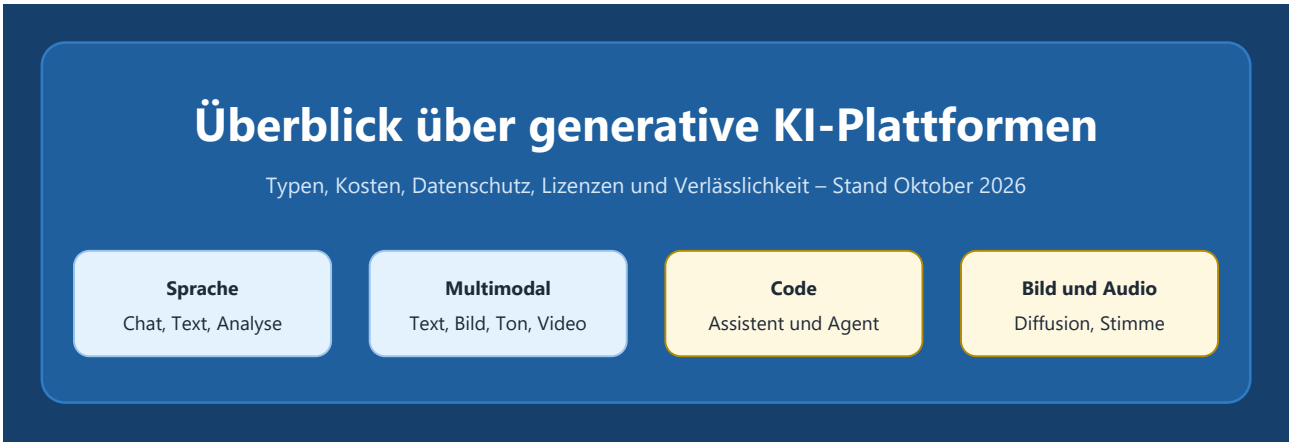
Überblick über generative KI-Plattformen



Wissen	Beratung	Schulung
Welche Sprachmodelle gerade vorn liegen und was das für Unternehmen heißt: › Q1-2026-Update der großen LLMs – Chancen, Risiken, Konsequenzen für Unternehmen	Plattformauswahl, Datenschutzprüfung und Einführung – so sieht ein begleitetes Projekt aus: › Copilot-Einführung bei der Harry Hase AG	Generative KI sicher nutzen – für IT, Fachbereiche und Führungskräfte, mit dem Leitfaden zur Einführung: › KI im Unternehmen erfolgreich einführen – Praxisleitfaden

Montagmorgen, Lenkungskreis bei einem Maschinenbauer mit gut 1.200 Leuten. Auf der Agenda steht „KI-Strategie“, auf dem Tisch liegen drei Angebote: Microsoft 365 Copilot vom Systemhaus, ChatGPT Business vom Marketing, das es ohnehin schon mit Firmenkreditkarte nutzt, und ein Angebot über zwei GPU-Server, weil der Entwicklungsleiter „unsere Daten nicht in die Cloud geben will“. Der Geschäftsführer fragt: „Welches davon ist die beste KI?“ Die ehrliche Antwort lautet: Die Frage ist falsch gestellt. Alle drei können auf demselben oder einem gleich guten Modell laufen. Sie unterscheiden sich in fast allem anderen – in Preis, Datenschutz, Integration, Haftung und Betriebsaufwand.

Genau darum geht es in diesem Beitrag. **Generative KI-Plattformen** sind Dienste und Produkte, die KI-Modelle zum Erzeugen von Texten, Bildern, Code, Audio oder Video bereitstellen – als Chat im Browser, als Assistent in Office und Entwicklungsumgebung, als Programmierschnittstelle oder als Modell zum Selbstbetrieb. Du bekommst hier eine Einordnung nach Typ und Einsatzzweck, die Kriterien, die im Unternehmen wirklich zählen, die Abrechnungsmodelle mit einer Beispielkalkulation und das, was bei Datenschutz, Urheberrecht, Lizenzen und Verlässlichkeit schiefgehen kann. Am Ende steht ein ausführliches FAQ.

Fakten: Generative KI-Plattformen auf einen Blick (Stand Oktober 2026)

Vier Modellfamilien: Sprachmodelle (Text, Analyse), multimodale Modelle (Text, Bild, Ton, Video in einem Modell), Code-Assistenten (spezialisierte Plattformen auf Basis von Sprachmodellen) und Diffusionsmodelle für Bilder, Video und Audio.

Marktführer bei Sprachmodellen: OpenAI (GPT-5.x), Google (Gemini 3.x), Anthropic (Claude), dazu Open-Weight-Modelle wie Llama 4, Mistral, Qwen und gpt-oss.

Kontextfenster: Bei den großen Modellen sind rund eine Million Token inzwischen verbreitet,

das sind mehrere tausend Seiten Text in einer Anfrage.

Abrechnung: Abo pro Person (etwa 20 bis 30 US-Dollar im Monat für Business-Tarife), Token-basierte API, Credits für Bilder und Video, Eigenbetrieb nach Hardware und Personal.

Recht: Die Transparenzpflichten nach Artikel 50 EU AI Act gelten seit dem 2. August 2026; für die maschinenlesbare Kennzeichnung bereits vermarkteter Systeme läuft eine Übergangsfrist bis Anfang Dezember 2026.

Generative KI-Plattformen einordnen: Modelltypen, Plattformen, Einsatzgebiete

Bevor du Produkte vergleichst, brauchst du zwei Unterscheidungen: nach dem, was ein Modell erzeugt, und nach der Schicht, auf der ein Produkt sitzt. Wer beides vermischt, vergleicht am Ende einen Motor mit einem Autohaus.

Vier Modellfamilien und was sie erzeugen

Sprachmodelle (Large Language Models, LLMs) sind auf riesigen Textmengen trainiert und sagen Token für Token voraus, wie ein Text plausibel weitergeht. Daraus entstehen Antworten, Zusammenfassungen, Übersetzungen, Analysen und – weil Programmcode auch nur Text ist – Code. Die aktuellen Spitzenmodelle sind dabei längst nicht mehr reine Textmodelle: GPT-5.x, Gemini 3.x und Claude nehmen Bilder, PDFs und Tabellen entgegen, viele auch Audio. Man spricht dann von **multimodalen Modellen**. Gemini verarbeitet nativ Text, Bild, Audio und Video, OpenAI erzeugt Bilder mit eigenen Bildmodellen direkt im Chat.

Code-Assistenten sind kein eigener Modelltyp, sondern spezialisierte Plattformen: GitHub Copilot, die Assistenten in JetBrains-IDEs, Amazon Q Developer, Cursor oder Claude Code setzen auf große Sprachmodelle auf und ergänzen sie um das, was im Entwicklungsalltag zählt – Zugriff auf das Repository, Ausführen von Tests, Vorschläge direkt im Editor und zunehmend agentische Abläufe, die selbstständig Änderungen über mehrere Dateien vorbereiten und als Pull Request einreichen.

Diffusionsmodelle erzeugen Bilder, Video und Audio, indem sie aus zufälligem Rauschen schrittweise ein Ergebnis herausrechnen, gesteuert durch einen Text-Prompt. Dazu gehören Stable Diffusion, FLUX, Midjourney, Adobe Firefly und Googles Imagen; bei Video etwa Veo, Sora oder Runway. Für Sprache gibt es eigene Text-zu-Sprache-Modelle wie die von ElevenLabs, die Stimmen erzeugen und klonen.

Typ	Erzeugt	Beispiele (Stand Oktober 2026)	Typischer Unternehmenseinsatz
Sprachmodell	Text, Analysen, Code	GPT-5.x, Gemini 3.x, Claude, Llama 4, Mistral, Qwen	Entwürfe, Zusammenfassungen, Wissensfragen, Auswertung von Dokumenten
Multimodales Modell	Text plus Bild, Ton, Video als Ein- oder Ausgabe	Gemini 3.x, GPT-5.x mit GPT Image, Claude (Bild- und PDF-Eingabe)	Belege und Fotos auswerten, Besprechungen protokollieren, Schaubilder erzeugen
Code-Assistent	Code, Tests, Erklärungen, Pull Requests	GitHub Copilot, Amazon Q Developer, Cursor, Claude Code, Gemini Code Assist	Vervollständigung, Refactoring, Testerstellung, Migrationen
Diffusionsmodell Bild/Video	Bilder, Bildbearbeitung, Videoclips	Midjourney V8, FLUX, Stable Diffusion 3.5, Adobe Firefly, Imagen, Veo	Marketingmotive, Produktvarianten, Konzeptgrafik, Storyboards

Typ	Erzeugt	Beispiele (Stand Oktober 2026)	Typischer Unternehmenseinsatz
Audio-Modell	Sprache, Stimmen, Musik	ElevenLabs, OpenAI- und Google-Sprachmodelle	Vertonung, Schulungsvideos, Telefonansagen, Barrierefreiheit

Modell und Plattform sind zwei verschiedene Dinge

Das Missverständnis aus dem Lenkungskreis begegnet mir in fast jedem Projekt. Ein **Modell** ist die trainierte Rechenvorschrift – die Gewichte, die aus einer Eingabe eine Ausgabe machen. Eine **Plattform** ist alles, was drumherum gebaut wird: Anmeldung, Berechtigungen, Anbindung an deine Daten, Sicherheitsfilter, Protokollierung, Verträge, Abrechnung und die Oberfläche, in der Menschen arbeiten. GPT-5 steckt in ChatGPT, in Microsoft 365 Copilot, in GitHub Copilot und in Microsoft Foundry. Claude läuft bei Anthropic direkt, in Amazon Bedrock, in Google Vertex AI und inzwischen auch als Option in Microsoft 365 Copilot. Dasselbe Modell, vier völlig unterschiedliche Datenschutz- und Lizenzsituationen.

