

Vom Pilotprojekt in den Produktivbetrieb:

Was KI wirklich kostet

Alex Flindt · IT System Engineer
Com In GmbH Computersystemhaus

KI-Konferenz Wismar 2026 · 13. Oktober 2026

Wer ich bin & Wer wir sind



Alex Flindt

IT System Engineer

- Seit 2019 bei Com In, davor 13 Jahre IT-Betrieb bei der Bundeswehr
- Azure, Virtualisierung, Kubernetes, Terraform, hybride Umgebungen
- Zuletzt: lokal betriebene Modelle, KI-Agenten, Prozessautomatisierung
- B.Sc. Cyber Security · IHK Certified IT Systems Manager

com.in Computersystemhaus

IT-Systemhaus in Schwerin – seit über 35 Jahren

- 14 Mitarbeiter
- Kunden: Mittelstand, öffentliche Einrichtungen und Organisationen
- IT-Strategie, Infrastruktur und Cloud, Managed Services, IT-Sicherheit
- Beschaffung und Lizenzmanagement
- KI-Beratung: Potenzialanalyse, Pilotprojekte, Automatisierung, Schulung

Com In GmbH Computersystemhaus · Graf-Schack-Allee 11 · 19053 Schwerin · www.comin.info · a.flindt@comin.de

Warum sich die Rechnung gerade jetzt lohnt

1 Die Preise sind gefallen

- Kleine Modelle kosten ab 0,10 \$ pro Million Eingabe-Token
- Kleine Modelle erledigen, wofür früher die großen nötig waren
- Projekte werden wirtschaftlich, die es vor zwei Jahren nicht waren
- GPT-3.5-Niveau: Ende 2022 noch 20 \$, Ende 2024 nur 0,07 \$ pro Million Token

2 Die Auswahl ist größer geworden

- Abo, API und eigener Betrieb stehen gleichwertig nebeneinander
- Auch kleine Unternehmen können selbst betreiben
- Viel mehr Modelle auf dem Markt

3 Die Fragen haben sich verschoben

- Nicht mehr: Kann die Technik das überhaupt?
- Sondern: Dürfen unsere Daten dorthin?
- Und: Wer betreibt das bei uns dauerhaft?

Über welche KI-Projekte reden wir?

Drei beispielhafte Typen von Vorhaben – mit sehr unterschiedlichen Kosten und Aufwänden.

1 Chat-Assistent

- Texte, E-Mails und Angebote entwerfen
- Übersetzen, zusammenfassen, recherchieren
- **Typischer Weg: Abo pro Nutzer**

2 Wissensmanagement

- Fragen an Handbücher, Verträge und Richtlinien
- Suche über Tickets, Protokolle und Ablagen
- **Typischer Weg: API oder eigene Hardware**

3 KI-Agenten für Routineaufgaben

- E-Mail-Eingang vorsortieren und beantworten
- Rechnungen und Belege auslesen
- Prozessautomatisierung
- **Typischer Weg: API oder eigene Hardware**

Von links nach rechts wächst die Integration – und damit der Aufwand in Personentagen.

Eine Aufgabe. Drei Preisschilder.

Beispiel: eingehende E-Mails lesen, einordnen und beantworten – dreimal dieselbe Aufgabe, dreimal ein anderer Weg.

1 Abo pro Nutzer

- Feste Gebühr je Mitarbeitenden / Monat
- Mensch sitzt davor und tippt
- Sofort startklar, keine IT nötig

ab ≈ 14 € / Nutzer / Monat

2 API, nutzungsabhängig

- Abrechnung nach verarbeiteter Textmenge
- Läuft im Hintergrund, ohne Mensch
- Braucht eine Anbindung

ab 0,10 \$ / Mio. Token

3 Eigene Hardware

- Einmalige Investition plus Betrieb
- Daten bleiben im Haus
- Braucht Betriebskompetenz

ab ≈ 2.900 € Investition

Das Ergebnis ist in allen drei Fällen vergleichbar. Der Unterschied liegt in der Menge – und in der Frage, wo die Daten liegen dürfen.

