

Consulting Briefing

29.09.2026 · boddenberg.de

AI/COPILOT

Claude Sonnet 5.5 verfügbar

Executive Summary

Sechs Tage nach Opus 5.5 hat Anthropic am 28. September 2026 nachgelegt: Claude Sonnet 5.5 ist da, das zweite Modell der 5.5-Familie. Die Kurzfassung für alle, deren Aufmerksamkeitsspanne inzwischen auf Token-Länge geschrumpft ist: Sonnet 5.5 kostet pro Token genau so viel wie Sonnet 5, also 2 Dollar pro Million Eingabe- und 10 Dollar pro Million Ausgabe-Token, und damit die Hälfte von Opus 5.5. Es schreibt über 30 Prozent schneller als sein Vorgänger, braucht deutlich weniger Token und Tool-Aufrufe und kommt deshalb laut Anthropic auf bis zu 30 Prozent niedrigere Kosten pro erledigter Aufgabe.

Der eigentliche Knaller steckt in den Benchmarks. Auf Terminal-Bench 4.0 schlägt Sonnet 5.5 mit 70,6 Prozent sogar den großen Bruder Opus 5.5 (66,4 Prozent). Bei der Büro-Wissensarbeit liegen beide praktisch gleichauf. Für dich heißt das: Sonnet 5.5 wird bei den meisten Unternehmen das neue Standardmodell, Opus 5.5 der Spezialist für die wirklich haarigen Fälle. Dazu kommt etwas, das bei einem Mittelklassemodell neu ist: Anthropic liefert Sonnet 5.5 mit denselben Cyber-Schutzmechanismen aus wie das Flaggschiff. Wer damit Angriffe basteln will, landet sichtbar beim alten Sonnet 5. Ein Downgrade als Strafe, quasi der Schulverweis in die Parallelklasse.

FAKTEN · Das Wichtigste in drei Zeilen

Verfügbar seit 28.09.2026 in den Claude-Apps, über die API als claude-sonnet-5.5 sowie bei AWS, Google Cloud, Microsoft Azure und in GitHub Copilot. Preis 2 \$ / 10 \$ pro Million Token, halb so teuer wie Opus 5.5. Über 30 Prozent schneller, bis zu 30 Prozent günstiger pro Aufgabe, Haiku 5.5 folgt in einigen Wochen.

Worum geht es im Detail?

Zum Hintergrund: Anthropic sortiert seine Modelle in Stufen. Haiku ist das flinke Leichtgewicht, Sonnet das Arbeitspferd für den Alltag, Opus die schwere Artillerie und darüber thront Fable als Luxusklasse. Das Muster der letzten zwei Jahre war immer gleich: Erst kommt das große Modell mit Pauken und Trompeten, ein paar Wochen später das mittlere, das dann überraschend nah an den großen Bruder heranrückt. Bei Sonnet 5.5 ist aus dem Heranrücken ein Überholen geworden, zumindest in einzelnen Disziplinen.

Anthropic positioniert Sonnet 5.5 ausdrücklich als schnellere und günstigere Ergänzung zu Opus 5.5. Seine Stärken sind gut umrissene Alltagsaufgaben, Fehlersuche in Code und das Erstellen sauberer Dokumente, Folien und Tabellen. Das Modell soll sogar Folienvorlagen so treu befolgen, dass kaum Nacharbeit nötig ist. Wer schon einmal eine KI-generierte Präsentation im Corporate Design gesehen hat, weiß, dass das ungefähr so selten ist wie ein pünktlicher ICE am Freitagabend.

BENCHMARKS ZUM START AM 28.09.2026

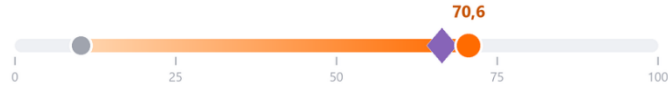
Der kleine Bruder holt auf

Herstellerangaben in Prozent. Grau: Sonnet 5, Orange: Sonnet 5.5, Lila: Opus 5.5.

