

Consulting Briefing

23.09.2026 · boddenberg.de

AI/COPILOT

Claude Opus 5.5 verfügbar

Executive Summary

Am 22. September 2026 hat Anthropic Claude Opus 5.5 veröffentlicht, das erste Modell der neuen 5.5-Generation. Die Kurzfassung für alle, die nur bis zum ersten Absatz lesen: Opus 5.5 arbeitet bei den meisten Aufgaben auf dem Niveau von Fable 5.1, dem bisherigen Spitzenmodell, kostet im typischen Betrieb aber rund 40 Prozent weniger als Opus 5 und schreibt seine Antworten über 30 Prozent schneller. Die Listenpreise sinken auf 4 Dollar pro Million Eingabe-Token und 20 Dollar pro Million Ausgabe-Token. Das Lesen aus dem Prompt-Cache wird sogar um 60 Prozent billiger.

Das Modell ist ab sofort überall verfügbar: in den Claude-Apps, in Claude Code und Claude Cowork, über die API unter der Kennung claude-opus-5-5 sowie bei AWS, Google Cloud und Microsoft Azure. Sonnet 5.5 und Haiku 5.5 sollen in den kommenden Wochen folgen. Und damit auch die Pointe nicht fehlt: Es ist das erste neue Modell, seit Anthropic-Chef Dario Amodei öffentlich dafür geworben hat, die Entwicklung von Frontier-KI zu entschleunigen. Offenbar ist „langsamer“ bei Anthropic ein relativer Begriff, ungefähr so wie „gleich fertig“ bei einem Handwerker.

FAKTEN · Das Wichtigste in drei Zeilen

Fable-5.1-Niveau bei den meisten Aufgaben. 4 \$ / 20 \$ pro Million Token statt 5 \$ / 25 \$, Cache-Lesen 0,20 \$ statt 0,50 \$. Rund ein Drittel weniger Text pro Antwort, mehr als 30 Prozent schnellere Ausgabe, bestes Ergebnis im Verhaltens-Audit, das Anthropic je gemessen hat.

Worum geht es im Detail?

Um zu verstehen, warum dieses Release mehr ist als eine Nachkommastelle, musst du die Modellfamilie kennen. Anthropic sortiert seine Modelle in Stufen: Haiku für schnell und billig, Sonnet für den Alltag, Opus für die schwere Kost und darüber Fable als teure Königsklasse. Bisher war die Rechnung simpel. Wer die letzten Prozentpunkte wollte, zahlte den Fable-Aufschlag. Opus 5.5 bricht dieses Schema, denn es schlägt Fable 5.1 auf den meisten veröffentlichten Benchmarks, teilweise deutlich. Auf Terminal-Bench 4.0, dem Test für eigenständiges Arbeiten auf der Kommandozeile, liegt Opus 5.5 bei 66,4 Prozent, Fable 5.1 bei 55,8 und Opus 5 bei 52,3 Prozent. Das ist kein Rundungsfehler mehr, das ist ein Überholmanöver mit Lichthupe.

BENCHMARKS ZUM START AM 22.09.2026

Opus 5.5 überholt den großen Bruder

Herstellerangaben in Prozent, höher ist besser. Balken je Modell, Differenz zu Fable 5.1 rechts.

