



Modul 1

Grundlagen „Künstliche Intelligenz“

Lektion 2

Wie funktionieren Sprachmodelle am Beispiel von ChatGPT?

Überblick

Machine Learning, Deep Learning ...

Künstliche Intelligenz und die verschiedenen Teilbereiche

Künstliche Intelligenz

Was verstehen wir noch mal genau darunter?

Der Begriff „Künstliche Intelligenz“

Teilbereich der Informatik, der sich mit der Schaffung von Maschinen oder Programmen beschäftigt, die in der Lage sind, Aufgaben durchzuführen, die menschliche Intelligenz erfordern, wie etwa Lernen, Urteilen und Problemlösen.

Betrachtet man zunächst die theoretischen Grundlagen, so stehen hinter den KI-Systemen komplexe mathematische und statistische Theorien, darunter:

- Wahrscheinlichkeitstheorie und Statistik
- Optimierungstheorie
- Theoretische Modelle des Lernens
 - PAC-Lernen (Probably Approximately Correct Learning)
 - Bayes'schen Modell

Die Entwicklung und Anwendung von KI-Modellen und -Technologien folgen verschiedenen Nutzungsmodellen und Herangehensweisen, die sich über die Jahre hinweg entwickelt haben. Diese Modelle reichen von der symbolischen KI, die auf fest definierten Regeln basiert, bis hin zu den neuesten Durchbrüchen im Bereich des maschinellen Lernens und der tiefen neuronalen Netze, die es Maschinen ermöglichen, aus Daten zu lernen und sich anzupassen, ohne explizit programmiert zu werden.



Künstliche Intelligenz

Was verstehen wir noch mal genau darunter?

Teilgebiete der „Künstlichen Intelligenz“

Künstliche Intelligenz

- Systeme, die ihr Umfeld wahrnehmen, schlussfolgern, Entscheidungen treffen und entsprechend handeln

Machine Learning

- Algorithmen, die anhand von Eingabedaten ein Modell trainieren, das dann anhand neuer oder noch nie zuvor gesehener Daten nützliche Vorhersagen treffen kann. ML gibt dem Computer auch die Fähigkeit zu lernen, ohne explizit zu programmieren.
- Neuronale Netze (NN) sind vom menschlichen Gehirn inspiriert und ahmen die Art und Weise nach, wie sich biologische Neuronen Signale verarbeiten und senden. NN bestehen aus miteinander verbundenen künstlichen Neuronen, die in Schichten organisiert sind.

Deep Learning

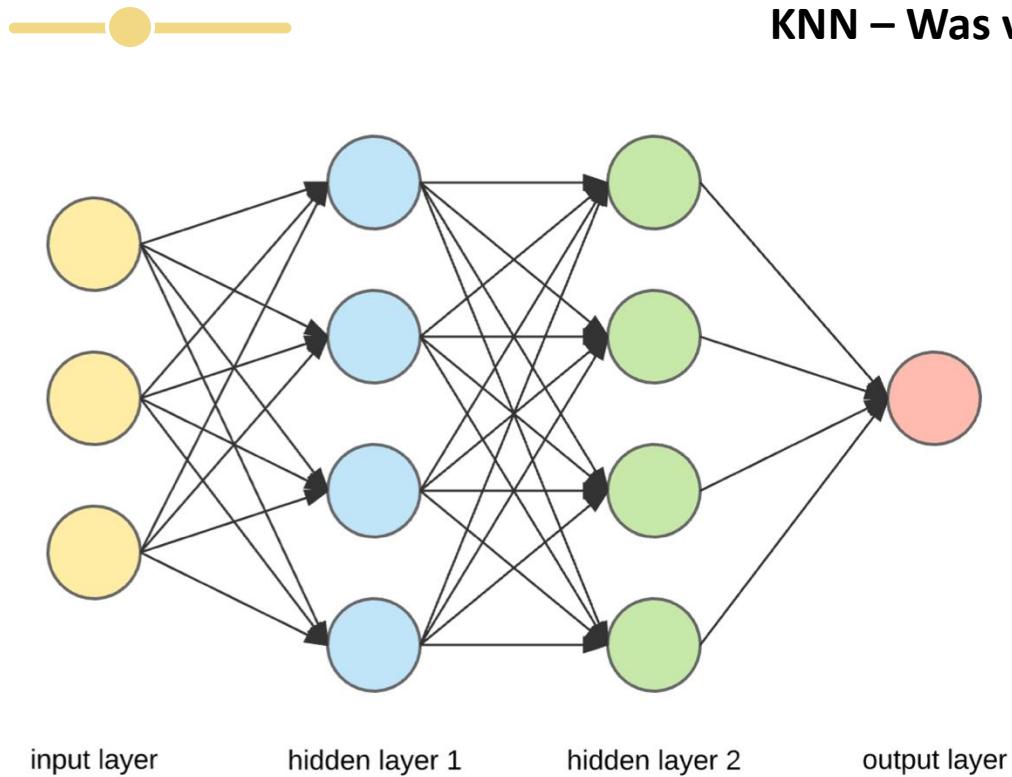
- Deep Learning (DL) bezieht sich auf jedes neuronale Netz mit mehreren verborgenen Schichten. Mit einer höheren Anzahl versteckter Schichten sind DL-Modelle gut geeignet, um komplexe reale Probleme anzugehen.

Generative KI

Erzeugung neuer Inhalte
(Text, Code, Bilder, Audio, Video, ...)

Künstliche Intelligenz

KNN – Künstliche Neuronale Netze



KNN – Was verstehen wir darunter?

Künstliche Neuronale Netzwerke (KNN)

KNNs sind die derzeit meist verwendete Technologie im Bereich der Künstlichen Intelligenz. KNN bestehen aus künstlichen Neuronen und sind von den neuronalen Netzwerken im Gehirn von Menschen und Tieren inspiriert.

KNN werden in einem „Learning Prozess – Maschinelles Lernen bzw. Deep Learning - mit großen Datenmengen trainiert und erlangen dadurch Fähigkeiten, die schwer wären, durch normale Programmierung herzustellen:

- das Erkennen von Objekten
- die Fähigkeit, Texte zu generieren, die Texten menschlichen Ursprungs ähneln
- die Fähigkeit, Bilder oder Videos zu generieren



Künstliche Intelligenz

LLM (Large Language Model) – Große Sprachmodelle

Diverse GIFs erstellen

Was sind LLM(s)?

The screenshot shows a conversation in ChatGPT 4.0. The user asks for an example of compound interest. ChatGPT responds by explaining what compound interest is and provides a detailed example of investing 1,000 EUR over 3 years at a 5% interest rate, showing how the balance grows to 1,050 EUR. It also includes a section on how to ask questions.

ChatGPT 4.0

Erkläre mir die Berechnung des Zinseszins anhand eines Beispiels mit einer Investition von 1.000 EUR in einfacher Sprache

Klar! Ich erkläre dir den Zinseszins ganz einfach mit einem Beispiel.

💡 Was ist Zinseszins?

Beim Zinseszins bekommst du Zinsen nicht nur auf dein Anfangskapital, sondern auch auf die Zinsen, die du in den Jahren davor bekommen hast.

📘 Beispiel:

Du investierst 1.000 EUR für 3 Jahre mit einem Zinssatz von 5% pro Jahr. Wir rechnen jedes Jahr neu, und die Zinsen bleiben auf dem Konto.

Jahr 1:

- Startkapital: 1.000 EUR
- Zinsen: 5% von 1.000 EUR = 50 EUR
- Neues Guthaben: 1.050 EUR

Stelle irgendeine Frage

+ ⚙ Tools

ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe [Cookie-Voreinstellungen](#).

Ein Large Language Model (LLM) oder großes Sprachmodell ist eine Version eines sehr intelligenten Computerprogramms, das darauf trainiert wurde, menschliche Sprache zu verstehen und zu erzeugen - wie ein extrem belesenes Gehirn, das Milliarden von Texten (Bücher, Webseiten ...) – gelesen hat.

"Large" (Groß): Das bezieht sich auf die enorme Menge an Daten, mit denen diese Modelle trainiert werden, und auf die riesige Anzahl an Parametern (quasi den "Verbindungen" im Modell), die es ihnen ermöglichen, komplexe Muster in der Sprache zu erkennen.

"Language" (Sprache): Ihr Hauptzweck ist die Verarbeitung und Generierung von Sprache. Sie können Texte schreiben, zusammenfassen, übersetzen oder Fragen beantworten.

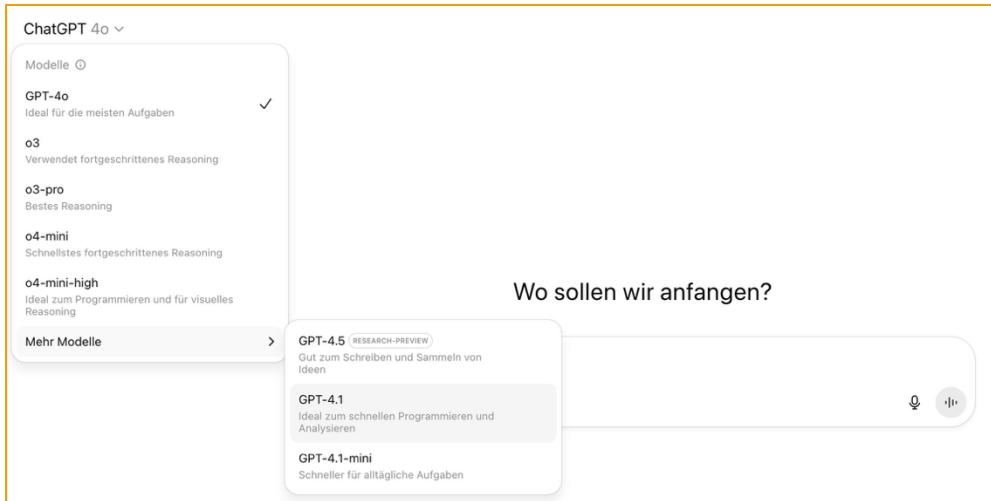
"Model" (Modell): Es ist ein mathematisches Modell, das Vorhersagen trifft. Wenn Sie ihm einen Satzanfang geben, versucht es vorherzusagen, welches Wort als Nächstes am wahrscheinlichsten kommt, basierend auf allem, was es gelernt hat.



Künstliche Intelligenz

GPT (Generative Pretrained Transformer) – GPT-4o (OpenAI)

Was steckt hinter dem Begriff GPT u.a. bei der Firma OpenAI für deren Modelle?



GPT steht für **Generative Pre-trained Transformer**. Das ist der spezifische Name, den das Unternehmen **OpenAI** für seine Familie von Large Language Models verwendet.

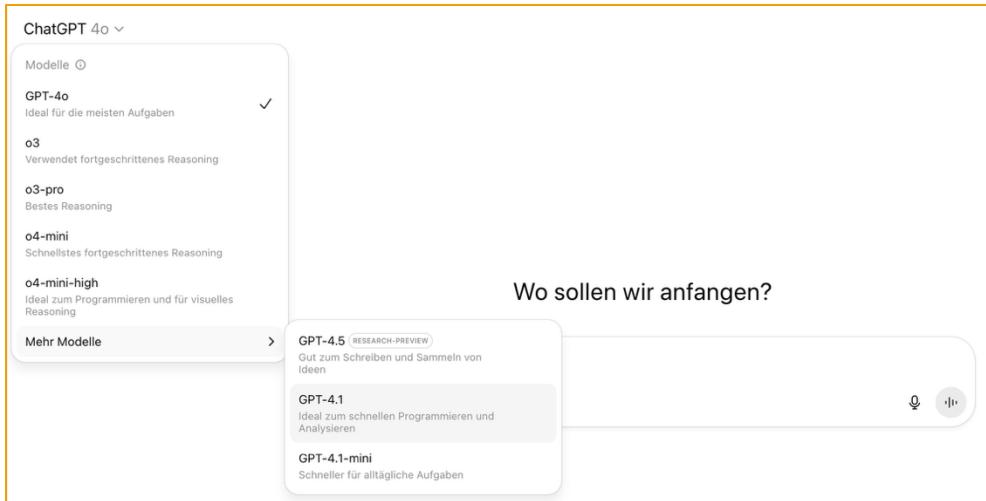
GPT ist eine Art von LLM: Stellen Sie sich vor, "LLM" ist wie die Kategorie "Auto". "GPT" wäre dann wie ein bestimmtes Modell oder eine Marke innerhalb dieser Kategorie, zum Beispiel "Mercedes C-Klasse". Es ist ein konkretes Beispiel für ein großes Sprachmodell.

"Generative": Das "G" in GPT ist entscheidend. Es bedeutet, dass diese Modelle darauf ausgelegt sind, neuen, originellen Text zu **generieren**. Sie können nicht nur Bestehendes verstehen, sondern auch kreativ neue Inhalte erstellen, sei es eine E-Mail, ein Gedicht oder eine Antwort auf eine Frage.

Künstliche Intelligenz

GPT (Generative Pretrained Transformer) – GPT-4o (OpenAI)

Was steckt hinter dem Begriff GPT u.a. bei der Firma OpenAI für deren Modelle?



"Pre-trained": Das "P" bedeutet, dass diese Modelle auf einer riesigen Menge von Textdaten **vortrainiert** wurden. Dieses Vortraining gibt ihnen ein breites Wissen über Sprache, Fakten und Zusammenhänge, bevor sie für spezifische Aufgaben feinabgestimmt werden.

"Transformer": Das "T" bezieht sich auf die spezifische neuronale Netzwerkarchitektur, die diesen Modellen zugrunde liegt. Diese Architektur ist besonders gut darin, Beziehungen zwischen Wörtern in langen Texten zu erkennen, was für das Verständnis von Kontext und Bedeutung entscheidend ist.

Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

1. Warum die Vorhersage so gut ist: Das Prinzip des „Nächsten Wortes“

The screenshot shows the ChatGPT 4.0 interface. At the top, there is a text input field with the instruction: "Erstelle ein Gedicht über die Entwicklung von LLMs - von den Anfängen der Theorie über Perceptron und Neuronale Netze bis hin zu GPT-3". Below this, the AI has generated a poem:

```
Von Neuronen, Denken und Maschine –  
Ein Gedicht zur Reise der LLMs  
  
Im Geiste einst, vor Jahren schon,  
da träumte man vom Denk-Karton:  
Maschinen, die verstehen, was wir sagen,  
die selbst Gedanken wagen, fragen.  
  
Ein Perceptron, so schlicht, so klein,  
begann als erster Licht zu sein.  
Ein Neuron – künstlich, doch inspiriert,  
vom Hirn, das uns zum Denken führt.  
Mit Zahlen, Vektoren, schlichtem Reiz  
kam erstes Lernen – stolz, wenn auch heiß.  
  
Doch bald man sah: zu starr, zu flach,  
die Grenzen kamen früh und schwach.  
So wuchs der Wunsch, man schichtete mehr,  
ein Netz entstand, tief, dicht und schwer.  
  
Backpropagation – der neue Schwung,  
ließ Netze lernen, frisch und jung.  
Sie sahen Bilder, sprachen Wort,  
↓  
| Stelle irgendeine Frage  
+ ⚙ Tools  
  
ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe Cookie-Voreinstellungen.
```

Ein LLM ist im Grunde ein Meister darin, das **nächste Wort (oder genauer: das nächste "Token", was auch Satzzeichen oder Wortteile sein können)** in einer Sequenz vorherzusagen. Es ist wie ein extrem guter Textergänzer auf Ihrem Smartphone, nur milliardenfach komplexer.

Warum ist es so gut?