● Sonnet 5 ● Sonnet 5.5 ◆ Opus 5.5

Terminal-Bench 4.0

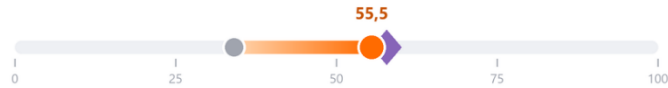
agentisches Arbeiten im Terminal



+4,2 vor Opus

CursorBench 4.0

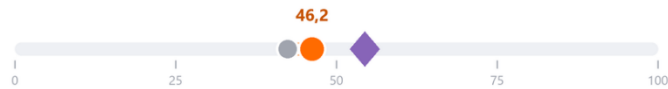
Coding im Editor-Alltag



2,3 hinter Opus

FrontierCode 1.1

harte Programmieraufgaben (Max)



8,2 hinter Opus

OSWorld 2.1

Computerbedienung am Desktop



1,7 hinter Opus

Humanity's Last Exam

Expertenwissen quer durch alle Fächer



3,2 hinter Opus

Wissensarbeit in Elo-Punkten · Dokumente, Tabellen, Office-Alltag

GDPval-AA v2.1



1844 vs. 1846 Opus

AA-Briefcase v1.1



1811 vs. 1822 Opus

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

Benchmark	Sonnet 5.5	Sonnet 5	Opus 5.5
Terminal-Bench 4.0	70,6 %	10,3 %	66,4 %
CursorBench 4.0	55,5 %	34,1 %	57,8 %
FrontierCode 1.1 (Max)	46,2 %	42,4 %	54,4 %
OSWorld 2.1	80,1 %	57,0 %	81,8 %
Humanity's Last Exam	64,5 %	54,9 %	67,7 %
GDPval-AA v2.1 (Elo)	1844	1449	1846
AA-Briefcase v1.1 (Elo)	1811	1359	1822

Die Zahlen muss man richtig lesen. Der Sprung auf Terminal-Bench von 10,3 auf 70,6 Prozent sieht aus wie ein Tippfehler, ist aber vor allem ein Zeichen dafür, dass Sonnet 5 mit der neuen, deutlich härteren Version dieses Tests schlicht überfordert war. Wo Opus 5.5 vorne bleibt, ist ebenfalls aufschlussreich: bei FrontierCode, also bei offenen, schwierigen Programmieraufgaben, und bei

Humanity's Last Exam. Anthropic sagt es selbst ganz offen: Für komplexe, offene Arbeit mit langer Reflexion bleibt Opus das bessere Werkzeug. Sonnet 5.5 punktet dort, wo die Aufgabe klar ist und Tempo zählt.

Die Anekdoten aus den Vorab-Tests sind fast unterhaltsamer als die Balken. Bei Base44 brauchte Sonnet 5.5 im Schnitt 3,6 Anläufe pro App-Bau, Opus 5 noch 7,7. Lovable meldet ein Drittel weniger Tool-Aufrufe und nur noch etwa halb so viele Shell-Befehle. Zendesk arbeitet Tickets 20 Prozent schneller ab, der Slackbot erzielt mit 14 Prozent weniger Ausgabe-Token bessere Ergebnisse. Bei Epic Games hielt das Modell bei System-Design-Audits und Datenfluss-Reviews das Niveau eines teureren Modells. Und weil jede Produktankündigung inzwischen ein Videospiel braucht: Sonnet 5.5 ist das erste Sonnet, das Pokémon Rot nur anhand von Screenshots durchspielt. Deine Mitarbeiter, die das 1998 auf dem Game Boy geschafft haben, dürfen sich jetzt offiziell als Frontier-Intelligenz bezeichnen.

WO DIE 30 PROZENT HERKOMMEN

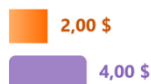
Halber Opus-Preis, weniger Token

Links: Listenpreise in US-Dollar pro Million Token. Rechts: Effekte aus den Vorab-Tests.

Listenpreis je Bestandteil

Sonnet 5.5 (unverändert zu Sonnet 5) vs. Opus 5.5

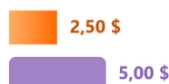
Input



Output



Cache schreiben



Cache lesen



Token-Diät im Praxistest

Gleiche Aufgabe, Token-Verbrauch (Balyasny)



Sonnet 5: 497.000 Token



Sonnet 5.5: 121.000 Token

1 Kästchen ≈ 10.000 Token

1/3

weniger Tool-Calls
Lovable

1/2

so viele Shell-Runs
Lovable

-14 %

Output-Token
Slack

+20 %

schneller bei Tickets
Zendesk

3,6

statt 7,7 Iterationen
Base44 vs. Opus 5

2,4×

schneller verarbeitet
Box

Gleicher Token-Preis, bis zu 30 % weniger pro Aufgabe

Abbildung 2: Preisvergleich und Effizienzgewinne aus den Vorab-Tests.

