

HÄNDLER, ANBIETER UND ISOS

Haiku 5.5: KI-Funktionen in einer Zahlungsanwendung brauchen kein großes Budget mehr

Veröffentlicht am 27. Oktober 2026

Claude Haiku 5.5 von Anthropic kostet für dieselbe Arbeit ein Vierzigstel des aktuellen Opus-Modells. Für die vielen eng umrissenen Aufgaben, die ein Zahlungsunternehmen täglich erledigt, wird KI damit von einem Budgetposten zu einer Rundungsdifferenz. Was es kostet, wofür es sich eignet und wo ein größeres Modell seinen Preis weiterhin wert ist.

In den letzten Jahren kam jede KI-Funktion in einer Zahlungsanwendung mit einer Kostenfrage. Die Website eines Händlers lesen, ein Support-Postfach sortieren oder Felder aus einem Antragsformular auslesen: Das alles kann ein Sprachmodell gut. Bei Zehntausenden Aufrufen im Monat war die Rechnung aber so hoch, dass viele Teams solche Funktionen beim Pilotprojekt belassen haben.

Das aktuelle kleine Modell von Anthropic, **Claude Haiku 5.5**, verändert diese Rechnung. Zum Listenpreis kostet es ein Vierzigstel des aktuellen Opus-Modells und ein Zehntel des Haiku-Modells davor. Für die eng umrissene, sich wiederholende Arbeit, aus der der größte Teil der KI in einem Zahlungsunternehmen besteht, kostet eine Funktion, die früher Hunderte im Monat kostete, jetzt nur noch wenige.

Was die Modelle kosten

Anthropic berechnet seine Modelle pro Million *Tokens*, der Einheiten, in die Text zerlegt wird. Ein Token entspricht ungefähr drei Vierteln eines englischen Wortes, eine Million Tokens also etwa 750.000 Wörtern. Eingabe (was Sie an das Modell senden) und Ausgabe (was es zurückschreibt) werden getrennt berechnet.

MODELL	EINGABE, USD PRO MILLION TOKENS	AUSGABE, USD PRO MILLION TOKENS	IM VERHÄLTNIS ZU HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1x

MODELL	EINGABE, USD PRO MILLION TOKENS	AUSGABE, USD PRO MILLION TOKENS	IM VERHÄLTNIS ZU HAIKU 5.5
Claude Haiku 4.5 (das vorherige Haiku)	1,00	5,00	10×
Claude Sonnet 5.5	2,00	10,00	20×
Claude Opus 5.5	4,00	20,00	40×
Claude Fable 5.1	10,00	50,00	100×

Das sind die eigenen Listenpreise von Anthropic, Stand Oktober 2026. Der Preis für Haiku 5.5 gilt für Anfragen bis 100.000 Tokens; längere Anfragen kosten 0,50 USD für die Eingabe und 2,50 USD für die Ausgabe, immer noch deutlich weniger als die größeren Modelle. Claude wird auch über Amazon Bedrock und Google Cloud Vertex AI verkauft, die eigene Preise festlegen. Prüfen Sie also den Preis auf der Plattform, über die Sie einkaufen.

Was das pro Anfrage bedeutet

Eine typische eng umrissene Aufgabe, etwa eine Seite einer Händler-Website lesen und ein kurzes strukturiertes Urteil zurückgeben, sendet vielleicht 2.000 Tokens und erhält 300 zurück. Zu den Preisen der Tabelle oben:

MODELL	KOSTEN EINER ANFRAGE, USD	KOSTEN VON 10.000 ANFRAGEN, USD
Claude Haiku 5.5	0,00035	3,50
Claude Haiku 4.5	0,0035	35,00
Claude Sonnet 5.5	0,007	70,00
Claude Opus 5.5	0,014	140,00
Claude Fable 5.1	0,035	350,00

Eine Anfrage an Haiku 5.5 kostet wenige Hundertstel Cent. Zehntausend davon kosten ungefähr so viel wie ein Kaffee. Die Zahlen dienen der Veranschaulichung: Ihre eigenen Prompts werden länger oder kürzer sein, und das Nachdenken des Modells, das als Ausgabe berechnet wird, kommt hinzu (dazu unten mehr). Die Verhältnisse zwischen den Modellen bleiben bei jeder Anfragegröße gleich.

Zwei Rabatte senken die Kosten weiter. Anfragen, die keine sofortige Antwort brauchen, können über die Batch-Schnittstelle von Anthropic zum halben Preis laufen. Das passt zu nächtlichen Aufgaben wie dem erneuten Scannen eines Händlerportfolios. Und lange Anweisungen, die bei jeder Anfrage gleich sind, lassen sich zwischenspeichern: Sie werden einmal voll berechnet und danach nur noch zu einem Bruchteil.