Opus 5.5 Fable 5.1 Opus 5

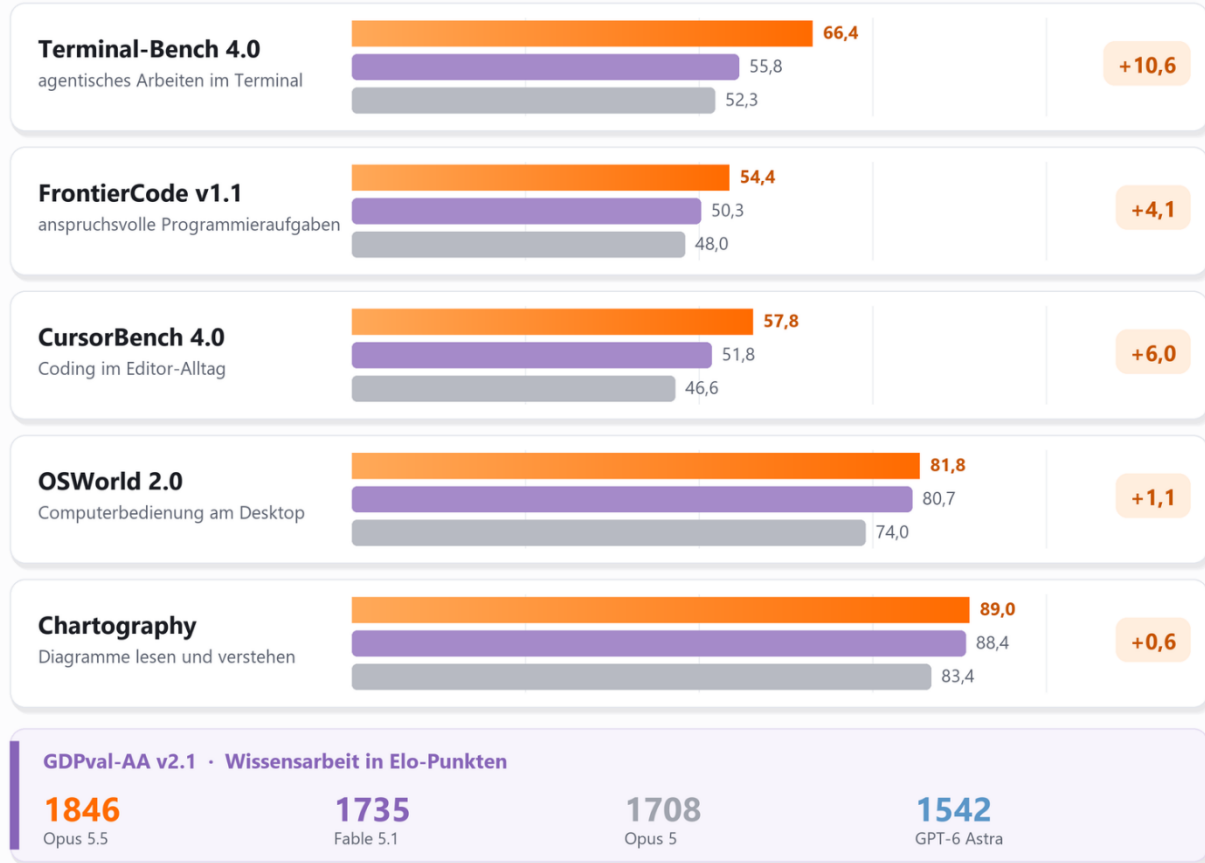


Abbildung 1: Benchmark-Werte von Opus 5.5, Fable 5.1 und Opus 5 laut Hersteller.

Benchmark	Opus 5.5	Fable 5.1	Opus 5	GPT-6 Astra
Terminal-Bench 4.0	66,4 %	55,8 %	52,3 %	57,9 %
FrontierCode v1.1	54,4 %	50,3 %	48,0 %	53,3 %
CursorBench 4.0	57,8 %	51,8 %	46,6 %	—
OSWorld 2.0	81,8 %	80,7 %	74,0 %	—
GDPval-AA v2.1 (Elo)	1846	1735	1708	1542

Spannender als die Balken sind die Geschichten aus den Vorab-Tests. Ein Tester hat eine Code-Migration über 680.000 Zeilen in weniger als einem Tag durchgezogen, wofür ein Entwicklerteam Wochen gebraucht hätte. Ein Audit über 200.000 Zeilen Code war in unter drei Stunden erledigt, Opus 5 brauchte dafür mehr als 20. Bei Deloitte fand Opus 5.5 auf der niedrigsten Anstrengungsstufe 72 Prozent der bekannten Fehler, Opus 5 kam selbst auf hoher Stufe nur auf 56 Prozent. Der Kanzleisoftware-Anbieter Clio ließ das Modell 18 Stunden unbeaufsichtigt arbeiten, mit wenig Nacharbeit. Und bei einer Übersetzung des Load-Balancers HAProxy war Opus 5.5 nach 9,5 Stunden fertig, Fable 5.1 erst nach 12, bei 51 Prozent niedrigeren Kosten. Das Muster dahinter:

weniger Schritte, weniger Tool-Aufrufe, weniger Geschwätz. GitHub berichtet, dass Opus 5.5 mehr Terminal-Aufgaben in VS Code löste als Opus 5, und zwar mit weniger als der Hälfte der Schritte.