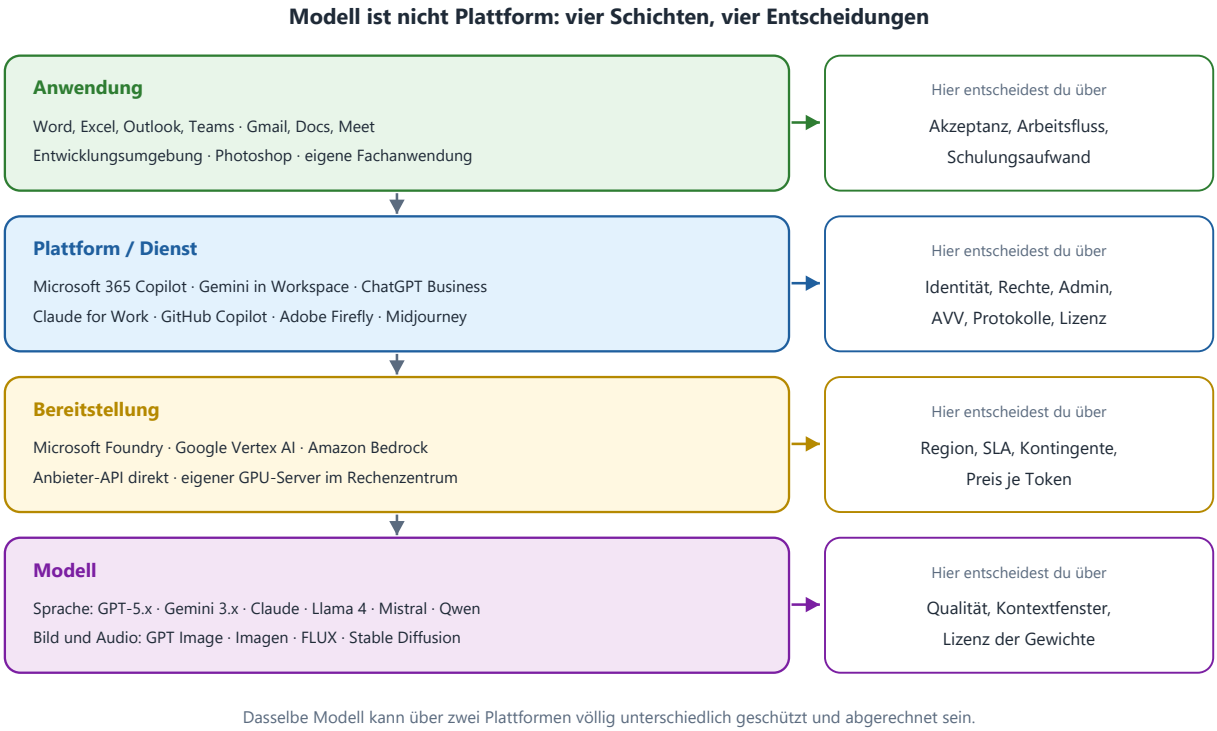


Abbildung 1: Die vier Schichten generativer KI. Die meisten Unternehmensentscheidungen fallen auf der Plattform- und der Bereitstellungsebene, nicht beim Modell.

Für die Praxis heißt das: Die Modellfrage ist die am schnellsten veraltende Frage im ganzen Projekt. Die Spitzenmodelle der großen Anbieter liegen in den meisten Büroaufgaben so dicht beieinander, dass der Unterschied im Alltag kleiner ist als der Unterschied zwischen einem guten und einem schlechten Prompt. Ob aber die Plattform deine Berechtigungen respektiert, wo sie Daten speichert und ob du sie zentral verwalten kannst, entscheidet über Erfolg oder Datenschutzvorfall. Wie sich die großen Sprachmodelle zuletzt entwickelt haben, ordnet der Beitrag Q1-2026-Update der großen LLMs – Chancen, Risiken, Konsequenzen für Unternehmen ein.

Die wichtigsten Plattformen im Überblick

Der Markt hat sich in den letzten zwei Jahren kräftig sortiert. Google hat Bard in Gemini umbenannt und „Duet AI“ ebenfalls, Bing Chat heißt Microsoft Copilot, Amazon CodeWhisperer ist

in Amazon Q Developer aufgegangen, und das eigenständige Bildmodell DALL-E 3 hat OpenAI in ChatGPT durch die GPT-Image-Modelle ersetzt, die auch per API verfügbar sind. Midjourney ist längst nicht mehr auf Discord angewiesen, sondern läuft vollwertig im Browser. Die folgende Tabelle zeigt die Plattformen, die dir im Unternehmen am häufigsten begegnen.

Plattform	Typ	Stärken	Schwächen	Abrechnung (Stand Oktober 2026)
ChatGPT (OpenAI)	Multimodaler Chat, Agenten, API	breite Fähigkeiten, Bild- und Datenanalyse, Projekte, eigene GPTs	Microsoft- und Google-Integration nur über Konnektoren; Consumer-Tarife für Firmendaten ungeeignet	Plus 20 USD, Pro 200 USD; Business 20 USD je Platz bei Jahreszahlung; Enterprise nach Vereinbarung
Google Gemini	Multimodaler Chat, Workspace-Assistent, API	sehr große Kontextfenster, Video und Audio nativ, tief in Gmail, Docs, Meet	volle Wirkung nur in Google Workspace; Funktionsumfang je Edition unterschiedlich	in Workspace Business und Enterprise enthalten; API nach Token
Claude (Anthropic)	Chat, Analyse, Code-Agent, API	lange Dokumente, Schreibqualität, Programmierung, Werkzeugnutzung	kein eigenes Office-Paket; Consumer-Tarife standardmäßig mit Trainingsnutzung	Pro und Max pro Person; Team und Enterprise pro Platz; API nach Token
Microsoft 365 Copilot	Assistent in Office, Teams, Agenten	arbeitet mit Graph-Daten unter den Rechten der Person, Purview-Schutz, zentrale Verwaltung	deckt Überberechtigungen gnadenlos auf; voller Umfang nur mit Zusatzlizenz	30 USD je Person (Enterprise), 21 USD in Copilot Business bis 300 Personen; Copilot Chat ohne Aufpreis
GitHub Copilot	Code-Assistent und Code-Agent	in allen gängigen IDEs, Modellauswahl, Agent für Pull Requests	Kontext bei großen Repositories begrenzt; Mehrverbrauch bei Agenten wird zusätzlich abgerechnet	Pro 10 USD, Pro+ 39 USD, Business 19 USD, Enterprise 39 USD je Person
Llama 4 (Meta)	Open-Weight-Sprachmodell, multimodal	Selbstbetrieb möglich, anpassbar, keine Lizenzgebühr	keine Open-Source-Lizenz, EU-Einschränkung bei multimodalen Modellen, hoher Hardwarebedarf	frei; Kosten für Hardware oder Cloud-Hosting
Midjourney	Bildgenerator (V8.x), Video	herausragende Ästhetik, Stilreferenzen, Web-App	keine Freistellung von Urheberrechtsansprüchen; ab 1 Mio. USD Jahresumsatz Pro- oder Mega-Tarif nötig	Abo von rund 10 bis 120 USD im Monat nach GPU-Zeit
Adobe Firefly	Bild, Vektor, Video, in Creative Cloud	eigene Modelle auf lizenziertem Material, direkt in Photoshop und Illustrator	Partnermodelle (etwa GPT Image, FLUX, Gemini) haben nicht automatisch dieselbe Rechtslage	in Creative Cloud enthalten; Firefly-Abos mit Credits für Premium-Funktionen
Stable Diffusion / FLUX	Open-Weight-Bildmodelle	Selbstbetrieb, LoRA-Anpassung, volle Prozesskontrolle	Lizenz je Modell sehr unterschiedlich; Filter und Moderation liegen bei dir	Hardware oder API; ab 1 Mio. USD Umsatz Enterprise-Lizenz für Stability-Modelle
ElevenLabs	Sprachsynthese, Stimmklon, Synchronisation	sehr natürliche Stimmen in vielen Sprachen	Stimmklone nur mit Einwilligung; Kosten steigen mit	Free-Tier, Abos nach Zeichen oder Minuten, API

Plattform	Typ	Stärken	Schwächen	Abrechnung (Stand Oktober 2026)
			Zeichenvolumen	

Wichtig: Preise und Modellstände veralten schnell

Alle Preise in diesem Beitrag sind US-Listenpreise netto mit Stichtag Oktober 2026. Die Anbieter ändern Tarife, Kontingente und Modellnamen teils im Quartalsrhythmus. Für eine Kaufentscheidung zählt das aktuelle Angebot – und vor allem der Vertragstext, nicht die Produktseite.

Einsatzgebiete: Welcher Plattformtyp wofür taugt

Rankings à la „die fünf besten KI-Tools fürs Marketing“ veralten schneller, als man sie lesen kann. Belastbarer ist die Frage, welcher Plattformtyp zu welchem Einsatzgebiet passt und worauf du dort besonders achten musst.

Einsatzgebiet	Passender Plattformtyp	Worauf es ankommt
Büro und interne Prozesse	Assistent in Microsoft 365 oder Google Workspace	Berechtigungen, Vertraulichkeitslabels, Protokollierung, Akzeptanz
Marketing und Content	Multimodaler Chat plus Bildgenerator mit klarer Rechtelage	Markenstimme, Freigabe, Bildrechte, Kennzeichnung
Softwareentwicklung	Code-Assistent in der IDE, Code-Agent	Code-Review, Lizenzfilter, statische Analyse
Kundenservice	API oder Agentenplattform mit RAG auf geprüftem Wissen	Quellenbindung, Übergabe an Menschen, Chatbot-Hinweis nach AI Act
Recht, Steuer, Compliance	Fachplattform mit kuratierten Quellen oder eigene RAG-Lösung	Zitiergenauigkeit, Mandatsgeheimnis, Vier-Augen-Prinzip
Gesundheitswesen	Spezialisierte Lösungen, Dokumentationsassistenz	Patientendaten, Medizinprodukterecht, Haftung
Finanzwesen	Abgeschottete Unternehmensplattform oder Eigenbetrieb	Aufsichtsrecht, Auslagerungsregeln, Modellrisikomanagement
Personalwesen	Office-Assistent, HR-Software mit KI-Funktion	Mitbestimmung, Diskriminierungsrisiko, Hochrisiko-Einstufung nach AI Act
Bildung und Schulung	Chat mit Lernmodus, Autorenwerkzeuge	Faktentreue, Altersgrenzen, Datenschutz
Design und Medien	Bild-, Video- und Audiogeneratoren	Stilkonsistenz, Auflösung, Persönlichkeits- und Markenrechte

Vergleichskriterien und Abrechnungsmodelle

Benchmark-Tabellen mit Prozentwerten für Mathematik-Olympiaden sind nett für die Fachpresse. Im Unternehmen entscheiden andere Dinge darüber, ob eine Plattform nach sechs Monaten noch genutzt wird oder als teure Leiche in der Lizenzverwaltung liegt.