Mustererkennung auf riesiger Skala: Das Modell hat so viele Texte gelesen, dass es unzählige Muster in der Sprache gelernt hat. Es weiß, welche Wörter oft zusammenstehen, welche Sätze typisch sind, welche Grammatikregeln gelten und sogar, welche Informationen oft aufeinander folgen.

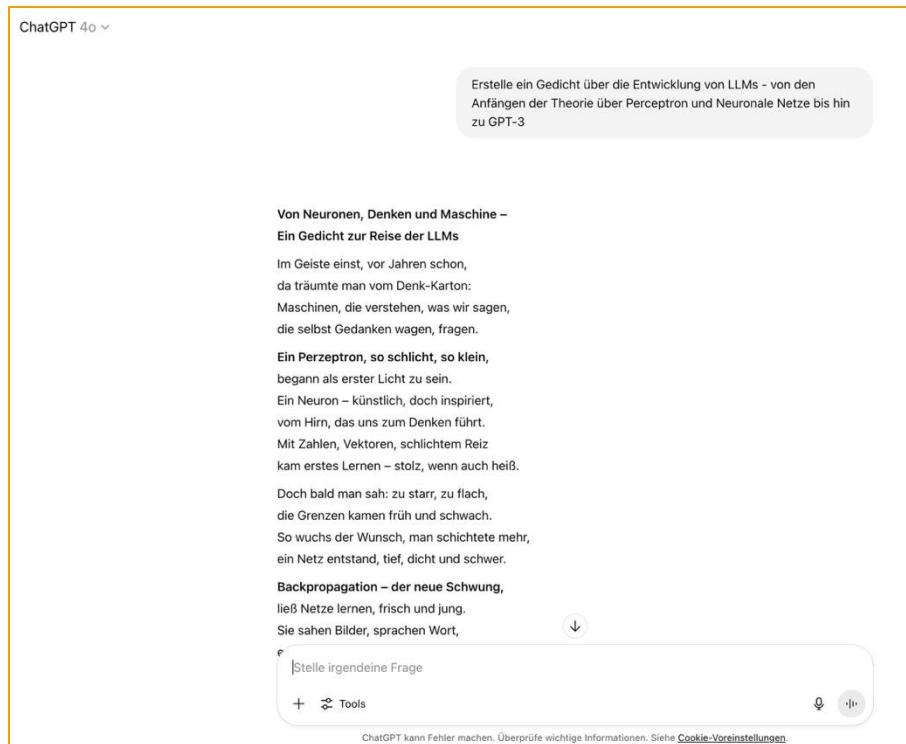
Statistische Wahrscheinlichkeit: Es berechnet für jedes mögliche nächste Wort, wie wahrscheinlich es ist, dass es an dieser Stelle im Satz oder Text erscheint, basierend auf dem bisherigen Kontext. Es wählt dann das wahrscheinlichste Wort aus.



Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

2. Wie funktioniert genau die Vorhersage des nächsten Wortes?



The screenshot shows the ChatGPT interface. At the top, there is a text input field with the instruction: "Erstelle ein Gedicht über die Entwicklung von LLMs - von den Anfängen der Theorie über Perceptron und Neuronale Netze bis hin zu GPT-3". Below this, the generated poem is displayed:

Von Neuronen, Denken und Maschine –
Ein Gedicht zur Reise der LLMs

Im Geiste einst, vor Jahren schon,
da träume man vom Denk-Karton:
Maschinen, die verstehen, was wir sagen,
die selbst Gedanken wagen, fragen.

Ein Perzeptron, so schlicht, so klein,
begann als erster Licht zu sein.
Ein Neuron – künstlich, doch inspiriert,
vom Hirn, das uns zum Denken führt.
Mit Zahlen, Vektoren, schlichtem Reiz
kam erstes Lernen – stolz, wenn auch heiß.

Doch bald man sah: zu starr, zu flach,
die Grenzen kamen früh und schwach.
So wuchs der Wunsch, man schichtete mehr,
ein Netz entstand, tief, dicht und schwer.

Backpropagation – der neue Schwung,
ließ Netze lernen, frisch und jung.
Sie sahen Bilder, sprachen Wort,

F Stelle irgendeine Frage

At the bottom of the interface, there is a footer note: "ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe [Cookie-Voreinstellungen](#)".

Stellen Sie sich vor, das Modell bekommt den Satzanfang: "Der Himmel ist..."

Kontext verstehen: Das Modell "liest" "Der Himmel ist". Es aktiviert sein Wissen über Himmel, Farben, Zustände usw.

Mögliche Fortsetzungen: Es überlegt, welche Wörter hier statistisch am wahrscheinlichsten folgen könnten: "blau", "grau", "klar", "bewölkt", "rot" (bei Sonnenuntergang), "hoch", "weit" usw.

Wahrscheinlichkeiten zuweisen: Basierend auf dem gelernten Wissen und dem Kontext weist es jedem dieser Wörter eine Wahrscheinlichkeit zu. Vielleicht ist "blau" zu 80% wahrscheinlich, "grau" zu 10%, "bewölkt" zu 5% und so weiter.

Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

2. Wie funktioniert genau die Vorhersage des nächsten Wortes?

ChatGPT 4.0 ✓

Erstelle ein Gedicht über die Entwicklung von LLMs - von den Anfängen der Theorie über Perceptron und Neuronale Netze bis hin zu GPT-3

**Von Neuronen, Denken und Maschine –
Ein Gedicht zur Reise der LLMs**

Im Geiste einst, vor Jahren schon,
da träumte man vom Denk-Karton:
Maschinen, die verstehen, was wir sagen,
die selbst Gedanken wagen, fragen.

Ein Perzeptron, so schlcht, so klein,
begann als erster Licht zu sein.
Ein Neuron – künstlich, doch inspiriert,
vom Hirn, das uns zum Denken führt.
Mit Zahlen, Vektoren, schlichtem Reiz
kam erstes Lernen – stolz, wenn auch heiß.

Doch bald man sah: zu starr, zu flach,
die Grenzen kamen früh und schwach.
So wuchs der Wunsch, man schichtete mehr,
ein Netz entstand, tief, dicht und schwer.

Backpropagation – der neue Schwung,
ließ Netze lernen, frisch und jung.
Sie sahen Bilder, sprachen Wort,

↓

„
+ Tools

ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe [Cookie-Voreinstellungen](#).

Auswahl treffen: Es wählt das Wort mit der höchsten Wahrscheinlichkeit (oder manchmal ein leicht weniger wahrscheinliches, um den Text natürlicher und weniger vorhersehbar zu machen). In diesem Fall wäre es wahrscheinlich "blau".

Nächster Schritt: Jetzt ist der Satz "Der Himmel ist blau". Das Modell nimmt diesen neuen, längeren Kontext und wiederholt den Prozess, um das nächste Wort vorherzusagen. So entsteht Schritt für Schritt ein ganzer Text.



Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

3. Was hat das mit den Trainingsdaten zu tun?

The screenshot shows the ChatGPT interface. At the top, there is a text input field with the instruction: "Erstelle ein Gedicht über die Entwicklung von LLMs - von den Anfängen der Theorie über Perceptron und Neuronale Netze bis hin zu GPT-3". Below this, the generated poem is displayed:

Von Neuronen, Denken und Maschine –
Ein Gedicht zur Reise der LLMs
Im Geiste einst, vor Jahren schon,
da träume man vom Denk-Karton:
Maschinen, die verstehen, was wir sagen,
die selbst Gedanken wagen, fragen.
Ein Perceptron, so schlicht, so klein,
begann als erster Licht zu sein.
Ein Neuron – künstlich, doch inspiriert,
vom Hirn, das uns zum Denken führt.
Mit Zahlen, Vektoren, schlichtem Reiz
kam erstes Lernen – stolz, wenn auch heiß.
Doch bald man sah: zu starr, zu flach,
die Grenzen kamen früh und schwach.
So wuchs der Wunsch, man schichtete mehr,
ein Netz entstand, tief, dicht und schwer.
Backpropagation – der neue Schwung,
ließ Netze lernen, frisch und jung.
Sie sahen Bilder, sprachen Wort,
↓

At the bottom of the interface, there is a footer note: "ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe [Cookie-Voreinstellungen](#)".

Die Trainingsdaten sind der absolute Schlüssel!

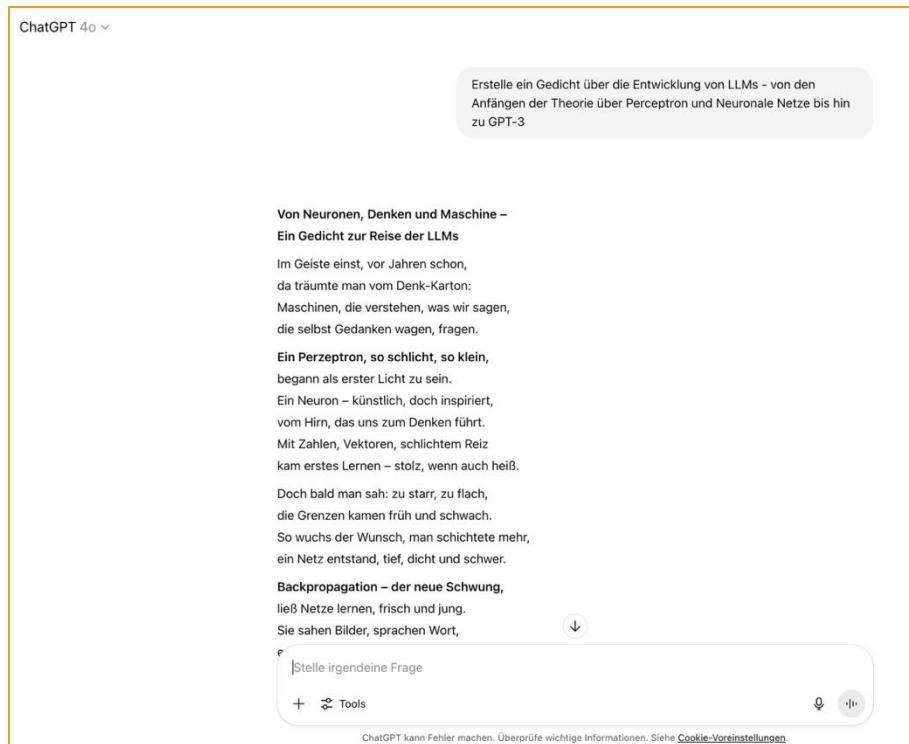
Die "Bibliothek" des Modells: Stellen Sie sich vor, das Modell ist ein Schüler, und die Trainingsdaten sind seine gesamte Bibliothek. Diese Bibliothek enthält Milliarden von Büchern, Artikeln, Webseiten, Konversationen – alles, was Menschen jemals geschrieben haben und öffentlich zugänglich ist.

Lernen durch Beobachtung: Das Modell "liest" diese riesige Bibliothek und lernt dabei, wie Sprache funktioniert. Es sieht zum Beispiel, dass nach "Der Himmel ist" sehr oft "blau" kommt, aber selten "Tisch". Es lernt Grammatik, Fakten, Stil und sogar Nuancen, indem es einfach nur beobachtet, wie Wörter und Sätze in realen Texten verwendet werden.

Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

3. Was hat das mit den Trainingsdaten zu tun?



The screenshot shows the ChatGPT interface. At the top, there is a text input field with the instruction: "Erstelle ein Gedicht über die Entwicklung von LLMs - von den Anfängen der Theorie über Perceptron und Neuronale Netze bis hin zu GPT-3". Below this, the AI has generated a poem:

Von Neuronen, Denken und Maschine –
Ein Gedicht zur Reise der LLMs

Im Geiste einst, vor Jahren schon,
da träume man vom Denk-Karton:
Maschinen, die verstehen, was wir sagen,
die selbst Gedanken wagen, fragen.

Ein Perzeptron, so schlicht, so klein,
begann als erster Licht zu sein.
Ein Neuron – künstlich, doch inspiriert,
vom Hirn, das uns zum Denken führt.
Mit Zahlen, Vektoren, schlichtem Reiz
kam erstes Lernen – stolz, wenn auch heiß.

Doch bald man sah: zu starr, zu flach,
die Grenzen kamen früh und schwach.
So wuchs der Wunsch, man schichtete mehr,
ein Netz entstand, tief, dicht und schwer.

Backpropagation – der neue Schwung,
ließ Netze lernen, frisch und jung.
Sie sahen Bilder, sprachen Wort,
↓
F Stelle irgendeine Frage
+ Tools

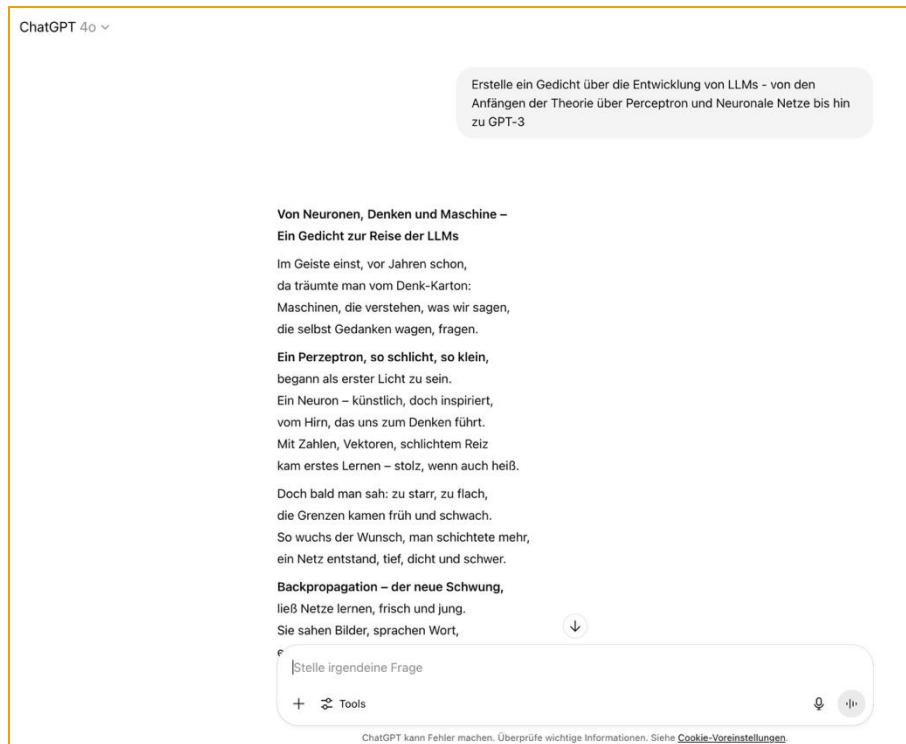
At the bottom of the interface, there is a note: "ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe [Cookie-Voreinstellungen](#)".

Qualität und Quantität: Je größer und vielfältiger die Trainingsdaten sind, desto besser kann das Modell Muster erkennen und desto genauer werden seine Vorhersagen. Wenn es nur Kinderbücher gelesen hätte, könnte es keine komplexen Finanzberichte verstehen. Wenn es nur Fachtexte gelesen hätte, könnte es keine kreativen Geschichten schreiben. Die Breite der Daten macht es so vielseitig.

Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

4. Was hat das mit der Art des Trainings zu tun?



The screenshot shows the ChatGPT interface. At the top, there is a text input field with the instruction: "Erstelle ein Gedicht über die Entwicklung von LLMs - von den Anfängen der Theorie über Perceptron und Neuronale Netze bis hin zu GPT-3". Below this, the AI has generated a poem:

```
Von Neuronen, Denken und Maschine –  
Ein Gedicht zur Reise der LLMs  
  
Im Geiste einst, vor Jahren schon,  
da träume man vom Denk-Karton:  
Maschinen, die verstehen, was wir sagen,  
die selbst Gedanken wagen, fragen.  
  
Ein Perzeptron, so schlicht, so klein,  
begann als erster Licht zu sein.  
Ein Neuron – künstlich, doch inspiriert,  
vom Hirn, das uns zum Denken führt.  
Mit Zahlen, Vektoren, schlichtem Reiz  
kam erstes Lernen – stolz, wenn auch heiß.  
  
Doch bald man sah: zu starr, zu flach,  
die Grenzen kamen früh und schwach.  
So wuchs der Wunsch, man schichtete mehr,  
ein Netz entstand, tief, dicht und schwer.  
  
Backpropagation – der neue Schwung,  
ließ Netze lernen, frisch und jung.  
Sie sahen Bilder, sprachen Wort,  
↓  
F Stelle irgendeine Frage  
+ ⚙ Tools  
ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Siehe Cookie-Voreinstellungen.
```

1. Vortraining (Pre-training): Das große "Lesen"

Wie es funktioniert: In dieser Phase wird das Modell mit den riesigen Datenmengen "gefüttert". Die Hauptaufgabe des Modells ist es, die fehlenden Wörter in Sätzen zu erraten. Man gibt ihm zum Beispiel den Satz "Der Himmel ist [LÜCKE]" und es muss "blau" erraten. Oder man gibt ihm einen Text und es muss das nächste Wort vorhersagen.

Das Ziel: Durch dieses ständige "Lückentext-Raten" lernt das Modell die statistischen Beziehungen zwischen Wörtern und Sätzen. Es lernt, wie Sprache strukturiert ist, welche Wörter zusammenpassen und welche Bedeutungen sie haben. Es ist ein unüberwachtes Lernen, da keine menschliche Korrektur nötig ist – das Modell vergleicht seine Vorhersage einfach mit dem tatsächlichen Wort im Originaltext.

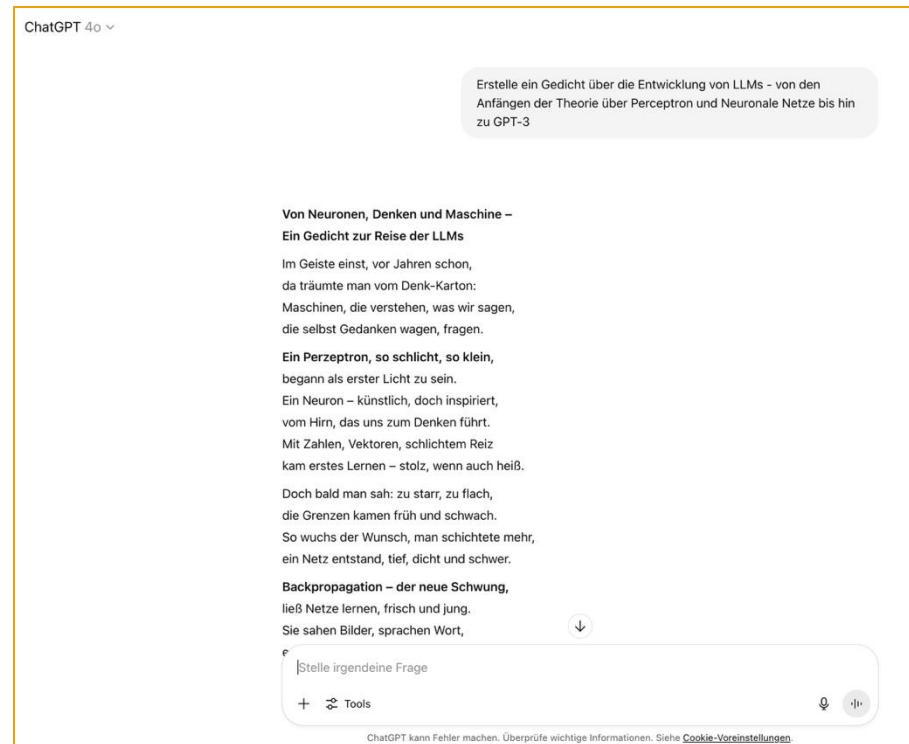
Ergebnis: Nach diesem Vortraining hat das Modell ein sehr breites, aber noch etwas rohes Verständnis von Sprache und Wissen.



Künstliche Intelligenz

Warum sind diese KI's und Sprachmodelle so faszinierend „gut“?

4. Was hat das mit der Art des Trainings zu tun?



2. Feinabstimmung (Fine-tuning): Die "Feinjustierung" durch Menschen

Wie es funktioniert: Nach dem Vortraining wird das Modell oft noch einmal mit kleineren, spezifischeren Datensätzen trainiert, die von Menschen bewertet wurden. Hier kommen menschliche Trainer ins Spiel, die dem Modell beibringen, was eine "gute" oder "hilfreiche" Antwort ist. Wenn das Modell zum Beispiel eine unhöfliche oder falsche Antwort gibt, wird es "bestraft", und wenn es eine gute Antwort gibt, wird es "belohnt".

Das Ziel: Diese Phase macht das Modell nützlicher und sicherer für den Endnutzer. Es lernt, Anweisungen besser zu befolgen, weniger "Unsinn" zu reden, sich höflicher auszudrücken und relevantere Antworten zu geben. Hier wird es zum Beispiel auch darauf trainiert, Fragen zu beantworten, Zusammenfassungen zu erstellen oder Übersetzungen anzufertigen, anstatt nur das nächste Wort zu erraten.

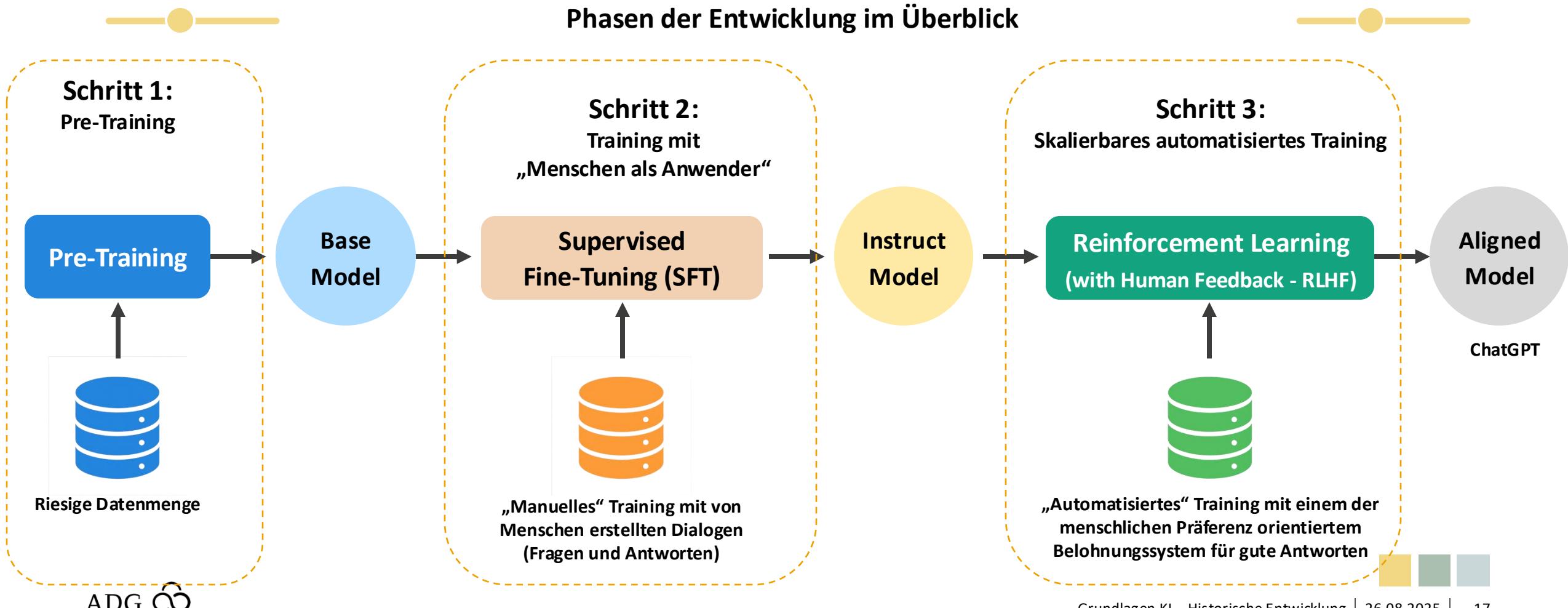
LLMs

ChatGPT – wie schauen hinter die „Kulissen“

Deep Dive ChatGPT

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



LLMs

ChatGPT – wie schauen hinter die „Kulissen“

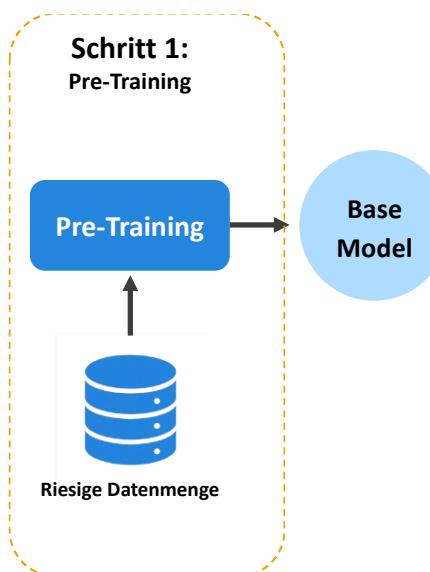
Pre-Training Phase: Datensätze für das Training, Texte und Wörter und Tokens ...

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen

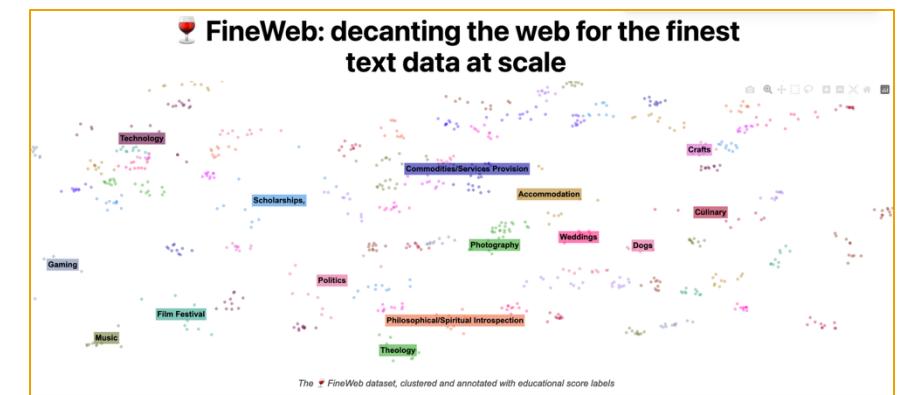


Pre-Training Datensätze:

Die Leistungsfähigkeit eines großen Sprachmodells (LLM) hängt stark von der Qualität und Größe seines Pretraining-Datensatzes ab.

Die Pretraining-Datensätze für hochmoderne LLMs wie z.B. ChatGPT sind nicht öffentlich verfügbar, aber es gibt einige Annahmen über die Herkunft. Ein typisches Beispiel für die Quelle dieser Datensätze ist die Plattform HuggingFace und dort der Fine-Web Datensatz.

Ziel für die Entwicklung von LLMs ist es in der Pre-Training Phase eine riesige Menge an qualitativ hochwertigen Textdaten zu erhalten. Eine Quelle dafür sind natürlich das Internet und internetbasierende Dienste wie Newsforen, Soziale Medien, Plattformen wie Reddit, digitale Bibliotheken usw.



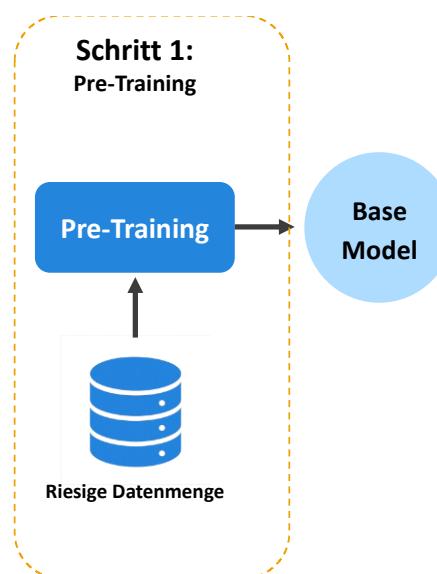
<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



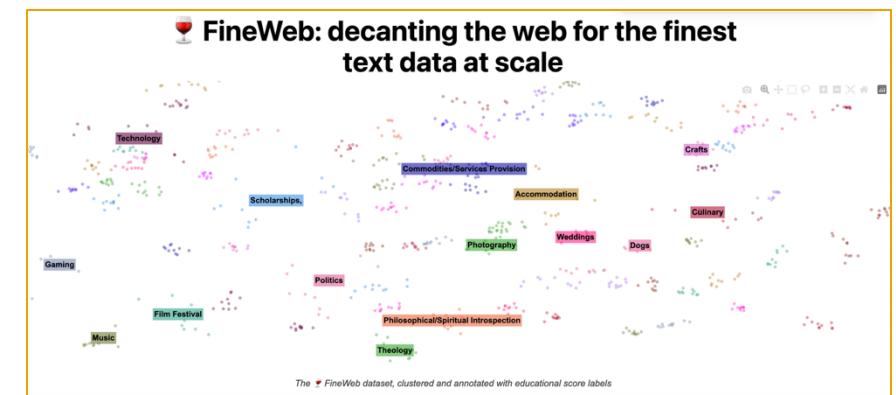
Schritt 1: Pre-Training mit riesigen Datenmengen



Wichtige Merkmale für die Datensätze

- Große Menge an Textdaten
- Qualitativ hochwertige Textdaten
- Diversität der Themen der hochwertigen Textdaten
- Der nebenstehende verlinkte Datensatz von FineWeb basiert auf „Rohdaten“ aus dem Internet, der dann in mehreren Schritten aufbereitet wurde und schließlich zu einer 44 Terrabyte großen Datei mit Textdaten als Ergebnis zur Verfügung gestellt wird.

=> Wir schauen uns diese Einzelschritte kurz an:



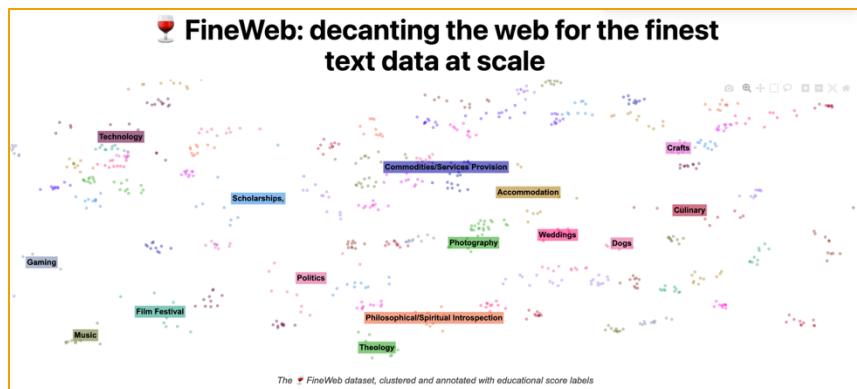
<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen



<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

CommonCrawl: Daten aus dem Internet recherchieren, filtern, ... aufbereiten:

Um den FineWeb Datensatz zu erstellen, wird CommonCrawl (CC) als Ausgangspunkt verwendet.

