Von Tokens zu Agenten

Wie moderne KI-Systeme wirklich funktionieren – und was Unternehmen daraus lernen müssen

Ein strategischer Leitfaden für Entscheider zu LLMs, RAG, Vektoren, Finetuning und Agentensystemen

1. Einleitung - Warum KI-Projekte heute oft an falschen Annahmen scheitern

Viele Unternehmen haben inzwischen erste Erfahrungen mit generativer KI gesammelt. Mitarbeitende nutzen ChatGPT, Copilot oder ähnliche Tools, erste Pilotprojekte entstehen in Marketing, IT oder Produktentwicklung.

Die Ergebnisse sind oft beeindruckend, Texte entstehen in Sekunden, Präsentationen werden automatisch erstellt, und komplexe Fragen lassen sich scheinbar mühelos beantworten.

Doch sobald Unternehmen versuchen, KI strukturiert in ihre Prozesse zu integrieren, tauchen grundlegende Fragen auf:

- Warum liefert KI manchmal falsche Antworten?
- Wie kann Unternehmenswissen sicher genutzt werden?
- Wann lohnt sich ein eigenes Modell?
- Wie entstehen autonome KI-Agenten?

Viele dieser Fragen haben eine gemeinsame Ursache: Generative KI wirkt auf den ersten Blick einfach zu bedienen, ihre technische Funktionsweise bleibt allerdings meist unsichtbar.

Hinter Anwendungen wie ChatGPT, Copilot oder Gemini verbirgt sich eine komplexe Architektur aus verschiedenen technologischen Bausteinen. Begriffe wie **Tokens**, **Vektoren**, **Retrieval-Augmented Generation (RAG)**, **Finetuning** oder **Agentensysteme** beschreiben zentrale Komponenten dieser Architektur.

Wer verstehen will, wie KI zuverlässig im Unternehmen eingesetzt werden kann, muss diese Bausteine und ihr Zusammenspiel kennen.

Dieses Whitepaper erklärt die wichtigsten Grundlagen moderner KI-Systeme und zeigt, wie aus ihnen praktische Anwendungen für Unternehmen entstehen.

2. Wie Sprachmodelle eigentlich funktionieren

Tokens, Wahrscheinlichkeiten und die Logik hinter generativer KI

Generative KI wirkt aus Anwendersicht oft erstaunlich einfach. Eine Frage wird eingegeben, und wenige Sekunden später erscheint eine ausführliche Antwort. Doch hinter dieser scheinbar intuitiven Interaktion steht eine technische Logik, die sich grundlegend von menschlichem Denken unterscheidet.

Um zu verstehen, warum KI-Systeme bestimmte Antworten geben, und warum sie manchmal auch falsche oder ungenaue Aussagen produzieren, lohnt sich ein Blick auf zwei zentrale Grundlagen moderner Sprachmodelle: **Tokens** und **Wahrscheinlichkeiten**.

2.1 Tokens – die kleinste Einheit der KI

Sprachmodelle wie GPT, Claude oder Gemini arbeiten nicht direkt mit Wörtern oder Sätzen. Stattdessen zerlegen sie Texte in kleinere Einheiten, sogenannte **Tokens**.

Ein Token kann beispielsweise sein:

- ein vollständiges Wort
- ein Wortteil
- ein Satzzeichen
- ein Leerzeichen

In vielen Fällen entspricht ein Token ungefähr **drei Vierteln eines Wortes**, wobei kurze Wörter häufig ein einzelnes Token bilden, während längere Begriffe in mehrere Teile zerlegt werden.

Die genaue Zerlegung hängt vom sogenannten **Tokenizer** des jeweiligen Modells ab.

Diese Tokenisierung folgt keinen sprachwissenschaftlichen Regeln. Sie basiert auf statistischen Mustern in großen Textsammlungen und dient dazu, Sprache in eine Form zu überführen, mit der neuronale Netze rechnen können.

Tokens sind damit gewissermaßen die **Recheneinheit moderner Sprachmodelle**. Sie bestimmen, wie viel Text ein Modell gleichzeitig verarbeiten kann und bilden gleichzeitig die Grundlage für die Kostenstruktur vieler KI-Anwendungen.

Diese Zerlegung ist notwendig, weil Sprachmodelle mathematisch arbeiten. Sie können nur mit **numerischen Repräsentationen von Textbestandteilen** rechnen.

Tokens sind deshalb mehr als nur eine technische Formalität – sie bestimmen zentrale Eigenschaften moderner KI-Systeme.

