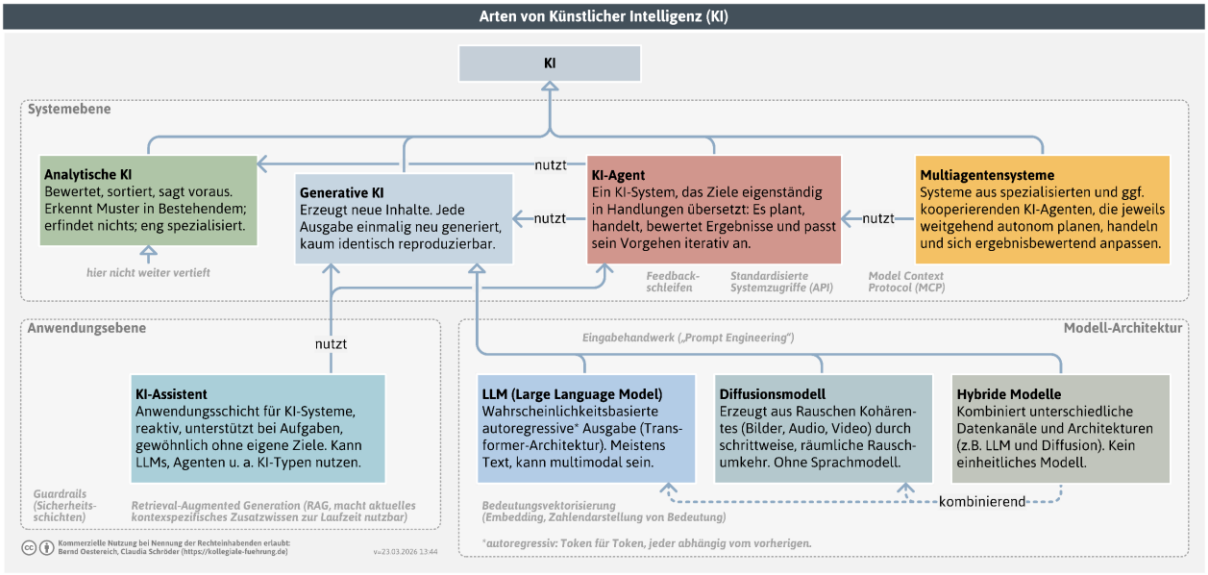

Bernd Oestereich

Arten von KI

Es gibt verschiedene Arten von KI (analytische, generative, agentische), Sprach- und Diffussionsmodelle als Architekturen, spezifische Antwortstrategien und noch ein paar weitere wichtige Begriffe und Unterscheidungen, die im folgenden Text kurz erläutert werden.

Nichts Spektakuläres, nur ein paar begriffliche Abgrenzungen.
Selbstorganisierte Teams sind dafür anfällig

Analytische KI: Erkennen und Vorhersagen

Analytische KI lernt Muster aus großen Datenmengen und wendet sie auf neue Eingaben an – sie erfindet dabei nichts. Im Kern beantwortet sie zwei Fragen: Was ist das? (Klassifikation) oder Was wird wahrscheinlich passieren? (Vorhersage).

Spam-Filter, Betrugserkennung, Qualitätskontrolle per Bilderkennung, Umsatzprognosen – das alles ist analytische KI. Sie ist in vielen Organisationen längst Alltag, wird aber oft nicht als KI wahrgenommen.

Generative KI: Erzeugen statt Erkennen

Generative KI erzeugt neue Inhalte – Texte, Bilder, Töne, Code. Jede Ausgabe wird neu generiert, sie sind nur bedingt identisch reproduzierbar. Das unterscheidet sie grundlegend von der analytischen KI.

Die zwei wichtigsten Unterformen: Große Sprachmodelle und Diffusionsmodelle.

Große Sprachmodelle (LLM)

Große Sprachmodelle (LLM) sind hochkomplexe Muster-Erkennungs- und Muster-Erzeugungssystem, für die zwei zeitlich entkoppelte Vorgänge relevant sind:

- Das Training eines LLM ist ein einmaliger, extrem aufwändiger Prozess: Das Modell liest (vereinfacht gesagt) einen Großteil des öffentlich verfügbaren Internets, digitalisierter Bücher und weiterer Textquellen und justiert dabei Milliarden interner Parameter in einem künstlichen neuronalen Netz so lange, bis es zuverlässig plausible Fortsetzungen zu beliebigen Texteingaben erzeugen kann. Danach ist das Modell eingefroren – es lernt im Betrieb nicht weiter dazu.
- Was dann im Chat passiert, nennt sich Inferenz: Das Modell berechnet auf Basis des bisherigen Gesprächsverlaufs (des sogenannten Kontextfensters) Wort für Wort, was als nächstes am wahrscheinlichsten passt – und fügt dabei Überraschend-Treffendes, gelegentlich aber auch souverän Falsches ein. Ein LLM konfabuliert (dazu verbreiteter Begriff: halluziniert) – es erfindet, wenn es nicht weiß, statt zu sagen, dass es nicht weiß. Das ist keine Fehlfunktion, sondern eine strukturelle Eigenschaft.

