



GLOSSAR · MAI 2026

KI-Glossar für Geschäftsführer

50 Begriffe, die du als Chef kennen musst.

CEO-GPT · moritzmaaker.com · Mai 2026

50 Begriffe, die du als Chef kennen musst

Du musst nicht Tech-Sprache sprechen. Du musst nur verstehen, worüber dein Team, deine Berater und dein Wettbewerb gerade reden. Dieses Glossar gibt dir die 50 Begriffe, mit denen du in jedem KI-Gespräch mitkommst, ohne dich zu blamieren.

Die Auswahl ist bewusst pragmatisch. Keine vollständige Lexikon-Liste, sondern die Begriffe, die du tatsächlich im Arbeitsalltag hörst. Vom Modell-Namen über die wichtigsten Techniken bis zu den Konzepten, die deine Berater dir verkaufen wollen.

Pro Eintrag bekommst du drei Dinge: eine knappe Definition ohne Tech-Jargon, ein konkretes Business-Beispiel, und Verweise auf verwandte Begriffe. Wenn ein Begriff einen anderen voraussetzt, springst du dorthin und liest nach.

Wie du dieses Glossar nutzt

- **Als Nachschlagewerk.** Ein Begriff fällt in einem Meeting, du schlägst hier nach. Index am Ende hilft.
- **Als Einarbeitung.** Von A bis Z durchlesen dauert circa 45 Minuten. Danach kennst du das Vokabular der gesamten Branche.
- **Als Briefing für dein Team.** Schick es deinen Mitarbeitern, bevor ihr den ersten KI-Anbieter trifft.

Quellen am Ende, alle Definitionen sind an die offiziellen Docs der Hersteller angelehnt. Stand Mai 2026.

Lese-Hinweis

Englische Begriffe bleiben Englisch, wenn sie als Fachterminus etabliert sind. Niemand sagt "Kontextfenster" zu "Context Window". Du sollst die Begriffe wiedererkennen, wenn sie fallen.

A

KONZEPT

Agent

Ein KI-System, das nicht nur antwortet, sondern eigenständig Aufgaben ausführt, Tools aufruft und mehrere Schritte hintereinander erledigt, bis ein Ziel erreicht ist.¹⁹

Beispiel: Du sagst dem Agent "buch mir einen Flug nach Berlin am Dienstag". Er sucht selbst nach Flügen, vergleicht Preise, ruft die Buchungs-API auf und schickt dir die Bestätigung. Klassischer Chatbot würde nur eine Liste vorschlagen.

Siehe auch: Tool Use, Workflow, Subagent, Claude Code

ANBIETER

Anthropic

KI-Unternehmen aus San Francisco, gegründet 2021 von ehemaligen OpenAI-Mitarbeitern. Hersteller der Claude-Modellfamilie. Fokus auf Sicherheit und Steuerbarkeit von KI.¹

Beispiel: Wenn du Claude in deinem Unternehmen einsetzt, ist dein Vertragspartner entweder Anthropic direkt oder ein Reseller wie AWS Bedrock und Google Cloud Vertex.

Siehe auch: Claude, Sonnet, Claude Code, OpenAI

TECHNIK

API

Application Programming Interface. Eine technische Schnittstelle, über die zwei Programme miteinander reden. Bei KI: der Zugang, über den deine eigene Software ein KI-Modell aufruft, statt es im Browser zu nutzen.^{1, 4}

Beispiel: Dein CRM ruft per API Claude auf, lässt jede eingehende Kundenanfrage zusammenfassen und priorisieren, schreibt das Ergebnis zurück. Kein Mensch sitzt dazwischen.

Siehe auch: Rate Limit, JSON-Mode, Streaming, Inferenz

C

TECHNIK

Chain of Thought (CoT)

Eine Prompt-Technik, bei der das Modell aufgefordert wird, seinen Lösungsweg Schritt für Schritt aufzuschreiben, bevor es die finale Antwort gibt. Erhöht die Genauigkeit bei komplexen Aufgaben deutlich.¹¹

Beispiel: Statt "Soll ich Projekt A oder B priorisieren?" schreibst du "Denk laut nach, wäge Kosten, Risiken und Umsatz beider Projekte ab, dann gib deine Empfehlung." Du bekommst eine nachvollziehbare Begründung statt nur ein Ergebnis.