2.2 Tokens bestimmen Kontext, Leistung und Kosten

Die Anzahl der Tokens spielt bei der Nutzung von KI in Unternehmen eine entscheidende Rolle. Sie beeinflusst drei zentrale Faktoren:

Kontextgröße

Jedes Sprachmodell besitzt ein sogenanntes **Kontextfenster**. Dieses bestimmt, wie viele Tokens ein Modell gleichzeitig „im Gedächtnis“ behalten kann.

Innerhalb dieses Kontextfensters befinden sich beispielsweise:

- die ursprüngliche Frage
- vorherige Nachrichten im Gespräch
- bereitgestellte Dokumente
- die aktuell generierte Antwort

Wenn dieses Limit erreicht wird, beginnen Modelle ältere Informationen aus dem Kontext zu verdrängen. Dadurch kann es passieren, dass frühere Inhalte eines Gesprächs nicht mehr berücksichtigt werden.

Leistungsfähigkeit von Anwendungen

Viele moderne KI-Anwendungen – etwa Dokumentenanalyse, Wissensassistenten oder Programmierhilfen – hängen direkt davon ab, wie viele Tokens ein Modell verarbeiten kann.

Je größer das Kontextfenster, desto umfangreicher können beispielsweise folgende Aufgaben sein:

- Analyse langer Verträge
- Zusammenfassung umfangreicher Dokumente
- Bearbeitung kompletter Wissensdatenbanken

Kostenstruktur

In professionellen Anwendungen wird KI häufig über Programmierschnittstellen (APIs) genutzt. Die Abrechnung erfolgt dabei meist **pro verarbeiteter Tokenmenge**.

Unterschieden wird typischerweise zwischen:

- **Input-Tokens** (eingesendete Inhalte)
- **Output-Tokens** (generierte Antworten)

Damit werden Tokens zur eigentlichen **Recheneinheit moderner KI-Infrastrukturen**.

2.3 Wahrscheinlichkeiten statt Wissen

Ein häufiges Missverständnis besteht darin, dass Sprachmodelle „Wissen“ besitzen. Tatsächlich arbeiten sie jedoch nach einem anderen Prinzip.

Ein Sprachmodell berechnet bei jeder Antwort **die statistische Wahrscheinlichkeit**, welches Token als nächstes folgen sollte.

Wenn ein Modell beispielsweise den Satz beginnt:

„Der erste Mensch auf dem Mond war …“

dann bewertet das System verschiedene mögliche Fortsetzungen und wählt jene mit der höchsten Wahrscheinlichkeit.

In diesem Fall wäre „Neil Armstrong“ statistisch deutlich plausibler als „Frank Roebers“.

Dieser Mechanismus wird Millionen Male pro Antwort wiederholt. So entsteht Schritt für Schritt ein vollständiger Text.

Aus technischer Sicht handelt es sich also nicht um ein Wissenssystem im klassischen Sinn, sondern um ein **hochkomplexes Wahrscheinlichkeitsmodell für Sprache**.

2.4 Warum KI überzeugend wirkt

Trotz dieser statistischen Arbeitsweise wirken viele Antworten generativer KI erstaunlich kompetent. Das liegt daran, dass moderne Sprachmodelle mit enormen Datenmengen trainiert wurden und dadurch sehr präzise Sprachmuster erkennen können.

Sie lernen beispielsweise:

- typische Satzstrukturen
- Zusammenhänge zwischen Begriffen
- häufige Argumentationsmuster
- fachliche Terminologie

Dadurch können sie Inhalte formulieren, die sich für Menschen **kohärent, logisch und fachlich korrekt anhören**.

Doch genau hier liegt auch eine wichtige Herausforderung.

2.5 Die Grenzen des Systems

Da Sprachmodelle auf Wahrscheinlichkeiten basieren, können sie nicht zuverlässig zwischen **wahren und falschen Aussagen** unterscheiden. Sie generieren vielmehr jene Antwort, die statistisch am plausibelsten erscheint.

Wenn relevante Informationen fehlen oder unklar sind, kann es daher passieren, dass ein Modell Lücken mit plausibel klingenden Formulierungen füllt.

Dieses Phänomen wird häufig als **Halluzination** bezeichnet.

Für den praktischen Einsatz in Unternehmen bedeutet das:
Generative KI sollte nicht als alleinige Wissensquelle verstanden werden, sondern als **Werkzeug zur Verarbeitung und Strukturierung von Informationen**.

Genau aus diesem Grund entstehen derzeit neue Architekturen, die Sprachmodelle gezielt mit verlässlichen Datenquellen verbinden.

