Handreichung für Lehrkräfte

Vom Wort zur Antwort:
Wie ein Sprachmodell Schritt für Schritt arbeitet

Roadmap-Übersicht für die KI-Fortbildungsreihe (Stufe 2) · Begleitmaterial zu App und Arbeitsblättern

---

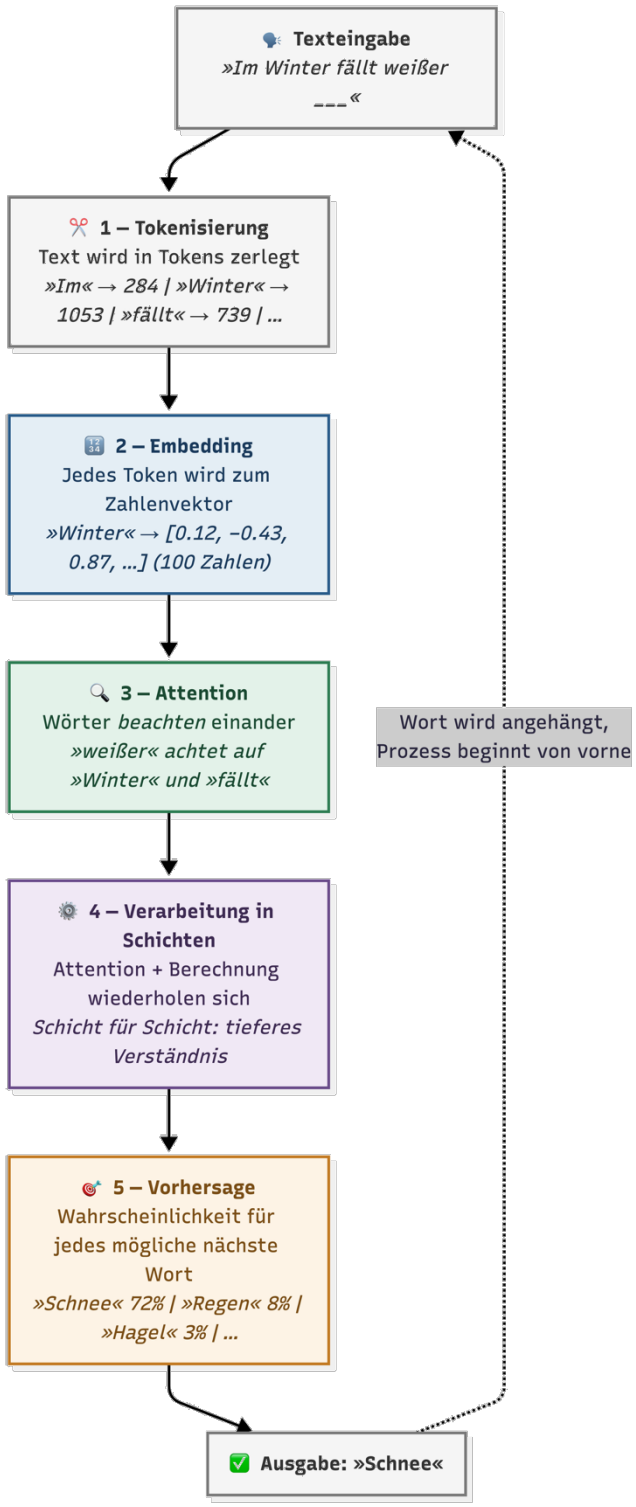
Worum geht es?

Wenn Sie in ChatGPT eine Frage tippen und eine Antwort erhalten, passieren im Hintergrund mehrere klar unterscheidbare Schritte. Diese Handreichung zeigt den **gesamten Weg vom eingegebenen Wort bis zur erzeugten Antwort** und ordnet ein, welche Teile davon in den Fortbildungsmaterialien behandelt werden.

Das Ziel ist nicht, jedes technische Detail zu verstehen, sondern die **Grundlogik** zu begreifen: Ein Sprachmodell ist im Kern eine **Wort-für-Wort-Vorhersagemaschine**, die auf raffinierten mathematischen Tricks beruht — und jeder dieser Tricks lässt sich im Prinzip nachvollziehen.

Die Pipeline:
5 Schritte vom Wort zur Antwort

Das folgende Diagramm zeigt die Verarbeitungsschritte eines modernen Sprachmodells (LLM) in vereinfachter Form. Die farbigen Markierungen zeigen, welcher Schritt in welchem Material behandelt wird.



