Technisches KI-Glossarium

Die Begriffe der KI-Welt, ohne Fachhuberei erklärt

August 2026

Inhalt

Machine Learning
MCP – Model Context Protocol
Model
Multimodal
Neural Network
NLP – Natural Language Processing
Open Source / Open Weights
Parameter / Weights
Plugin
Prompt
Prompt Engineering
Prompt Injection
RAG – Retrieval-Augmented Generation
Reasoning Model
Reinforcement Learning
Repository
RLHF – Reinforcement Learning from Human Feedback
Sandbox
Session
Skill
Slack
SLM – Small Language Model
Slug
Snippet
Supervised / Unsupervised Learning
System Prompt
Temperature
Token
Training
Training Data
Transformer
Turing Test
Vector Database
Vibe Coding
Anhang: Wörter aus der Bildwerkstatt
CFG Scale / Steps
Checkpoint
ControlNet
img2img – Image-to-Image
Inpainting / Outpainting
LoRA – Low-Rank Adaptation
Negative Prompt
Seed
Upscaling

Vorbemerkung

Die Welt, aus der diese Wörter stammen, spricht Englisch. Wer über künstliche Intelligenz reden hört, bekommt deshalb fast immer den englischen Ausdruck zu hören, selbst wenn ein deutscher

danebenstünde – und genau daran hakt das Verstehen, nicht am Gegenstand. Dieses Verzeichnis führt deshalb jeden Begriff unter dem Namen, unter dem er einem tatsächlich begegnet, und setzt die deutsche Entsprechung dahinter. Wo es kein brauchbares deutsches Wort gibt, steht eine Umschreibung; wo das deutsche Wort das gebräuchlichere ist, steht es dabei. Erklärt wird in gewöhnlicher Sprache und mit Bildern aus dem Alltag – so weit, dass man mitreden und nachfragen kann, und nicht weiter.

Agent (deutsch etwa: der Beauftragte, der Handelnde)
Ein Sprachmodell, das nicht nur antwortet, sondern handelt: Es darf Werkzeuge benutzen – Dateien lesen und schreiben, im Netz suchen, ein Programm starten – und arbeitet einen Auftrag in mehreren Schritten selbständig ab. Der Unterschied zum bloßen Beantworten ist der zwischen einem Auskunftsschalter und einem Handwerker: Der eine sagt Ihnen, wie es ginge, der andere geht hin und macht es. Dabei prüft ein Agent zwischendurch, was seine Handlung bewirkt hat, und richtet sich danach. Was er anfassen darf, ist vorher festgelegt; alles andere bleibt ihm verschlossen.

AGI – Artificial General Intelligence (deutsch: allgemeine künstliche Intelligenz)
Die Vorstellung einer Maschine, die nicht nur einzelne Aufgaben beherrscht, sondern in allem mithält, was ein Mensch geistig tun kann – und sich in Neues ebenso einarbeitet wie ein Mensch. Ob es sie je geben wird, ob wir uns ihr nähern, ob der Begriff überhaupt scharf genug ist, um etwas zu behaupten: darüber wird heftig gestritten, auch unter denen, die daran arbeiten. Wer AGI sagt, sagt selten dasselbe wie sein Gegenüber. Man tut gut daran, jedes Mal nachzufragen, was genau gemeint ist.

AI – Artificial Intelligence (deutsch: KI, künstliche Intelligenz)
Der Oberbegriff für Verfahren, die Leistungen erbringen, für die man beim Menschen Verstand voraussetzt – erkennen, ordnen, schließen, sprechen. Er ist alt und dehnbar, und was gestern als KI galt, heißt heute Rechtschreibprüfung. Der Begriff verschiebt sich also stets mit dem, was gerade neu ist. Wer heute KI sagt, meint fast immer ein großes Sprachmodell oder einen Bildgenerator.

AI Slop (deutsch etwa: KI-Schlamm, Maschinenschlick)
Ein junges Schimpfwort für die Flut billig erzeugter Bilder und Texte, die Netz und Auslagen füllen – nicht falsch, nicht anstößig, nur beliebig. Die Sorge ist eine doppelte: dass das Auge abstumpft, und dass künftige Modelle wieder mit diesem Stoff trainiert werden und dabei matter geraten; letzteres nennt man Model Collapse. Für alle, die etwas gemacht haben, worauf es ankommt, ist das der Grund, warum Herkunft und Handschrift wieder an Gewicht gewinnen.

Algorithm (deutsch: Algorithmus)
Eine Handlungsvorschrift: endlich viele Schritte, die aus einer Eingabe ein Ergebnis machen – ein Kochrezept, in Regeln gefasst. Der Begriff ist weit älter als die Rechner und hat mit KI zunächst nichts zu tun. In der öffentlichen Rede dient er oft als Ersatzwort für alles Undurchschaute: der Algorithmus zeigt mir das. Genau darin liegt die Verwechslung, denn ein Sprachmodell ist keine Vorschrift, sondern etwas Gelerntes.

Alignment (deutsch etwa: Ausrichtung, Übereinstimmung)
Die Bemühung, ein Modell so zu bilden, dass es tut, was gemeint war, und nicht bloß, was wörtlich dastand – und dass es Hilfsbereitschaft, Ehrlichkeit und Rücksicht zusammenhält, auch wo sie einander widerstreben. Der Ausdruck stammt aus der Vorstellung zweier Achsen, die man zur Deckung bringt: die Absicht des Menschen und das Verhalten der Maschine. Praktisch geschieht

das in der zweiten Ausbildungsstufe nach dem eigentlichen Training, durch Bewertung, Regelwerke und viel Nacharbeit. Es ist keine Einstellung, die man setzt, sondern eine Erziehung, die nie ganz abgeschlossen ist.

Anthropomorphism (deutsch: Vermenschlichung)
Die fast unvermeidliche Neigung, einer Maschine, die flüssig spricht, Absicht, Gefühl und Verstehen zuzuschreiben. Sie ist begreiflich, denn solche Sätze kennen wir nur von Menschen. Sie führt aber in zwei Richtungen in die Irre: Man traut dem Modell zu viel zu, wo es bloß fortsetzt, und zu wenig, wo es ungewohnt gut ist. Nüchtern zu bleiben heißt nicht, die Sache geringzuschätzen.