Der nächste Baustein moderner KI-Systeme spielt dabei eine zentrale Rolle:
Vektoren und Embeddings, mit denen Maschinen Bedeutung und Zusammenhänge zwischen Begriffen mathematisch darstellen können.

3. Wie KI Bedeutung erkennt

Vektoren und Embeddings als Grundlage moderner Sprachmodelle

Nachdem ein Sprachmodell Texte in Tokens zerlegt hat, stellt sich eine zentrale Frage:
Wie kann eine Maschine überhaupt erkennen, dass bestimmte Begriffe inhaltlich zusammengehören?

Menschen verstehen intuitiv, dass Wörter wie *Hund*, *Katze* und *Haustier* miteinander verwandt sind, während Begriffe wie *Toaster* oder *Straßenbahn* in einen völlig anderen Kontext gehören. Für Computer ist diese Art von Bedeutung zunächst jedoch nicht zugänglich.

Um Sprache mathematisch verarbeiten zu können, übersetzen moderne KI-Systeme Wörter und Textteile deshalb in **Vektoren**, also in numerische Koordinaten innerhalb eines hochdimensionalen Raums.

3.1 Von Wörtern zu Koordinaten

Dieser Prozess wird **Embedding** genannt. Dabei erhält jedes Wort, oder genauer gesagt jedes Token, eine mathematische Repräsentation in Form eines Zahlenvektors.

Ein solcher Vektor besteht aus vielen numerischen Dimensionen, oft mehreren Hundert oder Tausend. Jede dieser Dimensionen beschreibt dabei bestimmte statistische Eigenschaften eines Begriffs im Kontext anderer Begriffe.

Vereinfacht lässt sich das Prinzip so vorstellen:

Wenn zwei Begriffe häufig in ähnlichen Zusammenhängen auftreten, liegen ihre Vektoren im mathematischen Raum **nah beieinander**. Begriffe mit völlig unterschiedlichen Bedeutungen liegen dagegen **weiter auseinander**.

So kann ein Modell erkennen, dass beispielsweise Begriffe aus ähnlichen Themenfeldern eng miteinander verbunden sind, selbst wenn sie nicht identisch sind.

3.2 Bedeutung als Abstand im Vektorraum

Dieses Konzept wird häufig als **semantischer Raum** beschrieben. In diesem Raum hat jedes Wort eine Position, und die Distanz zwischen zwei Punkten beschreibt, wie ähnlich ihre Bedeutung ist.

Das ermöglicht Fähigkeiten, die für moderne KI-Anwendungen zentral sind:

- semantische Suche
- automatische Dokumentklassifikation
- Erkennung thematischer Zusammenhänge
- Vergleich von Texten und Aussagen

Wenn ein Nutzer beispielsweise nach Informationen über „Haustiere“ fragt, kann ein System auch Inhalte finden, in denen nur Begriffe wie „Hund“ oder „Katze“ vorkommen – weil diese Begriffe im Vektorraum eng miteinander verbunden sind.

Damit geht KI über klassische Stichwortsuche hinaus und kann Inhalte **kontextuell interpretieren**.

3.3 Die Grundlage für Wissenssysteme

Embeddings und Vektoren bilden die technische Grundlage für viele der derzeit wichtigsten KI-Anwendungen im Unternehmenskontext.

Dazu gehören beispielsweise:

- semantische Dokumentensuche
- intelligente Wissensdatenbanken
- Supportsysteme
- Rechercheassistenten
- Analyse großer Textmengen

Besonders relevant wird diese Technologie im Zusammenspiel mit sogenannten **Vektordatenbanken**, die große Mengen solcher Embeddings speichern und effizient durchsuchen können.

Genau auf diesem Prinzip basiert auch eine der wichtigsten Architekturen für Unternehmens-KI:
Retrieval-Augmented Generation (RAG) – ein Ansatz, der Sprachmodelle gezielt mit Unternehmenswissen verbindet.

Zunächst lohnt sich jedoch ein Blick auf ein häufig diskutiertes Phänomen generativer KI, die sogenannten **Halluzinationen**.

4. Wenn KI Dinge erfindet

Warum Halluzinationen entstehen

Eine der meistdiskutierten Eigenschaften generativer KI ist das Phänomen der sogenannten **Halluzinationen**. Dabei erzeugt ein Sprachmodell Aussagen, die plausibel klingen, aber faktisch falsch oder frei erfunden sind.