1

Tokenisierung — Text wird in Stücke zerlegt

Der eingegebene Text wird in kleine Einheiten zerlegt — sogenannte **Tokens**. Ein Token ist oft ein Wort, kann aber auch ein Wortteil sein.

Zum Beispiel wird „Eisverkäufer“ möglicherweise in „Eis“ + „verkäufer“ zerlegt. Jedes Token bekommt eine eindeutige Nummer (z. B. „Hund“ = 4827).

2

Embedding — Jedes Token wird zum Zahlenvektor

Jede Token-Nummer wird in einen **Vektor aus vielen Zahlen** umgewandelt (in der App: 50 oder 100 Zahlen, bei echten Modellen: Tausende). Diese Zahlenliste kodiert die **Bedeutung** des Wortes. Wörter mit ähnlicher Bedeutung haben ähnliche Zahlenmuster — deshalb liegen „Hund“ und „Katze“ im Vektorraum nah beieinander.

Material: App „Wörter im Vektorraum“ + Handreichung „Warum Analogien scheitern“

3

Attention — Wörter „beachten“ einander

Jetzt passiert der entscheidende Schritt: Jedes Wort „schaut sich um“ und prüft, welche anderen Wörter im Satz für seine Bedeutung wichtig sind. Bei „Der Hund bellt, weil **er** Angst hat“ muss das Modell erkennen, dass sich „er“ auf „Hund“ bezieht — nicht auf irgendetwas anderes. Dies geschieht über **mehrere parallele Aufmerksamkeits-Köpfe** (Multi-Head Attention), die verschiedene Arten von Beziehungen gleichzeitig erkennen können.

U0001F4CE Material: Arbeitsblatt „Satzlücken-Orakel“, Block D

4

Verarbeitung in Schichten — Attention und Berechnung wiederholen sich

Die Schritte Attention und anschließende Berechnung wiederholen sich in **vielen aufeinanderfolgenden Schichten** (bei GPT-4: fast 100 Schichten). Jede Schicht verfeinert das Verständnis: Frühe Schichten erkennen einfache Muster (Grammatik, Wortnähe), spätere Schichten erfassen komplexere Zusammenhänge (Ironie, logische Schlüsse, Weltwissen).

5

Vorhersage — Das nächste Wort wird ausgewählt

Am Ende steht eine **Wahrscheinlichkeitsverteilung über alle möglichen nächsten Wörter**. Das Modell sagt z. B.: „Nach ›Im Winter fällt weißer‹ kommt mit 72 % ›Schnee‹, mit 8 % ›Regen‹, mit 3 % ›Hagel‹ …“. Dann wird ein Wort ausgewählt, an den Text angehängt — und **der gesamte Prozess beginnt von vorne**, bis die Antwort fertig ist.

U0001F4CE Material: Arbeitsblatt „Satzlücken-Orakel“, Blöcke A–C

Die Schleife: So entsteht ein ganzer Text

Ein Sprachmodell erzeugt **immer nur ein einziges Wort auf einmal**. Um einen ganzen Satz zu schreiben, wiederholt es die Schritte 1–5 für jedes neue Wort: Das gerade erzeugte Wort wird an die Eingabe angehängt, und alles läuft erneut durch.

Welches Material behandelt welchen Schritt?

Die folgende Übersicht zeigt, wie die vorhandenen Materialien den einzelnen Pipeline-Schritten zugeordnet sind:

Schritt	Material	Lernziel	Sozialform
2 – Embedding	App „Wörter im Vektorraum“ (Modus: Wort-Detektiv, Analogie-Labor, Wort-Karte)	Wörter sind Zahlenvektoren; ähnliche Bedeutung = ähnliche Zahlen	Einzelarbeit / Plenum
2 – Embedding	Handreichung „Warum Analogien scheitern“	Grenzen von Embeddings verstehen; Polysemie erkennen	Lehrkraft-Hintergrund
3 – Attention	Arbeitsblatt „Satzlücken-Orakel“, Block D	Wortbezüge erkennen; verstehen, warum Kontext entscheidend ist	Partnerarbeit
5 – Vorhersage	Arbeitsblatt „Satzlücken-Orakel“, Blöcke A–C	Next-Token-Prediction intuitiv erleben	Einzel / Klasse