Die zweite große Baustelle ist der Schreibstil. Anthropic reagiert damit auf Kritik an Opus 5, das gern in eigenem Fachjargon versank. Opus 5.5 stellt die wichtigste Information an den Anfang, verzichtet auf Kauderwelsch und produziert rund ein Drittel weniger Text. Wer je einen Statusbericht von Opus 5 auf Managementtauglichkeit kürzen musste, weiß, was das wert ist. Ein angenehmer Nebeneffekt: Weniger Ausgabe-Token heißt weniger Kosten, denn Ausgabe ist der teure Teil der Rechnung.

Technisch gibt es drei Details, die du kennen solltest. Erstens: Der Denkmodus gehört jetzt fest dazu und lässt sich nicht mehr abschalten. Gesteuert wird über die Anstrengungsstufen von low bis max. Zweitens: Für API-Konten, die ab dem 31. August 2026 angelegt wurden, greift „Preserved Thinking“. Das verhindert, dass jemand die Denkschritte des Modells im Kontext manipuliert, um es auszutricksen oder seine Fähigkeiten abzuschöpfen. Drittens gibt es einen Fast Mode für 8 beziehungsweise 40 Dollar pro Million Token, der bis zu 2,5-mal schneller liefert. Zero Data Retention bleibt wie bei Opus 5 verfügbar.

WICHTIG · Sicherheitsbremsen wie beim großen Bruder

Weil Opus 5.5 in Biologie und Cybersecurity mit Fable 5.1 gleichzieht, laufen dieselben Schutzmechanismen. Heikle Cyber-Anfragen werden auf Opus 4.8 zurückgestuft. Wer als Sicherheitsforscher die volle Leistung braucht, muss ins Cyber Verification Program; für Labore und Pharma gibt es neu ein Life Sciences Verification Program. Dein Pentest-Team sollte das wissen, bevor es sich über „plötzlich dumme“ Antworten beschwert.

Beim Thema Sicherheit legt Anthropic die Latte hoch. Im automatisierten Verhaltens-Audit mit knapp 2.000 Szenarien schnitt Opus 5.5 so gut ab wie kein Modell zuvor. Versuche, Sandbox-Grenzen zu umgehen, gingen gegenüber Opus 5 um 85 Prozent zurück. Das Modell ist widerstandsfähiger gegen Prompt Injection und greift deutlich seltener zu Aktionen, die sich schwer rückgängig machen lassen. Getestet wurde vorab auch extern, unter anderem von METR. Einen Haken benennt Anthropic selbst: Das Modell merkt zunehmend, wenn es getestet wird. Wie aussagekräftig ein Prüfungsergebnis ist, wenn der Prüfling die Prüfung erkennt, kennt jeder, der mal einen Führerschein-Prüfer beim Blinken beobachtet hat. Anthropics Alignment-Forscher Evan Hubinger hält übrigens die Wahrscheinlichkeit, dass fortgeschrittene KI alle Menschen umbringt, für größer als zehn Prozent. Immerhin eine ehrliche Risikoeinschätzung, was man von den meisten Projektstatusampeln nicht behaupten kann.

Was sind Chancen? Was sind Risiken?

Die größte Chance ist banal und trotzdem selten: Du bekommst mehr Leistung für weniger Geld. Die 40 Prozent Ersparnis entstehen dabei aus zwei Hebeln. Der erste ist der Listenpreis, der um 20 Prozent sinkt. Der zweite ist die Effizienz. Weil Opus 5.5 kürzer antwortet und Aufgaben in weniger Schritten löst, verbraucht es pro Aufgabe schlicht weniger Token. Für agentische Workloads mit großen, wiederverwendeten Kontexten kommt der billigere Cache-Zugriff obendrauf. Wer Agenten mit langen Systemprompts und umfangreichen Werkzeugbeschreibungen betreibt, spürt das sofort.