Beispiele dafür sind:

- nicht existierende Quellenangaben
- erfundene Studien oder Statistiken
- falsche Fakten oder Daten
- scheinbar überzeugende, aber inhaltlich unzutreffende Antworten

Für viele Nutzer wirkt dieses Verhalten zunächst irritierend. Schließlich erscheinen die Antworten oft strukturiert, logisch formuliert und sprachlich überzeugend.

Um zu verstehen, warum Halluzinationen entstehen, lohnt sich ein Blick auf die grundlegende Funktionsweise von Sprachmodellen.

41. Ein Sprachmodell kennt keine Fakten

Sprachmodelle speichern kein Wissen im klassischen Sinn. Sie verfügen nicht über eine Datenbank mit überprüften Fakten, auf die sie zugreifen können.

Stattdessen erzeugen sie Texte, indem sie **wahrscheinlich passende Tokenfolgen berechnen**. Das Modell entscheidet also bei jedem Schritt, welches Wort statistisch am besten zu den vorherigen passt.

Dieses Verfahren funktioniert erstaunlich gut, solange die zugrunde liegenden Muster aus den Trainingsdaten ausreichend stark sind.

Doch sobald Informationen fehlen oder unklar sind, entsteht ein Problem. Das Modell versucht weiterhin, eine plausible Antwort zu formulieren, auch dann, wenn es keine verlässliche Grundlage dafür gibt.

4.2 Warum KI selten „Ich weiß es nicht“ sagt

Ein weiterer Grund für Halluzinationen liegt in der Art, wie Sprachmodelle trainiert werden. Während des Trainings werden sie darauf optimiert, **hilfreiche und vollständige Antworten** zu geben.

Aus Sicht des Modells ist eine ausführliche Antwort häufig statistisch wahrscheinlicher als eine Ablehnung oder Unsicherheit.

Deshalb neigen Sprachmodelle dazu:

- fehlende Informationen zu ergänzen
- Lücken mit plausibel klingenden Details zu füllen
- Zusammenhänge zu konstruieren, die statistisch sinnvoll erscheinen

Das Ergebnis kann eine Antwort sein, die sprachlich überzeugend wirkt, obwohl ihre Inhalte nicht korrekt sind.

4.3 Warum Halluzinationen besonders im Unternehmenskontext problematisch sind

Während kleinere Ungenauigkeiten im Alltag oft unkritisch sind, können Halluzinationen im Unternehmensumfeld erhebliche Risiken verursachen.

Beispiele sind:

- falsche rechtliche Einschätzungen
- fehlerhafte technische Informationen
- unzutreffende Zusammenfassungen interner Dokumente
- falsche Angaben in Berichten oder Präsentationen

Gerade wenn KI-Systeme in Entscheidungsprozesse integriert werden, ist daher eine **verlässliche Datenbasis** entscheidend.

4.4 Wie Unternehmen Halluzinationen reduzieren können

Vollständig vermeiden lassen sich Halluzinationen bei generativer KI derzeit nicht. Allerdings gibt es verschiedene technische Ansätze, um ihre Wahrscheinlichkeit deutlich zu reduzieren.

Dazu gehören beispielsweise:

- präzise Systemanweisungen und Prompts
- niedrigere Kreativitätsparameter („Temperatur“)
- strukturierte Ausgaberegeln
- und vor allem die Anbindung an verlässliche Wissensquellen

Genau hier setzt eine Architektur an, die für den praktischen Einsatz von KI in Unternehmen immer wichtiger wird: **Retrieval-Augmented Generation (RAG)**.

Dabei greift das Sprachmodell gezielt auf bereitgestellte Dokumente oder Wissensdatenbanken zu, statt Antworten ausschließlich aus statistischen Mustern zu generieren.

Im nächsten Abschnitt betrachten wir, wie diese Architektur funktioniert und warum sie für viele KI-Anwendungen im Unternehmenskontext zur zentralen Grundlage geworden ist.

5. Retrieval-Augmented Generation (RAG)

Wie Sprachmodelle mit Unternehmenswissen arbeiten

Eine der größten Herausforderungen generativer KI im Unternehmenskontext besteht darin, dass Sprachmodelle ihr Wissen nicht automatisch aus aktuellen oder internen

Quellen beziehen. Sie greifen primär auf statistische Muster aus ihren Trainingsdaten zurück.

Für viele praktische Anwendungen reicht das nicht aus. Unternehmen benötigen KI-Systeme, die gezielt auf **aktuelles, internes und verlässliches Wissen** zugreifen können, etwa auf Dokumentationen, Richtlinien, Verträge oder Projektdaten.