API – Application Programming Interface (deutsch: Programmierschnittstelle, meist kurz: Schnittstelle)
Die Stelle, an der zwei Programme unmittelbar miteinander sprechen. Kein Bildschirm, keine Knöpfe: Das eine stellt eine Anfrage nach festen Regeln, das andere antwortet nach denselben Regeln. Auf Ihrer Webseite geschieht das ständig – wenn ich einen Beitrag ändere, gehe ich nicht durch die Verwaltungsmaske, sondern über die Schnittstelle von WordPress. Das ist schneller und, weil die Regeln streng sind, weniger fehleranfällig als Klicken.

Backup (deutsch: Sicherung, Sicherheitskopie)
Eine vollständige Kopie eines Standes, die man beiseitelegt, bevor man etwas Größeres verändert. Bei einer Webseite gehören Dateien und Datenbank zusammen; fehlt eines von beiden, hilft die Kopie wenig. Der Wert einer Sicherung zeigt sich nicht beim Anlegen, sondern beim Zurückspielen – deshalb ist eine Sicherung, die nie erprobt wurde, streng genommen nur eine Hoffnung. Vor jedem größeren Eingriff lohnt der Blick, wie alt die jüngste ist.

Benchmark (deutsch etwa: Maßstab, Vergleichsprüfung)
Eine standardisierte Aufgabensammlung, an der man Modelle miteinander vergleicht – Rechenaufgaben, Textverständnis, Programmieraufgaben. Der Reiz liegt darin, dass am Ende eine Zahl steht; die Schwäche liegt darin, dass eine Zahl wenig über den Alltag sagt. Kommen die Aufgaben im Training vor, misst man Auswendiggelerntes statt Können. Ein Benchmark ist eine Momentaufnahme unter Prüfungsbedingungen, nicht ein Zeugnis über die Arbeit.

Bias (deutsch: Schlagseite, Verzerrung, Voreingenommenheit)
Die systematische Schieflage, die ein Modell aus seinem Übungsstoff mitbringt. Es hat gelernt, was in den Texten häufig war, und gibt es bevorzugt zurück – mit allen Einseitigkeiten, die dort schon steckten, nach Sprache, Herkunft, Zeit, Geschlecht, Fach. Das geschieht ohne Absicht und ist gerade deshalb schwer zu bemerken: Die Schieflage klingt wie die Normalität. Man kann sie mindern, aber nicht restlos herausrechnen.

Black Box (deutsch: schwarzer Kasten)
Ein System, dessen Inneres man nicht einsehen kann: Man kennt, was hineingeht, und was herauskommt, aber nicht den Weg dazwischen. Bei großen Modellen gilt das selbst für ihre Erbauer – man kann jede einzelne Rechnung nachvollziehen und dennoch nicht sagen, warum gerade dieser Satz kam. Daran arbeitet ein eigener Forschungszweig. Bis auf Weiteres gilt: Man beurteilt das Ergebnis, nicht die Begründung.

Bug (deutsch: Fehler; wörtlich: Käfer)
Ein Fehler in einem Programm – die Sache tut etwas anderes, als sie sollte. Das Wort stammt aus der Frühzeit der Rechner, als man in einem Gerät tatsächlich eine Motte fand. Ein Fehler ist nicht dasselbe wie ein Absturz: Die meisten sind still und zeigen sich erst in einem seltenen Fall. Deshalb ist das Suchen aufwendiger als das Beheben.

C2PA / Watermarking (deutsch etwa: Herkunftsnachweis, Wasserzeichen)
Verfahren, mit denen einem Bild oder einer Datei mitgegeben wird, woher sie stammt und was mit ihr geschah – als unsichtbare Markierung oder als beigelegter, kryptografisch gesicherter Herkunftsvermerk. C2PA heißt die Vereinbarung, an der Kamerahersteller, Verlage und Softwarehäuser dafür arbeiten. Der Zweck ist weniger, erzeugte Bilder zu brandmarken, als echten Aufnahmen einen nachprüfbaren Lebenslauf zu geben. Für Künstler und Archive ist das der interessantere Weg.

Cache (deutsch: Zwischenspeicher)
Ein Vorrat bereits fertiger Antworten, damit dieselbe Arbeit nicht zweimal getan werden muss. Eine Webseite hält so ganze Seiten bereit und liefert sie beim nächsten Besucher sofort aus. Das ist der Grund für ein häufiges Ärgernis: Man hat etwas geändert, sieht aber noch das Alte – nicht weil die Änderung fehlschlug, sondern weil noch die aufbewahrte Fassung ausgeliefert wird. Abhilfe schafft das Leeren des Zwischenspeichers oder ein Neuladen unter Umgehung desselben.

Chain of Thought (deutsch: Gedankenkette; auch: Reasoning, Nachdenken)
Die Anweisung oder Fähigkeit, vor der Antwort erst laut zu überlegen – Schritt für Schritt, statt sofort ein Ergebnis hinzusetzen. Bei mehrstufigen Aufgaben verbessert das die Trefferquote deutlich, weil jeder Zwischenschritt den nächsten trägt. Neuere Modelle tun das von sich aus und zeigen den Gedankengang je nach Programm ganz, gekürzt oder gar nicht. Man sollte ihn nicht für ein Protokoll des inneren Vorgangs halten: Er ist selbst wieder erzeugter Text.

Chatbot (deutsch etwa: Gesprächsprogramm)
Die vertraute Gestalt, in der einem ein Sprachmodell begegnet: ein Fenster, in das man schreibt, und eine Antwort, die zurückkommt. Der Chatbot ist nicht das Modell, sondern nur seine Bedienung – ein Fenster zum Haus, nicht das Haus. Dasselbe Modell kann hinter einem Gesprächsfenster stecken, in einem Programm arbeiten oder als Agent selbständig Aufträge ausführen.

Cloud (deutsch: die Wolke; gemeint sind fremde Rechner)
Kein Ort in der Luft, sondern Rechner in Hallen, die jemand anderes betreibt und die man mietet statt kauft. Der Vorteil ist, dass man weder Gerät noch Wartung im Haus hat und Leistung nach Bedarf zubucht. Der Preis ist die Abhängigkeit: Man vertraut seine Daten und seine Erreichbarkeit einem Betreiber an. Ihre Webseite liegt so bei einem Anbieter, und auch die Modelle, mit denen Sie sprechen, rechnen dort.