Die gemeinnützige Organisation Common Crawl durchsucht seit 2007 das Web und veröffentlicht in der Regel alle ein bis zwei Monate einen neuen Crawl, der 200 bis 400 TiB (TiB = Tebibyte ist etwas **größer** als TB = Terrabyte – ungefähr 1,1 TB = 1 TiB.) an Textinhalten umfasst, die durch automatisches Web-Crawling gewonnen wurden.

Als Beispiel enthält der neueste CC-Crawl (April 2024) 2,7 Milliarden Webseiten mit insgesamt 386 TiB unkomprimiertem HTML-Textinhalt.

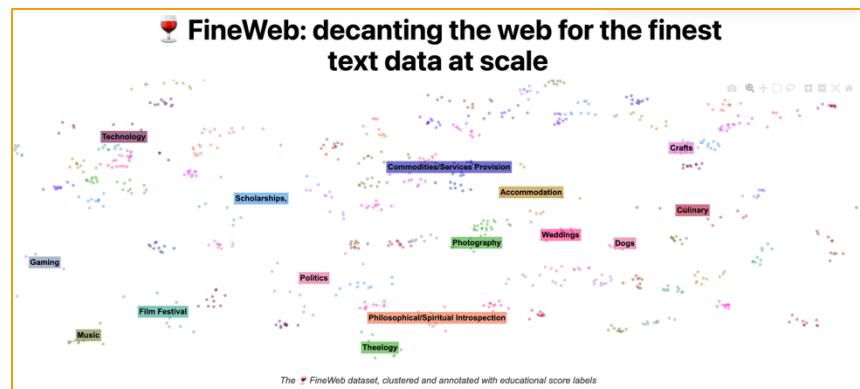


Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen



<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

CommonCrawl: Filtern von Internetdaten

Das Filtern ist ein wichtiger Bestandteil des Kurationsprozesses. Es besteht darin, Teile der Daten zu entfernen (sei es Wörter, Zeilen oder sogar ganze Dokumente), die die Leistung des Modells verringern und daher in dem evaluationsgetriebenen Datensatz-Erstellungsprozess als „niedrigere Qualität“ eingestuft werden.

Als Grundlage für das Filtering wird u.a.:

- URL-Filterung mithilfe einer Blockliste angewendet, um Inhalte für Erwachsene zu entfernen
- Ein Sprachklassifizierer eingesetzt, um nur englische Texte mit einem Score $\geq 0,65$ zu behalten



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

The screenshot shows the Hugging Face platform interface. At the top, there's a search bar and navigation links for Models, Datasets, Spaces, Community, Docs, Enterprise, and Pricing. Below that, it shows a dataset named "HuggingFaceFW/fineweb". The dataset has 2.2k likes, 572 followers, and 10B - 100B tokens. It was last updated on Nov 3, 2023. The main section is titled "Dataset card" and includes a "Dataset Preview" table. The table has columns for text, id, dump, and ux1. It lists several rows of data, each with a URL and a "View" button. To the right of the preview, there's a summary: "Downloads last month: 209,547", "Homepage and Repository: huggingface.co", "Point of Contact: please create a discussion on the Community tab", "License: Open Data Commons Attribution License (ODC-By) v1.0", "Size of downloaded dataset files: 51.3 TB", "Size of the auto-converted Parquet files: 51.3 TB", and "Number of rows: 24,989,712,068". Below the preview, there's a section for "Models trained or fine-tuned on HuggingFaceFW/fineweb".

<https://huggingface.co/datasets/HuggingFaceFW/fineweb>

FineWeb Datensatz:

Über den Link kann man auf die Detailseite des aktuellen FineWeb Datensatzes auf der Plattform HuggingFace gehen und dort den Datensatz auch herunterladen, um diesen dann z.B. für das Pre-Training eines eigenen "ChatGPTs" zu verwenden.

Man kann hier auch durch die Millionen von Zeilen der Datensätze scrollen und klicken, um sich die Inhalte anzuschauen, die dann in das Pre-Training einfließen werden.

OpenAI ChatGPT, Anthropic Claude, ...:

Die Hersteller veröffentlichen nicht die Datensätze, die für das Training der Modelle GPT-4o, O3, ... bzw. Claude 4.0 ... eingesetzt wurden, aber es werden natürlich viele Daten aus dem Internet verwendet und weitere digitale Datenbanken und Datenquellen.

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

How AP reported in all formats from tornado-stricken regions March 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take.
He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.
Ray was among the first journalists to arrive and he confirmed those reports -- in all formats. He shot powerful video, put victims on the phone with AP Radio and played back sound to an editor who transcribed the interviews and put the material on text wires. He then walked around the devastation with the Central Regional Desk on the line, talking to victims with the phone held so close that editors could transcribe his interviews in real time.
Ray also made a dramatic image of a young girl who found a man's prosthetic leg in the rubble, propped it up next to her destroyed home and spray-painted an impromptu sign: "Found leg. Seriously."
The following day, he was back on the road and headed for Georgia and a Super Tuesday date with Newt Gingrich's campaign. The drive would take him through a stretch of the South that forecasters expected would suffer another wave of tornadoes. To prevent running into THAT storm, Ray used his iPhone to monitor Doppler radar, zooming in on extreme cells and using Google maps to direct himself to safe routes. And then the journalist took over again.
"When weather like that occurs, a reporter must seize the opportunity to get the news out and allow people to see, hear and read the power of nature so that they can take proper shelter," Ray says.
So Ray now started to use his phone to follow the storms. He attached a small GoPro camera to his steering wheel in case a tornado dropped down in front of the car somewhere, and took video of heavy rain and hail with his iPhone. Soon, he spotted a tornado and the chase began on the scene of the storm's aftermath. Again, the tornadoes had struck a wickedly violent storm that made Tennessee, where our customers were.
"CLEVELAND, Tenn. (AP) — Fierce winds and heavy rain lashed the city Saturday, causing power outages and flooding. People were hospitalized Friday night. The byline? Robert Ray.
For being adept with technology, formats reporting that put the Associated Press in the States' prize.
© 2013 The Associated Press. All rights reserved. This material may not be published, broadcast, rewritten or redistributed without permission.
Have two little yellow, nine-volt batteries on top of your kidneys? Someone It seems as though your adrenal glands stand like statues in front of Brown, slippery beach that is you. However, if anything startling should happen, you'll be ready." - Robert Ray

Textdaten: Textdaten zum Training in ein Neuronales Netz „einfügen“.

Wenn man dann diesen kompletten Datensatz als Text zum Training verwenden möchte, hat man eine riesige Textdatei mit den Wörtern nacheinander, wie die nebenstehende Textdatei.

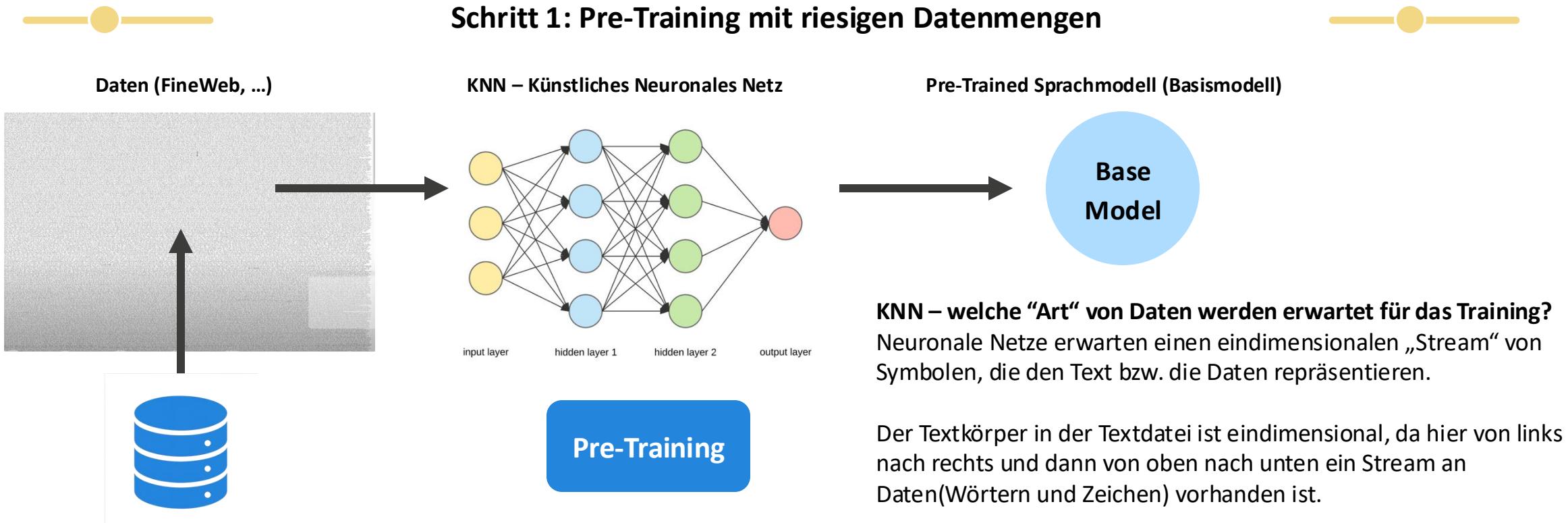
Hier wurden die Datensätze (Millionen von Zeilen) aus der FineWeb Datei in eine Textdatei überführt (nicht alle, nur ein Teil – zur Demonstration der nächsten Schritte).

Wenn man dann ein wenig mit der Lupe auf diese Textdatei „rauszoomed“, dann sieht man einen Korpus mit den Daten und wir wollen jetzt im nächsten Schritt, dass das zu entwickelnde Sprachmodell diese Daten und den Fluss der Texte (der Wörter nacheinander) „lernt“ und dabei natürlich auch „Muster“ erkennt.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

How AP reported in all formats from tornado-stricken regions March 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take.
He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.
Ray was among the first journalists to arrive and he confirmed those reports -- in all formats. He shot powerful video, put victims on the phone with AP Radio and played back sound to an editor who transcribed the interviews and put the material on text wires. He then walked around the devastation with the Central Regional Desk on the line, talking to victims with the phone held so close that editors could transcribe his interviews in real time.
Ray also made a dramatic image of a young girl who found a man's prosthetic leg in the rubble, propped it up next to her destroyed home and spray-painted an impromptu sign: "Found leg. Seriously."
The following day, he was back on the road and headed for Georgia and a Super Tuesday date with Newt Gingrich's campaign. The drive would take him through a stretch of the South that forecasters expected would suffer another wave of tornadoes. To prevent running into THAT storm, Ray used his iPhone to monitor Doppler radar, zooming in on extreme cells and using Google maps to direct himself to safe routes. And then the journalist took over again.
"When weather like that occurs, a reporter must seize the opportunity to get the news out and allow people to see, hear and read the power of nature so that they can take proper shelter," Ray says.
So Ray now started to use his phone to follow the storms. He attached a small GoPro camera to his steering wheel in case a tornado dropped down in front of the car somewhere, and took video of heavy rain and hail with his iPhone. Soon, he spotted a tornado and the chase was on. He followed an unmarked emergency vehicle to Cleveland, Tenn., where he was first on the scene of the storm's aftermath.
Again, the tornadoes had struck in locations that were hours from the nearest AP bureau. Damage and debris, as well as a wickedly violent storm that made travel dangerous, slowed our efforts to get to the news. That wasn't a problem in Tennessee, where our customers were well served by an all-formats report that included this text story.
"CLEVELAND, Tenn. (AP) - Fierce wind, hail and rain lashed Tennessee for the second time in three days, and at least 15 people were hospitalized Friday in the Chattanooga area."
The byline? Robert Ray.
For being adept with technology, chasing after news as it literally dropped from the sky and setting a standard for all-formats reporting that put the AP ahead on the most competitive news story of the day, Ray wins this week's \$300 Best of the States prize.
© 2013 The Associated Press. All rights reserved. Terms and conditions apply. See AP.org for details. Did you know you have two little yellow, nine-volt-battery-sized adrenal glands in your body, just chilling out, ~~maxin'~~, relaxin' all cool on top of your kidneys? Someone told me this and I checked it out. Turns out it's true.
It seems as though your adrenal glands are kind of like those British Royal Guards with the big, black fuzzy hats who stand like statues in front of Buckingham Palace. They just stand there quietly, not doing much really, just enjoying the brown, slippery beach that is your kidneys.
However, if anything startling should happen that requires your attention - like say you're about to give a speech at a

Text (Wörter, Zeichen, Sonderzeichen, Zahlen)

Herausforderung von normalem Text für KI:

Für die Nutzung dieser Daten in einem KNN – Künstlichem Neuronalem Netz – wird nach einem ausgewogenen Verhältnis von Zeichen, Symbolen, Silben als Repräsentation dieser Datenmenge an Text gefordert und einer optimalen Länge. Der aktuelle Datensatz von FineWeb z.B. ist in einem KNN so nicht verwendbar, da die mathematische Abbildung des Textes zu komplex wäre.

Daher benötigt es jetzt eine „Repräsentation“ der Texte (Wörter, Silben) in einem anderen – besser in der Mathematik nutzbarem – Format.

Dies kann man mit Unicode bzw. UTF-8 als Codierungsmethoden für Zeichen in allen Sprachen dieser Welt erreichen.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen



Transformation des Textes in UTF-8 Codierung

Unicode:

Unicode ist ein Zeichenkodierungsstandard, der versucht, jedem Zeichen jeder Sprache der Welt eine eindeutige Nummer (einen Code-Punkt) zuzuweisen.

UTF-8:

UTF-8 ist eine Möglichkeit, diese Unicode-Zeichen in einer Form zu kodieren, die von Computern verarbeitet werden kann. UTF steht für "Unicode Transformation Format".

Variable Länge:

Im Gegensatz zu Kodierungen mit fester Länge, bei denen jedes Zeichen immer die gleiche Anzahl von Bytes belegt, verwendet UTF-8 eine variable Länge. Zeichen, die im ASCII-Zeichensatz (die ersten 128 Zeichen) enthalten sind, werden mit einem Byte kodiert, während andere Zeichen, wie z.B. Umlaute oder Zeichen aus anderen Schriftsystemen, mehr Bytes benötigen.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen



Transformation des Textes in UTF-8 Codierung

(8 Bits entsprechen einem Zeichen, bei Sonderzeichen und speziellen Sprachen werden bis zu 32 Bits verwendet – also 4 Bytes)

Immer noch eine Herausforderung für KNNs:

Die Konvertierung in Bits und Bytes ist immer noch eine Herausforderung für KNNs. Jetzt werden die Texte über eine Sequenz von „0“ und „1“ repräsentiert. Das sind tatsächlich nur 2 Symbole und damit entsteht eine jetzt deutlich längere Kette als Repräsentanz des Textes als vorher – das ist ungünstig und muss weiter optimiert werden.