Woher kommen die 30 Prozent Ersparnis, wenn der Token-Preis gleich bleibt? Die Antwort ist banal und trotzdem wichtig: Das Modell redet weniger und arbeitet gezielter. Es bündelt Tool-Aufrufe effizienter, sucht nicht mehr bei jeder Gelegenheit im Web und kommt schneller zum Punkt. Beim Finanzhaus Balyasny brauchte eine Aufgabe 121.000 Token statt 497.000 mit Sonnet 5. Wer seine KI-Kosten bisher nur über den Listenpreis vergleicht, vergleicht also Äpfel mit Obstsalat. Entscheidend ist, was eine erledigte Aufgabe kostet, nicht was ein Token kostet.

WICHTIG · Denkmodus und Anstrengungsstufen

Sonnet 5.5 lässt sich über Anstrengungsstufen von Low über Medium, High und Xhigh bis Max steuern. Auf Low und Medium spielt es seine Kostenvorteile voll aus, auf High erreicht es manchmal Opus-Niveau. Wer von Modellen ohne Denkmodus migriert, muss laut Anthropic auf die Einstellung `between_tools` umstellen. Das Kontextfenster reicht bis zu einer Million Token, einzelne Produkte und Cloud-Anbieter können aber niedrigere Grenzen setzen.

Zur Verfügbarkeit: Sonnet 5.5 ist seit dem Start in den Claude-Apps, auf der Claude Platform unter `claude-sonnet-5-5` und bei AWS, Google Cloud und Microsoft Azure verfügbar, jeweils auch mit Zero Data Retention. GitHub hat es am selben Tag in Copilot Pro, Pro+, Max, Business und Enterprise freigeschaltet, nutzbar in VS Code, Visual Studio, JetBrains, Xcode, Eclipse, der Copilot CLI und auf `github.com`. Für Microsoft 365 Copilot gab es zum Start noch keine Ankündigung. Dort läuft seit Juli Sonnet 5 in Copilot Cowork und PowerPoint, das Update dürfte also eher eine Frage von Wochen als von Monaten sein.

Was sind Chancen? Was sind Risiken?

Die größte Chance ist ökonomisch: Du bekommst für die Hälfte des Opus-Preises in vielen Szenarien Opus-Qualität. Für Agenten, die täglich tausende Tickets, Dokumente oder Pull Requests anfassen, ist das der Unterschied zwischen einem netten Pilotprojekt und einem Business Case, den dir der Controller unterschreibt. Dazu kommt das Tempo: 30 Prozent schnellere Ausgabe merkt man im Alltag, vor allem bei Agenten, die viele Schritte hintereinander abarbeiten.

Die zweite Chance ist Vereinfachung. Viele Unternehmen betreiben heute einen Modell-Zoo, weil das günstige Modell zu dumm und das kluge zu teuer war. Mit Sonnet 5.5 als Standard und Opus 5.5 als Eskalationsstufe reichen oft zwei Modelle. Weniger Modelle heißt weniger Evaluationen, weniger Prompt-Varianten und weniger Wartung.

Aspekt	Chance	Risiko
Kosten	Halber Opus-Preis, bis zu 30 % weniger pro Aufgabe	Rechnung explodiert, wenn alle auf Max stellen
Qualität	Opus-Niveau bei klar umrissenen Aufgaben	Schwächer bei offenen, kniffligen Problemen
Security	Flaggschiff-Schutz gegen Missbrauch	Legitime Security-Teams landen auf Sonnet 5
Betrieb	Weniger Modelle, schnellere Agenten	Automatische Freigabe in GitHub Copilot
Migration	Gleiche Preise, gleiche API	Prompts und Thinking-Einstellung neu testen

Jetzt zu den Risiken, und hier wird es spannend. Das erste betrifft deine Security-Abteilung. Hintergrund: Die Cyber-Fähigkeiten von Sonnet 5.5 sind gegenüber Sonnet 5 so stark gestiegen, dass Anthropic dieselben Schutzmechanismen einsetzt wie bei Opus 5.5. Routine-Debugging bleibt erlaubt, riskantere Anfragen fallen sichtbar auf Sonnet 5 zurück. Das ist gut gegen Angreifer, aber ärgerlich für dein Red Team, das legitimerweise Exploits analysiert. Anthropic plant ein Cyber Verification Program mit gestaffeltem Zugang, bis dahin arbeitet dein Pentester eben mit dem Vorgänger. Sein Stundensatz bleibt übrigens derselbe.

WARNUNG · Der Distillation-Riegel

Erstmals blockieren Klassifikatoren das Extrahieren der Denkschritte, und gespeicherte Denkinhalte werden an das ursprüngliche Konto gebunden. Anlass: Forscher hatten aus 6.708 öffentlichen Agenten-Protokollen 315.320 Denkböcke verschiedener Anbieter entschlüsselt und darin API-Schlüssel und Passwörter gefunden. Wenn deine Agenten Zugangsdaten im Klartext durch den Prompt schieben, war das nie eine gute Idee. Jetzt ist es auch noch dokumentiert dumm.