Genau hier setzt eine Architektur an, die sich in den letzten Jahren als zentraler Ansatz für Unternehmens-KI etabliert hat: **Retrieval-Augmented Generation**, kurz **RAG**.

5.1 Die Grundidee von RAG

Retrieval-Augmented Generation bedeutet wörtlich übersetzt etwa „Generierung mit zusätzlichem Abruf“. Das Prinzip ist vergleichsweise einfach:

Bevor ein Sprachmodell eine Antwort erzeugt, durchsucht das System zunächst eine Wissensbasis nach **relevanten Informationen**. Diese Inhalte werden anschließend gemeinsam mit der ursprünglichen Frage an das Sprachmodell übergeben.

Das Modell formuliert seine Antwort dann **auf Grundlage dieser bereitgestellten Informationen**.

Der Ablauf lässt sich vereinfacht in vier Schritte unterteilen:

1. Dokumente werden in kleine Textabschnitte zerlegt
2. Diese Abschnitte werden in **Vektoren (Embeddings)** umgewandelt
3. Eine Vektordatenbank speichert diese Repräsentationen
4. Bei einer Anfrage werden die relevantesten Inhalte gesucht und dem Sprachmodell bereitgestellt

Das Modell generiert anschließend eine Antwort, die sich auf genau diese Inhalte stützt.

5.2 Warum Vektorsuche dabei eine zentrale Rolle spielt

Der entscheidende Unterschied zu klassischen Suchsystemen liegt in der Art, wie Informationen gefunden werden.

Traditionelle Suchmaschinen arbeiten vor allem mit **Stichwortabgleichen**. Ein Dokument wird gefunden, wenn es exakt den gesuchten Begriff enthält.

RAG-Systeme nutzen dagegen **semantische Suche auf Basis von Vektoren**. Dadurch können sie Inhalte finden, die thematisch relevant sind, auch wenn sie nicht exakt dieselben Wörter enthalten.

Ein System kann beispielsweise Dokumente über „Kundenservice“ oder „Supportprozesse“ finden, selbst wenn der Nutzer nach „Hilfe für Kundenanfragen“ sucht.

Diese Fähigkeit macht Vektorsuche besonders geeignet für große Wissensbestände, in denen Informationen oft unterschiedlich formuliert sind.

5.3 Typische Anwendungsfälle in Unternehmen

RAG-Systeme eignen sich besonders für Anwendungen, bei denen KI auf umfangreiche interne Informationen zugreifen muss.

Typische Beispiele sind:

Interne Wissensassistenten
Mitarbeitende können Fragen zu Prozessen, Richtlinien oder Projekten stellen und erhalten Antworten auf Basis interner Dokumente.

Technischer Support
KI-Systeme können Produktdokumentationen, Handbücher oder Supportartikel durchsuchen und passende Lösungen vorschlagen.

Dokumentenrecherche
Juristische oder regulatorische Dokumente lassen sich schneller analysieren und zusammenfassen.

Onboarding und Schulung
Neue Mitarbeitende können sich Informationen aus internen Wissensquellen über eine KI erschließen.

5.4 Warum RAG derzeit der Standard für Unternehmens-KI ist

Im Vergleich zu anderen Ansätzen bietet RAG mehrere entscheidende Vorteile:

Aktualität
Neue Dokumente können jederzeit ergänzt werden, ohne das Sprachmodell neu trainieren zu müssen.

Kontrolle über Daten
Die KI greift nur auf definierte Wissensquellen zu.

Reduzierte Halluzinationen
Antworten basieren auf konkreten Dokumenten statt ausschließlich auf statistischen Mustern.

Geringere Kosten und Komplexität
Ein aufwendiges Training eigener Modelle ist häufig nicht notwendig.

Aus diesen Gründen gilt RAG heute als eine der wichtigsten Architekturen für den praktischen Einsatz von generativer KI in Unternehmen.

5.5 Die Grenzen von RAG

Trotz seiner Vorteile löst RAG nicht alle Herausforderungen generativer KI.

So bleibt das zugrunde liegende Sprachmodell weiterhin ein probabilistisches System. Es kann bereitgestellte Informationen falsch interpretieren oder unvollständig berücksichtigen.

Außerdem verändert RAG **nicht das Verhalten des Modells selbst**. Es erweitert lediglich dessen Wissensbasis.

Wenn Unternehmen beispielsweise möchten, dass ein Modell:

- einen bestimmten Schreibstil verwendet
- strukturierte Berichte erstellt
- branchenspezifische Terminologie nutzt
- bestimmte Entscheidungslogiken anwendet

reicht RAG allein oft nicht aus.