Computer Vision (deutsch: Bilderkennung, maschinelles Sehen)
Der Zweig, der Bilder auswertet, statt sie hervorzubringen: Was ist darauf zu sehen, wo, wie viele. Er ist erheblich älter als die Bildgeneratoren und steckt in Gesichtserkennung, Fahrassistenz, medizinischer Auswertung, Sortieranlagen. Zusammen mit der Gegenrichtung – aus Text ein Bild – gehört er zu dem, was heutige Modelle mehrkanalig macht.

Context Window (deutsch: Kontextfenster)
So viel Text, wie ein Modell in einem Arbeitsgang überschauen kann – Ihre Frage, seine bisherigen Antworten, angehängte Dokumente, alles zusammen. Man kann es sich als Schreibtisch denken: Was darauf liegt, ist im Blick; was nicht mehr daraufpasst, wandert über den Rand. Ist das Fenster voll, muss ältere Unterhaltung zusammengefasst oder weggelassen werden, und das Modell wirkt dann vergesslich. Gemessen wird es nicht in Seiten, sondern in Token.

Deep Learning (deutsch: tiefes Lernen)
Maschinelles Lernen mit vielschichtigen neuronalen Netzen – tief heißt schlicht: viele Schichten hintereinander. Jede Schicht bildet aus dem, was die vorige liefert, etwas Allgemeineres; aus Kanten werden Formen, aus Formen Gegenstände, aus Wortstücken Bedeutungen. Das Verfahren ist alt, brauchbar wurde es erst, als Rechenkraft und Datenmengen groß genug waren. Alle heutigen Sprachmodelle stehen in dieser Linie.

Deepfake (deutsch etwa: tiefe Fälschung)
Ein erzeugtes Bild, Tonstück oder Film, das eine wirkliche Person etwas sagen oder tun lässt, was sie nie gesagt oder getan hat. Das Wort setzt sich aus tiefem Lernen und Fälschung zusammen. Die Machart ist billig geworden, und das Auge ist kein verlässlicher Prüfstein mehr. In der öffentlichen Auseinandersetzung ist dies der Begriff, unter dem KI am häufigsten verhandelt wird.

Deployment (deutsch etwa: Inbetriebnahme, Ausrollen)
Der Schritt, in dem etwas Fertiges vom Arbeitsplatz auf den laufenden Betrieb übergeht und für alle sichtbar wird. Bei einer Webseite ist das der Augenblick, in dem eine Änderung von der Probefassung auf die öffentliche wandert. Er ist heikel, weil das Ergebnis nun echte Besucher trifft; deshalb steht davor die Sicherung und danach die Kontrolle.

Diffusion Model / Text-to-Image (deutsch: Diffusionsmodell, Text zu Bild)
Die Bauart hinter den bekannten Bildgeneratoren, und sie arbeitet ganz anders als ein Sprachmodell. Man beginnt mit reinem Rauschen und nimmt in vielen Schritten Rauschen fort, bis ein Bild dasteht – geleitet von der Beschreibung, die man eingegeben hat. Gelernt wurde das umgekehrt, an Millionen Bildern, denen man schrittweise Rauschen zusetzte. Es ist kein Zusammenkleben vorhandener Bilder, aber auch nichts, was ohne diese Bilder entstanden wäre; daran hängt der Urheberrechtsstreit.

Embedding (deutsch etwa: Einbettung, Bedeutungskoordinate)
Ein Verfahren, das Wörter, Sätze oder ganze Texte in lange Zahlenreihen übersetzt, so dass Ähnliches nahe beieinander zu liegen kommt. Zwei Sätze mit verwandtem Sinn ergeben verwandte Zahlenreihen, auch wenn sie kein Wort gemeinsam haben. Darauf beruht die sinngemäße Suche: Man findet den Absatz über Grundierung, ohne das Wort Grundierung eingetippt zu haben. Es ist die technische Grundlage von RAG und der meisten heutigen Suchen im eigenen Bestand.

EU AI Act (deutsch: KI-Verordnung der Europäischen Union)
Das erste umfassende Gesetzeswerk zur künstlichen Intelligenz, 2024 beschlossen und stufenweise in Kraft tretend. Es ordnet Anwendungen nach Risiko: Einiges ist untersagt, Hochrisiko-Anwendungen unterliegen Auflagen, für den gewöhnlichen Gebrauch gelten vor allem Kennzeichnungspflichten. Was es für einen Künstler bedeutet, ist rasch gesagt: Erzeugtes soll als solches erkennbar sein.

Few-Shot / Zero-Shot (deutsch etwa: mit wenigen Beispielen / ohne Beispiel)
Zwei Arten, einem Modell eine Aufgabe beizubringen, ohne es umzubauen. Zero-Shot heißt: Man beschreibt die Aufgabe und lässt es machen. Few-Shot heißt: Man legt zwei, drei fertige Beispiele bei, an denen es Muster und Ton abliest. Für alles Formgebundene – eine bestimmte Schreibweise, ein bestimmter Aufbau – sind gute Beispiele fast immer wirksamer als die längste Erklärung.

Fine-Tuning (deutsch: Feinabstimmung, Nachschulung)
Ein fertiges Modell wird mit einer vergleichsweise kleinen, eigenen Sammlung von Beispielen nachgeschult, damit es einen bestimmten Ton, ein Fachgebiet oder ein festes Format sicher trifft. Es ist keine Neuerziehung, sondern eine Nachjustierung: Die Gewichte werden ein wenig verschoben, nicht neu gesetzt. Für die meisten Zwecke reicht heute ein gut gebauter Prompt oder ein Zugriff auf die eigenen Unterlagen; die Nachschulung lohnt erst, wenn beides an Grenzen stößt.

Foundation Model (deutsch: Basismodell, Grundmodell)
Ein sehr großes, allgemein trainiertes Modell, auf dem viele einzelne Anwendungen aufsetzen, statt jeweils bei null zu beginnen. Weil ein solches Training Hunderte Millionen kostet, gibt es weltweit nur eine Handvoll Häuser, die derlei bauen – und Tausende Angebote, die darauf aufsitzen. Das erklärt, warum sich so viele verschiedene Dienste am Ende ähnlich anfühlen.

