüpften Kamerapunkte). Jedes assoziative Element summiert die Signalstärken auf, wobei dämpfende erregende Signale annullieren. Falls die Summe der Signalstärken einen bestimmten Schwellenwert überschreitet, sendet das Element seinerseits ein Signal weiter an die Response-Ebene, oder es bleibt inaktiv. So etwas macht das natürliche Neuron eben auch: Es *feuert*, sendet also ein Signal einer bestimmten Stärke zu einem oder vielen anderen Neuronen, wenn die eingehenden erregenden zusammen mit den dämpfenden Signalen seine spezifische Schwelle überschreiten, andernfalls bleibt es stumm, weshalb hier auch von einem Alles-oder-Nichts-Prinzip die Rede ist: »Die Aktivität des Neurons ist ein ›Alles-oder-Nichts‹-Prozess.«⁸

Ob ein Signal eine dämpfende oder erregende Wirkung auf das assoziative Element ausübt, wird in Rosenblatts Perceptron mit einem numerischen Parameter, einem sogenannten Gewicht, eingestellt, ebenso, ob das ausgesendete Signal des assoziativen Elements und künstlichen Neurons seinerseits dämpfend oder erregend sein möge. Nehmen wir uns die Abbildung oben vor Augen, sollten wir uns also jede Linie mit einem Parameter versehen vorstellen, der verstärkt, wenn er größer Null ist oder dämpft, falls er negativ ist. Die A-Elemente addieren das Hereinkommende und feuern, wenn ihr Schwellwert erreicht ist.

Was das Ganze dann tut – und das ist der Kern der Sache – hängt von den konkreten Gewichtungen der Signale der Eingangszuleitungen, dem Schwellwert des Neurons und den Werten der Signalstärken der Leitungen, die zum Response-System führen, ab.⁹ Erst dann, wenn diese Werte (Parameter) passen, tut das Perceptron, was es soll. Rosenblatt gibt dafür eine Rückkopplungsregel an, die noch das Vorbild für die heutigen Lernalgorithmen ist: Stimmt beim Training der Output, bleiben die Parameter unverändert, weichen sie hingegen vom Gewünschten ab, werden sie in die jeweilige gewünschte Richtung verschoben, also entweder abgesenkt oder erhöht, je nachdem, ob der Response null oder eins betragen sollte.¹⁰ Die Justage aller dieser Parameter bildet also das

Prinzip der Programmierung dieses seltsamen Netzcomputers, und auch darin gleicht er den Large Language Models von heute.

Es geht aber eigentlich nicht darum, Nervengewebe nachzubauen, vielmehr handelt es sich hier um etwas viel Allgemeineres: Nämlich darum, ein System aus einfachen Elementen zu bauen, das aus beliebigen Eingabedaten statistische Muster extrahiert und durch Setzen von sehr vielen Parametern zu gewünschten Ausgabedaten transformiert. Die Folgen gegenüber der herkömmlichen Computation sind kategorial: An die Stelle der Logik tritt die Statistik, an die der Regeln die Simulation, an die der Programmierung das Training. Während bei der ab den Fünfzigerjahren dominierenden, sogenannten symbolischen Computerprogrammierung noch jeder Schritt nachvollzogen werden kann, wandert bei neuronalen Netzen alles in die Parameter und ist damit einer direkten Analyse entzogen. Wir können dann nur noch beobachten, was es tut, und uns um eine zielgerichtete Anpassung der Parameter bemühen, es trainieren, was so viel heißt wie es zu zwingen, »mit Ausgabe 1 für Form 1 und Ausgabe 2 für Form 2 zu antworten«.¹¹

Anders als bei Rosenblatts Perceptron werden es bei den LLM nicht mehr Umweltdaten, sondern Textanfänge sein, auf die das System geeignet ansprechen soll, sodass die Maschine – durch

Rückkopplung trainiert – mit demjenigen Wort antwortet, das den Text fortführt. Das Raster-Abbild von Welt wird durch Text ersetzt und erzeugt Text. Alles ist Schrift, nicht vorgeformt durch eine vorgängige Grammatik, sondern zunächst ebenso kontingent wie das Bild einer TV-Kamera, deren Muster erst noch erkannt werden muss. Aus Schrift wird Schrift. Diese hermetische Schleife auf einem geschlossenen Textkorpus, diese Abschottung vom Realen diktiert auch alles, was noch in diesem Büchlein hier folgt.

Die Parameter eines Perceptrons können in Form einer Tabelle, einer Matrix, aufgeschrieben werden. Ein positiver Tabellenwert bedeutet die Verstärkung der jeweiligen Verknüpfung, ein negativer die Abschwächung. Vor dem Training sind die richtigen Matrixwerte noch unbekannt, und das Perceptron benimmt sich zufällig und erratisch, das Training über Rückkopplung justiert dann die Werte bis zu ihrer erwünschten Konfiguration. Bis heute tun auch die Large Language Models nichts anderes, als massenhaft Matrizen miteinander zu multiplizieren, bis sie funktionieren, wie sie sollen.

Rosenblatt war sich darüber bewusst, dass sein statistischer Ansatz Folgen haben würde: »Der Preis, den wir für die Verwendung statistischer Prinzipien bei der Gestaltung des Systems zahlen, ist die Wahrscheinlichkeit, dass wir in einem

bestimmten Fall eine falsche Antwort erhalten – d. h. eine falsche Antwort, die in der Natur des Systems liegt und nicht auf eine Fehlfunktion einer seiner Komponenten zurückzuführen ist.«¹² Schon hier also treffen wir auf eine der Plagen, die selbst noch die aktuell so erfolgreichen Systeme heimsuchen: Sie können völlig richtig funktionieren und treffen dennoch falsche Aussagen. Heute hat sich für das unsinnige vor sich hin Reden der Begriff der »Halluzination« etabliert.¹³ Der Begriff stammt von Stephen L. Thaler, der sich anschaute, welche Resultate kleine künstliche neuronale Netze produzieren, wenn man ihre Input-Leitungen durchschneidet: Er benutzte dafür die neurologische Metapher des »brain damage«¹⁴ und bezeichnete die Aktivität des Netzes, einfach irgendetwas zu melden, was gar nicht da war, als »hallucination«¹⁵. Er fand heraus, dass ein Netz, das teilweise vom Input abgeschnitten ist, dennoch den vollen Output produzierte, für den es eigentlich den abgetrennten Input benötigt hätte. Thaler führte seine Untersuchungen noch ganz in der Domäne der Neurophysiologie durch, um physiologische Beschädigungen von Hirngewebe zu erklären. Wir werden im nächsten Kapitel klären, dass eine linguistische Sichtweise, jedenfalls auf die heutigen LLM, besseren Aufschluss bietet.

Diese Publikation wurde unterstützt aus Mitteln
der DFG-Kolleg-Forschergruppe MECS Medien-
kulturen der Computersimulation, Leuphana
Universität Lüneburg • (KFOR 1927).



Erste Auflage Berlin 2026
Copyright © 2026
MSB Matthes & Seitz Berlin
Verlagsgesellschaft mbH
Großbeerenstraße 57A | 10965 Berlin, Deutschland
info@matthes-seitz-berlin.de

Alle Rechte vorbehalten, insbesondere
die Nutzung des Werks für Text- und
Data-Mining im Sinne von §44b UrhG.

Satz: Monika Grucza-Nápoles, Cartagena
Druck und Bindung: Art-Druk, Szczecin
Printed in Poland
Umschlaggestaltung nach einer Idee von
Pierre Faucheux
ISBN 978-3-7518-3060-7
www.matthes-seitz-berlin.de