Die Kriterien, die sich im Alltag wirklich auswirken

Kriterium	Was es bedeutet	Warum es zählt	Prüffrage

Kriterium	Was es bedeutet	Warum es zählt	Prüffrage
Kontextfenster	Wie viel Text, Code oder Material das Modell in einer Anfrage berücksichtigt	Lange Verträge, ganze Handbücher, große Codebasen	Wie viel davon nutzt die Plattform wirklich – oder kürzt sie intern?
Multimodalität	Verarbeitung von Bild, Audio, Video, Tabellen	Belege, Fotos, Besprechungen, Diagramme	Geht das in der Business-Edition oder nur im Consumer-Produkt?
Anpassbarkeit	Eigene Anweisungen, Wissensquellen, Agenten, Feinabstimmung	Wiederkehrende Aufgaben, Fachsprache, Hausstil	Wer darf Agenten bauen, und wer gibt sie frei?
Integrationstiefe	Einbettung in Office, E-Mail, Chat, IDE, Fachanwendungen	Nutzung entsteht dort, wo ohnehin gearbeitet wird	Arbeitet der Assistent mit meinen Rechten auf meinen Daten?
Verwaltung	Zentrale Steuerung, Gruppen, Protokolle, Löschfristen	Ohne Admin-Steuerung keine Governance	Kann ich Funktionen pro Gruppe ein- und ausschalten?
Vertrag und Datenschutz	AVV, Speicherort, Subprozessoren, Trainingsausschluss	DSGVO, Geheimhaltung, Aufsicht	Steht der Trainingsausschluss im Vertrag oder nur im Blog?

Zum Kontextfenster eine Warnung aus der Praxis: Eine Million Token klingt nach „ich werfe einfach alles hinein“. Tatsächlich sinkt bei sehr langen Eingaben die Treffsicherheit für Details in der Mitte des Materials messbar, und jede Anfrage mit vollem Kontext kostet entsprechend. Ein gut gefilterter Ausschnitt schlägt den kompletten Aktenschrank fast immer – bei Qualität und bei Kosten.

Abrechnungsmodelle im Vergleich

Generative KI wird auf fünf Arten abgerechnet, und die Unternehmenspraxis mischt sie fast immer. Das Marketing hat Abos, die Entwicklung hat GitHub Copilot, der Kundenservice-Bot läuft über eine API, und irgendwo im Keller summt ein GPU-Server, den ein engagierter Kollege „mal zum Testen“ aufgebaut hat.

Modell	So wird abgerechnet	Wirtschaftlich, wenn …	Haken
Endnutzer-Abo pro Person	Fester Monatsbetrag, etwa 20 USD (ChatGPT Plus, Claude Pro)	einzelne Fachleute intensiv arbeiten und keine Firmendaten im Spiel sind	Consumer-Bedingungen, kein AVV, keine zentrale Verwaltung – Schatten-IT
Enterprise-Lizenz pro Nutzer	Fester Betrag je Platz, etwa 20 bis 30 USD, meist Jahresbindung	viele Menschen regelmäßig arbeiten und Integration in Office oder IDE wichtig ist	zahlt auch für Wenignutzer; Zusatzverbrauch für Agenten kommt obendrauf
Token-basierte API	Preis je Million Eingabe- und Ausgabe-Token, je nach Modell sehr verschieden	du eigene Anwendungen baust oder die Nutzung stark schwankt	Entwicklungs- und Betriebsaufwand; Kosten schwer vorhersagbar ohne Limits
Pay-per-Use und Credits	Credits je Bild, Videosekunde, Agentenaufruf	Bild- und Videoproduktion oder Agenten mit gelegentlicher Nutzung	Credit-Umrechnung intransparent; Verfall am Monatsende
Eigenbetrieb	Hardware oder gemietete GPU, Strom, Personal	hohe, gleichmäßige Last, strenge Datenhoheit, offene Modelle reichen qualitativ	Betrieb, Updates, Sicherheit und Moderation liegen komplett bei dir

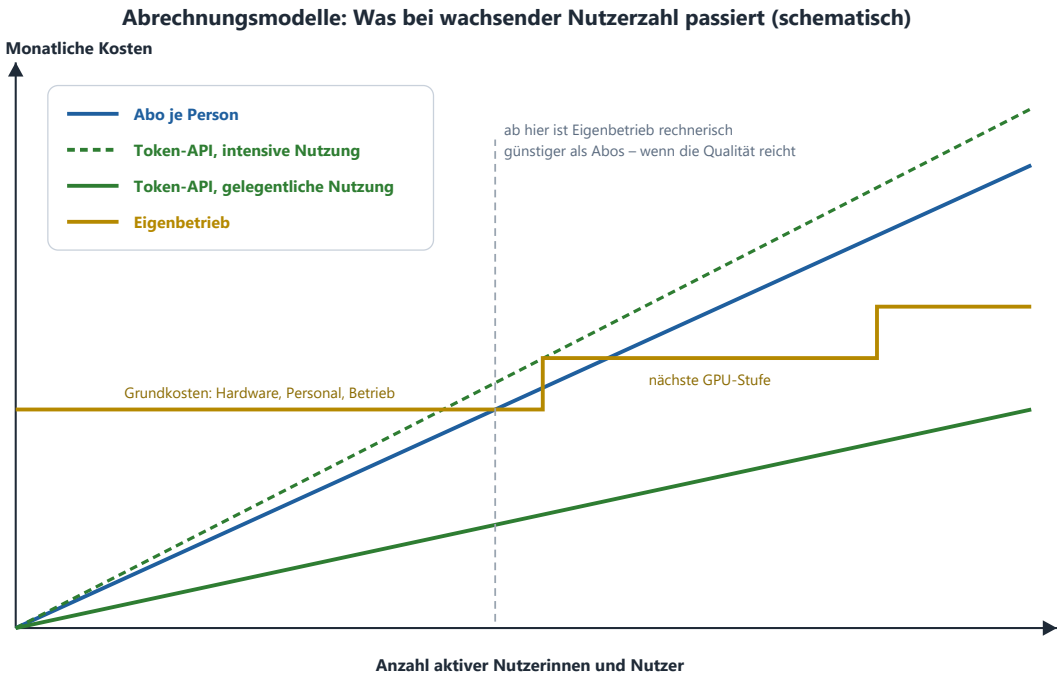


Abbildung 2: Schematischer Kostenverlauf. Abos wachsen linear mit der Zahl der Plätze, die API mit dem tatsächlichen Verbrauch, der Eigenbetrieb in Stufen mit der Hardware.

Nutzungsvolumen grob kalkulieren

Für eine API-Lösung brauchst du drei Zahlen: Anfragen pro Monat, durchschnittliche Eingabe-Token und durchschnittliche Ausgabe-Token je Anfrage. Als Faustregel entsprechen 1.000 Token im Deutschen etwa 600 bis 700 Wörtern; eine DIN-A4-Seite Fließtext liegt bei rund 600 bis 800 Token. Wichtig: Bei RAG-Lösungen steckt der Großteil der Eingabe nicht in der Frage, sondern in den mitgeschickten Dokumentpassagen und dem Systemprompt.

Ein Beispiel: Ein interner Wissensassistent für 500 Personen, jede stellt an 21 Arbeitstagen 20 Fragen. Pro Anfrage gehen 3.000 Token hinein (Frage, Anweisungen, gefundene Passagen) und 800 Token kommen heraus.

Rechenschritt	Wert
Anfragen pro Monat	500 × 20 × 21 = 210.000
Eingabe-Token pro Monat	210.000 × 3.000 = 630 Millionen
Ausgabe-Token pro Monat	210.000 × 800 = 168 Millionen
Kosten bei 2 USD / 12 USD je Million (Beispielpreis eines Pro-Modells)	1.260 USD + 2.016 USD ≈ 3.300 USD
Pro Person und Monat	rund 6,60 USD
Zum Vergleich: 500 Plätze à 30 USD	15.000 USD

Auf dem Papier gewinnt die API haushoch. In der Realität kommen Entwicklung, Hosting des Suchindex, Überwachung, Tests, Pflege der Wissensquellen und Support hinzu – schnell ein bis zwei Personenjahre im ersten Jahr. Und der API-Assistent kann keine E-Mail in Outlook zusammenfassen und keine Besprechung in Teams protokollieren. Der Vergleich ist also nicht „API gegen Abo“, sondern „spezialisierte Anwendung gegen allgemeinen Arbeitsplatzassistenten“. Beides hat seinen Platz.

Tipp: So kalkulierst du ohne böse Überraschungen

Miss vor dem Rollout zwei Wochen echte Nutzung in einer Pilotgruppe, statt Token zu schätzen. Setze in der API harte Budgetgrenzen und Warnschwellen je Anwendung. Leite einfache Anfragen auf ein kleines, günstiges Modell und nur schwierige auf das große – das senkt die Kosten oft um mehr als die Hälfte. Und nutze Prompt-Caching, wo der Anbieter es anbietet: Wiederkehrende Systemprompts werden dann deutlich günstiger berechnet.