> **Was bewusst weggelassen wurde**
>
> Zwei technische Details wurden aus didaktischen Gründen nicht in die Materialien aufgenommen: **Positional Encoding** (das Modell erhält zusätzlich Information darüber, an welcher Stelle im Satz ein Wort steht) und **Softmax** (ein mathematisches Verfahren, um die Ausgabewerte in Wahrscheinlichkeiten umzurechnen). Beides sind wichtige technische Bausteine, aber für das Konzeptverständnis der Zielgruppe nicht erforderlich. Falls Teilnehmende danach fragen: Positional Encoding sorgt dafür, dass das Modell die Wortstellung berücksichtigt; Softmax ist ein mathematischer Kniff, um große Wertebereiche auf Wahrscheinlichkeiten zwischen 0 und 1 zu stauchen.

Empfohlener Ablauf in der Fortbildung

Die Materialien bauen aufeinander auf. Hier ein bewährter Ablauf für eine 90-Minuten-Einheit:

Phase 1: „Ein Sprachmodell sagt das nächste Wort vorher“ (ca. 35 Min.)

Material: Arbeitsblatt „Satzlücken-Orakel“, Blöcke A–C

Die Teilnehmenden erleben Next-Token-Prediction intuitiv, indem sie selbst Lücken füllen (Block A), gemeinsam einen Satz Wort für Wort aufbauen (Block B) und reflektieren, welches Wissen für gute Vorhersagen nötig ist (Block C). Am Ende steht die Erkenntnis: *Ein Sprachmodell tut genau das — nur mit statistischen Mustern statt echtem Verständnis.*

Phase 2: „Aber woher weiß das Modell, was ein Wort bedeutet?“ (ca. 30 Min.)

Material: App „Wörter im Vektorraum“ (alle drei Modi)

Der Übergang von Phase 1 zu 2 ergibt sich organisch: In Phase 1 wurde klar, dass Wortbedeutung entscheidend ist. Jetzt zeigen die Teilnehmenden selbst, dass man Bedeutung als Zahlen darstellen kann. Der Wort-Detektiv macht Ähnlichkeit erfahrbar, das Analogie-Labor zeigt, dass man mit Bedeutung rechnen kann, und die Wort-Karte macht die Cluster sichtbar. Die Heatmap-Visualisierung unter dem Zielwort zeigt die konkreten Zahlenwerte.

Phase 3: „Wie versteht das Modell ganze Sätze?“ (ca. 20 Min.)

Material: Arbeitsblatt „Satzlücken-Orakel“, Block D

Block D führt den Attention-Mechanismus ein: Die Teilnehmenden zeichnen Wortbezüge ein und erleben, dass „Sie“, „sein“ und „er“ nur im Kontext eindeutig werden. Der Vergleich mit einer „blinden KI“ (ohne Attention) macht den Mehrwert greifbar.

Phase 4: Zusammenführung — Die Pipeline (ca. 5 Min.)

Material: Diese Roadmap (Seite 1)

Zum Abschluss zeigen Sie die Pipeline-Übersicht und ordnen gemeinsam ein: „Wir haben heute Schritt 2, 3 und 5 selbst erlebt. Schritte 1 und 4 passieren automatisch im Hintergrund.“ Das gibt den Teilnehmenden das befriedigende Gefühl, die Gesamtstruktur zu überblicken — ohne sich in jedem Detail verloren zu haben.

> **Didaktischer Hinweis zur Reihenfolge**
>
> Obwohl die Pipeline technisch mit Schritt 1 beginnt, empfehlen wir, in der Fortbildung mit **Schritt 5 (Vorhersage)** zu starten. Der Grund: Next-Token-Prediction ist das intuitivste Konzept und knüpft direkt an die ML-Vorerfahrung (Eisverkäufer-Regression) an. Von dort aus kann man rückwärts fragen: „Aber wie versteht das Modell, was ein Wort bedeutet?“ (Schritt 2) und „Wie versteht es ganze Sätze?“ (Schritt 3).

Erstellt als Übersichtsmaterial für die KI-Fortbildungsreihe (Stufe 2). Die Pipeline-Darstellung ist bewusst vereinfacht und auf das Konzeptverständnis für Lehrkräfte aller Fächer ausgerichtet.