

Nimm für jede Aufgabe das KI-Modell, das sie **am besten und günstigsten** löst

Neun KI-Modelle in vier Stufen, jedes mit Use Cases, Preis und Zugang. Dazu der Weg zum passenden Modell für deine Aufgabe.



Claude
Opus, Sonnet



GPT-6
Astra, Sol, Luna



Gemini
Flash, Argon



Grok
von xAI



Qwen
läuft lokal



Berkan Ceyhan · Stand Oktober 2026

Das Wichtigste auf einen Blick

- 01** Auf Platz eins steht Claude: **Opus 5.5 führt den Intelligenz-Index von Artificial Analysis an.** Sonnet 5.5 liegt zwei Punkte dahinter, kostet die Hälfte und läuft schon im kostenlosen Claude-Plan.
- 02** OpenAI deckt drei Rollen ab: **Astra bedient deinen Computer, Sol automatisiert günstig, Luna erledigt Massenarbeit.**
- 03** Grok 4.7 und Gemini 3.8 Flash passen für einfache Aufgaben. **Qwen 3.8 27B läuft ohne Cloud auf deinem Rechner.**
- 04** Gemini 4 Argon ist noch nicht frei, **Google nennt kein Datum.** Teste es gegen fünf eigene Aufgaben, sobald du es nutzen darfst.

Claude belegt **Platz eins** der Tier-List

Die Stufen sind meine Einordnung für Agenturen, Coaches und Dienstleister. Als Gegencheck dient der Intelligenz-Index von Artificial Analysis: Die Zahl am Modell ist sein Wert dort. Alle Preise stehen in Dollar pro Million Token, Stand Oktober 2026.

S

S-Tier

Die besten Ergebnisse



Claude Opus 5.5

Index 58



Claude Sonnet 5.5

Index 56

A

A-Tier

Computer bedienen,
günstig automatisieren



GPT-6 Astra

Index 53



GPT-6.1 Sol

Index 52

B

B-Tier

Schnell und billig für
einfache Aufgaben



Grok 4.7

Index 46



Gemini 3.8 Flash

Index 41



GPT-6 Luna

Index 38

L

Lokal

Läuft ohne Cloud auf
deinem Rechner



Qwen 3.8 27B

Index 34

?

Noch nicht frei

Kann das Ranking
drehen



Gemini 4 Argon

Index 53

S-Tier: Opus für das Schwere, Sonnet für alles andere

 Claude Opus 5.5 S-TIER · ANTHROPIC · INDEX 58	 Claude Sonnet 5.5 S-TIER · ANTHROPIC · INDEX 56
Anthropic baut Opus für lange, selbstständige Arbeit an Code und Wissensaufgaben. Im Index steht es auf Platz eins, vor dem teureren Fable 5.1 (10 \$ / 50 \$). Laut Anthropic erreicht Opus bei den meisten Aufgaben Fables Niveau.	Dein Standard. Bei Wissensarbeit liegt Sonnet fast gleichauf mit Opus (GDPval-AA: 1.844 gegen 1.846 Punkte) und kostet die Hälfte. Bei komplexer, offener Arbeit, die lange Urteilsvermögen verlangt, bleibt Opus laut Anthropic klar stärker. Scheitert Sonnet an der Tiefe, schick dieselbe Aufgabe an Opus.
✓ Recherche mit Quellen ✓ Konzepte und Angebote aus vielen Unterlagen ✓ Agenten, die lange allein arbeiten ✓ Größere Automatisierungen und Code	✓ E-Mails, Texte und Entwürfe ✓ Zusammenfassungen und Auswertungen ✓ Code und Automatisierungen ✓ Alltags-Agenten
Preis 4 \$ / 20 \$	Preis 2 \$ / 10 \$
Zugang Claude ab Pro (20 \$ im Monat), API <code>claude-opus-5-5</code>	Zugang Claude Free, API <code>claude-sonnet-5-5</code>

① KURZ ERKLÄRT

Was „pro Million Token“ heißt

Ein Token ist ein Wortstück, eine Million Token sind grob eine halbe Million Wörter. Der erste Preis gilt für das, was du schickst, der zweite für das, was die KI zurückschreibt.



A-Tier: Astra bedient deinen Computer, **Sol** automatisiert günstig



GPT-6 Astra

A-TIER · OPENAI · INDEX 53

OpenAI baut Astra, damit die KI Programme selbst bedient. Auf OSWorld 2.0, dem Test für Computerbedienung, erreicht es 72,6 %. Dafür zahlst du das Fünffache von Sonnet.