Wie du den Nutzen danach belastbar misst, statt nur Kosten zu zählen, beschreibt Metriken zur Messung des Erfolgs der Microsoft 365 Copilot KI-Agenten.

Datenschutz, Urheberrecht und Lizenzen

Hier passieren die teuersten Fehler, und sie passieren fast immer mit guter Absicht. Die Assistentin der Geschäftsführung, die den Entwurf des Aufhebungsvertrags „nur mal kurz glätten" lässt – im privaten Gratis-Konto. Der Entwickler, der den Stacktrace samt Datenbank-Zugangsdaten in den Chat kopiert. Niemand davon wollte etwas Böses. Alle haben Daten an einen Dienstleister übermittelt, mit dem es keinen Vertrag gibt.

Was mit Eingaben und Ausgaben technisch passiert

Jede Eingabe durchläuft dieselbe Kette: Die Plattform authentifiziert dich, prüft die Eingabe gegen Sicherheitsfilter, reichert sie gegebenenfalls mit Daten an (aus deinem Tenant, aus dem Web, aus hochgeladenen Dateien), schickt alles an das Modell, filtert die Ausgabe und zeigt sie an. Das Modell selbst lernt dabei nichts dazu – seine Gewichte ändern sich durch deine Anfrage nicht. Aber an mehreren Stellen entstehen Kopien: im Chatverlauf, in Missbrauchsprotokollen, gegebenenfalls in Trainingsdatensätzen und bei Subprozessoren.

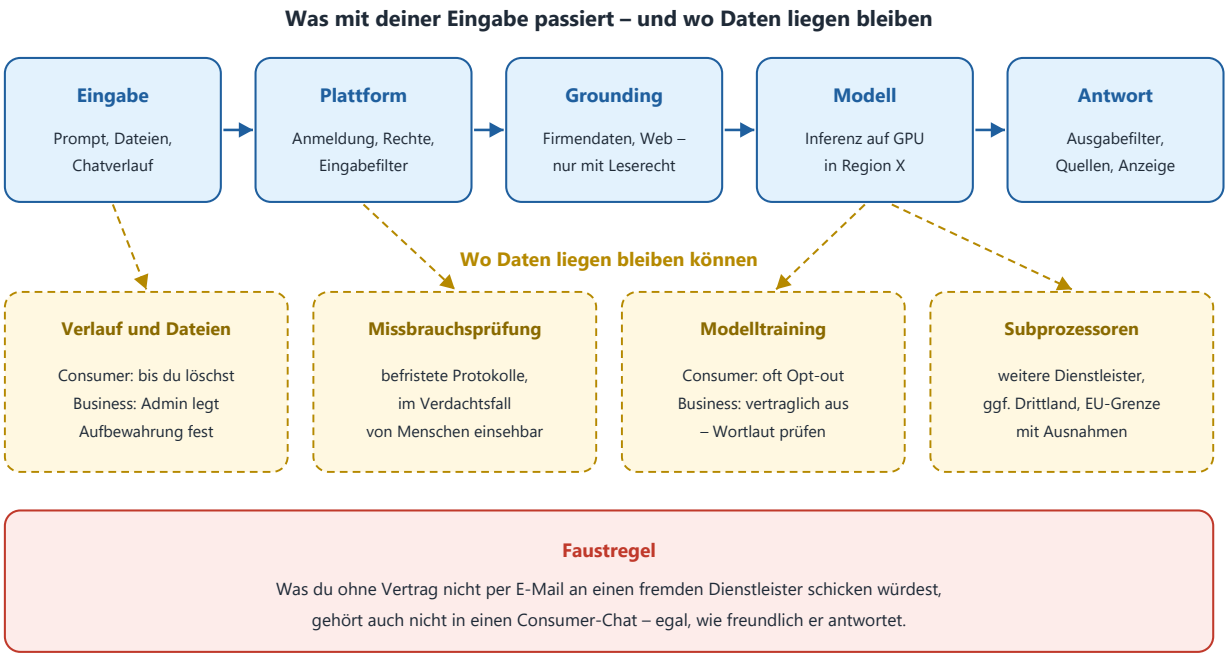


Abbildung 3: Der Weg einer Eingabe durch eine generative KI-Plattform. Entscheidend ist nicht das Modell, sondern was in den gestrichelten Kästen passiert.

Wie ernst man „gelöscht" nehmen darf, zeigt ein Fall aus dem Jahr 2025: Im Urheberrechtsverfahren der New York Times gegen OpenAI verpflichtete ein US-Gericht den Anbieter zeitweise, auch Chats von Consumer- und API-Kunden aufzubewahren, die eigentlich

gelöscht werden sollten. Ausgenommen waren Enterprise-Kunden und API-Verträge mit vereinbarter Null-Speicherung. Die Pflicht ist inzwischen eingeschränkt, die Lehre bleibt: Zusagen zur Speicherdauer gelten nur so lange, wie kein Gericht etwas anderes anordnet – und vertraglich abgesicherte Tarife sind deutlich besser geschützt als Gratiskonten.

Consumer- und Business-Tarife: der eigentliche Unterschied

Aspekt	Consumer (Free, Plus, Pro)	Business und Enterprise	Eigenbetrieb
Vertragspartner	die einzelne Person	das Unternehmen	niemand außer dir
Auftragsverarbeitung	nein	AVV beziehungsweise DPA als Vertragsbestandteil	entfällt, sofern intern
Training mit Eingaben	häufig Standard mit Opt-out	vertraglich ausgeschlossen	nur, wenn du es selbst tust
Verwaltung	keine	SSO, Gruppen, Richtlinien, Protokolle, Löschfristen	selbst zu bauen
Speicherort	meist USA	teils EU-Datenresidenz wählbar	dein Rechenzentrum
Haftung, Freistellung	praktisch keine	teils Freistellung bei Urheberrechtsansprüchen	keine

Ein Beispiel, warum du Trainingszusagen **prüfen statt voraussetzen** solltest: Anthropic hat im Herbst 2025 die Bedingungen der Consumer-Tarife geändert. Chats aus Claude Free, Pro und Max werden seitdem zum Training verwendet und bis zu fünf Jahre gespeichert, sofern die Person nicht widerspricht. Für die Business-Tarife und die API gilt das nicht. Wer seiner Belegschaft vor einem Jahr erklärt hat, „Claude trainiert nicht mit unseren Daten“, hatte damals recht – und hat es heute nicht mehr, wenn private Konten im Spiel sind. Ähnliche Änderungen gab es bei anderen Anbietern in beide Richtungen.

Auftragsverarbeitung, Speicherort und Drittlandtransfer

Sobald personenbezogene Daten in die Plattform gelangen – und das tun sie praktisch immer, schon über Namen in E-Mails –, brauchst du einen Vertrag zur Auftragsverarbeitung nach Artikel 28 DSGVO, eine Rechtsgrundlage für die Verarbeitung und bei Anbietern mit Sitz oder Subprozessoren in den USA eine Absicherung des Drittlandtransfers. Seit Juli 2023 gilt der Angemessenheitsbeschluss zum EU-US Data Privacy Framework; das Gericht der Europäischen Union hat ihn im September 2025 in erster Instanz bestätigt. Zertifizierte US-Anbieter dürfen damit Daten empfangen, ein Restrisiko einer späteren Aufhebung bleibt, wie bei den beiden Vorgängerabkommen.

Microsoft verarbeitet Daten von Microsoft 365 Copilot für europäische Kunden grundsätzlich innerhalb der EU-Datengrenze. Ausnahmen musst du kennen: Die Anthropic-Modelle, die Microsoft in Microsoft 365 Copilot, im Researcher, in Copilot Studio und in den Office-Agenten anbietet, sind ausdrücklich von der EU-Datengrenze ausgenommen; Anthropic arbeitet dabei als Subprozessor unter Microsofts Datenschutzbedingungen. Auch die Websuche über Bing läuft unter eigenen Bedingungen. OpenAI und Google bieten für Unternehmenskunden inzwischen ebenfalls europäische Datenresidenz an, allerdings nicht in jedem Tarif und nicht für jede Funktion.

Warnung: Neue Modelle, neue Datenflüsse

Prüfe nach jeder größeren Produktankündigung, ob neue Modelle oder Funktionen standardmäßig aktiviert werden und wo sie Daten verarbeiten. Bei Microsoft lassen sich die Anthropic-Modelle im Admin Center steuern, bei Google die Gemini-Funktionen in der Admin-

Konsole je Organisationseinheit. Dokumentiere die Entscheidung im Verzeichnis der Verarbeitungstätigkeiten – nicht nur im Kopf der zuständigen Admins.

Urheberrecht: Wem gehören generierte Texte, Bilder und Code?

Nach deutschem Recht schützt das Urheberrecht nur persönliche geistige Schöpfungen (§ 2 Absatz 2 UrhG). Ein Bild, das ein Generator auf einen Einzeiler hin ausspuckt, ist deshalb in aller Regel **nicht** urheberrechtlich geschützt – auch nicht zugunsten der Person, die den Prompt geschrieben hat. Das US Copyright Office sieht es genauso. Schutz kann erst entstehen, wenn ein Mensch das Ergebnis maßgeblich gestaltet: auswählt, kombiniert, überarbeitet, in ein eigenes Werk einbaut. Für dich heißt das: Du darfst KI-Ergebnisse meist nutzen, kannst sie aber oft nicht gegen Nachahmer verteidigen. Ein KI-generiertes Logo ist deshalb eine schlechte Idee – über das Markenrecht lässt es sich zwar eintragen, aber die Gestaltung selbst bleibt ungeschützt.