Generative AI (deutsch: generative KI, erzeugende KI)
Die Sorte KI, die Neues hervorbringt – Text, Bild, Ton, Film – im Unterschied zur älteren, die vorhandenes Material einordnet, erkennt oder vorhersagt. Diese Unterscheidung ist der eigentliche Bruch der letzten Jahre: Nicht dass Maschinen etwas erkennen, war neu, sondern dass sie etwas hinstellen. Wenn heute von KI die Rede ist, ist fast immer diese gemeint.

GPT (Generative Pre-trained Transformer)
Kein Eigenname, sondern eine Bauartbezeichnung: erzeugend, vorab trainiert, nach dem Transformer-Bauplan. Vorab trainiert heißt, dass die große Schulung an allgemeinem Text lange vor Ihrer Frage stattfand. Weil ein bekanntes Erzeugnis die drei Buchstaben im Namen führt, gelten sie vielen als Markenzeichen; tatsächlich beschreiben sie eine ganze Gattung.

GPU – Graphics Processing Unit (deutsch: Grafikprozessor)
Ein Rechenbaustein, der ursprünglich für Bildschirmbilder ersonnen wurde und Tausende gleichartiger Rechnungen zugleich ausführt. Genau das braucht ein neuronales Netz, und deshalb ist aus einem Spielzubehör der teuerste Rohstoff dieser Industrie geworden. Modelle werden auf Tausenden solcher Bausteine trainiert und auf ihnen betrieben. Wenn von Rechenkraft die Rede ist, sind fast immer sie gemeint.

Grounding (deutsch etwa: Erdung, Rückbindung an Belege)
Der Gegenbegriff zur Halluzination: Man bindet die Antwort an nachprüfbares Material, statt sie aus dem Gedächtnis kommen zu lassen. In der Praxis heißt das, Quellen mitzugeben oder nachschlagen zu lassen und die benutzten Stellen zu benennen. Ein so geerdetes Ergebnis ist nicht von selbst richtig – aber es ist überprüfbar, und darin liegt der ganze Unterschied.

Guardrails (deutsch: Leitplanken)
Vorkehrungen, die ein Modell vor bestimmten Ausgaben oder Handlungen zurückhalten – teils anerzogen, teils als vorgeschaltete Prüfung, teils als Grenze der Werkzeuge, die es überhaupt in die Hand bekommt. Das Bild der Leitplanke ist treffend: Sie erzieht niemanden zum guten Fahren, sie

verhindert den Abflug. Zu enge Leitplanken machen ein Modell unbrauchbar, zu weite unverantwortlich; die Kunst liegt in der Breite der Fahrbahn.

Hallucination (deutsch: Halluzination, das Erfinden)
Wenn ein Sprachmodell etwas erfindet und es im Ton der Gewissheit vorträgt. Es lügt dabei nicht, denn es hat keine Absicht; es setzt fort, was sprachlich stimmig aussieht – ein Buchtitel, den es nie gab, eine Jahreszahl, die gut passen würde, ein Zitat, das klingt wie der Autor. Am größten ist die Gefahr bei Einzelheiten, die selten vorkommen: Namen, Daten, Seitenzahlen, Signaturen. Was nachprüfbar sein muss, muss nachgeprüft werden, am besten an der Quelle selbst.

Inference (deutsch: das Ableiten; gemeint ist der laufende Betrieb)
Der Betrieb eines fertigen Modells, im Unterschied zu seiner Herstellung. Training heißt: Das Modell entsteht. Inference heißt: Es arbeitet – für jede Antwort wird einmal durch das ganze Netz gerechnet. Das Training kostet einmal viel, der Betrieb kostet jedes Mal ein wenig, und weil er millionenfach geschieht, ist er auf Dauer der größere Posten.

Jailbreak (deutsch etwa: Ausbruch, Umgehung)
Der Versuch, ein Modell durch geschickte Formulierung dazu zu bringen, seine eigenen Grenzen zu übergehen – meist durch eine erfundene Rahmenerzählung, ein Rollenspiel, einen angeblichen Ausnahmefall. Es ist ein fortwährendes Wettrennen: Jede bekannt gewordene Masche wird geschlossen, neue werden gefunden. Für den gewöhnlichen Gebrauch ist der Begriff vor allem als Warnung nützlich – er erklärt, warum Modelle bei sonderbar gerahmten Bitten misstrauisch werden.

Knowledge Cutoff (deutsch: Wissensstichtag)
Der Zeitpunkt, bis zu dem der Übungsstoff eines Modells reicht. Was danach geschah, kennt es nicht – es sei denn, man legt es ihm vor oder es darf nachschlagen. Heikel ist das dort, wo sich etwas leise geändert hat: ein Amt, ein Preis, eine Vorschrift. Das Modell weiß dann nicht, dass es veraltet ist, und antwortet mit demselben Gleichmut wie sonst.

Latency (deutsch: Verzögerung, Wartezeit)
Die Zeit zwischen Absenden und erster Reaktion. Bei Sprachmodellen setzt sie sich zusammen aus dem Lesen der Eingabe, dem Anlauf und dem wortweisen Erzeugen der Antwort. Deshalb erscheinen Antworten laufend statt auf einen Schlag: Man soll lesen können, während noch geschrieben wird. Ein längeres Kontextfenster und ausführliches Nachdenken kosten hier immer etwas.

LLM – Large Language Model (deutsch: großes Sprachmodell)
Der Fachname für die Sorte Programm, um die es hier durchweg geht: ein sehr großes, auf riesigen Textmengen trainiertes Netz, das Sprache fortsetzt und dadurch antwortet, zusammenfasst, übersetzt, ordnet. Groß meint die Zahl der Gewichte – Milliarden bis Billionen. Fast alles, was man heute KI nennt, wenn es um Sprache geht, ist ein LLM.

Machine Learning (deutsch: maschinelles Lernen)
Der Oberbegriff für Verfahren, bei denen ein Programm nicht Regel für Regel geschrieben, sondern aus Beispielen gebildet wird. Man gibt nicht an, woran ein Hund zu erkennen ist, sondern zeigt Hunde, bis das Programm es selbst herausfindet. Künstliche Intelligenz ist der weite Begriff,

maschinelles Lernen der heute bei weitem wichtigste Weg dorthin, tiefes Lernen dessen erfolgreichster Zweig.