Siehe auch: Reasoning, Prompt Engineering, Few-Shot

MODELL

Claude

KI-Modellfamilie von Anthropic. Konkurrent zu GPT und Gemini. Bekannt für lange Kontextfenster, sauberes Schreiben und gute Performance bei Coding und langen Dokumenten.¹

Beispiel: Du lädst einen 200-seitigen Vertrag in Claude, lässt dir die kritischen Klauseln markieren und einen Verhandlungsleitfaden schreiben. Im Browser unter claude.ai oder per API.

Siehe auch: Anthropic, Sonnet, GPT, Gemini, Claude Code

TOOL

Claude Code

Kommandozeilen-Tool von Anthropic, mit dem Claude direkten Zugriff auf deinen Rechner, deine Dateien und dein Git-Repository bekommt. Gedacht für Entwickler, aber zunehmend auch für Power-User, die Workflows automatisieren.²

Beispiel: Du sagst Claude Code "geh durch alle PDF-Rechnungen im Ordner Q2, extrahier die Beträge und schreib sie in eine Excel". Es liest, extrahiert, schreibt, alles lokal auf deinem Mac.

Siehe auch: Claude, Skill, Subagent, MCP, Vibe Coding

KONZEPT

Context Window

Die maximale Menge an Text, die ein KI-Modell auf einmal verarbeiten kann. Wird in Token gemessen. Größeres Fenster heißt: das Modell kann längere Dokumente, längere Konversationen oder mehr Code-Dateien gleichzeitig im Kopf behalten.¹

Beispiel: Bei einem Context Window von 200.000 Token passen circa 500 Seiten Text rein. Du kannst Claude einen ganzen Jahresbericht plus alle Quartalsberichte geben und nach Inkonsistenzen fragen.

Siehe auch: Token, LLM, Cowork

FEATURE

Cowork

Arbeitsweise, bei der du mit einem KI-Modell über längere Zeit gemeinsam an einem Projekt arbeitest, mit geteiltem Kontext, gespeicherten Dateien und Memory. Anthropic nennt das Feature "Projects", OpenAI "GPTs" und "Projects".¹⁷

Beispiel: Du legst ein Cowork-Projekt für deinen Quartalsabschluss an, lädst die Finanzdaten, deine Mitarbeiter-Listen und das alte Reporting hoch. Jedes Mal wenn du das Projekt öffnest, weiß die KI sofort den Kontext.

Siehe auch: Context Window, Skill, System Prompt

D

FEATURE

Deep Research

Modus oder Feature, bei dem die KI mehrere Minuten oder länger autonom im Web recherchiert, Quellen liest, vergleicht und einen strukturierten Bericht zurückgibt. Im Gegensatz zu einer normalen Chat-Antwort, die in Sekunden kommt.⁴

Beispiel: Du fragst "Wer sind die fünf größten Wettbewerber meines Kunden in der DACH-Region und wie ist deren Preis-Positionierung?". Deep Research klickt sich durch 30 bis 50 Seiten und liefert nach 10 Minuten eine fundierte Übersicht mit Quellen.

Siehe auch: Agent, Tool Use, Reasoning

E

TECHNIK

Embeddings

Zahlen-Repräsentation von Text. Jeder Satz wird in einen Vektor mit hunderten oder tausenden Dimensionen umgerechnet. Texte mit ähnlicher Bedeutung haben ähnliche Vektoren. Grundlage jeder Suche, die "nach Bedeutung" statt "nach Wort" arbeitet. ^{8, 10}

Beispiel: Du hast 10.000 Kundenanfragen in einer Datenbank. Statt nach exakten Wörtern zu suchen, suchst du "alle Anfragen, die Frust über Lieferzeit ausdrücken". Embeddings finden auch Texte, die nicht das Wort "Frust" enthalten.

Siehe auch: Vector Database, RAG, Token

F

TECHNIK

Few-Shot

Prompt-Technik, bei der du dem Modell ein paar Beispiele für die gewünschte Antwort gibst, bevor du die eigentliche Frage stellst. Das Modell ahmt das Muster nach. Gegenteil von Zero-Shot.¹²

Beispiel: Du willst Tickets klassifizieren. Statt zu erklären, gibst du drei Beispiele: "Lieferung kam zu spät" gleich "Logistik". "Rechnung falsch" gleich "Buchhaltung". Dann das echte Ticket. Trefferquote steigt deutlich.