Die Nutzungsrechte regeln die Plattformen in ihren Bedingungen: OpenAI, Google, Anthropic und Microsoft übertragen ihren Kunden die Rechte an den Ausgaben, soweit sie selbst welche haben. Midjourney erlaubt die kommerzielle Nutzung in bezahlten Tarifen, verlangt aber von Unternehmen mit mehr als einer Million US-Dollar Jahresumsatz den Pro- oder Mega-Tarif. Microsoft und Adobe gehen weiter und stellen Geschäftskunden unter Bedingungen von Urheberrechtsansprüchen Dritter frei – Microsoft mit dem Customer Copyright Commitment für Copilot und Azure OpenAI, sofern die eingebauten Schutzfilter aktiv sind.

Das **Restrisiko aus den Trainingsdaten** bleibt trotzdem. Das Landgericht München I hat im November 2025 im Verfahren der GEMA gegen OpenAI entschieden, dass das Memorieren von Liedtexten im Modell und deren Ausgabe Urheberrechte verletzt; das Training ist nicht pauschal durch die Text-und-Data-Mining-Schranke gedeckt, wenn Werke im Modell reproduzierbar gespeichert sind. Für dich als Anwender bedeutet das: Wenn ein Generator erkennbar ein bestehendes Werk, eine bekannte Figur oder einen geschützten Text wiedergibt, haftest du für die Nutzung – egal, wer das Modell trainiert hat.

Ausgabe	Schutz des Ergebnisses	Hauptrisiko	Gegenmaßnahme
Text	nur bei menschlicher Bearbeitung	wörtliche Übernahmen, falsche Fakten	Plagiatsprüfung bei Veröffentlichungen, redaktionelle Freigabe
Bild	meist keiner	Nähe zu bekannten Werken, Figuren, Marken, echten Personen	Prompts ohne Künstlernamen und Marken, Bildrecherche, Plattform mit Freistellung
Code	nur bei menschlicher Gestaltung	übernommene Fragmente unter Copyleft-Lizenz	Filter für öffentlichen Code, Lizenzscanner, Code-Review
Stimme	Stimme selbst als Persönlichkeitsrecht	Klon ohne Einwilligung	schriftliche Einwilligung, Kennzeichnung

Offene Gewichte sind keine Open-Source-Lizenz – das Beispiel Llama

„Llama ist doch Open Source“ – diesen Satz höre ich in fast jedem Workshop, und er ist falsch. Meta veröffentlicht die **Modellgewichte** von Llama zum Download, aber unter der eigenen Llama Community License mit Bedingungen, die keine Open-Source-Lizenz im Sinne der Open Source Initiative zulässt:

- **Nutzerschwelle:** Wer mit seinen Produkten zum Veröffentlichungszeitpunkt mehr als 700 Millionen monatlich aktive Nutzer hatte, braucht eine gesonderte Lizenz von Meta, die Meta nach eigenem Ermessen erteilt.
- **Acceptable Use Policy:** Bestimmte Einsatzzwecke sind verboten, und Meta kann die Richtlinie einseitig ändern.
- **EU-Einschränkung:** Für die multimodalen Modelle von Llama 3.2 und Llama 4 erhalten Unternehmen und Personen mit Sitz in der EU keine Nutzungsrechte aus der Lizenz. Ein europäisches Unternehmen darf diese Modelle also nicht selbst betreiben; die Nutzung eines Produkts, in das ein Nicht-EU-Anbieter sie eingebaut hat, ist dagegen erlaubt.
- **Namens- und Kennzeichnungspflichten:** Abgeleitete Modelle müssen „Llama“ im Namen tragen, Produkte den Hinweis „Built with Llama“.

Echte Open-Source-Lizenzen gibt es durchaus: OpenAI hat 2025 mit gpt-oss zwei Modelle unter Apache 2.0 veröffentlicht, Mistral und Qwen stellen viele Modelle ebenfalls unter Apache 2.0. Bei Bildmodellen ist es noch bunter: Stable Diffusion 3.5 steht unter der Stability AI Community License und ist für Organisationen mit weniger als einer Million US-Dollar Jahresumsatz kostenlos, darüber brauchst du eine Enterprise-Lizenz. FLUX.1 [schnell] ist Apache 2.0, FLUX.1 [dev] dagegen nur für nicht-kommerzielle Zwecke freigegeben – die erzeugten Bilder darfst du nutzen, das Modell in einem kommerziellen Dienst betreiben aber nicht. Lies also die Lizenz jedes einzelnen Modells, nicht die Überschrift auf der Downloadseite.

Kennzeichnungspflichten nach dem EU AI Act

Fakten: Artikel 50 EU AI Act – was seit August 2026 gilt

Anbieter generativer Systeme müssen synthetische Audio-, Bild-, Video- und Textinhalte maschinenlesbar als KI-erzeugt kennzeichnen (Wasserzeichen, Metadaten wie Content Credentials). Für Systeme, die vor dem 2. August 2026 bereits am Markt waren, gilt nach dem Digital-Omnibus eine Übergangsfrist bis zum 2. Dezember 2026.

Betreiber, also Unternehmen, die KI einsetzen, müssen Deepfakes – realistisch wirkende künstliche Bilder, Videos oder Tonaufnahmen von Personen, Orten oder Ereignissen – offenlegen. KI-generierte Texte, die die Öffentlichkeit über Angelegenheiten von öffentlichem Interesse informieren sollen, sind zu kennzeichnen, es sei denn, ein Mensch hat sie redaktionell geprüft und trägt die Verantwortung.

Chatbots müssen Menschen darauf hinweisen, dass sie mit einer KI sprechen, sofern das nicht offensichtlich ist.

Bußgelder für Verstöße gegen die Transparenzpflichten reichen bis 15 Millionen Euro oder 3 Prozent des weltweiten Jahresumsatzes. Seit Februar 2025 gilt außerdem die Pflicht zur KI-Kompetenz der Belegschaft (Artikel 4).

Für die Praxis heißt das: Der Werbe-Newsletter mit KI-Bild eines fiktiven Produkts ist kein Deepfake. Das realistische „Foto“ eures Geschäftsführers beim Spatenstich, der nie stattgefunden hat, sehr wohl. Und der Kundenservice-Bot braucht einen sichtbaren Hinweis. Wie wir die Regulierung insgesamt bewerten, steht im Beitrag Der EU AI-Act – Regulierung der Künstlichen Intelligenz. Viel Schaden, wenig Nutzen..

Verlässlichkeit: Halluzinationen, Code-Assistenten, Bildgeneratoren

Generative KI ist ein brillanter Praktikant mit fotografischem Gedächtnis, null Berufserfahrung und dem unerschütterlichen Selbstbewusstsein eines Menschen, der gerade sein erstes

Wochenendseminar besucht hat. Man kann mit so jemandem hervorragend arbeiten. Man sollte ihn nur nicht ungeprüft an Kunden schreiben lassen.

Warum Modelle plausibel falsche Fakten erzeugen

Ein Sprachmodell hat keine Datenbank mit Fakten, die es nachschlägt. Es berechnet, welches Token als nächstes am wahrscheinlichsten passt. Bei Wissen, das im Training tausendfach vorkam, ist das Wahrscheinlichste meist auch das Richtige. Bei seltenen Details – ein Aktenzeichen, eine Normnummer, die Garantiefrist eines bestimmten Produkts, eine Quellenangabe – ist das Wahrscheinlichste lediglich etwas, das so aussieht wie eine richtige Antwort. Dazu kommt, dass Modelle im Training dafür belohnt werden, hilfreich zu antworten, und „weiß ich nicht“ lange als wenig hilfreich galt. Die neuen Reasoning-Modelle halluzinieren bei vielen Aufgaben seltener, aber nicht null-mal. Wer erinnert sich nicht an die Anwälte in New York, die 2023 sechs frei erfundene Präzedenzfälle samt Aktenzeichen bei Gericht einreichten – das Muster wiederholt sich seither weltweit in Schriftsätzen, Gutachten und Behördenberichten.

Gegenmaßnahmen: was wirkt und was nicht reicht

Die wirksamste Maßnahme heißt **Retrieval-Augmented Generation** (RAG): Die Plattform sucht zuerst in geprüften Quellen nach passenden Passagen und gibt sie dem Modell mit der Anweisung, nur daraus zu antworten und die Quelle zu nennen. Microsoft 365 Copilot und Gemini in Workspace arbeiten nach diesem Prinzip mit deinen Dokumenten, eigene Assistenten baust du so auf einem Suchindex auf. RAG senkt Halluzinationen drastisch, ersetzt aber keine Prüfung: Ist die Quelle veraltet oder widersprüchlich, antwortet das Modell eben zuverlässig falsch.

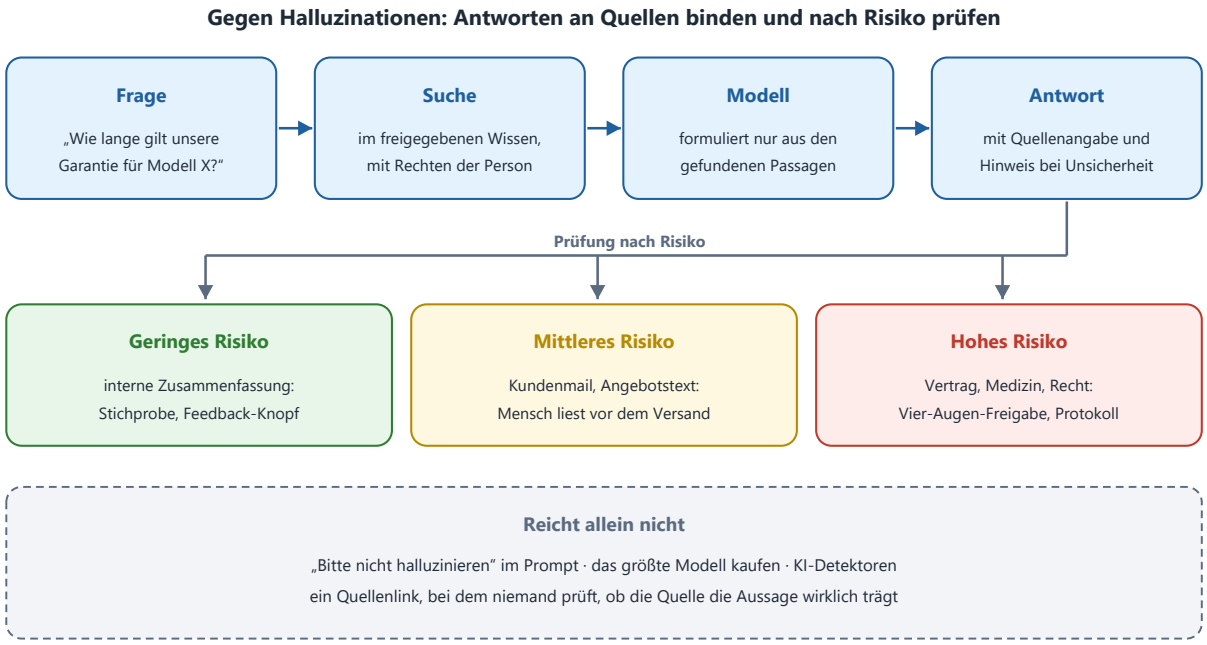


Abbildung 4: RAG bindet die Antwort an geprüfte Quellen, die Prüfung nach Risikoklasse fängt den Rest ab. Prompt-Tricks und größere Modelle allein ersetzen beides nicht.

