Hands-on LLM

Wie Sprachmodelle funktionieren und was wir damit machen können

KARS · Gais · 2. November 2026

Benjamin Hanimann

Was könnte als Nächstes kommen?

Der Kaffee ist heute besonders …

stark heiss gut

Welche Ergänzung passt?

Ein grosses Sprachmodell

LLM bedeutet Large Language Model.

Ein trainiertes neuronales Netz verarbeitet Sprache als Zahlen.

Es berechnet passende Fortsetzungen zum verfügbaren Kontext.

Tokens: die Einheiten des Texts

Ein Token kann ein Wort, ein Wortteil oder ein Zeichen sein.

tokenization

token ization

Beispiel einer Tokenisierung. Die Aufteilung hängt vom Tokenizer ab.

Wie eine Antwort entsteht

1	**Eingabe**	Text wird zu Tokens und ihren Zahlendarstellungen.
2	**Verarbeitung**	Das Netz verarbeitet Beziehungen im verfügbaren Kontext.
3	**Auswahl**	Das System wählt die nächste Einheit aus möglichen Fortsetzungen.
4	**Fortsetzung**	Die neue Einheit kommt dazu. Der Vorgang wiederholt sich.

Training: aus Beispielen lernen

1	**Vorhersage**	Das Modell erhält Text und sagt die nächste Einheit voraus.
2	**Vergleich**	Die Vorhersage wird mit dem tatsächlichen Folgetext verglichen.
3	**Anpassung**	Ein Lernverfahren verändert die Gewichte des Netzes.
4	**Wiederholung**	Viele Beispiele formen die Fähigkeiten des Modells.

Vom Grundmodell zum Assistenten

Vortraining

Das Modell lernt aus grossen Datenmengen sprachliche und andere Muster.

Weiteres Training

Beispiele und Rückmeldungen trainieren hilfreiches Antworten und das Befolgen von Aufträgen.

Attention: Beziehungen im Kontext

Ich sitze auf der Bank.

Ich überweise Geld an die Bank.

Die Wörter im Umfeld helfen, die Bedeutung einzuordnen.

Kontext: die aktuelle Arbeitsfläche

Anweisungen	Was soll das System tun?
Gespräch	Welche relevanten Nachrichten liegen vor?
Material	Welche Datei- oder Quellenausschnitte stehen bereit?
Werkzeuge	Welche Suchergebnisse oder Berechnungen kamen dazu?

Training und Nutzung sind verschieden

Training

Die Gewichte des Modells verändern sich durch ein Lernverfahren.

Nutzung

Das Modell berechnet eine Ausgabe mit den vorhandenen Gewichten und dem aktuellen Kontext.

Weshalb plausible Antworten falsch sein können

Eine passende Fortsetzung kann sachlich falsch sein.

Fakten, Namen und Quellen können erfunden oder verwechselt sein.

Wichtige Aussagen brauchen eine unabhängige Prüfung.

Was die Grundlagen für die Praxis bedeuten

Kontext	Relevantes Material und klare Rahmenbedingungen mitgeben.
Auftrag	Das gewünschte Ergebnis konkret beschreiben.
Prüfung	Fakten, Quellen und Verhalten am Ergebnis kontrollieren.
Überarbeitung	Beobachtete Schwächen mit einem Folgeauftrag ändern.

Modell, Anwendung und Werkzeuge

Modell

Verarbeitet Kontext und erzeugt Ausgaben.

Anwendung

Stellt Oberfläche, Gesprächsverlauf und Funktionen bereit.

Werkzeuge

Suchen, rechnen, lesen Dateien oder führen Aktionen aus.

Cloud und lokale Modelle

Cloud

Die Verarbeitung läuft auf Rechnern eines Dienstes. Funktionen und Datenverwendung hängen vom Angebot ab.

Lokal

Das Modell läuft auf dem eigenen Gerät. Hardware und Modell bestimmen, welche Aufgaben gut funktionieren.

Ein guter Arbeitsauftrag

Ziel	Was soll am Ende vorliegen, für wen und wozu?
Material	Welche Informationen und Beispiele sind verbindlich?
Ausgabe	Welche Form, Länge oder Funktion brauchen wir?
Prüfung	Woran erkennen wir, ob das Ergebnis brauchbar ist?