Siehe auch: Zero-Shot, Prompt Engineering, Chain of Thought

TECHNIK

Fine-Tuning

Ein vortrainiertes Modell wird mit deinen eigenen Daten weiter trainiert, damit es einen bestimmten Stil, ein bestimmtes Vokabular oder ein bestimmtes Verhalten lernt. Aufwendig und teuer, oft durch gutes Prompting und RAG ersetzbar.⁹

Beispiel: Eine Anwaltskanzlei trainiert ein Modell auf 20 Jahre eigene Vertragsentwürfe, damit alle KI-generierten Verträge der Hauskonvention folgen. Kosten: typisch fünfstellig, plus laufende Wartung.

Siehe auch: RAG, Prompt Engineering, LLM

G

MODELL

Gemini

Modellfamilie von Google. Tief in Google Workspace integriert, stark bei multimodalen Aufgaben mit Bild und Video. Direkter Wettbewerber zu Claude und GPT.⁶

Beispiel: Du lädst ein Foto deines Lagers in Gemini hoch, fragst "wie viele Paletten zähle ich hier und welche Marken stehen drauf". Gemini sieht das Bild, zählt und antwortet.

Siehe auch: Multimodal, GPT, Claude

KONZEPT

Generative AI (GenAI)

Sammelbegriff für KI-Systeme, die neue Inhalte erzeugen, statt nur zu klassifizieren oder vorherzusagen. Text, Bild, Audio, Video, Code. Alle aktuell gehypten Tools fallen darunter.⁹

Beispiel: ChatGPT schreibt einen Newsletter, Midjourney erzeugt ein Werbebild, ElevenLabs klonst eine Stimme. Alles Generative AI. Klassisches Machine Learning (Spam-Filter, Kreditscoring) ist es nicht.

Siehe auch: LLM, Multimodal, Voice Cloning

MODELL

GPT

Generative Pre-trained Transformer. Modellfamilie von OpenAI, die hinter ChatGPT steht. Aktuell GPT-4, GPT-4o, GPT-5. Faktisch Synonym für "das große Sprachmodell von OpenAI".^{4, 12}

Beispiel: Wenn du in ChatGPT schreibst, redest du mit einem GPT-Modell. Wenn deine Software per API openai.com aufruft, auch.

Siehe auch: OpenAI, LLM, Transformer, Claude

MODELL

Grok

KI-Modell von xAI, dem Unternehmen von Elon Musk. Direkt in X (vormals Twitter) integriert. Positioniert sich als weniger zensiertes Modell mit Echtzeit-Zugang zu X-Daten.⁷

Beispiel: Du willst wissen, was Wettbewerber in den letzten 48 Stunden auf X kommuniziert haben. Grok hat über X direkten Zugriff auf den Live-Feed, andere Modelle nicht.

Siehe auch: xAI, GPT, Claude

H

KONZEPT

Halluzination

Wenn ein KI-Modell etwas behauptet, das sachlich falsch ist, aber überzeugend klingt. Erfundene Zitate, falsche Zahlen, nicht existierende Studien. Das größte praktische Risiko beim Einsatz.²⁰

Beispiel: Du fragst nach einer rechtlichen Norm, das Modell erfindet einen Paragraphen, der nicht existiert. Wenn du das ungeprüft in einer Kundenmail verwendest, hast du ein Problem. Daher: kritische Fakten immer mit Quelle prüfen lassen, RAG einsetzen.

Siehe auch: RAG, Reasoning, LLM

TECHNIK

Inferenz

Der Vorgang, bei dem ein bereits trainiertes Modell eine Anfrage beantwortet. Trainieren ist einmalig und teuer. Inferenz passiert bei jedem einzelnen Aufruf und ist der Kostentreiber im laufenden Betrieb.⁹

Beispiel: Wenn du in Claude eine Frage tippst, läuft Inferenz auf den Servern von Anthropic. Pro 1.000 Token Input und Output rechnen sie ab. Bei API-Nutzung siehst du diese Kosten direkt in der Rechnung.

Siehe auch: API, Latency, Token, Rate Limit

J

TECHNIK

JSON-Mode

Modus, in dem das KI-Modell garantiert strukturiertes JSON zurückgibt statt freien Text. Wichtig, wenn die Antwort direkt von einem anderen Programm weiterverarbeitet werden soll.⁴

Beispiel: Du lässt die KI aus einer eingehenden E-Mail Name, Telefonnummer und Anliegen extrahieren. Im JSON-Mode bekommst du immer das gleiche Schema zurück, dein CRM kann es ohne Fehlerbehandlung übernehmen.