- ✓ Formulare und Portale ausfüllen
- ✓ Kundendaten im CRM pflegen
- ✓ Kalender und Postfach ordnen
- ✓ Recherche im Browser

Preis 10 \$ / 50 \$

Zugang API `gpt-6-astra`, ChatGPT ab Plus (in Plus über Work und Codex)



GPT-6.1 Sol

A-TIER · OPENAI · INDEX 52

OpenAI nennt Sol „near-Astra“: fast dieselbe Leistung für ein Fünftel des Preises. Es ist langsam, 57,7 Token pro Sekunde (Platz 137 von 224 im Tempo-Vergleich). Für Abläufe im Hintergrund stört das nicht.

- ✓ Wiederkehrende Abläufe
- ✓ Agenten im Hintergrund
- ✓ Coding-Aufgaben
- ✓ Aufgaben, für die Astra zu teuer wäre

Preis 2 \$ / 10 \$

Zugang API `gpt-6.1-sol`, ChatGPT Work und Codex ab Plus

⚠ ACHTUNG

Computerbedienung wählst du nicht nach einer Zahl

Auch Claude bedient Computer: Anthropic meldet 81,8 % für Opus 5.5 und 80,1 % für Sonnet 5.5 auf OSWorld 2.1. Das ist eine andere Testversion als die 72,6 % von Astra. Lass deinen eigenen Ablauf mit beiden laufen.

B-Tier: schnell und billig, solange die Aufgabe einfach bleibt



GPT-6 Luna

B-TIER · OPENAI · INDEX 38

Das günstigste Modell in der Cloud, zwanzigmal billiger als Sol. OpenAI baut es für große Mengen einfacher Arbeit. Es ist das einzige GPT-6-Modell im kostenlosen ChatGPT-Plan.

- ✓ Mails und Tickets sortieren
- ✓ Daten aus Dokumenten auslesen
- ✓ Texte in Kategorien einordnen
- ✓ Kurze Routine-Zusammenfassungen

Preis 0,10 \$ / 0,50 \$

Zugang ChatGPT Free und Go (Desktop-App), API `gpt-6-luna`



Gemini 3.8 Flash

B-TIER · GOOGLE · INDEX 41

Das schnellste der drei B-Tier-Modelle: 242 Token pro Sekunde, rund das 1,7-Fache von Luna und fast das Dreifache von Grok. Google nennt es sein intelligentestes Flash-Modell, gebaut auch für Agenten und lange Software-Aufgaben.

- ✓ Antworten ohne Wartezeit
- ✓ Entwürfe für Chat und Support
- ✓ Große Textmengen schnell auswerten

Preis 0,75 \$ / 3,75 \$, Einführungspreis bis 31.12.2026

Zugang Gemini API `gemini-3.8-flash`



Grok 4.7

B-TIER · XAI · INDEX 46

Grok 4.7 ist das stärkste Modell von xAI für Code, Agenten und Wissensarbeit, mit 500.000 Token Kontext. Es steht im B-Tier, weil es im Index sechs Punkte hinter Sol liegt. Im Index-Test schrieb es 240 Mio. Token, fast das Dreifache des Medians. Das macht es teurer, als der Listenpreis vermuten lässt.

- ✓ Code und Agenten
- ✓ Lange Dokumente
- ✓ Zweitmeinung zu Claude und GPT

Preis 2 \$ / 6 \$ (bis 200.000 Token Eingabe)

Zugang xAI API `grok-4.7`



Qwen 3.8 läuft auf deinem Rechner, deine Daten bleiben dort



Qwen 3.8 27B

LOKAL · ALIBABA · INDEX 34

Die Gewichte sind frei (Apache 2.0, auch für den Einsatz im Geschäft). In der 4-Bit-Fassung reichen etwa 16 GB Grafikspeicher. An Opus (58) kommt es nicht heran, dafür verlässt kein Kundendokument deinen Rechner.