Abo pro Nutzer: der einfachste Einstieg

Feste Gebühr je Mitarbeitenden, sofort nutzbar, keine Integration nötig.

Microsoft 365 Copilot Business

18,20 €

monatlich kündbar: 21,84 €
Add-on zur M365-Lizenz

ChatGPT Business

17,65 €

monatlich kündbar: 21,85 €
ab 2 Nutzern

Claude Team (Standard)

18 €

monatlich kündbar: 21,04 €
2 bis 150 Nutzer

Google Workspace Business Standard

13,60 €

Gemini inklusive –
mit Mail, Drive und Meet

Was im Abo enthalten ist

- Oberfläche, Nutzerverwaltung, Support
- Nutzung im Rahmen fairer Obergrenzen
- Anbindung an Office bzw. Workspace

Was nicht enthalten ist

- Anbindung an Fachanwendungen
- Verarbeitung ohne Mensch davor
- Schulung und Einführung im Betrieb

Listenpreise pro Nutzer und Monat bei Jahresvertrag, zzgl. MwSt. Stand 5. Oktober 2026, laut Herstellerseiten.

Wie viele Token braucht eine Aufgabe?

Richtwerte pro Vorgang und Kosten für 1.000 Vorgänge in US-Dollar – am Beispiel der drei OpenAI-Modelle.

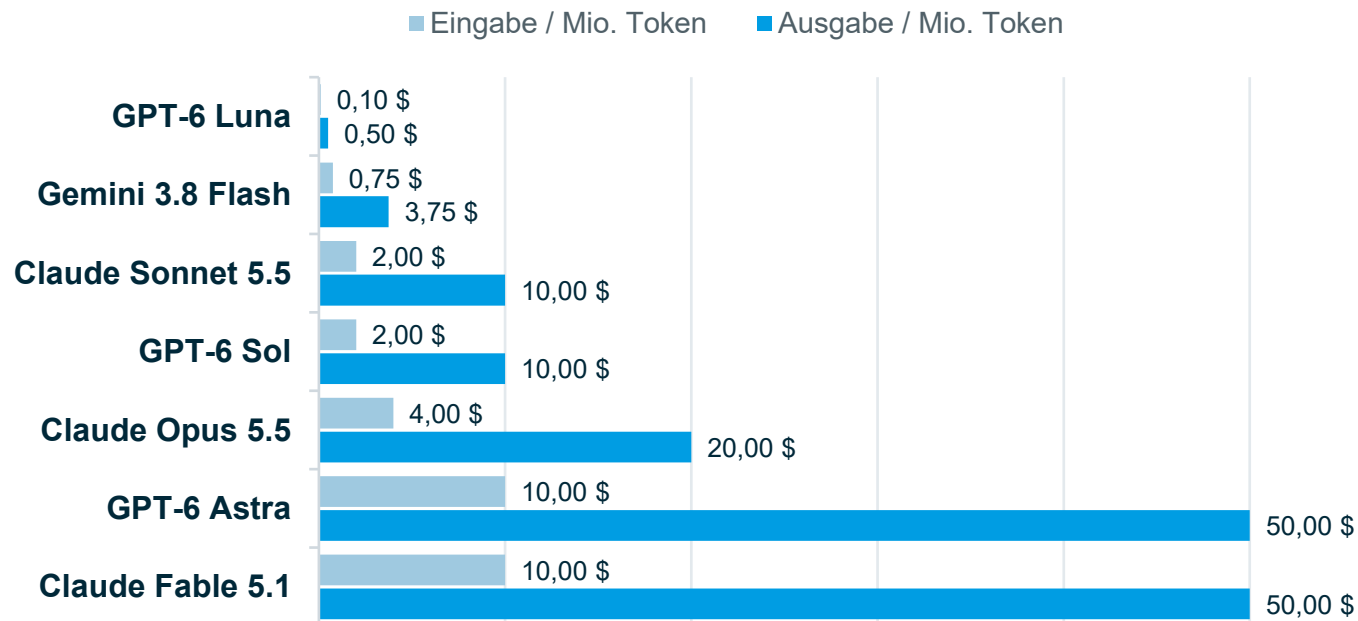
Tätigkeit	Eingabe	Ausgabe	klein GPT-6 Luna	mittel GPT-6 Sol	groß GPT-6 Astra
E-Mail einordnen	800	50	0,11 \$	2,10 \$	10,50 \$
Antwort auf E-Mail entwerfen	1.500	400	0,35 \$	7 \$	35 \$
Rechnung oder Beleg auslesen	2.000	300	0,35 \$	7 \$	35 \$
Frage an die Wissensdatenbank	6.000	500	0,85 \$	17 \$	85 \$
Protokoll aus 1 Std. Besprechung	12.000	1.500	1,95 \$	39 \$	195 \$
Vertrag mit 20 Seiten zusammenfassen	15.000	1.000	2 \$	40 \$	200 \$