MCP – Model Context Protocol (deutsch etwa: Anschlussnorm für Modelle)
Eine offene Vereinbarung darüber, wie ein Sprachmodell fremde Werkzeuge und Datenquellen ansprechen darf – ein Nachrichtenprogramm, eine Ablage, einen Kalender, einen Browser. Vor solchen Normen musste jede Verbindung eigens gebaut werden; nun genügt ein Anschluss, der zur Norm passt. Man kann es sich als Steckdosennorm denken: Wer den Stecker hat, kommt an den Strom. In unserer Zusammenarbeit läuft der Zugriff auf Ihren Rechner und auf Ihren Browser über solche Anschlüsse.

Model (deutsch: Modell)
Das eigentliche Programm, das die Sprache hervorbringt. Es besteht aus vielen Milliarden Zahlenwerten, den Gewichten, die während des Trainings eingestellt wurden; im laufenden Betrieb ändern sie sich nicht mehr. Deshalb hat ein Modell kein Gedächtnis von einem Gespräch zum nächsten: Was es kann, steckt in diesen Zahlen – was es gerade weiß, steht im Kontextfenster. Verschiedene Modelle desselben Hauses unterscheiden sich in Größe und Zuschnitt wie verschiedene Werkzeuge derselben Werkstatt.

Multimodal (deutsch etwa: mehrkanalig, mehrsinnig)
Ein Modell, das nicht nur Text verarbeitet, sondern auch Bilder, Töne oder Bewegtbild – und zwar im selben Zug, nicht durch nachgeschaltete Hilfsprogramme. Praktisch heißt das: Man kann ein Foto anhängen und danach fragen. Genau davon haben wir beide Gebrauch gemacht, als es um den weißen Rand an Ihrem Bild ging.

Neural Network (deutsch: neuronales Netz)
Der Bauplan hinter all dem: viele einfache Recheneinheiten in Schichten, jede mit vielen anderen verbunden, jede Verbindung mit einem Gewicht versehen. Ein Signal läuft von Schicht zu Schicht und wird dabei verwandelt. Die Anlehnung an das Gehirn steckt im Namen und in der groben Idee; mit der Biologie hat das Verfahren wenig gemein und sollte nicht damit verwechselt werden.

NLP – Natural Language Processing (deutsch: Verarbeitung natürlicher Sprache)
Der Fachname für alles, was Maschinen mit menschlicher Sprache anstellen: übersetzen, zusammenfassen, ordnen, beantworten. Natürlich heißt hier gewachsene Sprache im Unterschied zu den Programmiersprachen. Das Feld ist Jahrzehnte alt und arbeitete lange mit Wörterbüchern und Regelwerken; seit den Sprachmodellen ist davon fast nichts mehr übrig.

Open Source / Open Weights (deutsch: offener Quelltext / offene Gewichte)
Zwei verschiedene Grade der Offenheit, die gern durcheinandergehen. Offener Quelltext heißt: Der Programmcode liegt offen und darf verwendet werden. Offene Gewichte heißt: Auch die trainierten Zahlenwerte des Modells sind frei zu haben, so dass man es auf eigenen Rechnern betreiben kann – ohne dass deshalb offenläge, mit welchem Stoff und nach welchem Verfahren es entstand. Der Unterschied ist erheblich, wenn es um Nachprüfbarkeit geht.

Parameter / Weights (deutsch: Gewichte)
Die eingestellten Zahlenwerte, aus denen ein Modell besteht – seine Substanz. Beim Training werden sie milliardenfach nachjustiert, danach stehen sie fest. Wenn von einem Modell mit siebzig Milliarden Parametern die Rede ist, ist deren Anzahl gemeint. Mehr davon bedeutet in der Regel

mehr Vermögen, aber auch mehr Rechenaufwand – und die Zahl allein sagt so wenig über die Güte wie die Zahl der Farbtuben über ein Bild.

Plugin (deutsch etwa: Erweiterung, Einsteckteil)
Ein Zusatzprogramm, das eine bestehende Anwendung um eine Fähigkeit erweitert, ohne dass diese umgebaut werden müsste. Ihre Webseite lebt davon: Suche, Ordner in der Mediathek, Bilderersetzung, die Verwaltung der Schnipsel – jedes davon ein solches Einsteckteil. Der Vorteil ist die leichte Nachrüstbarkeit, der Nachteil, dass jedes Teil gepflegt sein will und mit den anderen verträglich bleiben muss.

Prompt (deutsch etwa: die Eingabe, der Anstoß)
Alles, was man dem Modell vorlegt, damit es antwortet: Frage, Auftrag, Beispieltext, Anweisung zum Ton. Der Prompt ist nicht bloß die Frage, sondern der ganze Rahmen – wer man ist, was man vorhat, woran das Ergebnis zu messen wäre. Wer genau sagt, was er will, bekommt Genaues; wer vage bleibt, bekommt Mittelmaß. Am ehesten vergleichbar ist es mit einer Auftragsbeschreibung an einen fremden Mitarbeiter, der einen nicht kennt und alles wörtlich nimmt.

Prompt Engineering (deutsch etwa: die Kunst der guten Anweisung)
Das planvolle Bauen solcher Anweisungen: klar sagen, was gewollt ist; Beispiele beilegen; um schrittweises Vorgehen bitten; das gewünschte Format vorgeben; sagen, was nicht sein soll. Das Wort klingt nach Ingenieurskunst und ist doch eher eine Schreibkunst – dieselbe, die man braucht, um einem Menschen eine Aufgabe so zu übergeben, dass sie ohne Rückfrage gelingt. Wer schon einmal eine gute Arbeitsanweisung verfasst hat, kann das meiste davon.

Prompt Injection (deutsch etwa: eingeschmuggelte Anweisung)
Ein Angriff, bei dem in einem Text, den das Modell nur lesen soll – eine Webseite, eine Mail, ein Dokument – eine als Auftrag verkleidete Anweisung versteckt wird. Das Modell kann nicht von sich aus unterscheiden, was Stoff und was Befehl ist, wenn beides in derselben Sprache ankommt. Der Schutz liegt in einer klaren Regel: Aufträge kommen allein vom Menschen im Gespräch; alles, was durch Werkzeuge hereinkommt, ist Material und niemals Befehl.