Beispiele für Alltag und Arbeit

Texte

Aus Notizen eine Einladung, Antwort oder kurze Zusammenfassung entwickeln.

Recherche

Informationen suchen, Quellen vergleichen und offene Fragen erkennen.

Daten

Eine Tabelle ordnen, Kategorien vorschlagen und Ergebnisse darstellen.

Dokumente

Ein PDF erklären und Aussagen mit passenden Fundstellen verbinden.

Bilder und Audio

Ein Motiv entwerfen, ein Storyboard planen oder Sprache transkribieren.

Kleine Anwendungen

Einen Timer, eine interaktive Erklärung oder ein Übungsspiel bauen.

Beispiel 1: ein Briefing, mehrere Texte

Fiktiver Reparaturtreff

14. November 2026, 14–17 Uhr
Werkraum, Dorfplatz 8
Anmeldung bis 12. November

Daraus: kurze Einladung, FAQ oder Checkliste.

Unser Test: Stimmen Datum, Zeit und Ort? Was fehlt?

Beispiel 2: Recherche mit Belegen

1	**Suchen**	Eine eingegrenzte Frage und passende Quellen festlegen.
2	**Lesen**	Die entscheidende Fundstelle tatsächlich öffnen.
3	**Abgleichen**	Prüfen, ob die Quelle die Aussage stützt.
4	**Einordnen**	Stand, Unsicherheit und Widersprüche benennen.

Beispiel 3: Bilder, Audio und Dokumente

Als Eingabe

Ein Bild beschreiben, ein Dokument auswerten oder Sprache in Text übertragen.

Als Ausgabe

Ein Motiv, ein Layoutentwurf oder eine Audiodatei erzeugen, wenn passende Funktionen bereitstehen.

Beispiel 4: eine kleine Anwendung

«Baue einen Pausentimer als einzelne HTML-Datei.»

«Dauer wählen. Starten. Anhalten. Zurücksetzen. Restzeit gross anzeigen.»

«Die Datei soll lokal im Browser funktionieren. Ohne Anmeldung oder externe Dienste.»

«Prüfe Start, Pause, Fortsetzen und das Ende bei null.»

Testen und gezielt verbessern

Beobachtung

«Beim Fortsetzen springt die Restzeit wieder auf den Anfang.»

Folgeauftrag

«Erhalte beim Anhalten die Restzeit. Fortsetzen soll von dort weiterzählen. Teste das Verhalten erneut.»

Ein Ergebnis wird brauchbar durch Prüfung

Inhalt	Stimmen Fakten, Quellen und Berechnungen?
Auftrag	Passt das Ergebnis zum gewünschten Zweck?
Funktion	Lässt es sich öffnen, bedienen und weiterverwenden?
Rahmen	Sind Material, Datenzugriff und Veröffentlichung geklärt?

Eine Idee für die Pause

Ich möchte … vereinfachen oder ausprobieren.

Dafür wäre … als Ergebnis hilfreich.

Eine erste kleine Version könnte …

Ich würde sie prüfen, indem …

Pause bis 14:30

Was möchten wir mit diesen Werkzeugen ausprobieren oder bauen?

Eine konkrete Aufgabe. Ein kleines erstes Ergebnis.

Nach der Pause sammeln wir Ideen und setzen einen Vorschlag gemeinsam um.

Unsere gemeinsame Bauidee

Welcher Vorschlag ist für mehrere Personen interessant?

Welchen kleinen Ausschnitt können wir heute zeigen?

Woran erkennen wir, ob er funktioniert?

Was nehmen wir aus dem Versuch mit?

Was hat das System gut gemacht?

Was mussten wir korrigieren?

Wofür würden wir das Ergebnis tatsächlich verwenden?

Grundlagen und Quellen

OpenAI Docs: Tokens und Kontext

Vaswani et al. (2017): Attention Is All You Need

Li et al. (2024): Mechanics of Next Token Prediction

Anthropic: Training und Grenzen von Sprachmodellen

Produktfunktionen: offizielle Seiten von ChatGPT, Gemini und Manus