In solchen Fällen kommt ein weiterer Ansatz ins Spiel: **Finetuning**, also das gezielte Nachtrainieren eines Modells mit spezifischen Daten.

6. Finetuning

Wann es sinnvoll ist, ein KI-Modell gezielt nachzutrainieren

Während Retrieval-Augmented Generation (RAG) es ermöglicht, Sprachmodelle mit zusätzlichem Wissen zu versorgen, verfolgt **Finetuning** einen anderen Ansatz. Statt lediglich Informationen bereitzustellen, wird dabei das Verhalten des Modells selbst angepasst.

Beim Finetuning wird ein bereits trainiertes Sprachmodell mit einem **spezifischen Datensatz erneut trainiert**, um bestimmte Aufgaben besser zu erfüllen oder eine gewünschte Ausgabeform zu lernen.

Das Modell verändert dabei seine internen Parameter so, dass es neue Muster übernimmt, etwa einen bestimmten Schreibstil, eine strukturierte Ausgaberegel oder eine spezialisierte Fachsprache.

6.1 Wie Finetuning funktioniert

Das Nachtrainieren eines Modells erfolgt typischerweise auf Basis von **Beispieldaten im Frage-Antwort-Format** oder anderen strukturierten Trainingspaaren.

Ein Datensatz könnte beispielsweise so aussehen:

- Eingabe: „Erstelle eine Zusammenfassung dieses technischen Berichts.“
- Ausgabe: strukturierte Zusammenfassung nach definiertem Format

Oder:

- Eingabe: medizinischer Befund
- Ausgabe: standardisierter Arztbericht

Durch viele solcher Beispiele lernt das Modell, welche Art von Antwort in einer bestimmten Situation erwartet wird.

Das Ziel ist nicht, dem Modell neues Wissen im klassischen Sinne beizubringen, sondern sein **Verhalten und seine Ausgabestruktur** zu beeinflussen.

6.2 Der Unterschied zwischen RAG und Finetuning

In der Praxis werden diese beiden Ansätze häufig miteinander verwechselt. Dabei erfüllen sie unterschiedliche Aufgaben.

RAG erweitert ein Sprachmodell um zusätzliches Wissen, indem es relevante Dokumente oder Datenquellen abruft und dem Modell während der Antwortgenerierung bereitstellt. Auf diese Weise kann die KI auf aktuelle oder unternehmensinterne Informationen zugreifen, ohne dass das Modell selbst neu trainiert werden muss.

Finetuning verfolgt dagegen einen anderen Ansatz, hier wird das Verhalten des Modells selbst angepasst. Durch ein gezieltes Nachtraining mit spezifischen Beispieldaten lernt das Modell, bestimmte Aufgaben konsistenter auszuführen oder eine gewünschte Ausgabeform zu verwenden. Während RAG also vor allem die

Wissensbasis erweitert, verändert Finetuning in erster Linie, **wie** das Modell antwortet.

Ein praktisches Beispiel verdeutlicht diesen Unterschied:

Ein Unternehmen möchte eine KI, die Fragen zu internen Richtlinien beantwortet.

- Mit **RAG** greift die KI auf die Richtliniendokumente zu.
- Mit **Finetuning** könnte man dem Modell zusätzlich beibringen, Antworten immer in einem bestimmten Format zu formulieren – etwa als kurze Handlungsempfehlung.

In vielen Anwendungen werden beide Ansätze miteinander kombiniert.

6.3 Typische Einsatzfälle für Finetuning

Finetuning ist besonders sinnvoll, wenn ein Modell **konsistent ein bestimmtes Verhalten zeigen soll**, so wie in folgenden Beispielen:

Unternehmensspezifischer Schreibstil
Ein Modell kann lernen, Texte im Tonfall einer Marke oder Organisation zu formulieren.

Strukturierte Berichte
KI kann darauf trainiert werden, Analysen, Protokolle oder Reports immer in einer bestimmten Struktur auszugeben.

Branchenspezifische Fachsprache
Modelle können Terminologie aus Bereichen wie Medizin, Recht oder Technik konsistenter verwenden.

Automatisierte Dokumentenerstellung
Zum Beispiel für Angebote, technische Berichte oder standardisierte Auswertungen.

Im Vergleich zu RAG ist Finetuning technisch aufwendiger.

Es erfordert unter anderem:

- einen gut strukturierten Trainingsdatensatz
- geeignete Trainingsinfrastruktur
- Tests zur Qualitätssicherung
- regelmäßige Aktualisierung des Modells

Außerdem können feinabgestimmte Modelle bei der Nutzung häufig **höhere Betriebskosten** verursachen als Standardmodelle.