Siehe auch: API, Tool Use, Workflow

L

KONZEPT

Large Language Model (LLM)

Großes Sprachmodell. Eine KI, die auf riesigen Textmengen trainiert wurde und Sprache versteht, generiert und transformiert. Claude, GPT, Gemini, Llama sind alle LLMs.⁸

Beispiel: Wenn jemand sagt "wir bauen das auf einem LLM", meint er meistens Claude, GPT oder Gemini im Hintergrund. Das ist die Schicht, auf der die meisten KI-Anwendungen aktuell aufsetzen.

Siehe auch: Transformer, GPT, Claude, Token

KONZEPT

Latency

Die Zeit zwischen Anfrage und Antwort. Bei KI-Modellen typisch 0,5 bis 30 Sekunden je nach Modell und Aufgabe. Hohe Latency killt User-Experience, niedrige Latency erfordert oft kleinere oder schnellere Modelle.⁴

Beispiel: Bei einem Voice-Bot, der mit Kunden telefoniert, darfst du keine 5 Sekunden Pause haben. Da nimmst du ein schnelleres, kleineres Modell. Bei einem Bericht, der nachts läuft, ist Latency egal.

Siehe auch: Streaming, Inferenz, Sonnet

MODELL

Llama

Open-Source-Modellfamilie von Meta. Du kannst Llama-Modelle herunterladen und auf eigenen Servern betreiben. Wichtig für Unternehmen mit hohen Datenschutz-Anforderungen oder Cloud-Verbot.¹⁴

Beispiel: Eine Klinik will Patientendaten nicht zu OpenAI schicken. Sie hostet Llama auf einem eigenen Server, läuft alle KI-Anfragen lokal. Datenschutz-konform, aber höherer Aufwand für Betrieb und Updates.

Siehe auch: Open Source, Mixtral, LLM

M

TECHNIK

MCP (Model Context Protocol)

Offener Standard von Anthropic, mit dem KI-Modelle einheitlich auf externe Tools, Datenbanken und Services zugreifen. Vergleichbar mit USB-C für KI. Wird von immer mehr Anbietern unterstützt.³

Beispiel: Du installierst einen MCP-Server für deinen Google Drive. Ab dem Moment kann Claude direkt auf alle deine Drive-Dokumente zugreifen, sie lesen, durchsuchen, bearbeiten. Ohne dass du Code schreibst.

Siehe auch: Tool Use, Claude Code, Skill

MODELL

Mixtral

Open-Weight-Modellfamilie von Mistral AI aus Frankreich. Bekannt für gute Performance bei kleiner Größe und europäischer DSGVO-Compliance. Kann selbst gehostet werden.¹⁵

Beispiel: Ein europäischer Mittelständler will KI einsetzen, aber keine US-Anbieter nutzen. Mixtral läuft auf eigenen Servern in Deutschland, Datenflüsse bleiben in der EU.

Siehe auch: Llama, Open Source, LLM

KONZEPT

Multimodal

Ein Modell, das mehr als nur Text versteht. Bilder, Audio, Video, PDFs. Moderne Modelle wie Claude, GPT-4 und Gemini sind multimodal, ältere reine Text-Modelle nicht.⁶

Beispiel: Du fotografierst eine handgeschriebene Notiz aus einem Meeting, schickst sie an Claude. Claude liest die Handschrift, fasst die Notiz zusammen und erstellt To-Dos. Reine Text-Modelle könnten das nicht.

Siehe auch: Gemini, Claude, Voice Cloning, Whisper

N

TOOL

n8n

Open-Source-Automatisierungs-Plattform, vergleichbar mit Zapier, aber selbst hostbar. Visuelle Workflow-Engine, in der du Trigger, Aktionen und KI-Aufrufe per Drag-and-Drop verbindest.¹⁶

Beispiel: Eingehende E-Mail mit Anhang triggert einen n8n-Workflow. Der Workflow schickt den Anhang an Claude zur Analyse, schreibt das Ergebnis in Notion, benachrichtigt dein Team per Slack. Alles ohne eine Zeile Code.

Siehe auch: Workflow, Agent, MCP

O

KONZEPT

Open Source

Bei KI-Modellen: die Modell-Gewichte sind frei verfügbar, du kannst das Modell herunterladen, anpassen und auf eigener Hardware laufen lassen. Gegenteil sind "Closed Source"-Modelle wie GPT-4 und Claude, die nur per API nutzbar sind.¹⁴

Beispiel: Llama, Mixtral, DeepSeek sind Open Source. Du kannst sie kostenlos nutzen, musst aber die Rechenleistung selbst stellen. Anthropic und OpenAI sind Closed Source, du zahlst pro Aufruf, dafür kein Infrastruktur-Aufwand.