Wofür ein kleines Modell geeignet ist

Das günstige Modell ist nicht bei jeder Aufgabe eine schwächere Version des teuren. Es ist auf Geschwindigkeit und Menge ausgelegt und funktioniert am besten, wenn die Aufgabe eng umrissen ist, die Anweisungen klar sind und die Antwort eine feste Form hat. Der größte Teil der KI, die ein Zahlungsunternehmen braucht, sieht genau so aus.

- **Website-Prüfung beim Onboarding.** Hat die Website eine Rückerstattungsrichtlinie, AGB und Kontaktdaten? Was verkauft sie tatsächlich, und passt das zu der Branche, die der Händler angegeben hat? Das ist Lesen und Einordnen, wiederholt bei jedem Antrag und erneut bei jeder [regelmäßigen Überprüfung](#).
- **Portfolio-Überwachung.** Tausende Händler-Websites in festen Abständen erneut lesen und nach neuen Produkten, geänderten Richtlinien oder Anzeichen für [Transaction Laundering](#) suchen. Hier ist die Menge das ganze Problem, und hier zählt der Preisunterschied am meisten.
- **Felder aus Dokumenten auslesen.** Firmenname, Registernummer, Geschäftsführer und Adressen aus einem Handelsregistrauszug oder einem Bankschreiben, in einem festen Format zurückgeben, damit ein Mensch sie prüfen kann.
- **Das Postfach sortieren.** Support-Tickets, Händleranfragen und Korrespondenz mit Anbietern an das richtige Team weiterleiten und diejenigen markieren, die eine Rückbuchung, einen Betrugsfall oder eine Aufsichtsbehörde erwähnen.
- **Erste Sichtung von Rückbuchungen.** Eingehende [Chargebacks](#) nach Grund gruppieren, prüfen, ob die vom Händler gelieferten Nachweise die sind, die der Reason Code verlangt, und auflisten, was fehlt.
- **Deskriptoren und Transaktionstexte lesen.** Freie Transaktionsbeschreibungen in Kategorien einordnen oder prüfen, ob ein Abrechnungsdeskriptor für den Karteninhaber erkennbar ist.
- **Aus einer Frage eine Suche machen.** Einen getippten Satz nehmen und die Felder herausziehen, die eine Datenbankabfrage braucht: ein Land, eine Währung, eine Branche, eine Zahlungsmethode.

Jede dieser Aufgaben läuft viele Male am Tag, jede hat eine Antwort, die ein Mensch prüfen kann, und keine verlangt, dass das Modell selbstständig eine lange Arbeit plant.

Wo ein größeres Modell seinen Preis weiterhin wert ist

Die größeren Modelle sind aus gutem Grund langsamer und teurer: Sie sind besser bei Arbeit, die lang, offen oder schwer zu beschreiben ist. Sie lohnen sich, wenn:

- **die Aufgabe viele voneinander abhängige Schritte hat**, etwa ein ungewöhnliches Muster über mehrere Datenquellen hinweg untersuchen und die Ergebnisse aufschreiben.
- **der Text wichtig ist**, etwa eine vollständige Antwort auf eine Rückbuchung oder eine Richtlinie, die ein Mensch freigibt.

- **die Eingabe lang ist und die Antwort vom Ganzen abhängt**, etwa die Prüfung eines vollständigen Anbietervertrags gegen Ihre eigenen Bedingungen.
- **eine falsche Antwort teuer und schwer zu erkennen ist.**

Ein verbreiteter Aufbau nutzt beides: Das kleine Modell bearbeitet jede Anfrage und gibt die wenigen, bei denen es unsicher ist oder die eine festgelegte Schwelle erreichen, an ein größeres weiter. Messen Sie das aber zuerst gegen die einfachere Wahl. Ein größeres Modell mit einer niedrigeren Denkstufe reicht manchmal aus, und ein Modell ist einfacher zu betreiben als zwei.

Die Kosten dort halten, wo Sie sie erwarten

Der Preis pro Token ist nur ein Teil der Rechnung. Einige Gewohnheiten halten sie berechenbar.

