

KI-Kosten senken: Modell und Effort

Was eine Aufgabe wirklich kostet und mit welchen zwei Reglern du das steuerst

Ein **Token** ist die Grundeinheit, in der KI-Modelle Text verarbeiten und abrechnen: Eingabe, Ausgabe und bei Reasoning-Modellen oft unsichtbare Denk-Tokens, die etwa OpenAI und Google als bezahlte Ausgabe berechnen. Was zählt, ist der **Preis der ganzen Aufgabe**: wie viele Tokens das Modell braucht, bis ein brauchbares Ergebnis vorliegt.

DIE STUDIE ZUM PREIS PRO AUFGABE

- **Die Studie:** Forscher von Stanford, UC Berkeley, Microsoft Research und Carnegie Mellon verglichen 8 Reasoning-Modelle an 12 Aufgaben.
- **Das Ergebnis:** In **32 Prozent** der paarweisen Vergleiche war das laut Preisliste billigere Modell insgesamt teurer, im Extremfall um das **28-Fache**.
- **Die Ursache:** unterschiedlich viele Denk-Tokens und Runden.
- **Die Einschränkung:** getestet wurden ältere Modelle wie GPT-5.4 und Gemini 3 Flash.

DIE ZWEI REGLER

REGLER 1: DAS MODELL
Große Modelle kosten pro Token mehr, brauchen bei schweren Aufgaben unter Umständen weniger Anläufe. Kleine reichen oft für Routine.

REGLER 2: DER EFFORT
Wie intensiv das Modell arbeitet: Nachdenken, Antwortlänge, Tool-Aufrufe. Niedrig spart Tokens, Low schränkt laut Anthropic die Leistung leicht ein.

API-LISTENPREISE PRO 1 MILLION TOKENS IN US-DOLLAR (STAND: 10. OKTOBER 2026)

Modell	Eingabe	Ausgabe	Hinweis
Claude Opus 5.5	4	20	
Claude Sonnet 5.5	2	10	
Claude Haiku 5.5	0,10	0,50	bis 100.000 Tokens Prompt, darüber 0,50 und 2,50
GPT-6.1 Sol	2	10	bei langem Kontext 4 und 15
GPT-6 Luna	0,10	0,50	bei langem Kontext 0,20 und 0,75
Gemini 3.8 Flash	0,75	3,75	bis 31.12.2026, ab 1. Jänner 2027 1,50 und 7,50

WELCHE EFFORT-STUFE FÜR WELCHE AUFGABE: FAUSTREGEL ZUM AUSPROBIEREN

Stufe	Geeignet, wenn ...	Beispiel
Low	du das Ergebnis sofort selbst beurteilen kannst	Ein Angebot freundlicher formulieren
Medium	eine normale Aufgabe sorgfältig erledigt werden soll	Stichpunkte zu einem Protokoll ausformulieren
High / Extra high	Fehler teuer sind und leicht übersehen werden	Rechnungen auf Unstimmigkeiten prüfen
Max	ein Agent lange ohne Kontrolle arbeitet	Dokumente selbstständig sortieren und umbenennen

Voreingestellt ist meist Medium, bei Claude Sonnet 5.5 über die API High. Gemini 3.8 Flash kennt nur Low bis High.

MODELL NICHT MITTEN IM CHAT WECHSELN

- **Bei Claude hat jedes Modell seinen eigenen Cache.** Nach einem Wechsel liest die neue KI den ganzen Verlauf frisch ein, der Preisvorteil verfällt.
- **Cache-Lesen kostet 5 Prozent** des regulären Eingabepreises, bei Claude Opus 5.5, Claude Sonnet 5.5 und GPT-6.1 Sol.
- **Besser:** zwischen zwei Arbeitsschritten wechseln, also neuer Chat mit dem passenden Modell.
- **Effort anpassen geht:** bei Claude Opus 5.5, Sonnet 5.5 und Haiku 5.5 bleibt der Cache erhalten.

SO FINDEST DU DIE BESTE EINSTELLUNG

- 1 **Beginne auf Low** und prüfe, ob die Qualität reicht.
- 2 **Eine Stufe höher**, wenn Präzision fehlt, dann vergleichen.
- 3 **Verbrauch beobachten:** Abo-Anzeige, bei der API Tokens und Dollar.
- 4 **Notieren**, welche Kombination aus Modell und Effort verlässlich funktioniert.

MERKSATZ

Erst auf Low testen, dann eine Stufe höher.