Siehe auch: Llama, Mixtral, Anthropic, OpenAI

ANBIETER

OpenAI

KI-Unternehmen aus San Francisco. Hersteller von ChatGPT, der GPT-Modellfamilie, Whisper und DALL-E. Größter Player im KI-Markt nach Bewertung und Nutzerzahl.⁴

Beispiel: Wenn dein Team ChatGPT Plus oder Enterprise nutzt, ist OpenAI dein Anbieter. Wenn dein Entwickler eine eigene App mit GPT-4 baut, auch.

Siehe auch: GPT, Whisper, Anthropic

P

KONZEPT

Prompt

Die Eingabe, die du dem KI-Modell gibst. Frage, Anweisung, Kontext, Beispiele, alles zusammen. Qualität des Prompts entscheidet zu 80 Prozent über Qualität der Antwort.¹²

Beispiel: "Fass das zusammen" ist ein schlechter Prompt. "Fass diesen Text in 5 Bulletpoints für meinen CFO zusammen, Fokus auf finanzielle Risiken, max. 200 Wörter" ist ein guter.

Siehe auch: Prompt Engineering, System Prompt, Few-Shot

KONZEPT

Prompt Engineering

Die Disziplin, Prompts so zu formulieren, dass sie konsistent gute Ergebnisse liefern. Mischung aus klarer Anweisung, gutem Kontext, Beispielen und Struktur-Vorgaben.¹²

Beispiel: Statt jedes Mal neu zu formulieren, baust du eine Prompt-Vorlage für "Angebot schreiben". Mit festen Sektionen, Tonalität, Beispiel-Outputs. Dein Team füllt nur noch die Variablen. Reproduzierbare Qualität.

Siehe auch: Prompt, System Prompt, Chain of Thought, Few-Shot

R

TECHNIK

RAG (Retrieval Augmented Generation)

Technik, bei der die KI vor der Antwort erst in einer eigenen Datenbank nachschaut, relevante Stellen findet und nur diese als Kontext nutzt. Macht Antworten faktenbasiert und reduziert Halluzinationen.¹⁰

Beispiel: Du baust einen Chatbot, der Fragen zu deinen 500 internen Richtlinien beantwortet. Statt alle Richtlinien jedes Mal in den Prompt zu packen, durchsucht RAG die Datenbank, holt die 3 relevanten Absätze, antwortet auf deren Basis.

Siehe auch: Embeddings, Vector Database, Halluzination

KONZEPT

Rate Limit

Obergrenze, wie viele Anfragen oder wie viele Token pro Minute oder Tag dein API-Account verarbeiten darf. Schützt den Anbieter vor Überlast, kann aber dein Produkt blockieren.^{1, 4}

Beispiel: Du baust einen Massen-Versand mit personalisierten Mails. Bei 100.000 Empfängern rennst du in das Rate Limit, der Anbieter blockt. Lösung: Tier upgraden, Anfragen über mehrere Stunden verteilen oder Batch-API nutzen.

Siehe auch: API, Inferenz, Latency

KONZEPT

Reasoning

Fähigkeit eines Modells, schlussfolgernd zu denken, nicht nur zu antworten. Bei modernen Reasoning-Modellen (GPT o3, Claude mit Extended Thinking, DeepSeek R1) denkt das Modell sichtbar länger nach, bevor es antwortet.¹¹

Beispiel: Du legst dem Modell eine komplexe Lieferkette mit fünf Engpässen vor und fragst nach dem optimalen Bestand. Ein Standard-Modell rät. Ein Reasoning-Modell rechnet, vergleicht Szenarien, kommt nach 30 Sekunden mit einer durchdachten Antwort.

Siehe auch: Chain of Thought, Deep Research, LLM

S

TECHNIK

Sampling

Der Vorgang, bei dem das Modell Wort für Wort entscheidet, was es als nächstes ausgibt.

Parameter wie Temperature und Top-P steuern, wie kreativ oder konservativ es dabei vorgeht.⁴

Beispiel: Beim Schreiben eines Werbetextes willst du mehr Kreativität, also höhere Temperature. Bei der Extraktion von Daten aus einer Rechnung willst du maximale Präzision, also Temperature auf 0. Sampling-Einstellung macht den Unterschied.