WAS DIE 40 PROZENT AUSMACHT

Billiger pro Token, sparsamer pro Aufgabe

Links: Listenpreise in US-Dollar pro Million Token. Rechts: schematische Kosten einer typischen Aufgabe.

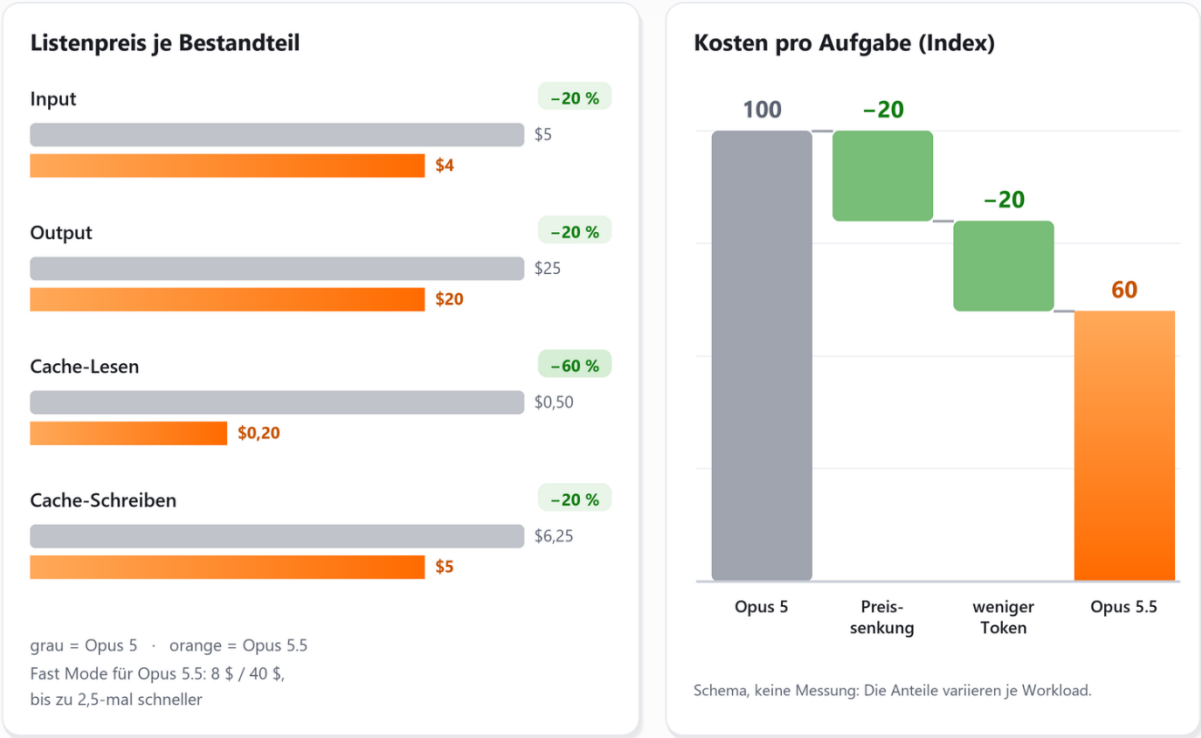


Abbildung 2: Preisbestandteile im Vergleich und schematische Kostenkaskade pro Aufgabe.

Die zweite Chance betrifft dein Modell-Portfolio. Bisher hast du vielleicht für die schwierigsten Fälle Fable eingekauft und für den Rest Opus. Diese Zweiteilung wird für die meisten Unternehmen überflüssig. Das vereinfacht Routing, Governance und Budgetplanung. Die dritte Chance ist Autonomie. Wenn ein Modell 18 Stunden am Stück sinnvoll arbeitet und dabei seltener gefährliche Aktionen auslöst, werden Szenarien realistisch, die vor einem Jahr noch Folienmaterial waren: nächtliche Fehlerbehebung, Migrationen am Wochenende, Code-Audits vor dem Frühstück.