Buchstabe y: 01111001 UTF-8 Hexadezimal: 79

Daher kann man im nächsten Schritt von der Darstellung der einzelnen Zeichen in Bits die Gruppierung von Bits in ein Bytes überführen und dann die Darstellung von Bytes als Hexadezimal Zahlen nutzen. Jedes Byte kann als zwei hexadezimale Ziffern ausgedrückt werden. Zum Beispiel kann das ASCII-Zeichen 'A' in UTF-8 als 41 (hexadezimal) dargestellt werden, was einem einzelnen Byte entspricht.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen



54	68	48	61	77	20	41	50	20	72	65	70	61	72	74	65	64	20	69	6e	20	61	6c	6c	20	66	6f	72	6d	61	74	73	20	66	72	6f	6d	20	74	6f	72	6e	61						
64	61	6d	2d	73	74	69	63	6b	65	6e	20	72	65	67	69	61	6e	73	4d	61	72	63	68	20	38	2c	20	32	30	31	32	0a	57	68	65	6e	20	74	68	65	20	66						
69	72	73	74	20	73	65	72	69	6f	75	73	20	62	75	74	65	64	20	69	6e	20	61	6c	6c	20	66	6f	72	6d	61	74	73	20	66	72	30	32	20	62	6c	65	77						
20	74	65	68	48	6f	77	20	41	50	20	72	65	70	6f	72	74	65	64	20	69	6e	20	61	6c	6c	20	66	6f	72	6d	61	74	73	20	66	72	6f	61	74	68	65	20	66					
65	20	64	6f	2d	73	74	69	63	6b	65	6e	20	72	65	67	69	61	6e	73	4d	61	72	63	68	20	38	2c	20	32	31	32	0a	57	68	65	6e	20	74	68	65	20	66						
66	72	69	72	73	20	73	65	72	69	6f	75	73	20	62	6f	75	74	20	61	6e	20	66	6f	72	6d	61	74	66	73	20	66	72	30	31	32	20	62	6c	65	77								
66	72	69	72	73	20	73	65	72	69	6f	75	73	20	62	6f	75	74	20	61	6e	20	66	6f	72	6d	61	74	66	73	20	66	72	30	31	32	20	62	6c	65	77								
61	6c	6c	6c	6c	20	74	68	72	61	75	67	68	20	6d	69	64	64	6c	65	20	41	6d	65	72	69	63	61	20	69	6e	20	74	68	65	20	61	66	20	74	68	65	20	66					
70	70	66	67	62	6f	6d	20	61	6e	79	20	41	50	20	62	75	72	65	61	75	2e	20	4f	75	72	20	63	6e	73	65	73	74	20	66	69	64	65	6e	20	66	72	6f	61	75	70			
65	72	26	61	6c	69	73	74	20	77	61	73	20	43	68	69	63	61	67	61	2d	62	61	73	65	4d	20	52	61	72	74	20	52	61	79	2e	20	77	68	61	20	64	72	6f	61	75	63		
69	74	70	65	64	20	68	69	73	20	70	6c	61	6e	73	20	74	61	76	65	2c	20	47	65	6f	72	67	69	61	20	66	61	72	70	53	75	70	6f	72	61	74	66	20	63					
20	74	66	75	72	20	54	75	65	73	64	61	79	2c	20	62	6f	6b	65	64	20	73	65	76	65	72	61	6c	6c	20	66	6e	69	67	68	74	73	20	74	6f	70	65	63						
68	20	69	74	69	73	20	73	65	73	64	70	74	20	61	74	65	62	70	73	4c	20	69	6e	20	61	6e	61	20	68	65	61	64	65	64	20	66	68	61	20	70								
6c	61	20	74	68	65	20	61	69	72	70	6f	72	74	2e	20	48	65	82	80	99	64	20	64	65	64	63	65	20	61	6e	63	65	20	74	68	65	20	65	72	67	68	69	63					
48	61	78	68	20	66	6c	69	67	68	74	20	74	61	20	74	61	6b	65	2e	0a	48	65	20	6e	65	76	52	20	67	61	74	20	6f	66	20	62	61	61	72	64	20	61	70					
67	65	74	68	61	72	69	73	62	75	72	67	2e	20	49	65	6c	6e	2c	20	77	68	65	72	65	20	66	6e	69	64	79	61	6e	20	72	65	71	67	74	20	73	75	67						
74	75	76	65	73	74	65	64	20	61	70	74	61	77	66	20	77	61	73	20	64	65	73	74	62	71	69	65	62	6e	20	54	68	61	74	20	64	65	63	69	73	69	61	6e	20	70			
54	77	74	75	72	66	65	64	20	61	75	74	20	74	61	20	62	65	65	20	61	20	6c	75	63	6b	79	20	62	72	65	61	20	66	61	72	70	24	68	65	20	41	50	20					
64	71	24	54	77	69	63	65	2e	0a	52	61	79	20	77	61	73	20	61	6d	6f	66	72	70	74	68	65	2d	20	61	6e	61	72	64	20	73	74	73	70	72	2d								
20	69	64	71	61	20	61	72	72	69	76	65	20	61	6e	64	20	68	65	20	63	61	6e	66	79	62	65	64	20	74	68	65	20	73	72	67	74	20	73	70	72	74	2d						
74	20	72	60	69	66	2e	0a	61	66	76	62	67	61	74	73	2e	20	48	65	80	23	70	63	68	67	20	74	67	65	73	20	76	65	70	67	65	20	75	72	67								
61	79	74	20	72	66	63	74	69	63	73	20	61	6f	2e	20	74	68	65	20	70	65	20	77	69	74	68	20	41	50	20	52	61	64	69	61	20	70	6c	62									
64	20	71	69	75	64	20	62	61	63	6b	63	20	73	65	73	65	64	20	74	65	62	70	61	20	65	64	69	64	20	71	67	63	72	69	62	65	20	66	65	20								
64	20	74	68	65	20	69	69	64	74	65	72	66	73	20	61	6e	64	20	70	75	74	20	74	68	65	20	6d	61	74	65	72	66	20	61	74	66	20	74	65	72	66	20						
65	78	75	78	74	20	77	69	72	65	73	62	7e	20	48	65	20	74	68	65	6e	20	77	61	6c	6b	65	64	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70				
74	61	77	74	69	63	65	2e	0a	52	61	79	20	77	61	73	20	61	6d	6f	66	72	70	74	68	65	2d	20	61	6e	61	72	64	20	73	74	73	70	72	2d									
20	69	64	71	61	20	61	72	72	69	76	65	20	61	6e	64	20	68	65	20	63	61	6e	66	79	62	65	64	20	74	68	65	20	73	72	67	74	20	73	70	72	74	2d						
74	20	72	60	69	66	2e	0a	61	66	76	62	67	61	74	73	2e	20	48	65	80	23	70	63	68	67	20	74	67	65	73	20	76	65	70	67	65	20	75	72	67								
65	74	75	76	65	73	20	69	63	6b	65	72	66	73	20	61	6e	65	20	70	75	74	20	74	68	65	20	6d	61	74	65	72	66	20	61	74	66	20	74	65	72	66	20						
66	78	75	78	74	20	77	69	72	65	73	62	7e	20	48	65	20	74	68	65	6e	20	77	61	6c	6b	65	64	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70				
74	61	74	71	64	69	63	65	2e	0a	52	61	79	20	77	61	73	20	61	6d	6f	66	72	70	74	68	65	2d	20	61	6e	61	72	64	20	73	74	73	70	72	2d								
65	20	65	20	66	69	63	65	2e	0a	52	61	79	20	77	61	73	20	61	6d	6f	66	72	70	74	68	65	2d	20	61	6e	61	72	64	20	73	74	73	70	72	2d								
65	20	66	69	63	65	2e	0a	52	61	79	20	77	61	73	20	61	6d	6f	66	72	70	74	68	65	2d	20	61	6e	61	72	64	20	73	74	73	70	72	2d										
66	69	78	76	65	73	20	69	63	6b	65	72	66	73	20	61	6e	65	20	70	75	74	20	74	68	65	20	6d	62	71	68	65	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70
66	69	78	76	65	73	20	69	63	6b	65	72	66	73	20	61	6e	65	20	70	75	74	20	74	68	65	20	6d	62	71	68	65	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70
66	69	78	76	65	73	20	69	63	6b	65	72	66	73	20	61	6e	65	20	70	75	74	20	74	68	65	20	6d	62	71	68	65	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70
66	69	78	76	65	73	20	69	63	6b	65	72	66	73	20	61	6e	65	20	70	75	74	20	74	68	65	20	6d	62	71	68	65	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70
66	69	78	76	65	73	20	69	63	6b	65	72	66	73	20	61	6e	65	20	70	75	74	20	74	68	65	20	6d	62	71	68	65	20	66	69	72	73	20	66	72	6f	61	6c	6b	65	73	74	73	70
66	69	78	76	65	73	20	69	63	6b	65	72	66	73	20	6																																	

Identifizierung von aufeinanderfolgenden Sequenzen von Hexa-Zahlen

Schon besser – aber immer noch Optimierungspotenzial vorhanden

Im nächsten Schritt kann man jetzt nach aufeinanderfolgenden Sequenzen suchen und diese wiederum abstrahieren.

Abstrahieren bedeutet, dass man für diese aufeinanderfolgenden Sequenzen (nebenstehend als Beispiel: 61 6c) ein neues „Symbol“ definiert (generiert). Damit würde man die Menge der Symbole für die Repräsentation des gesamten Textkörpers dieses riesigen Datensatzes von Text (FineWeb Datensatz) für die spätere Berechnung weiter optimieren.

Diesen Prozess – die Konvertierung von Text in eine Repräsentation dieses Textes nennt man **TOKENISIERUNG**.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

Tiktokenizer

The screenshot shows a user interface for the TikTok Tokenizer. At the top, there's a dropdown menu labeled "cl100k_base". Below it, a "Token count" field shows the value "145". The main area contains a text input box with a long paragraph about tornadoes in 2012. The text is color-coded with various words highlighted in different colors (e.g., green, blue, yellow) to represent tokens. Below the text input, the corresponding token IDs are listed as a long sequence of numbers.

```
4438, 10314, 5068, 304, 682, 20447, 505, 45272, 42728, 9890, 13918, 28623, 220, 23, 11, 220, 679, 17, 198, 45, 99, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504, 813, 6787, 3, 11, 5944, 311, 16272, 369, 7445, 7742, 11, 34070, 389, 2, 25952, 311, 279, 9919, 18585, 311, 279, 23170, 323, 19946, 369, 279, 17149, 13, 1283, 7070, 10491, 3131, 1070, 902, 11213, 311, 1935, 627, 1548, 2646, 2751, 38, 9, 4580, 264, 11277, 13, 12361, 11, 568, 9670, 709, 1043, 9017, 21750, 10481, 11, 12799, 2637, 1405, 2926, 6821, 12090, 264, 6424, 574, 14763, 13, 3011, 5597, 66, 56, 704, 311, 387, 264, 18069, 1464, 369, 279, 18314,
```

https://tiktokenizer.vercel.app/?model=cl100k_base

ChatGPT verwendet für die Englische Sprache 100.277 Tokens

Die GPT-Modelle von ChatGPT verwenden eine Menge von 100.277 Tokens als Repräsentanz von Texten (Zeichen, Zahlen, Sonderzeichen).

Mit der Tokenizer Anwendung auf der linken Seite kann man sich dies auch anschauen. Die farbig markierten Worte / Wortteile/ Silben des Textes sind jeweils ein Token.

Token ist sozusagen die neue Maßeinheit, die jetzt für die Eingabe bzw. Ausgabe von Text als Berechnungsgrundlage verwendet wird.

Basierend auf einer Menge von Tokens, die ein LLM verarbeitet (Input / Eingabe bzw. Output / Ausgabe) wird u.a. auch ein Preis von den Anbietern von LLMs pro 1 Mio. Tokens als Kosten für Endkunden definiert.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

The image contains two side-by-side screenshots of a 'Tiktokenizer' interface. Both screenshots show the input text 'Guten Morgen' and the resulting tokens. The top screenshot shows a token count of 4, with tokens 38, 13462, 386, 40663. The bottom screenshot shows a token count of 10, with tokens 38, 13462, 386, 40663, 459, 37844, 4538, 77, 6569, 126. Both screenshots have a dropdown menu set to 'cl100k_base'.

ChatGPT verwendet für die Englische Sprache **100.277 Tokens**

Nebenstehend ein paar Beispiele ..., hier sieht man wie ein deutscher Text in Tokens repräsentiert wird.

A screenshot of the 'Tiktokenizer' interface showing the input text 'GUTEN MORGEN AN DIESEM SONNIGEN TAG' and the resulting tokens. The token count is 13, with tokens 38, 1406, 965, 72413, 11332, 2147, 72506, 84839, 7751, 2, 45, 1953, 965, 22216. The dropdown menu is set to 'cl100k_base'. The screenshot highlights the difference between 'Guten Morgen' and 'GUTEN MORGEN'.

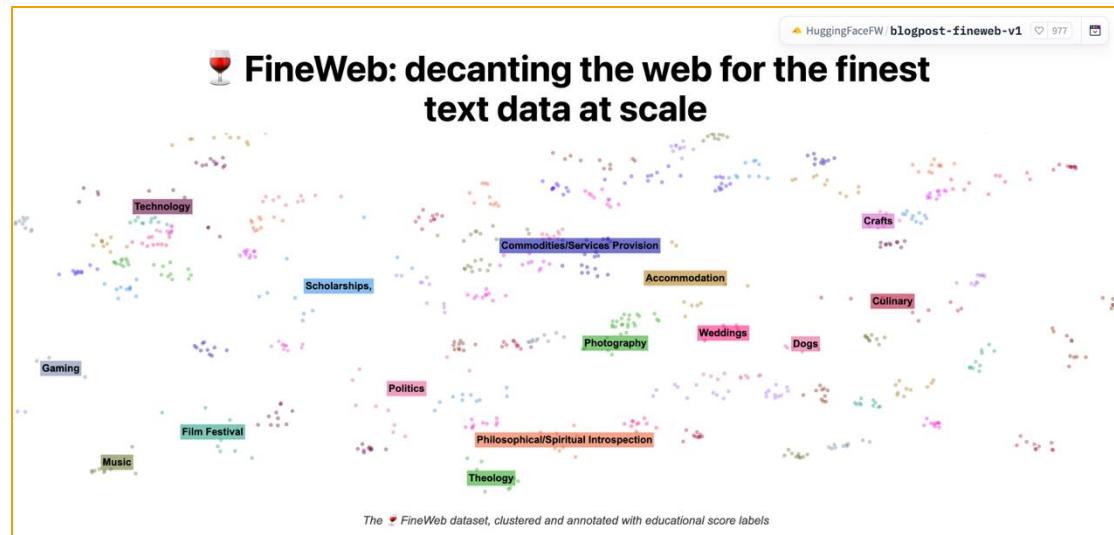
Hier sieht man auch den Unterschied zwischen Groß- und Kleinschreibung, der über Tokens ebenfalls abgebildet wird.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen



<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

FineWeb Datensatz:

Noch mal kurz zurück zu dem FineWeb Datensatz, den man für das Training von LLMs verwenden kann:

Dieser Datensatz – also ein Stream von unendlich viel Text umfasst:

**15-trillion Tokens
44TB Speicherplatz**



LLMs

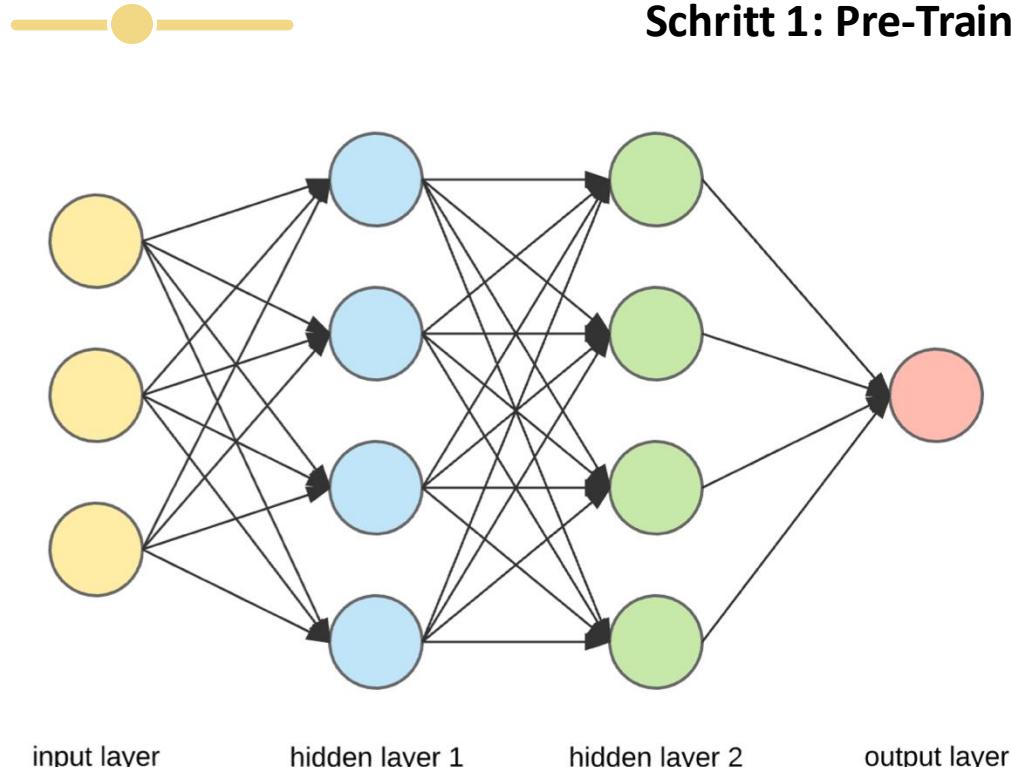
ChatGPT – wie schauen hinter die „Kulissen“

Phase 1:

Pre-Training Phase: Neuronales Netz, Algorithmen, Mathematik, Statistik

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training mit riesigen Datenmengen

FineWeb Datensatz für das Training mit einem Neuronalen Netz:

Das Neuronale Netz wird im ersten Schritt mit dem „Datensatz“ der Trainingsdaten (z.B. dem FineWeb Datensatz) „gefüttert“. D.h. die gesamten 15-trillion Tokens bzw. 44TB werden dem Netz als Eingabe bereitgestellt.