- ✓ Vertrauliche Kundendaten
- ✓ Verträge und interne Dokumente
- ✓ Arbeiten ohne Internet
- ✓ Bilder und Dokumente auslesen

Preis Keine Token-Kosten, nur Hardware und Strom

Zugang Hugging Face `Qwen/Qwen3.8-27B`, Ollama, llama.cpp, vLLM

⚠ ACHTUNG

Mit „Qwen 3.8 lokal“ ist nur die 27B-Fassung gemeint

Zu Qwen 3.8 gehört auch ein Modell mit 2,4 Billionen Parametern. Seine Gewichte sind offen, aber es braucht ein Rechenzentrum und steht unter einer eigenen Lizenz.

Mehr Intelligenz kostet **nicht immer mehr**

Sonnet liegt zwei Punkte unter Opus und kostet die Hälfte. Astra kostet das Fünffache von Sonnet und liegt drei Punkte darunter: Bei Astra bezahlst du für die Computerbedienung.

MODELL	INDEX-PUNKTE	PREIS PRO MIO. TOKEN
 Opus 5.5		58 4 \$ / 20 \$
 Sonnet 5.5		56 2 \$ / 10 \$
 GPT-6 Astra		53 10 \$ / 50 \$
 Gemini 4 Argon		53 2 \$ / 10 \$ Einführungspreis, noch nicht frei
 GPT-6.1 Sol		52 2 \$ / 10 \$
 Grok 4.7		46 2 \$ / 6 \$
 Gemini 3.8 Flash		41 0,75 \$ / 3,75 \$ Einführungspreis bis 31.12.2026
 GPT-6 Luna		38 0,10 \$ / 0,50 \$
 Qwen 3.8 27B		34 0 \$ läuft auf eigener Hardware

Intelligence Index v4.3.2 von Artificial Analysis, höchste gemessene Stufe je Modell, Abruf 6. Oktober 2026. Die Tests laufen bei Grok (xhigh) und Gemini (high) auf anderen Stufen als bei den übrigen. Preise: Stand Oktober 2026.

Vier Fragen führen dich zum passenden Modell

1

Dürfen die Daten deinen Rechner nicht verlassen?

JA →



Qwen 3.8 27B

läuft lokal, ohne Cloud

Nein, dann weiter

2

Soll die KI Programme und den Browser bedienen?

JA →



GPT-6 Astra

wenn das Ergebnis zählt



GPT-6.1 Sol

wenn es günstig sein soll

Nein, dann weiter

3

Ist die Aufgabe einfach und läuft hundertfach?

JA →



GPT-6 Luna

am günstigsten



Gemini 3.8 Flash

am schnellsten

Nein, dann weiter

4

Ist die Aufgabe lang, offen und braucht Urteilsvermögen?

JA →



Claude Opus 5.5

für das Schwere

Nein, dann weiter

5

Keine Frage trifft zu.

DANN →



Claude Sonnet 5.5

dein Standard



Gemini 4 Argon kann das Ranking drehen, steht aber noch nicht zur Wahl



Gesichert

Was heute feststeht

Google hat Argon am 30. September 2026 vorgestellt. Zugang haben zunächst nur Cyber-Partner im Fairwind-Programm, danach zahlende API-Kunden und Google-AI-Ultra-Abos.



Offen

Was noch niemand weiß

Google nennt kein Datum für die breite Freigabe. Es sagt, Argon schlage Astra, Fable und Opus, und verweist auf den Vals-Index. Artificial Analysis misst 53 Punkte, gleichauf mit Astra und fünf hinter Opus. Der Preis von 2 \$ / 10 \$ ist ein Einführungspreis.

🔔 MEIN TIPP

Schreib dir fünf echte Aufgaben auf, bevor Argon kommt

Nimm fünf Aufgaben aus deinem Alltag und lass sie heute durch Sonnet und Opus laufen. Sobald Argon frei ist, schickst du dieselben fünf hinterher und entscheidest nach deinen Ergebnissen.



Automatisierungs- Call für Agenturen, Coaches und Dienstleister

In 30 Minuten schauen wir gemeinsam, wo KI-Agenten und automatisierte Prozesse in deinem Betrieb zuerst Zeit und Geld sparen. Kostenlos.

FÜR WEN IST DAS?

- ✓ Agenturen, Coaches und Dienstleister
- ✓ ab 10.000 € Monatsumsatz
- ✓ ein klarer Flaschenhals im Tagesgeschäft
- ✓ du willst KI im Betrieb einsetzen

Call buchen

[calendly.com/
berkanceyhan/
automatisierungscall](https://calendly.com/berkanceyhan/automatisierungscall)



Scannen & Termin wählen