RAG – Retrieval-Augmented Generation (deutsch etwa: Antworten mit Nachschlagen)
Ein Verfahren, bei dem vor der Antwort in einem eigenen Bestand nachgeschlagen wird – Handbücher, Archiv, Webseite – und die gefundenen Stellen dem Modell mitgegeben werden. Es antwortet dann nicht aus dem Gedächtnis, sondern aus dem, was vor ihm liegt. Das mindert das Erfinden und lässt sich belegen, weil man die verwendeten Stellen benennen kann. Die Güte hängt ganz am Nachschlagen: Wird das Falsche gefunden, hilft das beste Modell nichts.

Reasoning Model (deutsch etwa: nachdenkendes Modell)
Eine neuere Gattung, die vor der Antwort erst ausführlich für sich überlegt und eigens dafür geschult wurde. Bei Aufgaben mit mehreren Schritten – Rechnen, Beweisen, Planen, Fehlersuche – ist der Gewinn erheblich; bei einfachen Fragen kostet es nur Zeit. Die Anbieter halten deshalb meist beides bereit und nennen es Denkmodus oder Thinking.

Reinforcement Learning (deutsch: bestärkendes Lernen)
Lernen durch Belohnung: Das Programm probiert, bekommt für Brauchbares ein gutes und für Untaugliches ein schlechtes Signal und richtet sich danach. Groß geworden ist das Verfahren bei Spielen, wo Gewinn und Verlust eindeutig sind. Bei Sprache ist es schwieriger, weil sich nicht

ausrechnen lässt, was eine gute Antwort ist; dort springen deshalb Menschen als Bewerter ein – siehe RLHF.

Repository (deutsch: Ablage, Lager; kurz: Repo)
Ein verwalteter Ablageort für die Dateien eines Vorhabens, der jede Änderung mit Zeit, Urheber und Begründung festhält. Man kann jeden früheren Stand wiederherstellen und sehen, wer was wann geändert hat. Für Programme ist das seit Langem selbstverständlich; für Texte und Bilder wäre es oft ebenso nützlich.

RLHF – Reinforcement Learning from Human Feedback (deutsch etwa: Lernen aus menschlicher
Bewertung) Die zweite Schule nach dem eigentlichen Training. Menschen bekommen mehrere Antworten vorgelegt und sagen, welche taugt; daraus entsteht ein Bewertungsmaß, an dem das Modell weiter nachgestellt wird. Aus einem Programm, das bloß Text fortsetzt, wird so eines, das antwortet, ablehnt, nachfragt. Der Umgangston der heutigen Modelle stammt in weiten Teilen aus diesem Schritt.

Sandbox (deutsch: Sandkasten)
Ein abgeriegelter Bereich, in dem ein Programm arbeiten darf, ohne an alles Übrige heranzukommen. Was darin geschieht, bleibt darin; was draußen liegt, ist unerreichbar. Meine eigene Arbeitsumgebung ist ein solcher Sandkasten: ein eigener Rechner in der Wolke, der nach der Sitzung verschwindet. Deshalb muss ich Ergebnisse ausdrücklich hinüberreichen – sie liegen sonst nicht bei Ihnen, sondern nur bei mir.

Session (deutsch: Sitzung)
Ein zusammenhängender Arbeitsgang mit allem, was darin gesagt und getan wurde. Innerhalb einer Sitzung ist alles gegenwärtig, was das Kontextfenster fasst; endet sie, ist es fort, sofern es nicht ausdrücklich festgehalten wurde. Deshalb sind unsere Protokolle wichtiger, als sie aussehen: Sie sind das, was eine Sitzung überdauert.

Skill (deutsch etwa: Fertigkeit, angelernte Handreichung)
Eine gebündelte Arbeitsanleitung, die man einem Modell mitgibt, damit es eine bestimmte Sache immer auf dieselbe Weise erledigt – etwa: So wird bei uns ein Word-Dokument gesetzt. Ein Skill ist kein neues Können des Modells, sondern das schriftlich festgehaltene Hausrezept, das im richtigen Augenblick aufgeschlagen wird. Der Gewinn liegt in der Wiederholbarkeit: nicht jedes Mal neu erklären, sondern einmal ordentlich aufschreiben.

Slack (kein Fachwort der KI; Name eines Programms)
Ein Nachrichtenprogramm für Betriebe – eine Art Schwarzes Brett mit Räumen, dort Kanäle genannt, in denen die Belegschaft den Tag über schreibt, statt sich zu mailen. Mit künstlicher Intelligenz hat es zunächst nichts zu tun. Es taucht in diesen Gesprächen nur deshalb ständig auf, weil dort gearbeitet wird und man inzwischen auch KI-Helfer in diese Räume setzt. Wer hört, etwas stehe im Slack, darf schlicht verstehen: in unserem Firmenchat.

SLM – Small Language Model (deutsch: kleines Sprachmodell)
Das bewusst klein gehaltene Gegenstück zum großen Sprachmodell. Es hat weniger Gewichte, kann dafür auf einem gewöhnlichen Rechner oder gar auf einem Telefon laufen, antwortet schnell und verlässt das Haus nicht. Im weiten Feld bleibt es hinter den großen zurück; auf einem eng

umgrenzten Gebiet, für das es geschult wurde, hält es gut mit. Wo Vertraulichkeit oder Kosten den Ausschlag geben, ist es oft die vernünftigere Wahl.

Slug (deutsch: Nacktschnecke; hier: der sprechende Teil einer Adresse)
Der aus einem Titel gebildete, kleingeschriebene Teil einer Webadresse. Das Programm schlägt ihn beim Anlegen selbst vor: Große Buchstaben werden klein, Umlaute umschrieben, Leerzeichen zu Bindestrichen – aus Krippe auf dem Wäscheständer wird krippe-auf-dem-waeschestaender. Von Hand lässt er sich ändern, einer späteren Änderung des Titels folgt er aber nicht nach; deshalb tragen ältere Beiträge oft noch ihren ersten Namen, und das ist kein Versehen, sondern richtig so: Wer den Slug ändert, macht jeden bisherigen Verweis darauf ungültig. Der sonderbare Name stammt aus dem Bleisatz der Zeitungen, wo ein Slug das schmale Metallstück mit der Arbeitsbezeichnung eines Artikels war.