1.000 Vertragszusammenfassungen mit dem größten OpenAI-Modell kosten rund 200 \$.

Token-Werte sind Schätzungen für deutsche Texte; die tatsächliche Menge hängt von Anweisung, Kontext und Antwortlänge ab.

Was ein Token kostet

Zwischen dem kleinsten und dem größten Modell liegt Faktor 100. Anbieter aus der EU spielen preislich mit.



Aus der EU, mit AVV

Eingabe / Ausgabe je Mio. Token

IONOS · Deutschland

gpt-oss-120b

0,15 € / 0,65 €

Scaleway · Frankreich

GLM-5.2 (Z.ai, China)

1,80 € / 5,50 €

Mistral · Frankreich

Mistral Large 3

0,50 \$ / 1,50 \$

Azure, AWS, Google

GPT, Claude, Gemini in EU-Region

Liste + 10 %

Meist offene Modelle – bei komplexen Aufgaben schwächer.

Ausgabe ist fünfmal teurer als Eingabe. Wer Antworten kurz hält und Aufgaben, die keine Eile haben, im Stapel verarbeitet, spart bis zu 50 %.

Listenpreise der Hersteller bzw. über OpenRouter, ohne MwSt., Stand 5. Oktober 2026. 1.000 Token ≈ 750 englische Wörter, deutscher Text braucht etwas mehr. EU-Anbieter laut Anbieterseiten; DSGVO und AVV im Einzelfall prüfen.

Rechenbeispiel: E-Mail-Eingang automatisch vorsortieren

Annahmen

- 200 E-Mails pro Arbeitstag
- 1.500 Token Eingang, 400 Token Antwortentwurf
- 250 Arbeitstage – 50.000 Vorgänge im Jahr
- Kleine Modellklasse (GPT-6 Luna)
- Kein Mensch im Regelbetrieb, nur Stichprobe

Ergebnis

≈ 18 \$

pro Jahr mit dem kleinen Modell (GPT-6 Luna)

Mit dem mittleren Modell (GPT-6 Sol) ≈ 350 \$, aus einem deutschen Rechenzentrum (IONOS, gpt-oss-120b) ≈ 24 €.

Die Verarbeitung ist nicht das Teure. Teuer wird, was davor und danach passiert – und genau dazu kommen wir gleich.

Bitkom, September 2026: Nur 8 % der KI-Anwender nennen den Token-Verbrauch einen großen Kostenpunkt.

Eigene Hardware ist eine Datenschutzentscheidung

Keine Kostenentscheidung. In unseren Projekten gibt fast immer die Vertraulichkeit den Ausschlag.

Das gibt den Ausschlag

- Personal- und Bewerberdaten
- Mandanten-, Patienten- oder Klientendaten
- Verträge, Kalkulationen, Konstruktionsdaten
- Geheimhaltungspflichten gegenüber Kunden
- Betriebsvereinbarung oder Auflagen des Betriebsrats