Maßnahme	Wirkung	Bewertung
RAG auf gepflegten Quellen	Antworten basieren auf Dokumenten statt auf Trainingswissen	wirkt stark
Quellenangaben mit Link	macht Prüfung schnell möglich	wirkt, wenn jemand klickt
Stichprobenprüfung	misst die Fehlerquote, deckt systematische Probleme auf	wirkt, Pflicht im Betrieb

Maßnahme	Wirkung	Bewertung
Freigabeprozess nach Risiko	Mensch prüft, bevor etwas nach außen geht	wirkt, kostet Zeit
Testsets mit Sollantworten	erkennt Verschlechterungen nach Modell- oder Quellenwechsel	wirkt
„Erfinde nichts“ im Prompt	leicht positive Wirkung	reicht nicht
Größeres oder neueres Modell	weniger Fehler, aber nicht keine	reicht nicht
Zweites Modell prüfen lassen	findet manche Fehler, teilt manche Irrtümer	nur ergänzend
KI-Detektoren	erkennen Herkunft schlecht und Wahrheit gar nicht	wirkt nicht

Code-Assistenten im Entwicklungsalltag

Code-Assistenten sind die Plattformkategorie mit dem am besten belegten Produktivitätsgewinn – und mit den am meisten unterschätzten Risiken. Drei Punkte gehören in jede Entwicklungsrichtlinie:

- **Lizenzfragmente:** Modelle können Codepassagen aus dem Training nahezu wörtlich reproduzieren, auch solche unter GPL oder anderen Copyleft-Lizenzen. GitHub Copilot bietet einen Filter, der Vorschläge blockiert, die öffentlichem Code entsprechen (ab etwa 150 Zeichen), oder alternativ eine Code-Referenzierung, die Fundstelle und Lizenz anzeigt. Aktiviere das zentral in der Organisation, nicht pro Person. Ergänze einen Lizenzscanner in der Build-Pipeline.
- **Unsichere Muster:** Generierter Code enthält in Studien regelmäßig SQL-Injection-Lücken, fest kodierte Geheimnisse, veraltete Kryptografie oder fehlende Eingabeprüfung – weil das Trainingsmaterial genau so aussieht. Dazu kommt das „Slopsquatting“: Modelle erfinden Paketnamen, die Angreifer dann tatsächlich mit Schadcode registrieren.
- **Kontextgrenzen:** Auch mit großen Kontextfenstern sieht der Assistent bei großen Projekten nur einen Ausschnitt. Er kennt eure Architekturentscheidungen, internen Bibliotheken und Konventionen nur, wenn sie dokumentiert und auffindbar sind. Anweisungsdateien im Repository, wie sie GitHub Copilot und andere Assistenten auswerten, helfen erheblich.

Warnung: Code-Agenten brauchen dieselben Regeln wie neue Teammitglieder

Agenten, die selbstständig Pull Requests erstellen, bekommen keine Schreibrechte auf den Hauptzweig, keine Produktivgeheimnisse und keinen Weg an Code-Review, statischer Analyse und Tests vorbei. Branch-Schutz, verpflichtende Reviews, SAST, Secret-Scanning und Abhängigkeitsprüfung gelten für KI-Code ausnahmslos. Wer „der Agent hat es geschrieben, die Tests waren grün“ als Review akzeptiert, hat kein Review.

Bildgeneratoren in der Praxis

Bildgeneratoren liefern in Sekunden Motive, für die früher ein Fotoshooting nötig war. Im professionellen Einsatz stößt du schnell auf vier Themen:

Stilkontrolle und Konsistenz. Ein einzelnes schönes Bild ist leicht. Zwanzig Bilder für eine Kampagne, auf denen dieselbe Figur, dasselbe Produkt und dieselbe Farbwelt vorkommen, sind schwer. Midjourney bietet dafür Stil- und Charakterreferenzen, Firefly Stil- und Strukturreferenzen, die großen multimodalen Modelle bearbeiten ein Bild per Anweisung weiter. Wer echte Wiederholgenauigkeit braucht, landet bei offenen Modellen mit festem Seed, ControlNet für Posen und Komposition und eigenen Anpassungen.

Anpassung über Fine-Tuning und LoRA. Ein LoRA-Modul (Low-Rank Adaptation) ist eine kleine Zusatzdatei von wenigen bis einigen hundert Megabyte, die einem Basismodell einen Stil, ein Produkt oder eine Figur beibringt, ohne das ganze Modell neu zu trainieren. Für ein Produkt reichen oft 20 bis 40 gute Fotos. Das funktioniert mit Stable Diffusion und FLUX im Selbstbetrieb; Firefly bietet für Unternehmenskunden eigene Custom Models auf Basis der Firmenbilder.

Personen und Marken. Realistische Bilder echter Personen berühren das Recht am eigenen Bild und das allgemeine Persönlichkeitsrecht, bei Prominenten zusätzlich Werberecht. Fremde Logos, Markenprodukte und geschützte Figuren gehören nicht in Werbemotive, auch wenn der Generator sie bereitwillig erzeugt. Die großen Plattformen blockieren vieles, offene Modelle gar nichts.

Auflösung für Druckprodukte. Hier wird es für Poster und Plakate ernst. Die meisten Generatoren erzeugen nativ ein bis vier Megapixel. Für Druck rechnet man mit 300 dpi bei Betrachtung aus der Nähe und 150 dpi bei Postern, die aus einem Meter Abstand gesehen werden.

Kantenlänge in Pixeln	Druckbreite bei 300 dpi	Druckbreite bei 150 dpi	Reicht für
1.024	8,7 cm	17,3 cm	Social Media, Web, Briefmarkenformat im Druck
2.048	17,3 cm	34,7 cm	A4 randlos knapp, A3-Poster bei Abstand
4.096	34,7 cm	69,4 cm	A3 in Fotoqualität, A1-Poster bei Abstand
7.016 (A1 bei 300 dpi)	59,4 cm	118,8 cm	A1 in voller Qualität – nur mit Hochskalierung erreichbar

Tipp: Hochskalieren, aber richtig

KI-Upscaler erfinden beim Vergrößern Details hinzu. Das sieht bei Landschaften großartig aus und bei Schrift, Gesichtern und technischen Details oft grauenhaft. Prüfe hochskalierte Motive in 100-Prozent-Ansicht an den kritischen Stellen, bevor sie in die Druckerei gehen, und setze Texte grundsätzlich nachträglich im Layoutprogramm statt im Generator.

Betrieb und Integration: Selbstbetrieb, Microsoft 365 und Google Workspace

Selbstbetrieb gegenüber Cloud-Diensten

Der Wunsch, „die KI im eigenen Haus“ zu betreiben, ist verständlich. Offene Modelle wie Llama, Mistral, Qwen oder gpt-oss sind für viele Aufgaben gut genug, und die Datenhoheit ist unschlagbar. Die erste Frage lautet aber immer: Wie viel Grafikspeicher (VRAM) brauche ich? Die Faustregel: Anzahl der Parameter mal Bytes pro Parameter, plus 10 bis 30 Prozent für den Kontext (KV-Cache) – bei langen Kontexten deutlich mehr.

Modellgröße	16 Bit (FP16/BF16)	8 Bit	4 Bit	Typische Hardware
7–8 Mrd. Parameter	ca. 16 GB	ca. 9 GB	ca. 5 GB	eine Workstation-GPU mit 12–16 GB
13–14 Mrd.	ca. 28 GB	ca. 15 GB	ca. 9 GB	eine GPU mit 16–24 GB

Modellgröße	16 Bit (FP16/BF16)	8 Bit	4 Bit	Typische Hardware
30–32 Mrd.	ca. 64 GB	ca. 34 GB	ca. 20 GB	eine GPU mit 24–48 GB (quantisiert)
70 Mrd.	ca. 140 GB	ca. 75 GB	ca. 40 GB	eine 80-GB-GPU oder zwei 48-GB-GPUs (quantisiert)
120 Mrd. MoE (z. B. gpt-oss-120b)	–	–	ca. 65 GB (nativ 4 Bit)	eine 80-GB-GPU
400 Mrd. MoE (z. B. Llama 4 Maverick)	ca. 800 GB	ca. 400 GB	ca. 220 GB	Server mit vier bis acht großen GPUs
Bildmodell (SDXL / FLUX)	ca. 8 / 24 GB	– / ca. 13 GB	– / ca. 8 GB	eine GPU mit 12–24 GB

Quantisierung ist dabei das wichtigste Werkzeug: Die Gewichte werden mit weniger Bits gespeichert, statt 16 Bit etwa 8 oder 4. Bei 8 Bit ist der Qualitätsverlust kaum messbar, bei 4 Bit meist gering, darunter wird er deutlich. Achtung bei Mixture-of-Experts-Modellen (MoE): Sie rechnen pro Token nur mit einem Teil der Parameter und sind deshalb schnell, müssen aber vollständig in den Speicher passen.