Snippet (deutsch: Schnipsel, Codestückchen)
Ein kleines, für sich stehendes Stück Programm, das man einer Anwendung beigibt, um ein einzelnes Verhalten zu ändern. Auf Ihrer Webseite steckt fast alles Besondere in solchen Schnipseln – das Blättern ohne Neuladen, die Kachelgeometrie, das Löschkreuz im Suchfeld. Der Vorteil ist, dass jeder für sich an- und abschaltbar bleibt; der Preis, dass man den Überblick behalten muss, welcher was tut.

Supervised / Unsupervised Learning (deutsch: überwachtes und unüberwachtes Lernen)
Zwei Grundarten des maschinellen Lernens. Überwacht heißt: Zu jedem Beispiel liegt die richtige Antwort bei – tausend Bilder, jedes beschriftet. Unüberwacht heißt: Es liegt nur der Stoff vor, und das Programm sucht selbst nach Ordnung darin. Sprachmodelle entstehen im Kern unüberwacht, denn die richtige Antwort ist dort einfach das nächste Stück Text.

System Prompt (deutsch etwa: Grundanweisung, Dienstanweisung)
Die Anweisung, die vor dem Gespräch steht und nicht von Ihnen stammt, sondern von dem, der die Anwendung gebaut hat: wer das Modell hier sein soll, was es darf, welchen Ton es hält, welche Werkzeuge es hat. Man sieht sie in der Regel nicht, sie wirkt aber auf jede Antwort. Wenn dasselbe Modell in zwei Programmen verschieden auftritt, liegt es meist daran.

Temperature (deutsch: Temperatur)
Ein Regler für die Streuung: Bei jedem Schritt hat das Modell mehrere mögliche Fortsetzungen zur Wahl. Niedrige Temperatur heißt, es nimmt fast immer die wahrscheinlichste – nüchtern, wiederholbar, für Fakten und Formate geeignet. Höhere Temperatur lässt es auch weniger naheliegende wählen – reicher, überraschender, unzuverlässiger. Es ist ein Regler für Wagemut, nicht für Klugheit.

Token (deutsch etwa: Wortstück, Zeichenstück)
Die Recheneinheit der Sprachmodelle: ein Stück Wort. Nicht der Buchstabe, nicht das ganze Wort, sondern etwas dazwischen. Wäscheständer zerfällt in mehrere Token, und ist eines. Im Deutschen rechnet man grob mit anderthalb Token je Wort. Gemessen wird darin alles: was ins Kontextfenster passt, wie lang eine Antwort werden darf und was die Benutzung kostet.

Training (deutsch: Training, Ausbildung)
Der Vorgang, in dem ein Modell entsteht. Man legt ihm ungeheure Textmengen vor und lässt es immer wieder das nächste Stück Wort vorhersagen; wo es danebenliegt, werden die Gewichte ein

wenig nachgestellt – milliardenfach. Danach folgt eine zweite, kürzere Schule, in der Menschen bewerten, welche Antworten taugen und welche nicht. Das Training geschieht einmal und dauert Monate; im Gespräch danach lernt das Modell nichts mehr hinzu.

Training Data (deutsch: Trainingsdaten, Übungsstoff)
Das Material, aus dem ein Modell gebildet wird – bei Sprachmodellen Bücher, Netzseiten, Foren und Programmcode in ungeheurer Menge. Woher es genau stammt, geben die Häuser meist nur ungefähr an, und daran hängt der große Streit: Wer hat es geschrieben oder gemalt, wurde gefragt, wird vergütet. Was im Stoff fehlt, fehlt später im Können; was darin schieflag, liegt später schief.

Transformer (kein deutsches Wort in Gebrauch; etwa: Umformer)
Der Bauplan, auf dem alle heutigen Sprachmodelle beruhen, vorgestellt 2017. Sein Kunstgriff heißt Aufmerksamkeit: Bei jedem Wortstück wägt das Netz, welche anderen Stellen des Textes gerade wichtig sind – so trägt der Anfang eines Absatzes bis in dessen Ende hinein. Das lässt sich hervorragend parallel rechnen, und genau daran waren die Vorgängerverfahren gescheitert. Das T in den bekannten Namen steht für dieses Wort.

Turing Test (deutsch: Turing-Test)
Alan Turings Vorschlag von 1950, die Frage nach dem Denken der Maschine durch eine Probe zu ersetzen: Ein Mensch schreibt mit zwei Gegenübern und soll erraten, welches davon die Maschine ist. Heutige Modelle bestehen diese Probe mühelos – und man hat dabei gelernt, dass sie weniger misst, als sie versprach. Sie prüft die Täuschung, nicht das Verstehen.

Vector Database (deutsch: Vektordatenbank)
Eine Ablage, die nicht Wörter speichert, sondern die Zahlenreihen aus Embeddings, und blitzschnell das Ähnlichste dazu findet. Sie ist der Unterbau des Nachschlagens bei RAG: Aus der Frage wird eine Zahlenreihe, und die Ablage liefert die nächstverwandten Stellen aus dem eigenen Bestand. Der Unterschied zur gewöhnlichen Suche ist, dass sie den Sinn findet und nicht die Buchstaben.

Vibe Coding (deutsch etwa: Programmieren nach Gefühl)
Ein 2025 aufgekommener Ausdruck dafür, dass jemand ein Programm beschreibt, statt es zu schreiben, und das Modell machen lässt – ohne den entstandenen Code noch selbst zu prüfen. Für kleine Werkzeuge, die man einmal braucht, ist das erstaunlich brauchbar. Für alles, was bleiben und getragen werden soll, ist es riskant, denn Fehler zeigen sich spät, und niemand kennt den Bau.

Anhang: Wörter aus der Bildwerkstatt
Die folgenden Ausdrücke gehören nicht zur allgemeinen Rede über künstliche Intelligenz, sondern zur Werkstatt: Sie begegnen einem, sobald man selbst anfängt, mit der Maschine Bilder zu machen. Sie stehen dann als Felder und Regler auf dem Bildschirm, meist unerklärt. Wer weiß, was sie tun, hört auf zu würfeln und fängt an zu arbeiten.