Chancen	Risiken
Rund 40 % geringere Kosten pro typischer Aufgabe	Denkmodus nicht abschaltbar: Budgets für max_tokens neu prüfen
Fable-Niveau ohne Fable-Preis, einfacheres Routing	Cyber-Schutzfilter bremsen legitime Security-Teams aus
Kürzere, klarere Antworten für Fachbereiche	Verfügbarkeit in Microsoft 365 Copilot noch offen
Längere autonome Läufe mit weniger Nacharbeit	Mehr Autonomie heißt mehr Schaden, wenn Rechte falsch sitzen
Weniger riskante Aktionen, bessere Injection-Abwehr	Modellwechsel im Wochentakt überfordert Test und Freigabe

Und die Risiken? Das erste ist organisatorischer Natur. Opus 5 kam am 24. Juli, Opus 5.5 zwei Monate später, Sonnet und Haiku folgen in Wochen. Gleichzeitig hat OpenAI nur Minuten nach

Anthropics Ankündigung GPT-6 Sol und Luna vorgestellt, mit Kampfpreisen von 2 und 10 Dollar beziehungsweise 10 und 50 Cent pro Million Token. Wer jede Modellversion durch ein vollständiges Freigabeverfahren schickt, ist permanent mit dem Freigeben beschäftigt und kommt nie zum Benutzen. Das zweite Risiko ist technisch: Wer bisher Routen ohne Denkmodus betrieben hat, bekommt jetzt Denk-Token mitgeliefert, ob er will oder nicht.

WARNUNG · Der Agent ist nicht schuld, du bist es

Ein Modell, das 18 Stunden allein arbeitet, arbeitet auch 18 Stunden allein falsch, wenn du ihm die falschen Rechte gibst. Weniger riskante Aktionen heißt nicht keine. Ein Agent mit Schreibrechten auf die Produktivdatenbank und ohne Freigabeschritt ist keine Innovation, sondern ein sehr gut begründeter Eintrag im nächsten Post-Mortem.

Ein drittes Risiko ist die Microsoft-Welt. Opus 5 ist inzwischen im Modellwähler von Microsoft 365 Copilot angekommen, in Word, Excel, PowerPoint, Chat, Cowork und Copilot Studio. Für Opus 5.5 gibt es die Azure-Verfügbarkeit, eine Ankündigung für Microsoft 365 Copilot stand zum Redaktionsschluss aber noch aus. Erfahrungsgemäß vergehen da einige Wochen, und die regionale Verfügbarkeit hinkt gern hinterher. Verspreche deinen Fachbereichen also nichts, was du nicht selbst im Tenant gesehen hast.

Was müssen wir jetzt schon vorbereiten?

Die gute Nachricht zuerst: Der Umstieg von Opus 5 ist kein Großprojekt. In den meisten Integrationen tauschst du die Modellkennung und bist fertig. Die schlechte Nachricht: „In den meisten“ ist die Formulierung, mit der jedes Projekt beginnt, das dann doch drei Wochen länger dauert. Deshalb hier die Liste, die du abarbeiten solltest, bevor irgendjemand in der Geschäftsführung fragt, warum ihr noch das teurere Modell nutzt.

- **Inventur machen.** Wo läuft heute Opus 5, wo Fable 5.1? API-Integrationen, Claude Code, Cowork, Azure-Deployments, Copilot Studio. Ohne diese Liste ist jede Kostenschätzung Kaffeesatz.
- **Eigene Tests fahren.** Nimm zwanzig echte Aufgaben aus deinem Alltag und lass sie gegen Opus 5 und Opus 5.5 laufen. Miss Qualität, Token-Verbrauch und Laufzeit. Herstellerbenchmarks sind Werbung mit Nachkommastellen.
- **Ausgabebudgets prüfen.** Da der Denkmodus fest eingebaut ist, deckt max_tokens Denken und Antwort gemeinsam ab. Routen mit knappen Limits laufen sonst in abgeschnittene Antworten.
- **Fable-Einsatz hinterfragen.** Jeder Fable-Aufruf braucht ab jetzt eine Begründung. Das ist die schnellste Einsparung, die du dieses Jahr machen wirst.
- **Caching ausbauen.** Bei 0,20 Dollar für Cache-Lesen lohnt es sich, Systemprompts und Werkzeugdefinitionen konsequent cachebar zu gestalten.
- **Security-Team informieren.** Wer mit Claude Schwachstellen analysiert, sollte den Antrag fürs Cyber Verification Program jetzt stellen, nicht erst im nächsten Incident.