Siehe auch: Temperature, Token, Streaming

FEATURE

Skill

Wiederverwendbares Paket aus Anweisungen, Beispielen und Tools, das einem KI-Modell beigebracht wird, eine bestimmte Aufgabe besonders gut zu erledigen. In Claude Code und ChatGPT als Feature umgesetzt.²

Beispiel: Du baust einen "Angebots-Skill". Er kennt deine Preise, dein CI, deine Vertrags-Klauseln. Jedes Teammitglied kann ihn aufrufen, bekommt sofort ein druckreifes Angebot in deiner Hauskonvention.

Siehe auch: Claude Code, MCP, System Prompt, Subagent

MODELL

Sonnet

Mittlere Modellvariante in der Claude-Familie, zwischen dem schnellen Haiku und dem starken Opus. Im Alltag der häufigste Claude-Modus, weil er den besten Trade-off zwischen Geschwindigkeit, Kosten und Qualität bietet.¹

Beispiel: Für Standard-Aufgaben wie E-Mails schreiben, Texte zusammenfassen, einfache Analysen nimmst du Sonnet. Erst bei komplexen Aufgaben mit viel Reasoning lohnt sich das teurere Opus.

Siehe auch: Claude, Anthropic, Latency

TECHNIK

Streaming

Die KI-Antwort wird Token für Token sofort ausgespielt, statt am Ende auf einmal. Macht die wahrgenommene Wartezeit kürzer und das Nutzererlebnis flüssiger.^{1, 4}

Beispiel: In ChatGPT siehst du, wie der Text Wort für Wort erscheint. Das ist Streaming. Bei einer Batch-Analyse über Nacht brauchst du es nicht, bei einem Live-Chatbot ist es Pflicht.

Siehe auch: Latency, API, Token

FEATURE

Subagent

Ein zusätzlicher KI-Agent, den ein Haupt-Agent für Teilaufgaben aufruft. Erlaubt parallele Bearbeitung, klare Aufgaben-Trennung und sauberes Kontext-Management bei komplexen Workflows.^{2, 19}

Beispiel: Dein Haupt-Agent bekommt die Aufgabe, einen Markt-Report zu schreiben. Er startet einen Subagent für Wettbewerbs-Recherche, einen für Zahlen-Aufbereitung, einen für Schreiben. Jeder Subagent hat eigenen Kontext und arbeitet parallel.

Siehe auch: Agent, Claude Code, Workflow

TECHNIK

System Prompt

Anweisung, die zu Beginn jeder Konversation an das Modell geht und Rolle, Tonalität, Regeln und Format vorgibt. Vom Nutzer meistens nicht sichtbar, aber prägend für jede Antwort.^{1, 4}

Beispiel: Du baust einen Kundenservice-Bot. System Prompt: "Du bist Mitarbeiter von Firma X. Antworte in Du-Form. Maximal 100 Wörter. Wenn du etwas nicht weißt, leite an einen Menschen weiter." Damit hast du Verhalten zementiert.

Siehe auch: Prompt, Prompt Engineering, Skill, Cowork

T

TECHNIK

Temperature

Parameter zwischen 0 und 2, der steuert, wie kreativ oder berechenbar ein Modell antwortet. 0 heißt: bei gleicher Frage immer fast die gleiche Antwort. 1 heißt: variabler, kreativer. Über 1 wird es chaotisch.⁴

Beispiel: Für Datenextraktion aus Rechnungen Temperature auf 0, damit Beträge nicht halluziniert werden. Für Brainstorming einer neuen Kampagne Temperature auf 0,8, damit die Vorschläge nicht alle gleich klingen.

Siehe auch: Sampling, Halluzination, API

KONZEPT

Token

Die kleinste Einheit, in die ein LLM Text zerlegt. Ein Wort kann ein Token sein, manchmal sind es zwei oder drei. Faustregel: 1.000 Token entsprechen circa 750 deutschen Wörtern. Kosten und Limits werden in Token gerechnet.⁸

Beispiel: Dein Vertrag hat 20 Seiten, circa 8.000 Wörter, circa 10.000 Token. Bei einem Modell mit 200.000 Token Kontext passt der Vertrag locker rein, plus 95 Prozent freier Raum für Analyse.