Diffusionsmodelle

Diffusionsmodelle sind die Grundlage der meisten Bildgenerierungswerkzeuge (Midjourney, Stable Diffusion, DALL·E, FLUX). Das Prinzip: Ein Bild wird schrittweise in Rauschen aufgelöst (unscharf gemacht) – und das Modell lernt, diesen Prozess umzukehren. Diffusionsmodelle verstehen nicht, was eine Katze ist; sie haben gelernt, wie Katzen-Pixel in Bezug zu anderen Pixeln stehen. Stärke: Erzeugung. Nicht: Verstehen.

RAG – wenn das LLM aktuelle Quellen nutzt

Ein LLM kennt zunächst nur, was es beim Training gesehen hat – danach ist sein Wissen eingefroren. RAG (Retrieval-Augmented Generation, auf Deutsch: wissensangereicherte Generierung) löst das, indem das Modell zum Abfragezeitpunkt auf benutzerdefinierte Quellen zugreift (firmeninterne Dokumente, Datenbanken, ein Intranet), daraus abfragespezifisch eine Art Spickzettel anfertigt und diesen der eigentlichen Anfrage mitgibt.

RAG läuft in zwei Phasen ab:

- Phase 1 – Vorbereitung (Indexierung): Die Dokumente der Wissensbasis werden in kleine Blöcke (Chunks) zerlegt und in numerische Vektoren (Embeddings) umgewandelt. Diese Vektoren werden in einer speziellen Datenbank gespeichert, die schnelle Ähnlichkeitssuchen ermöglicht (Vektordatenbank).
- Phase 2 – Abruf und Antwort (zur Laufzeit): Stellt eine Nutzerin eine Frage, wird diese ebenfalls in einen Vektor umgewandelt und mit den gespeicherten Vektoren verglichen. Die ähnlichsten Abschnitte werden als Kontext an das LLM übergeben – zusammen mit der eigentlichen Frage. Das LLM antwortet dann auf Basis beider Quellen: seinem Trainingswissen und den kontextbezogenen gefundenen Inhalten.

RAG ist keine eigene KI-Architektur, sondern eine Erweiterungstechnik. In Produkten heißt das oft „Eigene Dokumente hochladen“ oder „KI auf euren Daten“. Wichtig für Datenschutz und Informationssicherheit: Die eigenen Inhalte fließen (je nach Anbieter und Einstellungen) nicht automatisch in fremde Trainingsdaten ein.

KI-Assistent, KI-Agent, Multiagentensystem – drei verschiedene Dinge

Diese drei Begriffe werden im Alltag durcheinandergeworfen. Eine einfache Abgrenzung:

- Ein KI-Assistent reagiert. Er ist in eine Benutzeroberfläche eingebettet, durch einen Hintergrundprompt auf eine Rolle konfiguriert und antwortet auf Eingaben – aber er wartet, bis jemand etwas fragt. Normale KI-Chats sind KI-Assistenten.
- Ein KI-Agent handelt. Er hat nicht nur Sprachfähigkeit, sondern auch Handlungsfähigkeit: Er kann Werkzeuge nutzen (Websuche, E-Mail, Code-Ausführung, Datenbankzugriff), Zwischenergebnisse bewerten und mehrstufige Aufgaben selbstständig abarbeiten – ohne für jeden Schritt eine menschliche Eingabe zu brauchen. Wer einem Agenten eine Aufgabe übergibt, übergibt auch Entscheidungsspielraum. Ein klar definierter Handlungsrahmen ist deshalb keine technische Nebensache, sondern eine Führungsaufgabe.
- Ein Multiagentensystem besteht aus mehreren spezialisierten KI-Agenten, die arbeitsteilig zusammenwirken. Ein koordinierender Agent (Orchestrator) zerlegt die Gesamtaufgabe, verteilt Teilaufgaben an Unteragenten und führt Ergebnisse zusammen. Die Agenten können parallel arbeiten, sich gegenseitig rückkoppeln und Ergebnisse anderer prüfen. Leistungsfähig bei komplexen, mehrstufigen Prozessen – und entsprechend anspruchsvoll in Gestaltung und Betrieb. Fehler früher Agenten pflanzen sich durch das System fort; Kontrollpunkte mit menschlicher Prüfung sind professionelle Praxis.

Weitere Begriffe

- Prompt: Ein Prompt ist (ursprünglich) das Signal eines IT-Systems, dass es zu einer Texteingabe bereit ist. Ein Prompt ist die Aufforderung des Systems an die Benutzerin, etwas einzugeben, beispielsweise eine Anweisung. Wenn im Zusammenhang mit KI von Prompts oder vom prompten gesprochen wird, dann ist damit jedoch (fälschlicherweise, aber mittlerweile üblich) der Eingabetext und die Eingabe der Benutzerin gemeint. Prompten meint nun also: ein KI-Modell gezielt mit einer Eingabe ansprechen.
- Kontextfenster: Die Menge an Text (Eingabe + bisheriger Gesprächsverlauf), die ein LLM in einem Durchgang im Blick hat. Was außerhalb liegt, existiert für das Modell nicht.
- Bedeutungsvektor (Embedding): Texte werden in Zahlen (Vektoren) umgewandelt, so dass ähnliche Bedeutungen auch ähnliche Zahlen bekommen. Die sind dann Grundlage für semantische Suche – also eine Suche, die Bedeutung versteht, nicht nur Stichwörter abgleicht.
- Block (Chunk): Ein Textabschnitt, in den Dokumente für die Indexierung zerlegt werden. Die Blockgröße beeinflusst die Qualität der Suchergebnisse.
- Vektordatenbank: Eine spezialisierte Datenbank, die Embeddings speichert und sehr schnell Ähnlichkeiten berechnen kann.