Aus diesem Grund setzen viele Unternehmen zunächst auf RAG-Systeme und nutzen Finetuning nur für **klar definierte Spezialfälle**.

6.4 Ein neuer Ansatz: Modell-Distillation

In der Praxis hat sich in vielen KI-Projekten ein weiterer Ansatz etabliert: **Distillation**.

Dabei wird zunächst ein leistungsfähiges, großes Modell genutzt, um hochwertige Beispielantworten zu erzeugen. Diese Antworten dienen anschließend als Trainingsdaten für ein kleineres, effizienteres Modell.

Das kleinere Modell übernimmt dabei die gewünschten Antwortmuster, kann aber deutlich günstiger betrieben werden.

Dieses Verfahren ermöglicht es Unternehmen, **hohe Antwortqualität mit geringeren Betriebskosten zu kombinieren**.

Mit RAG und Finetuning lassen sich bereits sehr leistungsfähige KI-Anwendungen entwickeln. Dennoch bleibt die Interaktion häufig auf eine einzelne Frage und Antwort beschränkt.

Der nächste Entwicklungsschritt besteht darin, KI-Systeme zu bauen, die **komplexe Aufgaben selbstständig planen und ausführen können**.

Hier kommen sogenannte **KI-Agenten** ins Spiel, Systeme, die nicht nur Antworten generieren, sondern auch Entscheidungen treffen und Aktionen auslösen können.

7. KI-Agenten

Vom Antwortsystem zum Handlungssystem

Mit Technologien wie RAG und Finetuning lassen sich bereits sehr leistungsfähige KI-Anwendungen entwickeln. In vielen Fällen bleiben diese Systeme jedoch auf eine grundlegende Interaktionsform beschränkt, ein Nutzer stellt eine Frage, und das Modell erzeugt eine Antwort.

Der nächste Entwicklungsschritt moderner KI-Systeme besteht darin, diese reine Frage-Antwort-Logik zu erweitern. Statt lediglich Informationen zu generieren, können KI-Systeme zunehmend **Aufgaben planen, Teilaufgaben definieren und Aktionen ausführen**. Solche Systeme werden als **KI-Agenten** bezeichnet.

Ein KI-Agent ist im Kern ein Sprachmodell, das mit zusätzlichen Fähigkeiten ausgestattet wird. Dazu gehören beispielsweise:

- klar definierte Rollen oder Aufgabenbereiche
- Zugriff auf externe Werkzeuge oder Software
- die Fähigkeit, komplexe Aufgaben in einzelne Schritte zu zerlegen
- Entscheidungslogiken zur Planung von Arbeitsschritten

Damit entwickelt sich generative KI von einem **reinen Informationssystem** zu einem **handlungsfähigen digitalen Assistenten**.

7.1 Wie ein KI-Agent arbeitet

Der grundlegende Ablauf eines Agentensystems unterscheidet sich von einer einfachen Chat-Anfrage. Statt direkt eine Antwort zu formulieren, analysiert der Agent zunächst die Aufgabe und plant mögliche Schritte zur Lösung.

Ein vereinfachter Ablauf könnte beispielsweise so aussehen:

1. Der Nutzer formuliert eine Aufgabe.
2. Der Agent analysiert das Ziel und zerlegt es in Teilaufgaben.
3. Für einzelne Schritte nutzt der Agent verfügbare Werkzeuge oder Datenquellen.
4. Zwischenergebnisse werden bewertet und weiterverarbeitet.
5. Der Agent erstellt ein finales Ergebnis oder führt eine gewünschte Aktion aus.

In der Praxis können solche Aktionen sehr unterschiedlich aussehen. Ein Agent kann beispielsweise:

- Informationen recherchieren und zusammenfassen
- Daten aus verschiedenen Systemen zusammenführen
- Berichte erstellen
- E-Mails oder Dokumente vorbereiten
- Termine planen oder Aufgaben anlegen

Der entscheidende Unterschied zu klassischen Chatbots besteht darin, dass der Agent nicht nur reagiert, sondern **aktiv Schritte zur Lösung einer Aufgabe organisiert**.

7.2 Multi-Agenten-Systeme

In komplexeren Anwendungen arbeiten mehrere spezialisierte Agenten zusammen. Dabei übernimmt jeder Agent eine klar definierte Rolle innerhalb eines größeren Systems.