Das zweite Risiko ist die automatische Freischaltung. In GitHub Copilot Business und Enterprise sind neue Modelle standardmäßig aktiv, solange kein Administrator den globalen Standard abgeschaltet oder das Modell gezielt gesperrt hat. Abgerechnet wird nutzungsbasiert zu Listenpreisen. Wer also keine Modellrichtlinie hat, erfährt von neuen Modellen gern über die Monatsrechnung. Das dritte Risiko ist der Benchmark-Rausch: Ein Modell, das auf Terminal-Bench glänzt, kann in deiner Legacy-Codebasis aus dem Jahr 2009 trotzdem gegen die Wand laufen. Herstellerbenchmarks ersetzen keine eigenen Tests, sie ersetzen höchstens die Ausrede, keine gemacht zu haben.

TIPP · Konsequenz eskalieren statt pauschal upgraden

Stelle nicht einfach alles auf Opus 5.5 oder alles auf Sonnet 5.5 um. Lass Sonnet 5.5 auf Medium als Standard laufen und eskaliere nur dann auf Opus, wenn der Agent zweimal scheitert oder die Aufgabe als offen markiert ist. Das spart erfahrungsgemäß mehr als jede Preisverhandlung.

Was müssen wir jetzt schon vorbereiten?

BETRIEBSMODELL AB OKTOBER 2026

Wer bekommt welchen Job?

Modell-Routing mit Sonnet 5.5 als Standard, inklusive der eingebauten Cyber-Weiche.

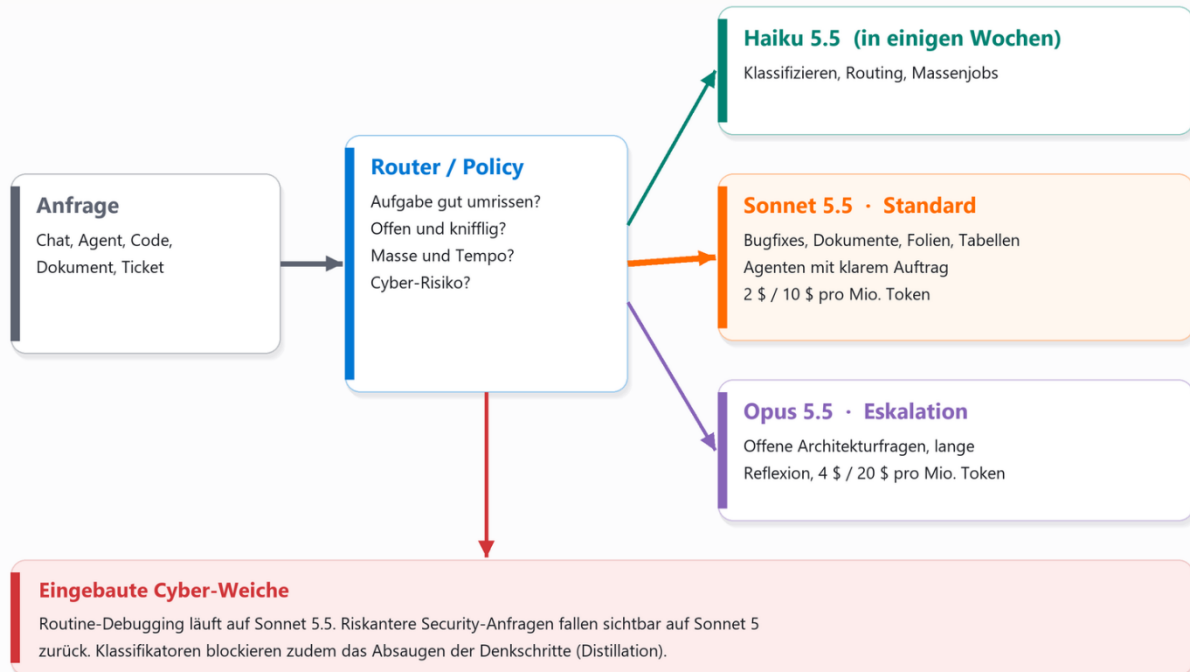


Abbildung 3: Empfohlenes Modell-Routing mit Sonnet 5.5 als Standard.