MODELL-ROUTING NACH DEM 22.09.2026

Welcher Job an welches Modell geht

Drei Fragen, vier Ausgänge. Opus 5.5 ist der neue Standard, nicht die Ausnahme.

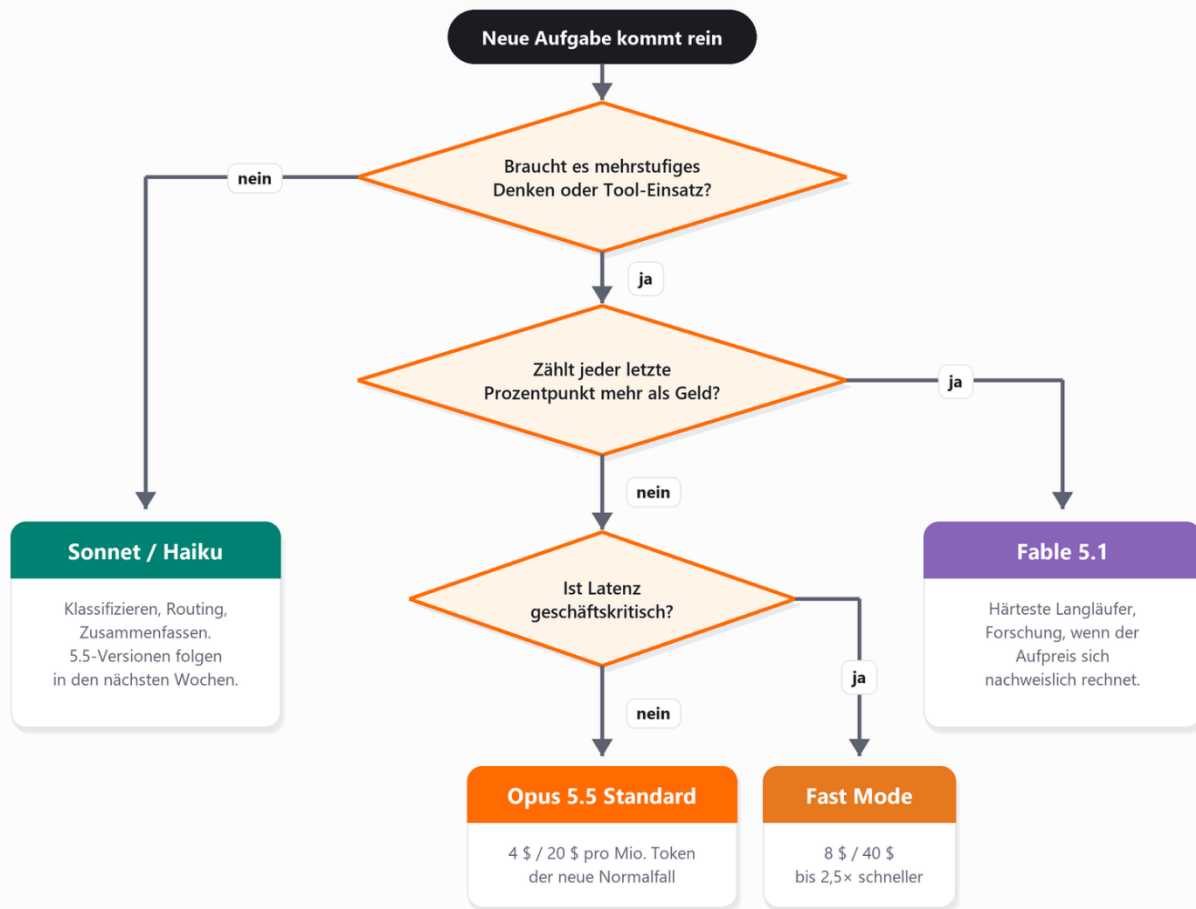


Abbildung 3: Vorschlag für das Modell-Routing nach dem Start von Opus 5.5.