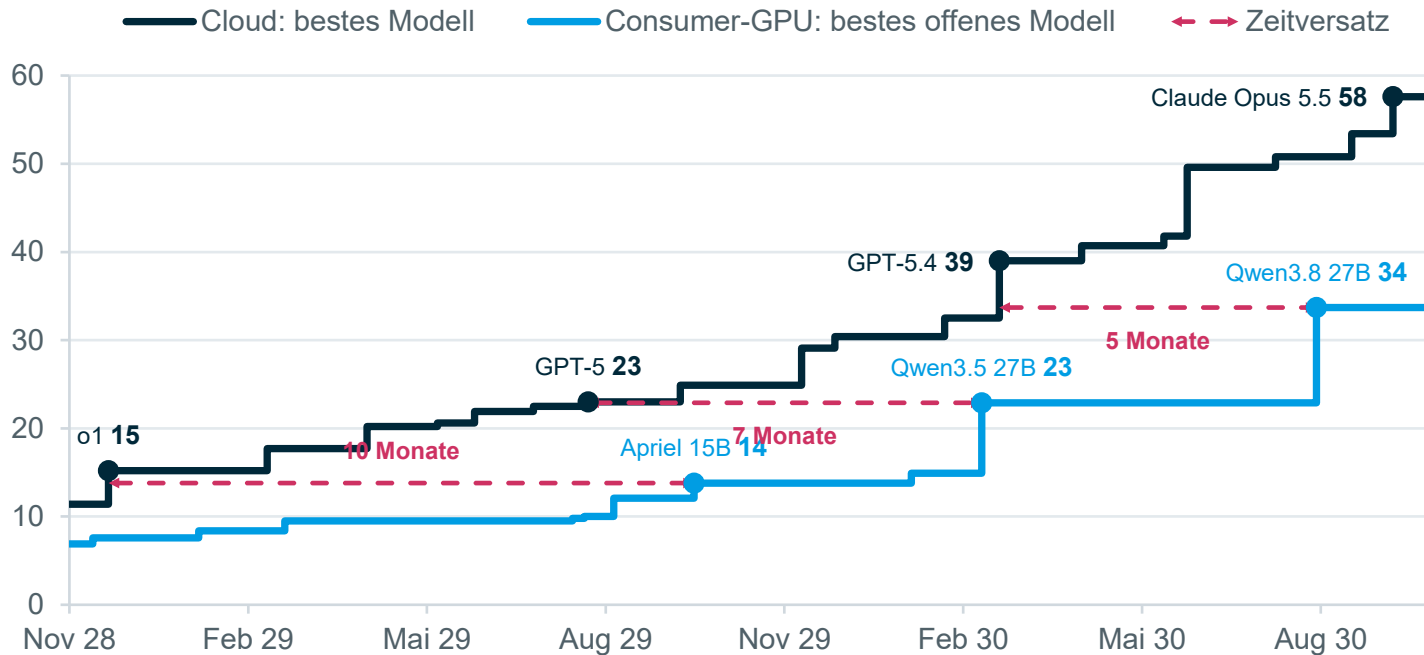
Das gibt selten den Ausschlag

- Der Preis pro Anfrage
- Die Hoffnung auf ein besseres Modell
- Unabhängigkeit vom Anbieter
- Der Wunsch, es selbst zu können

Erste Frage im Projekt: Dürfen diese Daten das Haus verlassen? Erst danach reden wir über Preise.

Lokal heißt nicht abgehängt: rund ein halbes Jahr Rückstand

Artificial Analysis Intelligence Index. Rote Linien: Monate, bis derselbe Wert auf einer einzelnen Gaming-Grafikkarte lief.



Stand August 2026

Qwen3.8 27B erreicht, was es in der Cloud erstmals im März gab – mit GPT-5.4.

Seit Anfang 2025

5 bis 10 Monate, bis das Spitzenniveau lokal verfügbar war.

Qwen3.8 27B läuft auf Stufe 1 der Kostenleiter: Spitzenniveau vom Frühjahr – die Daten bleiben im Haus.

Quelle: Artificial Analysis Intelligence Index v4.3.2, artificialanalysis.ai/leaderboards/models, abgerufen am 05.10.2026. Bestwert zum jeweiligen Datum; Werte vor 2026 teils von Artificial Analysis geschätzt. Consumer-GPU: offene Modelle bis 40 Mrd. Parameter (Abgrenzung wie Epoch AI).

Die Kostenleiter: vom Arbeitsplatz zum Cluster

Stufe 1 und 2 für Einzelplätze und kleine Teams – ab Stufe 3 ist es ein Rechenzentrumsprojekt.