Siehe auch: Context Window, Inferenz, LLM

TECHNIK

Tool Use (Function Calling)

Fähigkeit eines Modells, externe Tools selbständig aufzurufen. Statt nur Text zu generieren, ruft das Modell eine Funktion auf, zum Beispiel "such in der Datenbank", "schicke E-Mail", "ruf Wetter-API". Voraussetzung für jeden echten Agent.^{1, 4, 19}

Beispiel: Du fragst den Agent "wie ist das Wetter morgen in Berlin und ist mein Outdoor-Event-Slot frei?". Tool Use sorgt dafür, dass er die Wetter-API und deinen Kalender aufruft, statt zu raten.

Siehe auch: Agent, MCP, JSON-Mode, API

Transformer

Neuronale Netz-Architektur, die 2017 von Google publiziert wurde und seitdem das Fundament aller großen Sprachmodelle ist. Das "T" in GPT und in BERT. Kernidee: das Modell schaut auf alle Wörter eines Textes gleichzeitig statt nacheinander.¹³

Beispiel: Wenn dir jemand erklärt "wir trainieren eine eigene Transformer-Architektur", redet er von der zugrundeliegenden Bauweise. Für dich als Geschäftsführer relevant: praktisch jedes moderne KI-Sprachmodell ist ein Transformer.

Siehe auch: LLM, GPT, Fine-Tuning

V

TOOL

Vector Database

Spezialisierte Datenbank, die Embeddings speichert und nach Ähnlichkeit durchsucht. Pinecone, Weaviate, Qdrant, ChromaDB sind verbreitet. Rückgrat jedes RAG-Systems.¹⁰

Beispiel: Deine 50.000 Support-Tickets liegen in einer Vector Database. Bei einer neuen Anfrage findet die DB in Millisekunden die 10 ähnlichsten alten Fälle, die KI nutzt sie als Basis für die Antwort.

Siehe auch: RAG, Embeddings

KONZEPT

Vibe Coding

Begriff, der 2025 populär wurde. Beschreibt das Bauen von Software, indem du der KI in natürlicher Sprache sagst, was du willst, statt selbst Code zu schreiben. Die KI generiert, du testest, korrigierst per Sprache.²

Beispiel: Du sagst Claude Code "bau mir ein Dashboard, das meine Verkäufe nach Region zeigt, mit Filter nach Monat". Eine Stunde später hast du eine funktionierende Web-App. Du hast nie eine Zeile Code geschrieben.

Siehe auch: Claude Code, Agent, Skill

KONZEPT

Voice Cloning

Technik, bei der aus wenigen Minuten Sprachaufnahme eine künstliche Stimme erzeugt wird, die wie die Original-Person klingt. Anbieter: ElevenLabs, OpenAI, Play.ht.¹⁸

Beispiel: Du klonst deine eigene Stimme einmal. Danach lässt du alle deine Skript-Texte, Newsletter-Intros und Video-Voiceovers von deiner KI-Stimme einsprechen, in mehreren Sprachen. Zeitersparnis enorm, ethisch sauberer Einsatz vorausgesetzt.

Siehe auch: Multimodal, Whisper, Generative AI

W

MODELL

Whisper

Speech-to-Text-Modell von OpenAI. Wandelt gesprochene Sprache in Text um, unterstützt über 90 Sprachen, sehr robust auch bei schlechter Audio-Qualität. Open Source und per API nutzbar.⁵

Beispiel: Du nimmst dein Vertriebs-Meeting auf, schickst die Aufnahme an Whisper, bekommst ein sauberes Transkript zurück. Claude oder GPT fasst es danach in 5 Bulletpoints zusammen, schickt Action-Items per Mail.

Siehe auch: OpenAI, Multimodal, Voice Cloning

KONZEPT

Workflow

Eine vordefinierte Abfolge von Schritten, bei der mehrere Tools, Daten und oft auch KI-Aufrufe automatisch hintereinander ausgeführt werden. Unterschied zum Agent: Workflow folgt einem festen Plan, Agent entscheidet selbst.^{16, 19}

Beispiel: Workflow "Neuer Lead". Trigger: Formular ausgefüllt. Schritt 1: Lead in CRM anlegen. Schritt 2: KI schreibt Begrüßungsmail. Schritt 3: Termin-Link senden. Schritt 4: Slack-Notification. Läuft jedes Mal exakt gleich.

Siehe auch: n8n, Agent, Tool Use, Subagent

X

ANBIETER

xAI

KI-Unternehmen von Elon Musk, gegründet 2023. Hersteller des Grok-Modells. Direkt mit dem sozialen Netzwerk X verzahnt, dadurch Zugriff auf Echtzeit-Social-Daten.⁷

Beispiel: Wenn dein Marketing-Team einen Pulse-Check zu einer Marken-Krise braucht, ist Grok von xAI durch den X-Live-Feed oft am schnellsten am Puls. Andere Modelle haben diesen Daten-Zugang nicht.