Dabei werden die Gewichte der Schichten des Neuronalen Netzes per Zufall definiert – es handelt sich um ein untrainiertes Netz mit einem ersten „Abbild“ des Datensatzes in diesem mathematischen Gehirn – ein noch sehr unklares und diffuses eher chaotisches „Gehirn“.

Danach beginnt man mit dem Training, in dem man aus dem gesamten Datensatz einen Teil von Tokens auswählt mit einer beliebigen Länge und diese als Input / Eingabe dem Netz gibt, um dann die Ausgabe des nächsten Tokens zu erzeugen.



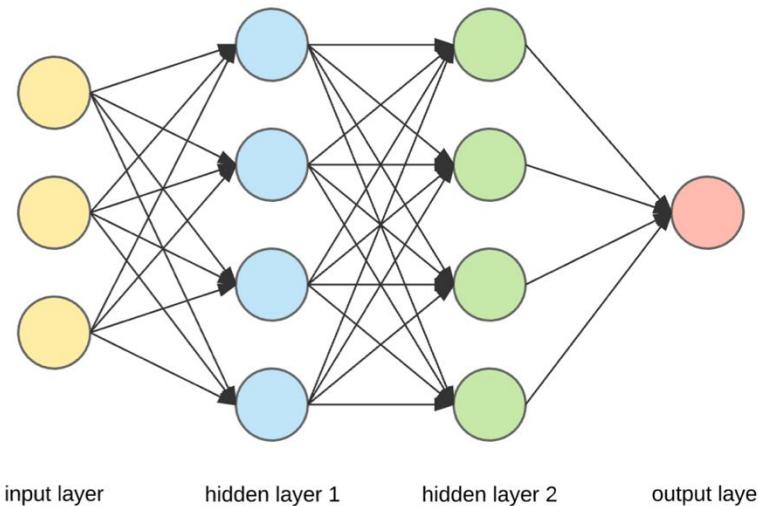
Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

egionsMarch 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take.
He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.

4438, 10314, 5068, 304, 682, 20447, 505, 45272, 42728, 9890, 13918, 28623, 220, 23, 11, 220, 679, 17, 198, 45 99, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504, 813, 6787, 3 11, 5944, 311, 16272, 369, 7445, 7742, 11, 34070, 389 2, 25952, 311, 279, 9919, 18585, 311, 279, 23170, 323, 19946, 369, 279, 17149, 13, 1283, 7070, 10491, 3131, 1 070, 902, 11213, 311, 1935, 627, 1548, 2646, 2751, 38 9, 4580, 264, 11277, 13, 12361, 11, 568, 9670, 709, 10 043, 9017, 21750, 10481, 11, 12799, 2637, 1405, 2926, 6821, 12090, 264, 6424, 574, 14763, 13, 3011, 5597, 66 56 704 311 387 264 18069 1464 369 279 10314



Trainingsablauf des KNNs

Die nachfolgend beschriebene Methode für das Training wird mehrfach (millionenfach) durchlaufen.

Was genau wird jetzt gemacht:

Ausschnitte aus dem gesamten Textkorpus werden ausgewählt (Tokenlänge ist unterschiedlich ... je mehr Tokens, um so mehr Rechenpower ist erforderlich) und dem Netz als Input zur Verfügung gestellt. Es erfolgt dann basierend auf den Algorithmen des Neuronalen Netzes eine Berechnung des statistisch am besten passenden nächsten Tokens als Ausgabe.

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped

4599, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504

regionsMarch 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take. He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.

4438, 10314, 5068, 304, 682, 20447, 505, 45272, 42728, 9890, 13918, 28623, 220, 23, 11, 220, 679, 17, 198, 4599, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504, 813, 6787, 31, 5944, 311, 16272, 369, 7445, 7742, 11, 34070, 3892, 25952, 311, 279, 9919, 18585, 311, 279, 23170, 323, 19946, 369, 279, 17149, 13, 1283, 7070, 10491, 3131, 1070, 902, 11213, 311, 1935, 627, 1548, 2646, 2751, 389, 4580, 264, 11277, 13, 12361, 11, 568, 9670, 709, 10043, 9017, 21750, 10481, 11, 12799, 2637, 1405, 2926, 6821, 12090, 264, 6424, 574, 14763, 13, 3011, 5597, 6656 704 311 387 264 18069 1464 369 279 10314

Trainingsdurchlauf

Hier sehen Sie jetzt einen Ausschnitt aus dem Text mit 145 Tokens, die in das KNN als Input eingegeben werden.

Da es sich ja um einen eindimensionalen Stream an Tokens für den gesamten Datensatz handelt, sind die darauf folgenden nächsten Tokens „bekannt“. Nebenstehend in einem roten Rahmen markiert.

D.h. wir wissen, welche Wörter bzw. Tokens die „richtigen“ sind, da wir ja den gesamten Textkorpus vorliegen haben.

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

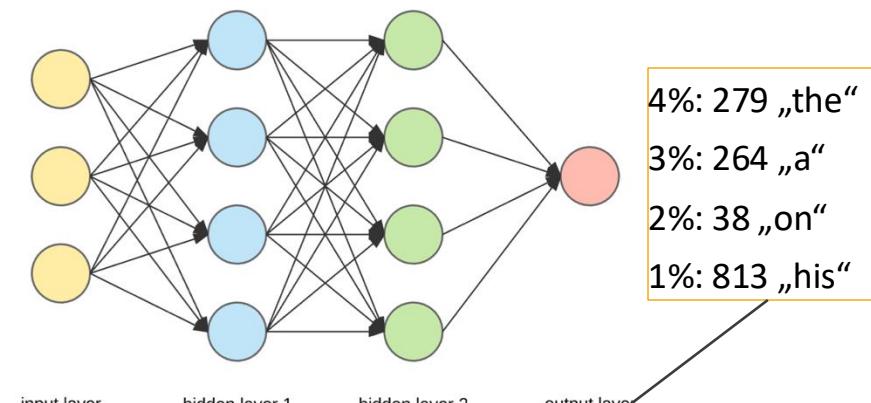


Schritt 1: Pre-Training mit riesigen Datenmengen



regionsMarch 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take.
He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.

4438, 10314, 5068, 304, 682, 20447, 505, 45272, 42728, 9890, 13918, 28623, 220, 23, 11, 220, 679, 17, 198, 45 99, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504, 813, 6787, 3 11, 5944, 311, 16272, 369, 7445, 7742, 11, 34070, 389 2, 25952, 311, 279, 9919, 18585, 311, 279, 23170, 323, 19946, 369, 279, 17149, 13, 1283, 7070, 10491, 3131, 1 070, 902, 11213, 311, 1935, 627, 1548, 2646, 2751, 38 9, 4580, 264, 11277, 13, 12361, 11, 568, 9670, 709, 10 043, 9017, 21750, 10481, 11, 12799, 2637, 1405, 2926, 6821, 12090, 264, 6424, 574, 14763, 13, 3011, 5597, 66 56 704 311 387 264 18069 1464 369 270 10314



In Summe werden hier 100.277 Tokens mit Wahrscheinlichkeiten berechnet, die als nächstes folgen könnten.

Trainingsdurchlauf - Fortsetzung

Das noch am Anfang des Trainings sich befindende Neuronale Netz berechnet jetzt erst Mal eine Wahrscheinlichkeit der nächsten sinnvollen Tokens basierend auf dem aktuellen „Wissensstand“ also dem aktuellen Abbild des gesamten Datensatzes in den Parametern und Gewichten des KNN.

Anbei eine Wahrscheinlichkeitsverteilung der nächsten möglichen Tokens, die der Eingabe folgen könnte.

Aus dem Datensatz erkennen wir, das das Wort „his“ hier folgen müsste, aber es wird „the“ mit 4% Wahrscheinlichkeit vorgeschlagen.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

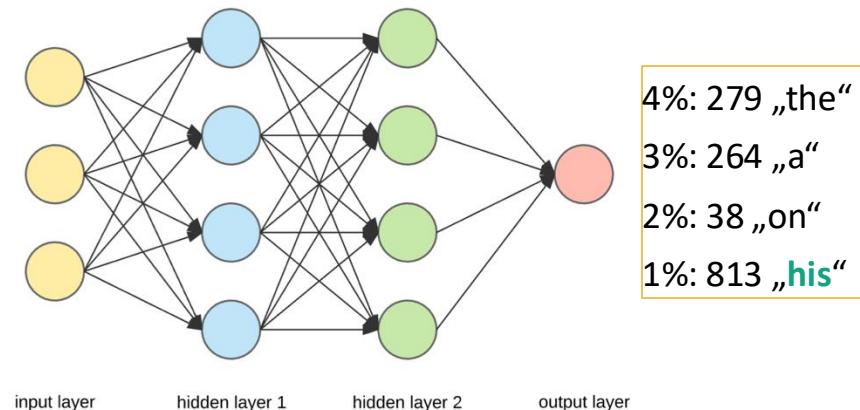


Schritt 1: Pre-Training mit riesigen Datenmengen



regionsMarch 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take. He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.

4438, 10314, 5068, 304, 682, 20447, 505, 45272, 42728, 9890, 13918, 28623, 220, 23, 11, 220, 679, 17, 198, 45 99, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504, 813, 6787, 3 11, 5944, 311, 16272, 369, 7445, 7742, 11, 34070, 389 2, 25952, 311, 279, 9919, 18585, 311, 279, 23170, 323, 19946, 369, 279, 17149, 13, 1283, 7070, 10491, 3131, 1 070, 902, 11213, 311, 1935, 627, 1548, 2646, 2751, 38 9, 4580, 264, 11277, 13, 12361, 11, 568, 9670, 709, 10 043, 9017, 21750, 10481, 11, 12799, 2637, 1405, 2926, 6821, 12090, 264, 6424, 574, 14763, 13, 3011, 5597, 66 56 704 311 387 264 18069 1464 369 270 10314



Trainingsdurchlauf – Update der Gewichte des Neuronalen Netzes:

Aus dem Datensatz wissen wir, dass der nächste Token (das nächste Wort) 813 also „his“ ist.

Wir können damit jetzt das gesamte Netz „updaten“, in dem wir die Gewichte so anpassen (dafür gibt es eine Rückwärts-Berechnung von der Ausgabe zurück bis zur Eingabe über alle Schichten und Gewichte), dass beim nächsten Durchlauf mit der gleichen Eingabe von Tokens als Input der Token 813 (his) eine höhere Wahrscheinlichkeit bekommt – bis sozusagen dieser Token der „wahrscheinlichste“ ist. Dafür sind natürlich mehrere Durchläufe durch das Netz mit Anpassungen notwendig.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

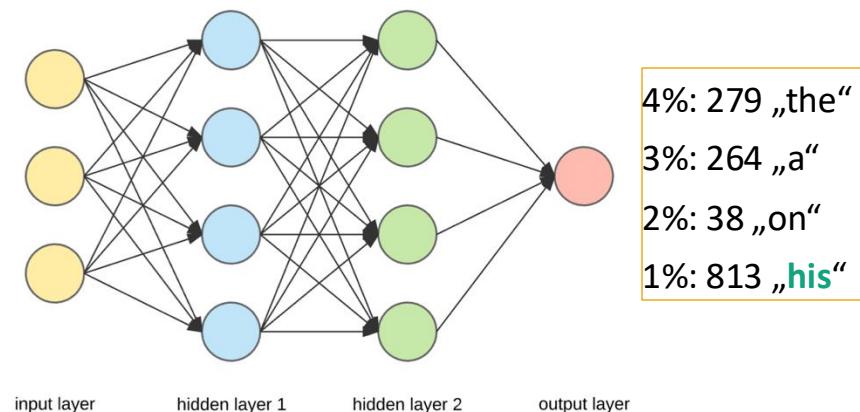


Schritt 1: Pre-Training mit riesigen Datenmengen



regionsMarch 8, 2012
When the first serious bout of tornadoes of 2012 blew through middle America in the middle of the night, they touched down in places hours from any AP bureau. Our closest video journalist was Chicago-based Robert Ray, who dropped his plans to travel to Georgia for Super Tuesday, booked several flights to the cities closest to the strikes and headed for the airport. He'd decide once there which flight to take. He never got on board a plane. Instead, he ended up driving toward Harrisburg, Ill., where initial reports suggested a town was destroyed. That decision turned out to be a lucky break for the AP. Twice.

4438, 10314, 5068, 304, 682, 20447, 505, 45272, 42728, 9890, 13918, 28623, 220, 23, 11, 220, 679, 17, 198, 45 99, 279, 1176, 6129, 25646, 315, 45272, 288, 315, 220, 679, 17, 42423, 1555, 6278, 5270, 304, 279, 6278, 315, 279, 3814, 11, 814, 24891, 1523, 304, 7634, 4207, 505, 904, 10314, 44402, 13, 5751, 18585, 2835, 23672, 574, 10780, 6108, 8563, 13558, 11, 889, 12504, 813, 6787, 3 11, 5944, 311, 16272, 369, 7445, 7742, 11, 34070, 389 2, 25952, 311, 279, 9919, 18585, 311, 279, 23170, 323, 19946, 369, 279, 17149, 13, 1283, 7070, 10491, 3131, 1 070, 902, 11213, 311, 1935, 627, 1548, 2646, 2751, 38 9, 4580, 264, 11277, 13, 12361, 11, 568, 9670, 709, 10 043, 9017, 21750, 10481, 11, 12799, 2637, 1405, 2926, 6821, 12090, 264, 6424, 574, 14763, 13, 3011, 5597, 66 56 704 311 387 264 18069 1464 369 270 10314



Training, Rückwärts-Anpassung, Tokens, ...

Die zuvor beschriebene Methode des Trainings findet jetzt millionenfach mit unterschiedlichsten Ausschnitten aus dem gesamten Textkorpus (also unterschiedliche Textlängen/ Tokenlängen) gleichzeitig statt.