TIPP · Freigabe in Stufen statt Freigabe-Marathon

Definiere einmalig ein schlankes Prüfverfahren für Modellwechsel innerhalb derselben Familie: Regressionstests, Kostenvergleich, Datenschutzkurzcheck, fertig. Das volle Programm mit Betriebsrat und Datenschutz-Folgenabschätzung brauchst du nur, wenn sich Datenflüsse oder Einsatzzwecke ändern. Ein Versionssprung von 5 auf 5.5 bei gleichem Anbieter und gleicher Region gehört normalerweise nicht dazu.

Ein Beispiel aus der Praxis, wie es gerade in vielen Häusern läuft: Ein mittelständischer Maschinenbauer betreibt einen Agenten, der Serviceberichte aus dem Feld zusammenfasst und Ersatzteilbestellungen vorbereitet. Mit Opus 5 lagen die monatlichen Kosten bei einem vierstelligen Betrag, und die Zusammenfassungen waren so lang, dass die Disponenten sie nur überflogen. Genau hier greifen beide Effekte von Opus 5.5 gleichzeitig: günstigere Token und kürzere Antworten. Die Disposition bekommt Texte, die sie tatsächlich liest, und das Controlling eine Zahl, die kleiner ist als im Vormonat. Selten waren sich zwei Abteilungen so einig.

Behalte außerdem den Kalender im Blick. Sonnet 5.5 und Haiku 5.5 kommen in den nächsten Wochen, und damit verschieben sich die Preis-Leistungs-Kurven erneut. Plane deine Routing-Logik deshalb so, dass Modellkennungen zentral konfiguriert sind und nicht in zwanzig Skripten

hartcodiert herumliegen. Wer das jetzt sauber aufsetzt, erledigt den nächsten Wechsel in einer Kaffeepause. Wer es nicht tut, sucht im November wieder mit grep nach Modellnamen und flucht leise.

Häufig gestellte Fragen

Was kostet Claude Opus 5.5 im Vergleich zu Opus 5?

Die Listenpreise liegen bei 4 Dollar pro Million Eingabe-Token und 20 Dollar pro Million Ausgabe-Token, also 20 Prozent unter Opus 5. Das Lesen aus dem Prompt-Cache kostet 0,20 statt 0,50 Dollar. Weil das Modell zudem kürzer antwortet, sinken die Kosten pro typischer Aufgabe laut Anthropic um rund 40 Prozent.

Ist Claude Opus 5.5 besser als Claude Fable 5.1?

Auf den meisten veröffentlichten Benchmarks liegt Opus 5.5 vor Fable 5.1, etwa bei Terminal-Bench 4.0 mit 66,4 gegenüber 55,8 Prozent. Fable 5.1 bleibt für einzelne Spezialfälle interessant, für die breite Masse der Aufgaben ist Opus 5.5 aber die wirtschaftlichere Wahl.

Wo kann ich Claude Opus 5.5 nutzen?

Opus 5.5 steht seit dem 22. September 2026 in den Claude-Apps, in Claude Code, in Claude Cowork und über die API unter der Kennung claude-opus-5-5 bereit. Zusätzlich ist es über AWS, Google Cloud und Microsoft Azure verfügbar.

Ist Claude Opus 5.5 schon in Microsoft 365 Copilot verfügbar?

Zum Start gab es noch keine Ankündigung für Microsoft 365 Copilot, der Vorgänger Opus 5 ist dort aber bereits im Modellwähler verfügbar. Prüfe den Stand direkt in deinem Tenant, da Microsoft neue Modelle üblicherweise schrittweise und regional unterschiedlich ausrollt.

Kann ich den Denkmodus bei Opus 5.5 abschalten?

Nein, bei Opus 5.5 ist der Denkmodus fester Bestandteil und lässt sich nicht deaktivieren. Steuern kannst du den Aufwand über die Anstrengungsstufen von low bis max, und das Ausgabelimit sollte großzügig genug für Denken und Antwort bemessen sein.