Siehe auch: Grok, OpenAI, Anthropic

Z

TECHNIK

Zero-Shot

Prompt-Technik, bei der du dem Modell ohne Beispiele direkt eine Aufgabe gibst. "Klassifiziere diese Mail als Reklamation oder Anfrage." Moderne Modelle sind oft schon ohne Beispiele gut. Gegenteil von Few-Shot.¹²

Beispiel: Du willst schnell testen, ob ein Modell deine Aufgabe versteht, ohne erst Beispiele zu kuratieren. Du gibst die Anweisung in einem Satz und schaust, wie gut das Ergebnis ist. Wenn schwach, baust du Few-Shot auf.

Siehe auch: Few-Shot, Prompt, Chain of Thought

Alle 50 Begriffe auf einen Blick

Alphabetisch sortiert, für den schnellen Nachschlag.

1. Agent
2. Anthropic
3. API
4. Chain of Thought (CoT)
5. Claude
6. Claude Code
7. Context Window
8. Cowork
9. Deep Research
10. Embeddings
11. Few-Shot
12. Fine-Tuning
13. Gemini
14. Generative AI (GenAI)
15. GPT
16. Grok
17. Halluzination
18. Inferenz
19. JSON-Mode
20. Large Language Model (LLM)
21. Latency
22. Llama
23. MCP (Model Context Protocol)
24. Mixtral
25. Multimodal
26. n8n
27. Open Source
28. OpenAI
29. Prompt
30. Prompt Engineering
31. RAG (Retrieval Augmented Generation)
32. Rate Limit

- 33. Reasoning
- 34. Sampling
- 35. Skill
- 36. Sonnet
- 37. Streaming
- 38. Subagent
- 39. System Prompt
- 40. Temperature
- 41. Token
- 42. Tool Use (Function Calling)
- 43. Transformer
- 44. Vector Database
- 45. Vibe Coding
- 46. Voice Cloning
- 47. Whisper
- 48. Workflow
- 49. xAI
- 50. Zero-Shot

Quellen

Alle Definitionen sind an die offiziellen Docs und Primär-Publikationen der Hersteller angelehnt. Wikipedia wurde bewusst nicht zitiert, stattdessen wurde auf die jeweilige Primärquelle verwiesen.

1. Anthropic · Claude Documentation · docs.anthropic.com/en/docs/intro
2. Anthropic · Claude Code Documentation · docs.claude.com/en/docs/claude-code
3. Anthropic · Model Context Protocol · modelcontextprotocol.io
4. OpenAI · Platform Documentation · platform.openai.com/docs
5. OpenAI · Whisper Research · openai.com/index/whisper
6. Google · Gemini API Documentation · ai.google.dev/gemini-api/docs
7. xAI · API Documentation · docs.x.ai/docs/overview
8. Hugging Face · NLP Course · huggingface.co/learn/nlp-course
9. NVIDIA · Glossary of AI Terms · nvidia.com/en-us/glossary
10. IBM · What is Retrieval Augmented Generation · ibm.com/topics/retrieval-augmented-generation
11. Wei et al. · Chain-of-Thought Prompting Elicits Reasoning in LLMs · arxiv.org/abs/2201.11903
12. Brown et al. · Language Models are Few-Shot Learners · arxiv.org/abs/2005.14165
13. Vaswani et al. · Attention Is All You Need · arxiv.org/abs/1706.03762
14. Meta · Llama Models · llama.com
15. Mistral AI · Models Overview · docs.mistral.ai
16. n8n · Documentation · docs.n8n.io
17. Anthropic · Projects (Memory & Context) · anthropic.com/news/projects
18. ElevenLabs · Voice Cloning Documentation · elevenlabs.io/docs/product-guides
19. Anthropic · Building Effective Agents · anthropic.com/engineering/building-effective-agents
20. OpenAI · Cookbook · Evaluation · cookbook.openai.com

MEHR DAVON? DIENSTAGS PER MAIL.

Mehr davon? Dienstags per Mail.

Jede Woche eine kurze Mail: was bei meinen Kunden funktioniert, welche neuen Skills dazukommen, ohne Bullshit.

moritzmaaker.com

Trag dich einmal ein, kriegst direkten Zugang zu allen Skills. Kein Spam, jederzeit abmelden.