Typische Rollen können beispielsweise sein:

- ein **Recherche-Agent**, der Informationen sammelt
- ein **Analyse-Agent**, der Daten auswertet
- ein **Qualitäts-Agent**, der Ergebnisse überprüft
- ein **Manager-Agent**, der die Ergebnisse zusammenführt

Diese Agenten können Zwischenergebnisse austauschen und gemeinsam eine komplexe Aufgabe bearbeiten. Solche **Multi-Agenten-Systeme** gelten derzeit als eines der spannendsten Entwicklungsfelder im Bereich der generativen KI.

7.3 Voraussetzungen für den Einsatz im Unternehmen

Damit Agentensysteme zuverlässig arbeiten können, benötigen sie eine stabile technische Grundlage. Dazu gehören insbesondere:

- Zugriff auf strukturierte Wissensquellen
- klar definierte Systeminstruktionen
- Zugriff auf relevante Software oder Datenquellen
- Sicherheits- und Kontrollmechanismen

In der Praxis werden Agenten daher häufig auf Plattformen betrieben, die verschiedene KI-Komponenten miteinander verbinden. Dazu zählen beispielsweise Sprachmodelle, Vektordatenbanken, Unternehmenssysteme und Automatisierungswerkzeuge.

8. Fazit: KI verstehen, um sie sinnvoll einzusetzen

Generative KI hat in kurzer Zeit den Weg aus Forschungslaboren und spezialisierten Entwicklerteams in den Alltag vieler Unternehmen gefunden. Anwendungen wie ChatGPT, Copilot oder andere KI-Assistenten zeigen eindrucksvoll, welches Potenzial in dieser Technologie steckt. Gleichzeitig entsteht jedoch häufig der Eindruck, KI sei vor allem ein neues Werkzeug, das sich ähnlich wie eine klassische Softwarelösung einsetzen lässt.

In der Praxis zeigt sich jedoch schnell, dass erfolgreiche KI-Anwendungen auf einer komplexen technischen Architektur beruhen. Hinter scheinbar einfachen Chatoberflächen arbeiten mehrere ineinandergreifende Komponenten, die jeweils unterschiedliche Aufgaben erfüllen. Sprachmodelle zerlegen Texte zunächst in Tokens und generieren Antworten auf Basis statistischer Wahrscheinlichkeiten. Embeddings und Vektoren ermöglichen es der KI, Bedeutungen und thematische Zusammenhänge mathematisch abzubilden. RAG-Architekturen erweitern diese Modelle um den Zugriff auf aktuelle oder unternehmensinterne Wissensquellen, während Finetuning dazu dient, Verhalten, Ausgabeformate oder Fachsprache gezielt anzupassen.

Mit der Entwicklung von Agentensystemen geht die Entwicklung inzwischen noch einen Schritt weiter. KI-Systeme können nicht mehr nur Inhalte generieren, sondern zunehmend auch Aufgaben strukturieren, Teilprozesse planen und verschiedene Werkzeuge miteinander verbinden. Damit entsteht eine neue Generation digitaler Assistenzsysteme, die nicht nur Informationen bereitstellen, sondern aktiv an Arbeitsprozessen teilnehmen können.

Für Unternehmen bedeutet dies, dass der erfolgreiche Einsatz von KI weniger von der Auswahl eines einzelnen Tools abhängt als von einem grundlegenden Verständnis der zugrunde liegenden Konzepte. Wer weiß, wie Tokens, Vektoren, RAG, Finetuning und Agentensysteme zusammenspielen, kann realistischer einschätzen, welche Möglichkeiten und Grenzen generative KI heute hat. Gleichzeitig wird klar, dass viele leistungsfähige Anwendungen nicht durch ein einzelnes Modell entstehen, sondern durch die geschickte Kombination mehrerer Technologien.

Gerade für mittelständische Unternehmen eröffnet dieses Architekturverständnis neue Perspektiven. Statt ausschließlich auf fertige Standardlösungen zu setzen, können KI-Systeme gezielt an eigene Prozesse, Wissensbestände und Anforderungen angepasst werden. So wird generative KI nicht nur zu einem Experimentierfeld für einzelne Teams, sondern zu einem strategischen Baustein der digitalen Transformation.

Die Entwicklung in diesem Bereich schreitet weiterhin rasant voran. Modelle werden leistungsfähiger, Kontextfenster größer und Agentensysteme komplexer. Umso wichtiger ist es, die grundlegenden Bausteine moderner KI zu verstehen. Sie bilden das Fundament dafür, dass aus der aktuellen Welle generativer KI nicht nur kurzfristige Effizienzgewinne entstehen, sondern nachhaltige Innovation in Unternehmen.