Dafür ist eine entsprechende Infrastruktur mit Rechenpower (Prozessoren, Rechenzentren,...) erforderlich, um diese parallel laufenden Berechnungen in dem KNN durchzuführen.

Somit wird das Künstliche Neuronale Netz im besser in der vorhersage des wahrscheinlichsten nächsten Tokens basierend auf dem Input an Tokens.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

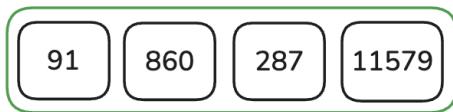


Schritt 1: Pre-Training mit riesigen Datenmengen

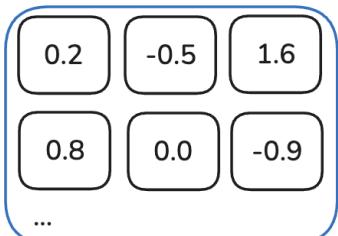


Neuronales Netzwerk (Innenleben)

Eingabe Sequenz von Tokens x
Zwischen 1 bis z.B. 8.000 tokens



Parameter (/ "Gewichte") w
Billionen in der Regel



Gigantische mathematische Formel

$$1/(1 + \exp(-(w_0 * (1/(1 + \exp(-(w_1 * x_1 + w_2 * x_2 + w_3)))) + w_4 * (1/(1 + \exp(-(w_5 * x_1 + w_6 * x_2 + w_7)))) + w_8 * (1/(1 + \exp(-(w_9 * x_1 + w_{10} * x_2 + w_{11})))) + w_{12})))$$

→ 100,277 mögliche Tokens



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

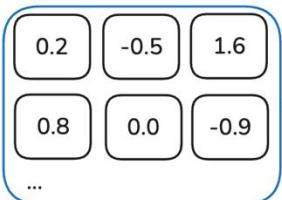
Schritt 1: Pre-Training mit riesigen Datenmengen

Neuronales Netzwerk (Innenleben)

Eingabe Sequenz von Tokens x
Zwischen 1 bis z.B. 8.000 tokens



Parameter ("Gewichte") w
Billionen in der Regel



Gigantische mathematische Formel

$$1/(1 + \exp(-(w_0 * (1/(1 + \exp(-(w_1 * x_1 + w_2 * x_2 + w_3)))) + w_4 * (1/(1 + \exp(-(w_5 * x_1 + w_6 * x_2 + w_7)))) + w_8 * (1/(1 + \exp(-(w_9 * x_1 + w_{10} * x_2 + w_{11})))) + w_{12})))$$

100,277 mögliche Tokens

Eingabe Sequenz – Token Länge – Kontext Fenster

Je nach Sprachmodell wird beim Training eine unterschiedliche maximale Länge von Tokens verwendet. Diese lag am Anfang bei **4.000 Tokens** und hat mittlerweile Größen von über 128.000 Tokens erreicht (sogar über 1 Mio. Tokens).

Wir sprechen dann hier auch von einem **Kontext-Fenster** mit einer bestimmten Tokenmenge, die das Sprachmodell in seinem Kurzzeitgedächtnis nutzen kann, also immer wieder der gigantischen mathematischen Formel einspeisen kann, um den nächsten Token zu berechnen.



Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

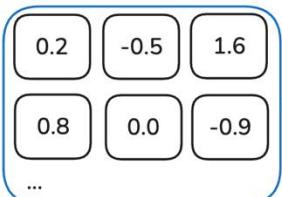
Schritt 1: Pre-Training mit riesigen Datenmengen

Neuronales Netzwerk (Innenleben)

Eingabe Sequenz von Tokens x
Zwischen 1 bis z.B. 8.000 tokens



Parameter (/ "Gewichte") w
Billionen in der Regel



Gigantische mathematische Formel

$$1/(1 + \exp(-(w_0 * (1/(1 + \exp(-(w_1 * x_1 + w_2 * x_2 + w_3)))) + w_4 * (1/(1 + \exp(-(w_5 * x_1 + w_6 * x_2 + w_7)))) + w_8 * (1/(1 + \exp(-(w_9 * x_1 + w_{10} * x_2 + w_{11})))) + w_{12})))$$

100,277 mögliche Tokens

Gewichte / Parameter

Die Gewichte der Parameter (wir reden hier in der Regel über eine Anzahl von mehreren Billionen bis zu hunderten von Billionen), die in den Schichten der neuronalen Netze eingestellt werden können, um die perfekte statistische Vorhersage basierend auf dem Trainingsdatensatz zu berechnen ... diese können Sie sich wie ein riesiges Mischpult in einem Tonstudio vorstellen, wo jeder Regler für eine bestimmte Funktion auf eine Range von Werten eingestellt werden kann, um den optimalen Sound zu generieren. In unserem Fall wird durch millionenfaches gleichzeitiges „Rückwärts-Rechnen“, die optimale Einstellung des Netzwerks „berechnet“.



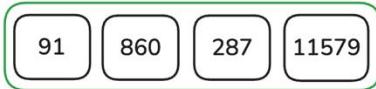
Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

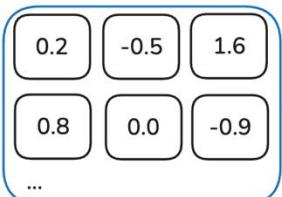
Schritt 1: Pre-Training mit riesigen Datenmengen

Neuronales Netzwerk (Innenleben)

Eingabe Sequenz von Tokens x
Zwischen 1 bis z.B. 8.000 tokens



Parameter (/ "Gewichte") w
Billionen in der Regel



Gigantische mathematische Formel
$$1/(1 + \exp(-(w_0 * (1/(1 + \exp(-(w_1 * x_1 + w_2 * x_2 + w_3)))) + w_4 * (1/(1 + \exp(-(w_5 * x_1 + w_6 * x_2 + w_7)))) + w_8 * (1/(1 + \exp(-(w_9 * x_1 + w_{10} * x_2 + w_{11})))) + w_{12})))$$

100,277 mögliche Tokens

Gigantische mathematische Formel (Perceptron – Abbildung des Gehirns in viele Schichten mit Neuronen)

Die Formel bzw. die Algorithmen sind spezifisch für die jeweiligen LLMs, im wesentlichen werden hier die Eingabe Werte (Tokens) mit den Werten für die Gewichte in der jeweiligen Schicht multipliziert, exponentielle Normalisierungen ... werden durchgeführt. Die Algorithmen sind so gewählt, dass Millionen von gleichzeitigen (parallelen) Berechnungen effizient durchgeführt werden können.

Das Ziel besteht darin, die Werte für die Gewichte so zu optimieren, das basierend auf der Eingabe von Tokens, die darauf folgenden Tokens bei der Aussage konsistent dem ursprünglichen Datensatz entsprechen.

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training mit riesigen Datenmengen

Inhaltsverzeichnis
Verbergen
(Anfang)

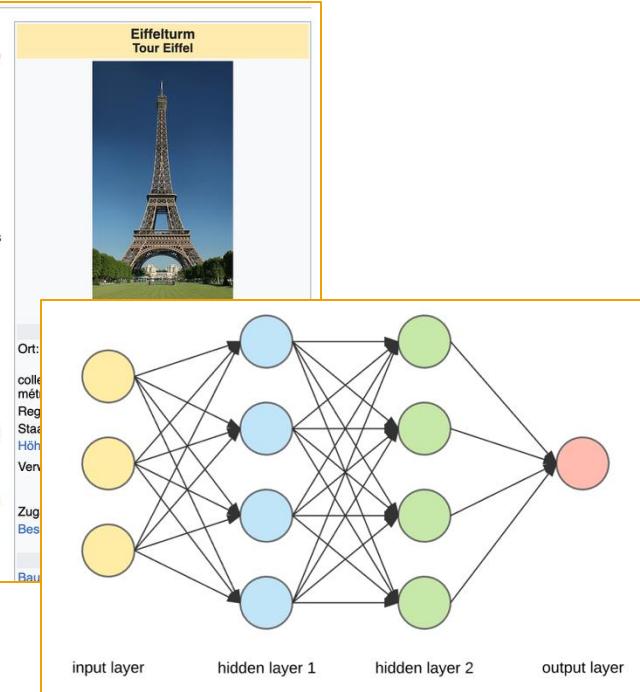
- > Geschichte
- > Beschreibung
- > Technik
- > Tourismus
- > Ereignisse
- > Rezeption und Wirkung
- > Frequenzen und Programme
- Film
- > Literatur
- > Weblinks
- Einzelnachweise

Der **Eiffelturm** (französisch *Tour Eiffel*, [tuʁ_ɛfɛl]) ist ein 330 Meter hoher Eisenfachwerkturm in Paris. Er steht im 7. Arrondissement am nordwestlichen Ende des Champ de Mars (Marsfeld), nahe dem Ufer der Seine. Das von 1887 bis 1889 errichtete, 10.100 Tonnen schwere Bauwerk wurde als monumentales Eingangstor und Aussichtsturm für die Weltausstellung zur Erinnerung an den 100. Jahrestag der Französischen Revolution errichtet. Der nach dem Erbauer Gustave Eiffel benannte und zum Errichtungszeitpunkt noch 312 Meter hohe Turm war von seiner Erbauung bis zur Fertigstellung des Chrysler Building 1930 in New York das höchste Bauwerk der Welt. Mit der Ausstrahlung des ersten öffentlichen Radioprogramms in Europa 1921 und des ersten französischen Fernsehprogramms 1935 trug das Bauwerk als Sendeturm zur Geschichte des Hörfunks und des Fernsehens bei. Der Fernsehturm ist die wichtigste Sendeanlage des Großraums Paris und beherbergt als Turmrestaurant das mit einem Michelin-Stern ausgezeichnete Restaurant *Le Jules Verne*.^[1]

Als höchstes Bauwerk von Paris prägt er das Stadtbild bis heute und zählt mit gut sechs Millionen zahlenden Besuchern pro Jahr zu den meistbesuchten Wahrzeichen der Welt. Der Turm ist eine der bekanntesten Ikonen der Architektur und der Ingenieurkunst. Der Eiffelturm ist das Vorbild vieler Nachahmungen und wird in Kunst und Kultur im Zusammenhang mit Paris und Frankreich vielfach aufgegriffen. Er gilt als nationales Symbol der Franzosen und avancierte zu einer weltweiten Ikone der Moderne. Seit 1964 ist der Eiffelturm als *monument historique* denkmalgeschützt, und 1986 nahm die American Society of Civil Engineers das Bauwerk in die Liste der historischen Meilensteine der Ingenieurbaukunst auf.

Eingabe

Als höchstes Bauwerk von Paris prägt er das Stadtbild bis heute und zählt mit gut sechs Millionen zahlenden Besuchern pro Jahr zu den



Ausgabe

Als höchstes Bauwerk von Paris prägt er das Stadtbild bis heute und zählt mit gut sechs Millionen zahlenden Besuchern pro Jahr zu den **meistbesuchten Sehenswürdigkeiten in der ganzen Welt**. Der **Eifelturm** ist eine der ...

Ergebnis des Pre-Trainings:

Das Ergebnis des Trainings ist ein sogenanntes Basis-Modell des LLMs (Base-Model) mit einer Repräsentation des gesamten Wissens aus dem Datensatz, der für das Training genutzt wurde (z.B. FineWeb). Repräsentation bedeutet ein Abbild des Datensatzes als Wissen im Neuronalen Netz mit dem letzten Stand der Gewichte nach dem Training.

Wenn man dann als Input dem Base-Modell einen Text aus Wikipedia gibt (Daten aus Wikipedia als Bestandteil von FineWeb Daten), dann werden Tokens für die Ausgabe generiert, die statistisch sehr ähnlich dem Inhalt des originären Satzes aus Wikipedia sind.

LLMs

ChatGPT – wie schauen hinter die „Kulissen“

Hardware für das Pre-Training in der Phase 1:

GPUs – Graphic Processing Units || CPUs – Central Processing Units

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training in der Cloud auf GPU's



GPUs – Graphic Processing Units in der Cloud für das Training

Eine GPU (Graphics Processing Unit) ist eine spezielle elektronische Schaltung, die ursprünglich für die schnellere Erstellung von Bildern und Videos entwickelt wurde.

Die bemerkenswerte Fähigkeit von Graphics Processing Units, eine Vielzahl von Berechnungen schnell auszuführen, hat dazu geführt, dass sie in verschiedenen Bereichen eingesetzt werden, darunter künstliche Intelligenz und wissenschaftliche Berechnungen, wo sie sich durch ihre Fähigkeit auszeichnen, dateiintensive und rechenintensive Aufgaben zu bewältigen.



NVIDIA GPU H100 (Kosten: ca. 35.000 EUR)

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training in der Cloud auf GPU's

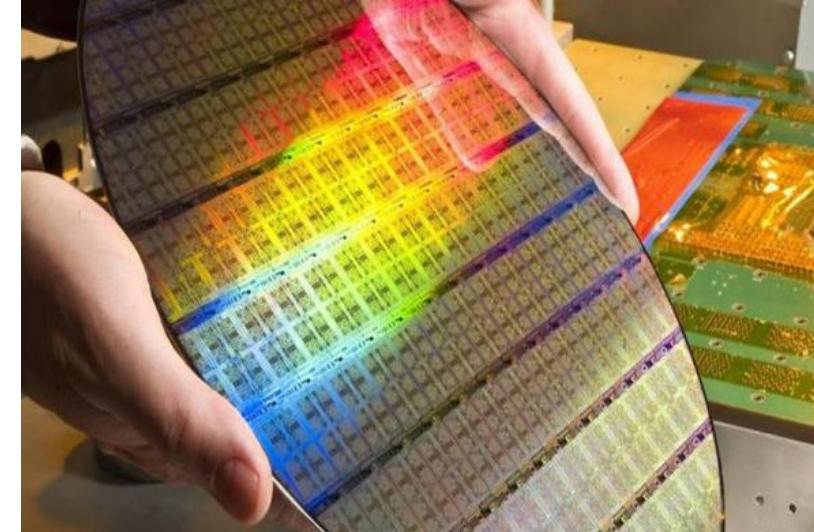


GPUs vs CPUs

GPUs werden auf einem Silikon-Wafer aufgebaut, einem dünnen, kreisförmigen Schnitt aus hochreinem Silizium.

Dieser Wafer dient als Basis, auf die Milliarden winziger Transistoren geätzt werden. Diese Transistoren fungieren als Miniatur-Gatter, die den Stromfluss steuern und die Bausteine der Logikschaltungen der GPU bilden.

Ein komplexes Netzwerk aus mikroskopisch kleinen Metalldrähten, sogenannten Interconnects, verbindet diese Transistoren, sodass sie miteinander kommunizieren und Berechnungen durchführen können. Schließlich wird die gesamte Baugruppe in ein Gehäuse aus Materialien wie Kunststoff, Keramik und Metall eingefasst, um die empfindlichen Schaltkreise zu schützen und die beim Betrieb entstehende Wärme abzuleiten.



Einer der komplexesten monolithischen Chips ist derzeit der GH100 von NVIDIA mit 80 Mrd. Transistoren.