1 Stufe 1

Workstation mit Consumer-GPU

≈ 6.400 €

netto, z. B. HP OMEN mit NVIDIA RTX 5090

- Ein Arbeitsplatz oder eine kleine Testgruppe
- Kleine bis mittlere Modelle

2 Stufe 2

Kompaktsystem für Büro oder Serverraum

≈ 2.900–4.900 €

netto, z. B. Mac mini M5 Pro oder NVIDIA DGX Spark

- Bis 128 GB Speicher – auch größere Modelle
- Langsamer bei vielen Nutzern

3 Stufe 3

GPU-Server für den Serverraum

≈ 117.000 €

netto, HPE DL380a, 2× NVIDIA H200 NVL

- Große Modelle, viele Nutzer parallel
- Braucht Rack, Klima und Notstrom

4 Stufe 4

GPU-Knoten für einen KI-Cluster

≈ 405.000 €

netto, HPE XD685, 8× NVIDIA H200

- Eigene Strom- und Kühlplanung
- Nur ein Knoten – ein Cluster hat mehrere

Zwischen Stufe 1 und einem einzelnen Cluster-Knoten liegt Faktor 60. Die Software ist dieselbe.

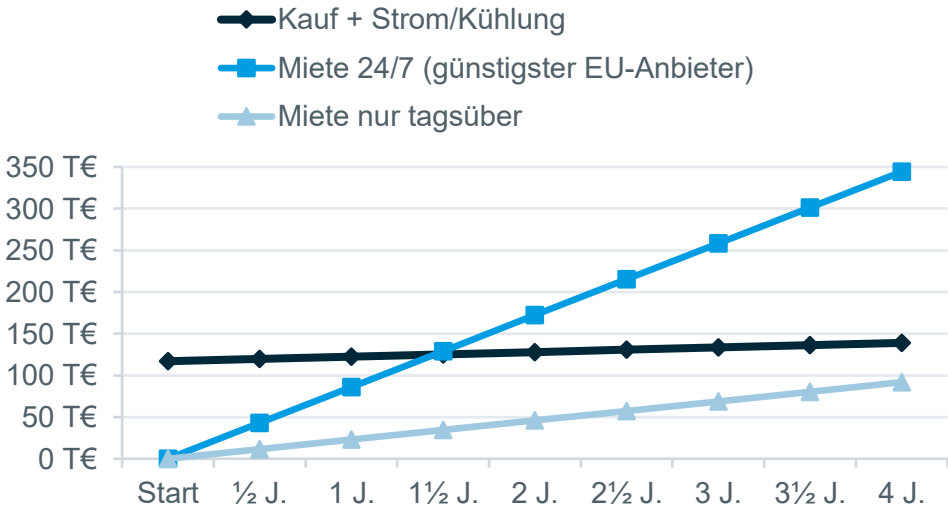
Netto, gerundet. Stufe 1–2 aus Händlerpreisen Oktober 2026; Stufe 3–4 Distributorpreise, Juni 2026, inkl. 3 Jahre Support. Stufe 3–4 nur der Server – ohne Switches, Storage, Rack, USV, Einrichtung, Strom und Kühlung.

Kaufen oder mieten: Die Auslastung entscheidet

Muss es ins Haus, ist Miete nur die Probefahrt vor dem Kauf. Dürfen die Daten in die EU, lohnt sich Miete statt API, wenn ein eigenes Modell gebraucht wird oder das Volumen dauerhaft hoch ist.

Stufe	Kauf, einmalig	Miete EU, Monat	Kauf lohnt, 24/7*	Kauf lohnt, nur tagsüber*
1 · Workstation RTX 5090, 32 GB	≈ 6.400 €	≈ 215 € Hetzner GEX45, 24 GB	≈ 3 J.	≈ 2½ J.
2 · Kompaktsystem Mac mini, DGX Spark	≈ 2.900–4.900 €	≈ 1.200 € Hetzner GEX131, 96 GB	≈ 2–4 Mon.	≈ 2–4 Mon.
3 · GPU-Server 2× H200, Serverraum	≈ 117.000 €	≈ 7.200–11.700 € Nebius · STACKIT	≈ 1–2 J.	≈ 4–8 J.
4 · Cluster-Knoten 8× H200, Rechenzentrum	≈ 405.000 €	≈ 28.400–30.700 € Nebius · OVHcloud, 8× H200	≈ 1–2½ J.	≈ 6–7 J.