KI-Antwortstrategien

KI-Systeme ermöglichen unterschiedliche Strategien und Vorgehensweisen ein, um anspruchsvolle Aufgaben gut zu lösen. Hier einige Beispiele:

- **Schlussfolgerungsprozesse** (Reasoning): Reasoning ist eine Modelleigenschaft, aus gegebenen Informationen durch logische Schlussfolgerungen neue Erkenntnisse abzuleiten – also nicht nur

Muster in Trainingsdaten abzurufen, sondern aktiv zu schlussfolgern. Dabei kommen deduktives Reasoning (von allgemeinen Regeln auf konkrete Fälle schließen), induktives Reasoning (aus Beispielen verallgemeinern) und abduktives Reasoning (die wahrscheinlichste Erklärung für eine Beobachtung finden) zum Einsatz.

Wichtig ist: Reasoning ist beim Sprachmodell kein separates Modul, sondern emergentes Verhalten – es entsteht aus der Tiefe und Breite des Trainings und wird durch bestimmte Anweisungsstrategien aktiviert oder verstärkt.

- **Explizite Denkschrittfolgen** (Chain-of-Thought – CoT): Explizite Denkschrittfolgen ist die Technik, bei der der Denkweg zunächst explizit formuliert wird, bevor es zur Antwort kommt. Wenn man dem Modell im Prompt zeigt, wie ein Gedankengang aussehen könnte (z. B. durch Beispiele mit Zwischenschritten), dann löst es auch neue, unbekannte Aufgaben deutlich besser. Werden komplexe Aufgaben zunächst in Teilprobleme zerlegt, die nacheinander gelöst werden, dann fließen Lösungen voriger Schritte oft in den jeweils nächsten Schritt ein.

  Selbst wenn man dem Prompt ganz schlicht nur ein „Think step by step“ zufügt (Zero-Shot CoT), beginnen die meisten Modelle bereits beginnt von sich aus, einen Gedankengang zu entwickeln, ohne dass man ihm Beispiele zeigen musste. Das klingt trivial, wirkt aber erstaunlich zuverlässig.

- **Explizite Aufgabenplanung**: Noch grundsätzlicher ist das Konzept des Aufgabenplanung. Während beim Konzept der Denkschrittfolge ein Gedankengang linear aufgebaut wird, geht es bei der Aufgabenplanung einen Schritt weiter: Das System entwirft zunächst einen Handlungsplan – eine Sequenz von Aktionen oder Teilzielen –, bevor es mit der eigentlichen Ausführung beginnt.

- **Denk-Handlungs-Schleifen** (ReAct-Ansatz): Beim ReAct-Ansatz (Reasoning + Acting) wechselt das Modell dabei zyklisch zwischen Denken (Reasoning-Schritt: Was weiß ich? Was brauche ich?) und Handeln (Action-Schritt: Werkzeug aufrufen, Suche starten, Code ausführen). So entsteht eine Schleife aus Nachdenken und Ausprobieren, die an menschliches Problemlösen erinnert.

- **Denkbaumverfahren** (Tree of Thoughts – ToT): Eine wichtige Erweiterung von Denkschrittfolgen ist das Denkbaumverfahren (Tree of Thoughts). Statt eines einzigen linearen Gedankengangs erzeugt das Modell hier mehrere Gedankenstränge parallel, bewertet sie und verfolgt die vielversprechendsten weiter – ähnlich wie ein Schachspieler, der verschiedene Zugfolgen im Kopf durchspielt. Das Modell baut also keinen Pfad, sondern einen Baum möglicher Überlegungen und kann dabei auch auf frühere Verzweigungen zurückspringen.

- **Selbstreflexion und Selbstkorrektur**: Ein weiteres Konzept ist die Selbstreflexion: Das Modell bewertet seine eigene Antwort, identifiziert Fehler oder Lücken und verbessert sie – ohne externen Eingriff. Das Modell führt nach einer Handlung eine Art Nachbesprechung mit sich selbst durch und nutzt diese Erkenntnisse im nächsten Versuch.

Verwandt damit ist das Konzept der Metakognition – das Nachdenken über das eigene Denken. Ob die Metakognition von Sprachmodellen mit der von Menschen vergleichbar ist, ist zwar wissenschaftlich umstritten, aber funktional zeigen sich ähnliche Muster.

Bernd Oestereich Werkstatt für kollegiale Führung