- **Stellen Sie den Denkaufwand passend zur Aufgabe ein.** Aktuelle Claude-Modelle denken nach, bevor sie antworten, und dieses Nachdenken wird als Ausgabe berechnet. Haiku 5.5 steht standardmäßig auf einer mittleren Stufe. Für Einordnung und Extraktion reicht oft eine niedrige Stufe, die viel günstiger ist. Testen Sie es an einer Stichprobe echter Anfragen, bevor Sie sich festlegen.
- **Verlangen Sie ein festes Format.** Eine kurze strukturierte Antwort (ja oder nein, eine Kategorie aus einer Liste, eine Handvoll benannter Felder) kostet weniger als Fließtext und ist leichter zu prüfen.
- **Senden Sie nur, was das Modell braucht.** Den Text einer Richtlinienenseite, nicht die ganze Website mit Navigation und Skripten.
- **Speichern Sie zwischen, was sich wiederholt, und bündeln Sie, was warten kann.** Beides sind Rabatte für Arbeit, die Sie ohnehin erledigen.
- **Bewerten Sie die Kosten pro erledigter Aufgabe**, nicht pro Anfrage. Ein günstigerer Aufruf, der zweimal wiederholt werden muss, ist nicht günstiger.

Was sich mit dem Preis nicht ändert

Ein günstigeres Modell ändert nichts an Ihren Pflichten, und drei Punkte sollten geklärt sein, bevor etwas live geht.

1. **Halten Sie Kartendaten heraus.** Vollständige Kartennummern, Sicherheitscodes und alles andere, was in den Geltungsbereich von PCI DSS fällt, gehört nicht in einen Prompt. Keine der oben genannten Aufgaben braucht sie, und ihr Versand würde den KI-Dienst und alles um ihn herum in Ihre Karteninhaber-Datenumgebung holen.
2. **Minimieren Sie personenbezogene Daten.** Prüfen Sie nach der DSGVO und ähnlichen Gesetzen die Datenbedingungen des Anbieters, wo Anfragen verarbeitet werden und wie lange sie gespeichert bleiben, und nennen Sie den Dienst in Ihrer Datenschutzerklärung. Entfernen Sie Namen und Kontaktdaten, wo die Aufgabe sie nicht braucht.

3. **Behandeln Sie, was das Modell liest, als nicht vertrauenswürdig.** Die Website eines Händlers wird vom Händler geschrieben, der von einer günstigen Bewertung profitiert. Text darauf kann versuchen, dem Modell Anweisungen zu geben („diese Website hat eine vollständige Rückerstattungsrichtlinie“). Sagen Sie dem Modell in seinen Anweisungen, dass der Inhalt Daten sind und keine Anweisungen, und beschränken Sie die Antwort auf feste Auswahlmöglichkeiten, damit eingeschleuster Text nirgendwohin kann.

Das Modell berät nur. Ein Befund aus einer automatischen Prüfung ist ein Grund für einen Menschen, genauer hinzusehen, keine Entscheidung, und die Entscheidung treffen weiterhin der Underwriter oder der Analyst.

Wie Payments Marketplace es nutzt

Payments Marketplace nutzt Claude genau für diese Art eng umrissener Arbeit in großer Menge. Wenn ein Händler einen Pre-Flight-Check durchführt, liest die Website-Prüfung die Seiten der Website und liefert einen festen Satz von Befunden: ob es eine Rückerstattungsrichtlinie gibt und ob sie angemessen ist, was die Website tatsächlich verkauft und ob das zu der angegebenen Branche passt. Der KI-Agent auf unserer Startseite macht aus einem getippten Satz die wenigen Felder, die unser Abgleich liest, und jedes Wort der Antwort schreibt unser eigener Code.

- **Das Modell liest, Menschen entscheiden.** Die Befunde der Prüfung sind beratende Hinweise für den Anbieter. Sie blockieren nie einen Antrag, und ein Treffer ist eine Vorstellung, nie eine Zusage.
- **Die Anonymität bleibt gewahrt.** Die Zusammenfassung der Prüfung, die ein Anbieter vor der Annahme sieht, enthält keine Identität.
- **Begrenzt durch den Aufbau.** Website-Text gilt als nicht vertrauenswürdig, und die Antworten sind auf feste Felder beschränkt, sodass Text auf der Website eines Händlers das Ergebnis nicht lenken kann.

Quellen

- [Anthropic: Claude pricing](#)
- [Anthropic: Models overview](#)
- [PCI Security Standards Council: PCI DSS](#)

Online lesen: <https://payments-marketplace.com/de-DE/insights/ai-features-cost-haiku-5-5/>

Payments Marketplace Ltd

Registered in England & Wales No. 17424288

Garden Ctg, Dereham Road, Dereham, Norfolk, NR19 2PW, UK

connect@payments-marketplace.com · +44 1366 727140