Beispiel Stufe 3: kumulierte Kosten in Tsd. €



Ab Stufe 3 entscheidet die Auslastung: Rund um die Uhr lohnt der Kauf nach 1–2 Jahren, nur tagsüber erst nach 4–8.

* Netto, gerundet, ohne Personalkosten. Kauf: Gerät bzw. Server inkl. 3 J. Support, Strom 25 ct/kWh, ab Stufe 3 plus 40–80 % Kühlung; obere Spanne mit +20 % für Rack, USV und Switches. Miete: Listenpreise von EU-Anbietern mit AVV, Stand 5.10.2026, inkl. Einrichtung, Speicher und Traffic. Stufe 1–2: Hetzner, Monatsmiete – Teilzeit spart nichts. Stufe 3–4: stundenweise bei Nebius (NL), OVHcloud (FR), STACKIT (DE, nur H100). Tagsüber = 8 Std. an 250 Tagen + 10 % für Start und Laden des Modells; der gekaufte Server läuft durch. 1 € = 1,1225 \$ (EZB, 2.10.2026).

Datenfrage vor Preisfrage

Erst klären, wo die Daten hin dürfen – dann rechnen. Der Preis entscheidet über die Größe der Lösung, nicht über den Weg.

Frage 1 Wo dürfen die Daten verarbeitet werden? – entscheidet sich an Personenbezug, Geschäftsgeheimnissen und Kundenverträgen

NUR IM EIGENEN HAUS

Eigene Hardware

Hardware-Stufe nach Auslastung.

Gemietete GPUs zählen nicht – auch beim EU-Anbieter verlassen die Daten das Haus. Miete höchstens als Probefahrt vor dem Kauf.

Frage 2 entfällt: Nutzer und Prozesse laufen auf derselben Hardware.

INNERHALB DER EU

EU-Anbieter oder EU-Region

IONOS, STACKIT, Mistral, Scaleway – oder EU-Regionen von Microsoft, Google, AWS (Daten in der EU, Konzern unter US-Recht).

AUCH AUSSERHALB DER EU

Globale Anbieter direkt

Drittlandtransfer absichern: Anbieter nach EU-US Data Privacy Framework zertifiziert oder Standardvertragsklauseln.

Frage 2 Arbeitet ein Mensch damit – oder ein Prozess?

MENSCH → Abo pro Nutzer

nur für die, die es wirklich nutzen

PROZESS → API nach Volumen

abgerechnet pro Token, mit Budgetdeckel

Gilt für jeden externen Anbieter: Auftragsverarbeitungsvertrag nach Art. 28 DSGVO, Daten auf das Nötige reduzieren, Zugriffe und Löschfristen regeln. Eine EU-Region ersetzt keine dieser Pflichten.

Was in keiner Preisliste steht

Ab hier rechnen wir nicht mehr in Euro, sondern in Personentagen.