Die Hardware ist der kleinere Teil der Rechnung. Ein produktiver Eigenbetrieb braucht eine Inferenz-Software wie vLLM oder einen vergleichbaren Server, Lastverteilung, Überwachung, Updates, Anbindung an die Benutzerverwaltung, Protokollierung und jemanden, der nachts angerufen wird. Und er braucht etwas, woran kaum jemand denkt: **Moderation und Sicherheitsfilter.** Bei Cloud-Anbietern filtern eingebaute Systeme gefährliche Anfragen, Hassrede, Anleitungen zu Straftaten und Versuche, Systemanweisungen auszuhebeln. Ein selbst betriebenes offenes Modell hat davon nur, was ins Training eingeflossen ist – und abgeleitete Varianten aus dem Netz oft gar nichts. Eingabe- und Ausgabefilter, Protokollierung und Missbrauchserkennung musst du selbst einbauen, etwa mit eigenen Klassifikationsmodellen. Die Verantwortung dafür kannst du an niemanden abgeben.

Tipp: Der pragmatische Mittelweg

Viele Unternehmen fahren mit einer Mischung am besten: Arbeitsplatzassistent aus der Cloud für die Breite, offene Modelle in der eigenen Cloud-Subscription (etwa in Microsoft Foundry, Vertex AI oder Bedrock) für Anwendungen mit hohem Volumen, und echter Eigenbetrieb nur für die wenigen Daten, die das Haus wirklich nicht verlassen dürfen.

Einbettung in Microsoft 365

Microsoft 365 Copilot ist die Plattform, die in deutschen Unternehmen am häufigsten auf dem Tisch liegt, und sie folgt einem klaren Prinzip: Copilot arbeitet **im Namen der angemeldeten Person** und sieht über Microsoft Graph genau die Mails, Dateien, Chats und Termine, auf die diese Person ohnehin Zugriff hat. Die Identität kommt aus **Microsoft Entra ID**; Bedingter Zugriff, Gerätekonformität und mehrstufige Authentifizierung gelten wie für jede andere Microsoft-365-Anwendung. Das ist Stärke und Risiko zugleich: Copilot erfindet keine neuen Berechtigungen, findet aber jede bestehende. Die Excel-Datei mit den Gehältern, die seit 2019 für „Jeder außer externen Benutzern“ freigegeben ist, war vorher nur theoretisch ein Problem.

Den Schutz sensibler Daten übernimmt **Microsoft Purview**: Vertraulichkeitsbezeichnungen mit Verschlüsselung wirken auch für Copilot, eine DLP-Richtlinie für den Speicherort „Microsoft 365 Copilot“ verhindert, dass Copilot Inhalte mit bestimmten Labels überhaupt verarbeitet, und Data

Security Posture Management for AI zeigt, welche KI-Dienste mit welchen Daten in Berührung kommen – auch Fremddienste im Browser. SharePoint Advanced Management hilft beim Aufräumen von Überfreigaben und kann einzelne Sites per Restricted Content Discovery von der Copilot-Suche ausnehmen. Wie man das in eine tragfähige Governance gießt, beschreibt der Beitrag Microsoft Copilot-Governance – Grundlagen, Umsetzung, Praxis ausführlich.

Bei der **Lizenzierung** gibt es zwei Ebenen. Copilot Chat mit Websuche, Datei-Upload und nutzungsbasiert abgerechneten Agenten steht allen Konten mit Entra-ID und berechtigtem Microsoft-365-Plan ohne Aufpreis zur Verfügung. Die volle Integration in Word, Excel, PowerPoint, Outlook und Teams mit Zugriff auf Graph-Daten kostet die Zusatzlizenz Microsoft 365 Copilot, als Listenpreis 30 US-Dollar pro Person und Monat bei Jahresbindung; für Organisationen bis 300 Personen gibt es Copilot Business für 21 US-Dollar. Verbrauch durch Agenten jenseits der Lizenz läuft über Copilot Credits. Die Details stehen in Microsoft 365 Copilot Lizenzierung. Zugewiesen werden Lizenzen über das Microsoft 365 Admin Center oder gruppenbasiert in Entra ID, gesteuert werden Copilot, Agenten, Webzugriff und die zusätzlichen Modelle zentral im Copilot-Bereich des Admin Centers. Für die Verwaltung von Agenten als eigene Identitäten hat Microsoft mit Agent 365 eine eigene Steuerungsebene geschaffen, die wir in Agent 365 – KI-Agenten als Steuerzentrale für Microsoft 365, Sicherheit und Automatisierung beschreiben. Den Überblick über die gesamte Copilot-Familie findest du auf der Themenseite Microsoft Copilot.

Einbettung in Google Workspace

Google geht seit Januar 2025 einen anderen Weg: Gemini ist in den Business- und Enterprise-Editionen von Google Workspace enthalten, ein separates Add-on gibt es nicht mehr; im Gegenzug hat Google die Preise der Editionen erhöht. Gemini arbeitet in Gmail, Docs, Sheets, Slides, Meet und Drive ebenfalls mit den Rechten der angemeldeten Person, als eigenständige Gemini-App mit Zugriff auf Workspace-Inhalte und mit NotebookLM für die Arbeit mit eigenen Quellen. Wer mehr will – Agenten, Konnektoren zu Fremdsystemen, unternehmensweite Suche –, landet bei Gemini Enterprise, das separat lizenziert wird.

Die Steuerung erfolgt in der **Admin-Konsole**: Gemini-App, Gemini in den einzelnen Anwendungen, NotebookLM und Funktionen wie die Websuche lassen sich je Organisationseinheit oder Gruppe ein- und ausschalten. Die Identität kommt aus Cloud Identity, auf Wunsch föderiert mit Entra ID. Für den Datenschutz gelten Workspace-DLP-Regeln, Vertrauensregeln für Drive und Labels; mit clientseitiger Verschlüsselung geschützte Inhalte sind für Gemini nicht lesbar. Google sagt vertraglich zu, Workspace-Daten nicht zum Training allgemeiner Modelle zu nutzen und Inhalte nicht ohne Erlaubnis außerhalb der Domain zu verwenden. Datenregionen für die Speicherung gibt es in den höheren Editionen.

Aspekt	Microsoft 365 Copilot	Gemini in Google Workspace
Identität	Entra ID, Bedingter Zugriff	Cloud Identity, optional Föderation mit Entra ID
Datenzugriff	Microsoft Graph mit Rechten der Person	Drive, Gmail, Kalender mit Rechten der Person
Schutz sensibler Daten	Purview-Labels, DLP für Copilot, DSPM for AI, SharePoint Advanced Management	Workspace-DLP, Drive-Labels, Vertrauensregeln, clientseitige Verschlüsselung
Lizenzierung	Copilot Chat ohne Aufpreis; Zusatzlizenz für volle Integration	in Business- und Enterprise-Editionen enthalten; Gemini Enterprise separat
Admin-Steuerung	Copilot-Bereich im Admin Center, gruppenbasierte Lizenzen,	Admin-Konsole je Organisationseinheit oder Gruppe

Aspekt	Microsoft 365 Copilot	Gemini in Google Workspace
	Agentenverwaltung	
Datengrenze	EU-Datengrenze, Ausnahmen für Anthropic-Modelle und Websuche	Datenregionen in höheren Editionen
Agenten	Agent Builder, Copilot Studio, Agent 365	Gems, Gemini Enterprise, Workspace Flows

In beiden Welten gilt: Der Assistent ist nur so sicher wie die Berechtigungsstruktur darunter. Plane vor dem Rollout einen Berechtigungs-Check der größten Dateiablagen ein. Das ist unspektakulär, kostet ein paar Wochen und erspart dir den Moment, in dem die Azubi-Gruppe im Teams-Chat fragt, warum Copilot die Bonusliste kennt.

Fazit: Erst die Plattform wählen, dann das Modell

Generative KI-Plattformen unterscheiden sich weniger in der Klugheit ihrer Modelle als in allem, was um das Modell herum gebaut ist: Identität und Berechtigungen, Vertrag und Speicherort, Lizenz der Ausgaben und der Gewichte, Abrechnung, Verwaltung und Integration in die Werkzeuge, mit denen deine Leute ohnehin arbeiten. Wer zuerst diese Fragen klärt, kann das Modell darunter austauschen, wenn im nächsten Quartal ein besseres erscheint – und das wird es.

Für die meisten Unternehmen sieht die tragfähige Architektur so aus: ein Arbeitsplatzassistent in der Office-Umgebung, die ohnehin im Haus ist, ein Code-Assistent mit sauberer Review-Kette, eine Bildplattform mit klarer Rechtelage, wenige spezialisierte Anwendungen über APIs mit RAG auf gepflegten Quellen und Eigenbetrieb nur dort, wo er sich wirklich begründen lässt. Dazu klare Regeln, welche Daten wohin dürfen, eine Prüfung nach Risikoklassen und eine Belegschaft, die weiß, was diese Werkzeuge können und was nicht. Wie du den Weg dorthin strukturierst, zeigt der Beitrag KI im Unternehmen erfolgreich einführen – Praxisleitfaden.

Tipp: Fünf Schritte für die nächste Woche

1. Inventur: Welche KI-Dienste werden heute schon genutzt, mit welchen Konten? 2. Datenklassen festlegen: Was darf in Consumer-Dienste, was nur in vertraglich abgesicherte, was gar nicht? 3. Eine freigegebene Plattform je Einsatzzweck benennen und Consumer-Konten für Firmendaten untersagen. 4. Berechtigungen der größten Ablagen prüfen. 5. Eine Pilotgruppe mit Messung starten statt einer Grundsatzdebatte.