Erstens: Baue dir ein Evaluationsset aus echten Aufgaben. Zwanzig bis fünfzig typische Tickets, Code-Änderungen oder Dokumentaufträge aus deinem Alltag, mit bekanntem Sollergebnis. Lass Sonnet 5, Sonnet 5.5 und Opus 5.5 darauf los und miss nicht nur die Qualität, sondern auch Token, Tool-Aufrufe und Laufzeit. Ein Kunde von uns hat das mit seinen Supportvorlagen gemacht und festgestellt, dass Sonnet 5.5 auf Medium besser abschnitt als sein bisheriges Opus-Setup auf High. Die Folie für den Lenkungskreis hat sich quasi selbst geschrieben, diesmal sogar im Corporate Design.

Zweitens: Kläre die Modellrichtlinien. In GitHub Copilot legst du fest, ob neue Modelle automatisch aktiv werden. In Azure und bei den anderen Clouds prüfst du, in welchen Regionen Sonnet 5.5 verfügbar ist und ob Zero Data Retention vertraglich sauber abgebildet ist. Das ist keine Formsache, denn ohne diese Klärung stoppt dich spätestens der Datenschutzbeauftragte, und zwar zu Recht.

Drittens: Sprich mit deinem Security-Team über den Cyber-Fallback. Wenn dort KI für Malware-Analyse oder Schwachstellenforschung genutzt wird, muss klar sein, dass manche Anfragen auf Sonnet 5 zurückfallen, und wer sich für das kommende Verifikationsprogramm bewirbt. Viertens: Räume deine Agenten auf. Keine Zugangsdaten im Prompt, keine Geheimnisse in Protokollen, die irgendwo öffentlich landen könnten. Fünftens: Plane Haiku 5.5 schon mit ein. Für Klassifikation und Routing in großen Mengen wird es in einigen Wochen die logische dritte Stufe.

Zeitraum	Maßnahme	Verantwortlich
Oktober 2026	Evaluationsset bauen, Sonnet 5.5 gegen Sonnet 5 und Opus 5.5 testen	KI-Team, Fachbereich
Oktober 2026	Modellrichtlinie in GitHub Copilot prüfen und dokumentieren	Plattform-Admin
Oktober 2026	Cyber-Fallback und Verifikationsprogramm mit Security klären	CISO, Red Team
November 2026	Routing Sonnet 5.5 als Standard, Opus als Eskalation einführen	Architektur
Ab Verfügbarkeit	Haiku 5.5 und Sonnet 5.5 in Microsoft 365 Copilot bewerten	M365-Team

TIPP · Schneller Gewinn für Entwicklerteams

Lass deine Entwickler Sonnet 5.5 in GitHub Copilot zwei Wochen lang als Standard für Bugfixes nutzen und vergleiche danach die Anzahl der Iterationen pro Pull Request. Das kostet dich nichts außer einer Umfrage und liefert belastbarere Zahlen als jedes Datenblatt.

Häufig gestellte Fragen

Was kostet Claude Sonnet 5.5 über die API?

Claude Sonnet 5.5 kostet 2 Dollar pro Million Eingabe-Token und 10 Dollar pro Million Ausgabe-Token, genauso viel wie Sonnet 5 und halb so viel wie Opus 5.5. Cache-Schreiben kostet 2,50 Dollar, Cache-Lesen 0,20 Dollar pro Million Token.

Ist Claude Sonnet 5.5 besser als Claude Opus 5.5?

Auf Terminal-Bench 4.0 liegt Sonnet 5.5 mit 70,6 Prozent vor Opus 5.5 mit 66,4 Prozent, bei Büro-Wissensarbeit sind beide nahezu gleichauf. Bei offenen, komplexen Aufgaben mit langer Reflexion bleibt Opus 5.5 laut Anthropic jedoch überlegen.

Ist Claude Sonnet 5.5 schon in Microsoft 365 Copilot verfügbar?

Zum Start am 28. September 2026 gab es dafür noch keine Ankündigung, in Microsoft 365 Copilot ist derzeit Sonnet 5 in Copilot Cowork und PowerPoint verfügbar. Über Microsoft Azure und in GitHub Copilot lässt sich Sonnet 5.5 dagegen bereits nutzen.

Warum antwortet Claude Sonnet 5.5 bei Security-Fragen mit Sonnet 5?

Sonnet 5.5 nutzt dieselben Cyber-Schutzmechanismen wie Opus 5.5: Routine-Debugging ist erlaubt, riskantere Security-Anfragen fallen sichtbar auf Sonnet 5 zurück. Für legitime Sicherheitsforschung plant Anthropic ein Cyber Verification Program mit gestaffeltem Zugang.

Wann erscheint Claude Haiku 5.5?

Anthropic hat Haiku 5.5 für die Wochen nach dem Start von Sonnet 5.5 angekündigt, ein genaues Datum gibt es noch nicht. Es soll als günstigste Stufe der 5.5-Familie kostensensible Massenaufgaben abdecken.