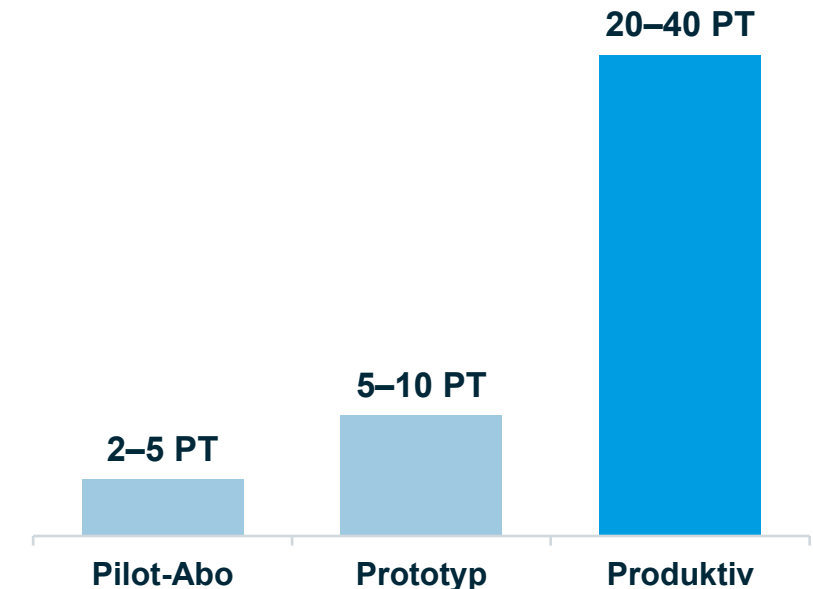
Erfahrungswerte aus Projekten. Die Lizenz- oder Token-Kosten stecken in keinem dieser drei Blöcke.

Bitkom, Sept. 2026: 50 % der KI-Anwender nennen Datenaufbereitung einen großen Kostenpunkt, 41 % Integration – Lizenzen nur 21 %, Token 8 %.

Zwischen „funktioniert“ und „läuft im Betrieb“

Realistische Aufwände aus Projekten – in Personentagen, nicht in Euro.

Vorhaben	Aufwand	Wer macht es?
Abo für eine Pilotgruppe ausrollen	2–5 PT	IT und Fachbereich
Prototyp einer Prozessautomatisierung	5–10 PT	IT oder Dienstleister
Derselbe Prozess produktiv	20–40 PT	IT und Dienstleister
Datenaufbereitung und Zugriffsklärung	10–30 PT	Fachbereich
Eigene Hardware aufbauen und absichern	10–20 PT	IT und Dienstleister
Laufender Betrieb nach Livegang	1–3 PT / Monat	IT



Paretoprinzip: Die ersten 80 % sind schnell gebaut – die letzten 20 % machen die Arbeit.

„Ein produktionsreifes GenAI-System kann um Größenordnungen teurer sein als ein Pilot.“ – Gartner, 2026 (übersetzt)

Fazit: Klein anfangen, ehrlich rechnen

„Wer nur eine Lizenz kauft, hat KI eingekauft, aber noch nichts verändert.“ – Dr. Ralf Wintergerst, Bitkom-Präsident, September 2026

1 Datenfrage vor Preisfrage

Erst klären, wohin die Daten dürfen. Der Preis bestimmt danach die Größe der Lösung – nicht den Weg.

2 In Personentagen rechnen

Lizenz und Token sind die kleinsten Posten. Daten, Integration und Betrieb bestimmen die Rechnung.

3 Betrieb von Anfang an mitplanen

Ein Pilot ohne Zuständigen wird nie produktiv. Zuständigkeit und Budget gehören in die erste Kalkulation.

✓ Jetzt tun

- Einen Prozess wählen, der sofortigen Mehrwert verspricht – nicht den spannendsten oder kompliziertesten
- Abo für eine kleine Gruppe, nach drei Monaten Nutzung messen
- Aufwand in Personentagen schätzen, Verantwortlichen benennen

! Noch warten

- Lizenzen für die gesamte Belegschaft auf einmal
- Eigene Hardware ohne konkreten Vertraulichkeitsgrund

Der günstigste nächste Schritt ist der kleinste, der echte Zahlen liefert – keiner davon kostet mehr als ein paar Personentage.

Vielen Dank. Fragen gern in der Pause oder per Mail.
Alex Flindt · IT System Engineer · a.flindt@comin.de



Com In Unternehmensgruppe | Graf-Schack-Allee 11
19053 Schwerin | GERMANY
Tel.: +49 385 59333-0 | E-Mail: info@comin.de | www.comin.de

Weiterlesen und selbst ausprobieren

Empfehlungen zum Vertiefen – alle Links geprüft am 6. Oktober 2026.