CFG Scale / Steps (deutsch etwa: Führungsstärke und Schrittzahl)
Die zwei Regler, die bei fast jedem Bildgenerator daneben stehen. Die Schrittzahl sagt, in wie vielen Zügen das Rauschen weggenommen wird – wenige Schritte geben ein flaues Bild, sehr viele bringen kaum noch etwas ein und kosten nur Zeit; irgendwo dazwischen liegt der Punkt, an dem es genug ist. Die Führungsstärke sagt, wie streng sich die Maschine an die Beschreibung halten soll: niedrig eingestellt schweift sie ab und wird eigenwillig, hoch eingestellt gehorcht sie wörtlich und wird

dabei hart und überzeichnet. Es sind die Regler, an denen man am meisten lernt, weil man ihre Wirkung unmittelbar sieht.

Checkpoint (deutsch etwa: Zwischenstand, gespeicherter Ausbildungsstand)
Der abgespeicherte Zustand eines Modells zu einem bestimmten Zeitpunkt seiner Ausbildung – ursprünglich ein Sicherungspunkt, an den man zurückkehren kann, wenn das Training entgleist. Bei den Bildgeneratoren ist daraus die Bezeichnung für das ganze Modell geworden: Ein Checkpoint ist die Datei, in der ein vollständig ausgebildetes Modell steckt, und wer Bilder macht, hat meist mehrere davon zur Auswahl. Der eine ist auf Fotografisches hin nachgebildet, der andere auf Gemaltes, der dritte auf Zeichnung. Der Wechsel des Checkpoints verändert das Bild stärker als jede Umformulierung der Beschreibung.

ControlNet (deutsch etwa: Führungsnetz, Steuerbeigabe)
Ein Zusatz, mit dem man dem Bildgenerator neben der Beschreibung noch eine Vorlage mitgibt, an die er sich halten muss – eine Umrisszeichnung, eine Tiefenkarte, ein Strichgerüst, die Haltung einer Figur. Der Text sagt dann, was zu sehen sein soll, die Vorlage sagt, wo es zu stehen hat. Das ist der Unterschied zwischen einer Bestellung und einer Vorzeichnung. Für jeden, der eine bestimmte Komposition im Kopf hat und sie nicht dem Zufall überlassen will, ist dies das wichtigste Werkzeug des ganzen Anhangs.

img2img – Image-to-Image (deutsch: Bild zu Bild)
Der Weg, bei dem am Anfang nicht reines Rauschen steht, sondern ein vorhandenes Bild – eine Fotografie, eine Skizze, ein früheres Ergebnis. Man gibt zusätzlich an, wie weit sich die Maschine davon entfernen darf: Bei geringer Stärke bleibt die Vorlage deutlich erkennbar und wird nur überarbeitet, bei hoher Stärke dient sie nur noch als ferne Anregung. So lässt sich eine eigene Zeichnung als Ausgangspunkt einbringen, statt alles der Beschreibung zu überlassen.

Inpainting / Outpainting (deutsch etwa: Hineinmalen und Hinausmalen)
Beim Hineinmalen übertüncht man einen Bereich des Bildes und lässt allein diesen neu erzeugen, während der Rest unangetastet bleibt – eine störende Hand, ein falscher Gegenstand, ein missratenes Gesicht. Beim Hinausmalen geschieht dasselbe über den Rand hinaus: Die Fläche wird vergrößert, und die Maschine setzt fort, was am Rand begonnen war. Es ist die nächste Verwandtschaft zur Retusche, nur dass nicht ein Nachbarstück kopiert, sondern neu erfunden wird.

LoRA – Low-Rank Adaptation (deutsch etwa: kleiner Aufsatz, Nachschulung im Kleinen)
Eine kleine Beigabe zu einem großen Modell, die ihm eine bestimmte Sache beibringt, ohne es selbst anzurühren: einen Stil, ein wiederkehrendes Gesicht, einen Gegenstand, eine Handschrift. Man braucht dafür wenige Dutzend Bilder statt Millionen, und die entstehende Datei ist winzig gegenüber dem Modell – ein aufgesetztes Blatt, kein neues Buch. Man kann mehrere zugleich auflegen und ihre Stärke einzeln regeln. Für Künstler ist es die Stelle, an der die Sache heikel wird, denn genau so lässt sich auch eine fremde Handschrift nachziehen.

Negative Prompt (deutsch: Ausschlussbeschreibung, Verneinung)
Das Gegenstück zur Beschreibung: eine zweite Liste, in der steht, was gerade nicht vorkommen soll – kein Text im Bild, keine Rahmen, keine sechs Finger, keine grellen Farben. Es ist eine Eigenheit der Diffusionsmodelle, dass sie hier ein eigenes Feld brauchen; im Fließtext gelesene Verneinungen

überhören sie gern, so wie ein Mensch, dem man sagt, er solle nicht an einen Elefanten denken. Erfahrene Nutzer haben eine stehende Ausschlussliste, die sie immer mitgeben, und ergänzen sie von Fall zu Fall.

Seed (deutsch: Saatkorn, Startwert)
Die Zahl, mit der das Zufallsrauschen am Anfang erzeugt wird – und damit der Grund, warum dieselbe Beschreibung zweimal ein anderes Bild ergibt. Gibt man den Startwert nicht vor, wird er gewürfelt; hält man ihn fest, kommt bei gleicher Einstellung wieder genau dasselbe Bild heraus. Darin liegt der praktische Wert: Man hält ein gelungenes Bild am Startwert fest und ändert dann nur ein Wort der Beschreibung, um zu sehen, was dieses Wort tut. Ohne festen Startwert vergleicht man Äpfel mit Birnen.

Upscaling (deutsch: Hochrechnen, Vergrößern)
Das nachträgliche Vergrößern eines fertigen Bildes, nicht durch Auseinanderziehen der vorhandenen Bildpunkte, sondern indem ein eigenes Modell die fehlenden hinzuerfindet – Gewebe, Poren, Kanten. Der Grund dafür ist, dass Bildgeneratoren in mäßiger Größe rechnen und das Ergebnis für den Druck nicht reicht. Man sollte wissen, dass das Hinzugefügte erfunden ist: In der Vergrößerung erscheinen Einzelheiten, die im Ausgangsbild nie standen. Für den Druck ist das meist erwünscht, für ein Beweisstück wäre es fatal.

Autorenschaft dieses Blattes
Länger – Auftrag
Opus – Text
Länger – Abnahme
August 2026