Weiterlesen: Mehr zu Copilot, Agenten und Automatisierung

› Q1-2026-Update der großen LLMs – Chancen, Risiken, Konsequenzen für Unternehmen

› Copilot-Agenten in der Praxis

› Automatisierung mit KI in der Praxis – Power Automate, n8n, Zapier und Make im Vergleich

› Microsoft Copilot-Governance – Grundlagen, Umsetzung, Praxis

› Copilot-Einführung bei der Harry Hase AG

FAQ: Häufige Fragen zu generativen KI-Plattformen

Was ist generative KI überhaupt?

Generative KI bezeichnet Modelle, die neue Inhalte erzeugen – Texte, Bilder, Code, Audio oder Video –, statt nur vorhandene Daten zu sortieren oder zu bewerten. Sie lernen aus großen Datenmengen statistische Muster und erzeugen daraus Ergebnisse, die es so vorher nicht gab, die aber den Mustern der Trainingsdaten folgen. Die bekanntesten Vertreter sind Sprachmodelle wie GPT, Gemini und Claude sowie Bildgeneratoren wie Midjourney oder Stable Diffusion.

Was unterscheidet generative von klassischer KI?

Klassische, sogenannte diskriminative KI ordnet ein: Sie erkennt eine Katze auf einem Foto, bewertet einen Kreditantrag oder sortiert Spam aus. Generative KI erzeugt: Sie schreibt die Antwortmail, malt das Katzenbild oder formuliert die Begründung. Technisch beruhen beide auf neuronalen Netzen, werden aber auf unterschiedliche Ziele trainiert. In der Praxis arbeiten sie oft zusammen – ein generatives Modell schreibt, ein klassifizierendes prüft auf Tonfall, vertrauliche Inhalte oder Regelverstöße.

Wie funktioniert ein großes Sprachmodell?

Ein Sprachmodell basiert auf der Transformer-Architektur und wird zunächst darauf trainiert, in riesigen Textmengen das jeweils nächste Token vorherzusagen. Dabei lernt es Grammatik, Fakten, Zusammenhänge und Argumentationsmuster. Anschließend wird es mit Beispielen und menschlichem Feedback darauf abgestimmt, hilfreich, ehrlich und sicher zu antworten. Neuere Reasoning-Modelle rechnen vor der Antwort zusätzlich interne Zwischenschritte durch, was bei Mathematik, Code und mehrstufigen Aufgaben deutlich hilft. Bei jeder Anfrage erzeugt das Modell seine Antwort Token für Token – es schlägt nichts nach, solange ihm keine Quellen mitgegeben werden.

Welche generative KI-Plattform ist die beste?

Die Frage nach „der besten“ hat keine allgemeine Antwort. Für Büroarbeit ist meist der Assistent der Office-Umgebung am wertvollsten, die du ohnehin nutzt – Microsoft 365 Copilot oder Gemini in Workspace –, weil er deine Daten mit deinen Rechten kennt. Für Programmierung zählen IDE-Integration und Review-Kette, für Bilder Ästhetik und Rechtelage, für eigene Anwendungen Preis, Region und API-Qualität. Teste mit deinen eigenen Aufgaben, nicht mit Benchmarks, und prüfe den Vertrag vor der Demo.

Darf ich KI-generierte Inhalte kommerziell nutzen?

In der Regel ja, wenn du einen Tarif nutzt, dessen Bedingungen das erlauben, was bei bezahlten Business-Tarifen der großen Anbieter der Fall ist. Du musst aber sicherstellen, dass das Ergebnis keine Rechte Dritter verletzt – keine erkennbaren fremden Werke, Figuren, Marken oder realen Personen ohne Einwilligung – und die Kennzeichnungspflichten des AI Act beachten. Denk daran, dass rein maschinell erzeugte Inhalte meist keinen Urheberrechtsschutz genießen, Wettbewerber sie also ebenfalls verwenden dürfen.

Lernt der Anbieter aus meinen Eingaben?

Das Modell lernt während deiner Sitzung nicht dauerhaft dazu. Ob deine Eingaben später in das Training neuer Modelle einfließen, hängt vom Tarif ab: Bei Business-, Enterprise- und API-Verträgen schließen die großen Anbieter das vertraglich aus, bei Consumer-Tarifen ist Training häufig Standard mit Widerspruchsmöglichkeit. Eingaben werden außerdem für Missbrauchsprüfung befristet gespeichert und können im Verdachtsfall von Menschen gesichtet werden. Verlass dich nicht auf Erinnerungen oder Blogartikel – die Bedingungen ändern sich, wie das Beispiel Anthropic im Herbst 2025 gezeigt hat.

Kann man KI-generierte Inhalte zuverlässig erkennen?

Nein, nicht zuverlässig. Text-Detektoren liefern viele Fehlalarme, besonders bei Menschen, die in einer Fremdsprache schreiben, und sind durch Umformulieren leicht auszutricksen. Bei Bildern helfen Wasserzeichen wie Googles SynthID und Herkunftsnachweise nach dem C2PA-Standard (Content Credentials), die Adobe, OpenAI, Google und Microsoft einbetten – sofern sie nicht entfernt wurden. Mit der Kennzeichnungspflicht nach Artikel 50 AI Act werden solche Markierungen zum Standard, eine Garantie sind sie nicht. Im Zweifel hilft klassische Quellenprüfung.

Sind Open-Weight-Modelle eine echte Alternative zu kommerziellen Plattformen?

Für viele Aufgaben ja. Offene Modelle wie gpt-oss, Qwen, Mistral oder Llama erreichen bei Zusammenfassungen, Klassifikation, Extraktion und Standardtexten ein Niveau, das vor zwei Jahren nur geschlossene Spitzenmodelle hatten. Bei schwierigem Reasoning, sehr langen Kontexten und Multimodalität liegen die kommerziellen Modelle meist noch vorn. Rechne den Betriebsaufwand ehrlich mit, prüfe die Lizenz jedes Modells einzeln und bedenke, dass du für Sicherheitsfilter und Moderation selbst verantwortlich bist.

Welche neuen Sicherheitsrisiken bringen generative KI-Plattformen?

Das wichtigste neue Risiko heißt Prompt Injection: Anweisungen, die in Dokumenten, Mails oder Webseiten versteckt sind und das Modell dazu bringen, Daten preiszugeben oder Aktionen auszuführen. Je mehr Werkzeuge ein Assistent oder Agent bedienen darf, desto gefährlicher wird das. Dazu kommen Datenabfluss über Schatten-IT, überzeugendere Phishing-Mails, Deepfake-Anrufe bei der Buchhaltung und unsicherer generierter Code. Gegenmittel sind minimale Rechte für Agenten, Bestätigung durch Menschen vor kritischen Aktionen, Protokollierung und Schulung. Wie Agenten in Microsoft 365 sicher gebaut werden, steht in Copilot-Agenten in der Praxis.

Ersetzt generative KI Arbeitsplätze?

Sie verändert vor allem Tätigkeiten. Routinen wie Standardkorrespondenz, Protokolle, erste Entwürfe, einfache Programmieraufgaben und Übersetzungen gehen deutlich schneller, in manchen Bereichen wie Übersetzung, Stockfotografie und Content-Produktion sinkt die Nachfrage bereits spürbar. Gleichzeitig wächst der Bedarf an Menschen, die Ergebnisse prüfen, Prozesse gestalten, Daten pflegen und Verantwortung übernehmen. Wer generative KI beherrscht, hat derzeit eher bessere Karten als schlechtere – ersetzt wird eher, wer sich dem Thema komplett verweigert.

Wie bereite ich die Belegschaft auf generative KI vor?

Mit Schulung, die am Arbeitsalltag ansetzt, nicht mit Hochglanzvorträgen. Seit Februar 2025 verlangt Artikel 4 AI Act ohnehin, dass alle, die im Unternehmen KI einsetzen, über ausreichende KI-Kompetenz verfügen. Bewährt haben sich kurze Grundlagenschulungen zu Möglichkeiten, Grenzen und Datenregeln, praxisnahe Übungen mit echten Aufgaben aus den Fachbereichen, ein Netz von Ansprechpartnerinnen und Ansprechpartnern in den Abteilungen und eine kurze, verständliche KI-Richtlinie. Den Betriebsrat holst du früh ins Boot – bei Systemen, die Verhalten oder Leistung auswerten können, ist er ohnehin mitzubestimmen.

Brauche ich eigene Infrastruktur, um generative KI zu nutzen?

Für Cloud-Plattformen wie Microsoft 365 Copilot, Gemini oder ChatGPT nicht – die Rechenleistung stellt der Anbieter. Du brauchst saubere Identitäten, aufgeräumte Berechtigungen und Netzwerkregeln, die die Dienste zulassen. Eigene Infrastruktur wird erst bei Selbstbetrieb offener Modelle nötig, dann aber richtig: GPUs mit ausreichend Speicher, Inferenz-Software,

Überwachung und Personal. Eine Zwischenstufe ist das Hosting offener Modelle in der eigenen Cloud-Subscription.

Wohin entwickeln sich generative KI-Plattformen?

Der klare Trend geht von Assistenten, die antworten, zu Agenten, die handeln: Sie recherchieren über viele Quellen, bedienen Anwendungen, starten Workflows und arbeiten über Stunden an einer Aufgabe. Standards wie das Model Context Protocol verbinden Modelle mit Werkzeugen und Daten, und die Plattformanbieter bauen Verwaltung, Identität und Kontrolle für Agenten aus. Für Unternehmen verschiebt sich die Frage damit von „Welches Modell?“ zu „Welche Aktionen erlauben wir, und wer kontrolliert sie?“. Was Microsoft dafür angekündigt hat, fasst Erwartete Neuerungen in Microsoft Copilot in 2026 zusammen; wie Agenten mit klassischer Automatisierung zusammenspielen, zeigt Automatisierung mit KI in der Praxis – Power Automate, n8n, Zapier und Make im Vergleich.