Künstliche Intelligenz

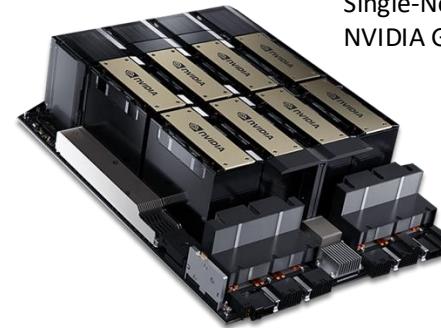
ChatGPT – Die Entwicklung eines Sprachmodells

Schritt 1: Pre-Training in der Cloud auf GPU's

GPUs vs CPUs

GPUs führen eine große Anzahl von Abläufen gleichzeitig aus. Sie erreichen dies durch eine große Anzahl von Prozessorkernen, die gleichzeitig an verschiedenen Teilen einer Aufgabe arbeiten können. Diese Architektur für parallele Verarbeitung ermöglicht es GPUs, Aufgaben zu bearbeiten, die CPUs viel länger brauchen würden.

Stellen Sie sich eine Aufgabe vor, die in Tausende kleinerer, unabhängiger Schritte unterteilt werden kann. Eine GPU kann diese Schritte auf ihre vielen Kerne verteilen, sodass sie gleichzeitig berechnet werden können. Diese Fähigkeit zur Mehrfachverarbeitung ist der entscheidende Vorteil von GPUs gegenüber CPUs, insbesondere in Bereichen wie Bild- und Videoverarbeitung, wissenschaftlichen Simulationen und vor allem im maschinellen Lernen, wo große Datenmengen und komplexe Algorithmen die Norm sind.



Single-Node mit 8x H100
NVIDIA GPUs



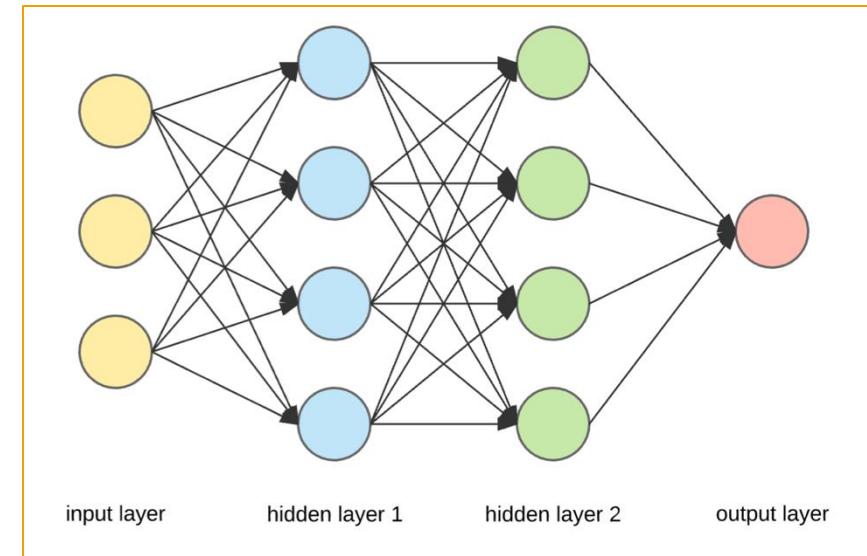
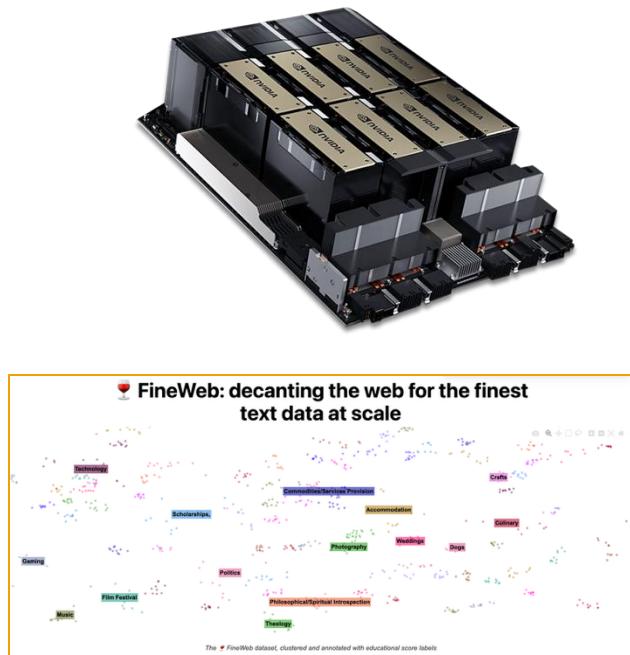
Datencenter mit multiplen H100 Nodes in Server Schränke
zusammengebaut

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training in der Cloud auf GPU's



Wettrennen: Daten + GPUs + Mathematik

Je mehr GPUs, je mehr Daten und je besser die Algorithmen / Mathematik ...

... um so besser werden die Sprachmodelle und die Vorhersage des wahrscheinlichsten nächsten Tokens, um dann:

Texte, Bilder, Musik oder Videos zu „generieren“.

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Schritt 1: Pre-Training in der Cloud auf GPU's



MODEL ↑	CREATOR ↑	CONTEXT WINDOW ↑
o3-pro	OpenAI	200k
Gemini 2.5 Pro	Google	1m
o3	OpenAI	128k
o4-mini (high)	OpenAI	200k
Gemini 2.5 Pro (Mar '25)	Google	1m
DeepSeek R1 0528 (May '25)	deepseek	128k
Gemini 2.5 Pro (May' 25)	Google	1m
Grok 3 mini Reasoning (high)	XAI	1m
o3-mini (high)	OpenAI	200k
Gemini 2.5 Flash (Reasoning)	Google	1m

MODEL ↑	CREATOR ↑	CONTEXT WINDOW ↑
Gemini 2.5 Flash (Reasoning)	Google	1m
Claude 4 Opus Thinking	ANTHROP\C	200k
MiniMax M1 80k	MINIMAX	1m
o3-mini	OpenAI	200k
Qwen3 235B (Reasoning)	Alibaba	128k
o1	OpenAI	200k
MiniMax M1 40k	MINIMAX	1m
Llama Nemotron Ultra Reasoning	NVIDIA	128k
Claude 4 Sonnet Thinking	ANTHROP\C	200k

Wettbewerb von diversen Herstellern von Sprachmodellen (USA, Europa, China, ...).



LLMs

ChatGPT – wie schauen hinter die „Kulissen“

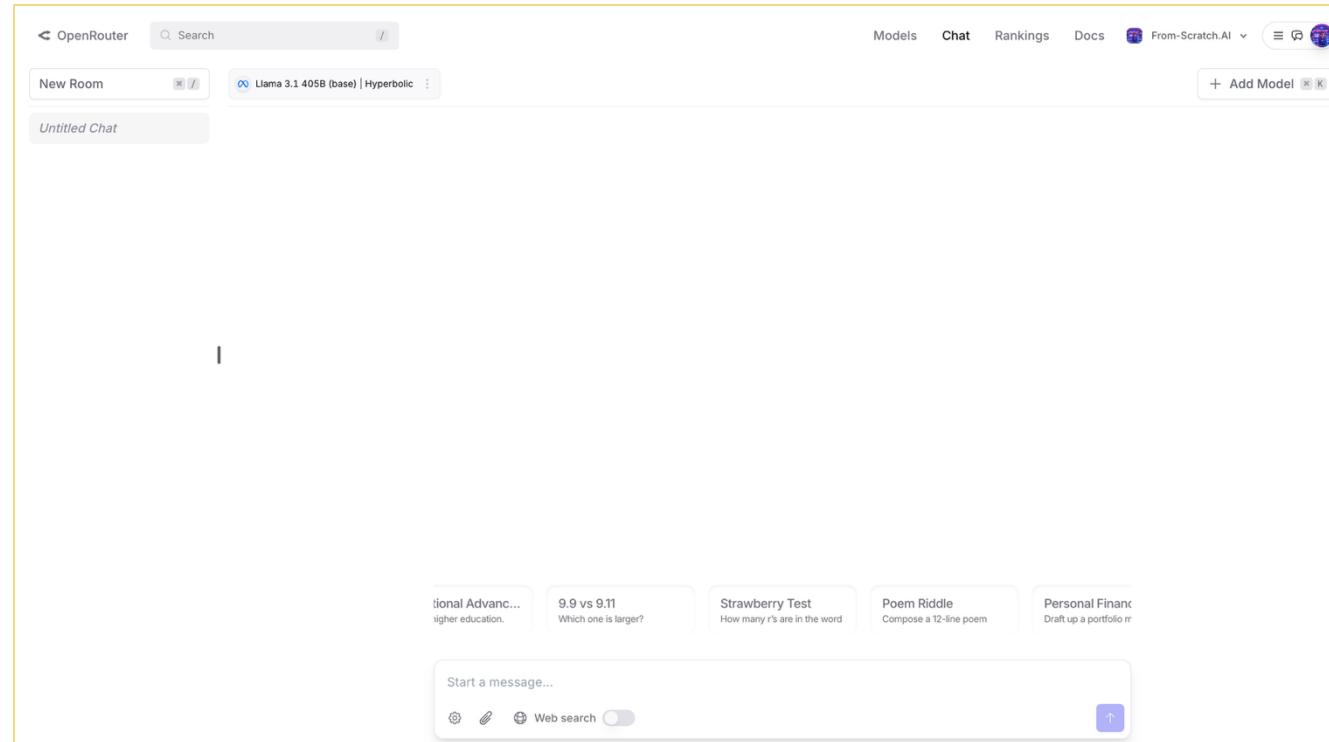
Ergebnis der erste Phase des Trainings – „Ein Basis-Sprachmodell“

Beispiel eines Basis-Sprachmodells: Meta Llama 3 (2024)

Künstliche Intelligenz

Die Entwicklung eines Sprachmodells – Basis Modell

Schritt 1: Ergebnis des Pre-Trainings „Basis Modell“



<https://openrouter.ai/>

Beispiel Meta Llama 3.1 405B (Base Model)

Nebenstehend haben wir über den Provide OpenRouter das Basismodell 3.1 405B von Meta (Facebook) aufgerufen.

Das Modell ist noch kein Chat-Modell, dass mit dem Kunden über ein beliebiges Thema „chatten“ kann. Es hat das Pre-Training durchlaufen und die Parameter „repräsentieren“ eine „Abbildung“ des gelernten Wissens.

Der gesamte Text aus den Trainingsdaten ist jetzt in den Parametern und den Gewichten sozusagen optimal „eingearbeitet“ worden.



Künstliche Intelligenz

Die Entwicklung eines Sprachmodells – Basis Modell



Schritt 1: Ergebnis des Pre-Trainings „Basis Modell“



Beispiel Meta Llama 3.1 405B (Base Model)

Das Modell kann z.B. mit seinem Wissen in den Parametern und Gewichten einen angefangenen Text „statistisch korrekt“ fortsetzen, siehe nebenstehendes Beispiel:

<https://openrouter.ai/>

LLMs

ChatGPT – wie schauen hinter die „Kulissen“

Hardware für das Pre-Training in der Phase 1:

GPUs – Graphic Processing Units || CPUs – Central Processing Units

Künstliche Intelligenz

ChatGPT – Die Entwicklung eines Sprachmodells



Ergebnis des Pre-Trainings mit riesigen Datenmengen



FineWeb Datensatz für das Training mit einem Neuronalen Netz:

Das Neuronale Netz wird im ersten Schritt mit dem „Datensatz“ der Trainingsdaten (z.B. dem FineWeb Datensatz) „gefüttert“. D.h. die gesamten 15-trillion Tokens bzw. 44TB werden dem Netz als Eingabe bereitgestellt.

Dabei werden die Gewichte der Schichten des Neuronalen Netzes per Zufall definiert – es handelt sich um ein untrainiertes Netz mit einem ersten „Abbild“ des Datensatzes in diesem mathematischen Gehirn – ein noch sehr unklares und diffuses eher chaotisches „Gehirn“.

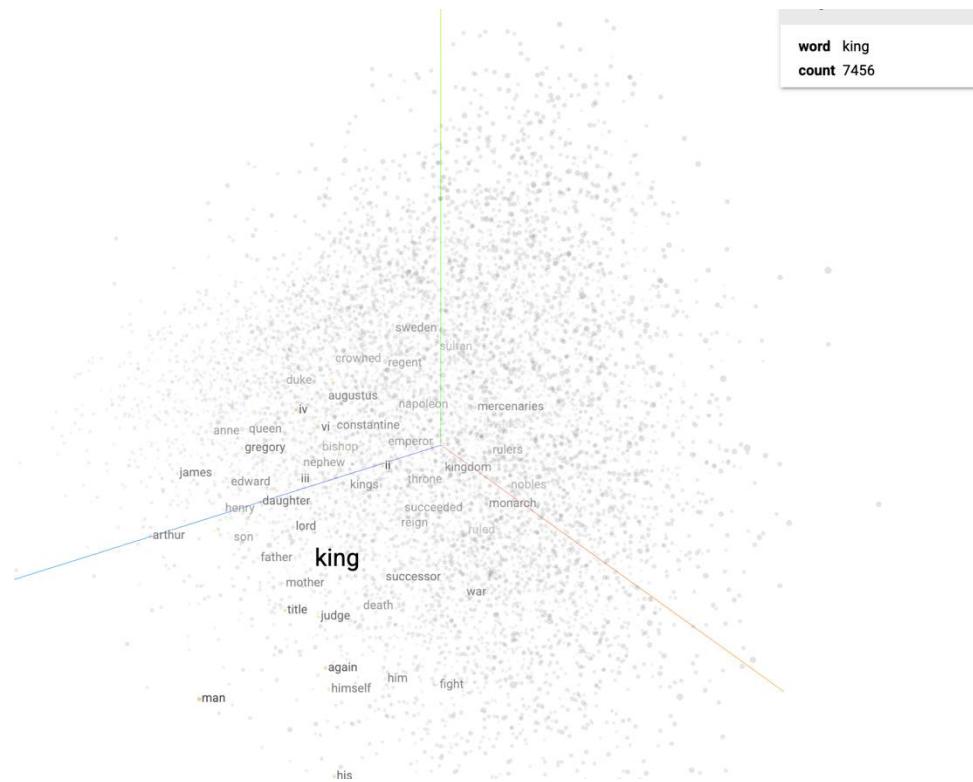
Danach beginnt man mit dem Training, in dem man aus dem gesamten Datensatz einen Teil von Tokens auswählt mit einer beliebigen Länge und diese als Input / Eingabe dem Netz gibt, um dann die Ausgabe des nächsten Tokens zu erzeugen.



Künstliche Intelligenz

LLMs – der Trainingsprozess im Detail

Embedding – Tokens in Vektoren



2. Feinabstimmung (Fine-tuning): Die "Feinjustierung" durch Menschen

Wie es funktioniert: Nach dem Vortraining wird das Modell oft noch einmal mit kleineren, spezifischeren Datensätzen trainiert, die von Menschen bewertet wurden. Hier kommen menschliche Trainer ins Spiel, die dem Modell beibringen, was eine "gute" oder "hilfreiche" Antwort ist. Wenn das Modell zum Beispiel eine unhöfliche oder falsche Antwort gibt, wird es "bestraft", und wenn es eine gute Antwort gibt, wird es "belohnt".

Das Ziel: Diese Phase macht das Modell nützlicher und sicherer für den Endnutzer. Es lernt, Anweisungen besser zu befolgen, weniger "Unsinn" zu reden, sich höflicher auszudrücken und relevantere Antworten zu geben. Hier wird es zum Beispiel auch darauf trainiert, Fragen zu beantworten, Zusammenfassungen zu erstellen oder Übersetzungen anzufertigen, anstatt nur das nächste Wort zu erraten.





McCulloch-Pitts-Neuron

Wir schauen genauer hin die Kapitel kommen noch im Laufe der Woche

Kontakt

Rückfragen, Ideen, Feedbacks, ...

philipp.diel@adg-campus.de