1 Lage und Studien

Bitkom KI-Studie 2026 – wie deutsche Unternehmen KI nutzen
<https://www.bitkom.org/Presse/Presseinformation/Erstmals-nutzt-Mehrheit-Unternehmen-KI>

Stanford AI Index – jährlicher Überblick zu Leistung, Kosten und Verbreitung
<https://hai.stanford.edu/ai-index>

Epoch AI – wie schnell die Preise je Leistungsniveau fallen
<https://epoch.ai/data-insights/llm-inference-price-trends>

2 Modelle und Preise vergleichen

Artificial Analysis – Leistung, Tempo und Preis je Modell
<https://artificialanalysis.ai/leaderboards/models>

OpenRouter – Preise vieler Modelle und Anbieter nebeneinander
<https://openrouter.ai/models>

OpenAI Tokenizer – zeigt, wie viele Token ein eigener Text hat
<https://platform.openai.com/tokenizer>

3 Selbst ausprobieren

Ollama – offene Modelle auf dem eigenen Rechner betreiben
<https://ollama.com>

LM Studio – Desktop-App für lokale Modelle, ohne Kommandozeile
<https://lmstudio.ai>

KI-Readiness-Check (DFKI) – den KI-Reifegrad im eigenen Betrieb messen
<https://werner.dfki.de/readiness-welcome>

4 Recht und Anlaufstellen

KI-Service Desk der Bundesnetzagentur – Infos zur KI-Verordnung
<https://www.bundesnetzagentur.de/ki>

EU-Kommission: KI-Kompetenz – FAQ zur Schulungspflicht
<https://digital-strategy.ec.europa.eu/de/faqs/ai-literacy-questions-answers>

Mittelstand-Digital Zentrum Rostock – kostenfreie Workshops
<https://www.digitalzentrum-rostock.de/kuenstliche-intelligenz/>

Quellen der Zahlen in diesem Vortrag

Preise ohne MwSt., Stand 5. Oktober 2026. Hardware-Kaufpreise aus Händler- und Distributorangeboten, daher ohne Link.

1 Abo und API

Microsoft 365 Copilot Business

<https://www.microsoft.com/de-de/microsoft-365/copilot/business>

ChatGPT Business

<https://chatgpt.com/de-DE/pricing>

Claude Team

<https://claude.com/de/pricing>

Google Workspace Business Standard

<https://workspace.google.com/pricing>

Claude-Modelle (API)

<https://claude.com/pricing>

GPT-6 und Gemini 3.8 Flash (API, über OpenRouter)

<https://openrouter.ai/api/v1/models>

2 EU-Anbieter und Miete

IONOS AI Model Hub

<https://cloud.ionos.de/managed/ai-model-hub>

STACKIT AI Model Serving

<https://stackit.com/en/products/data-ai/stackit-ai-model-serving>

Mistral

<https://mistral.ai/pricing>

Scaleway Generative APIs

<https://www.scaleway.com/en/pricing/model-as-a-service>

Hetzner GPU-Server

<https://www.hetzner.com/dedicated-rootserver/matrix-gpu/>

Nebius

<https://nebius.com/prices>

OVHcloud

<https://www.ovhcloud.com/de/public-cloud/prices/>

STACKIT Cloud-Preise

<https://stackit.com/de/preise/cloud>

3 Studien und Einordnung

Bitkom, KI-Studie, 14.09.2026

<https://www.bitkom.org/Presse/Presseinformation/Erstmal-s-nutzt-Mehrheit-Unternehmen-KI>

Stanford AI Index Report 2025

<https://hai.stanford.edu/ai-index/2025-ai-index-report>

Epoch AI, LLM-Preistrends

<https://epoch.ai/data-insights/llm-inference-price-trends>

Artificial Analysis Intelligence Index

<https://artificialanalysis.ai/leaderboards/models>

Gartner zu KI-Agenten, 25.06.2025

<https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>

Gartner Hype Cycle GenAI 2026 (The Register)

<https://www.theregister.com/ai/ml/2026/05/28/most-generative-ai-and-custom-model-projects-will-be-a-bust-gartner/5247633>

Gartner zu Pilot und Betrieb (Campus Technology)

<https://campustechnology.com/articles/2026/06/22/gartner-half-of-gen-ai-projects-could-exceed-budget-by-2028.aspx>