

MERCHANTS, PROVIDERS EN ISO'S

Haiku 5.5: AI-functies in een betaalapplicatie vragen geen groot budget meer

Gepubliceerd op 27 oktober 2026

Claude Haiku 5.5 van Anthropic kost voor hetzelfde werk een veertigste van het huidige Opus-model. Voor de afgebakende taken in grote volumes die een betaalbedrijf elke dag uitvoert, wordt AI daarmee van een begrotingspost een afrondingsverschil. Wat het kost, waarvoor het geschikt is en waar een groter model zijn prijs nog steeds waard is.

De afgelopen jaren kwam elke AI-functie in een betaalapplicatie met een kostenvraag. De website van een merchant lezen, een supportinbox sorteren of velden uit een aanvraagformulier halen: een taalmodel doet dat allemaal goed. Maar bij tienduizenden aanroepen per maand was de rekening zo hoog dat veel teams zulke functies bij een pilot lieten.

Het huidige kleine model van Anthropic, **Claude Haiku 5.5**, verandert die rekensom. Tegen de catalogusprijs kost het een veertigste van het huidige Opus-model en een tiende van het vorige Haiku-model. Voor het afgebakende, terugkerende werk waaruit het grootste deel van de AI in een betaalbedrijf bestaat, kost een functie die vroeger honderden per maand kostte nu nog maar een paar.

Wat de modellen kosten

Anthropic rekent zijn modellen af per miljoen *tokens*, de eenheden waarin tekst wordt opgedeeld. Een token is ongeveer driekwart van een Engels woord, dus een miljoen tokens is zo'n 750.000 woorden. Invoer (wat u naar het model stuurt) en uitvoer (wat het terugschrijft) worden apart geprijsd.

MODEL	INVOER, USD PER MILJOEN TOKENS	UITVOER, USD PER MILJOEN TOKENS	TEN OPZICHTE VAN HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1x
Claude Haiku 4.5 (de vorige Haiku)	1,00	5,00	10x

MODEL	INVOER, USD PER MILJOEN TOKENS	UITVOER, USD PER MILJOEN TOKENS	TEN OPZICHTE VAN HAIKU 5.5
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×

Dit zijn de eigen catalogusprijzen van Anthropic per oktober 2026. De prijs van Haiku 5.5 geldt voor verzoeken tot 100.000 tokens; langere verzoeken kosten 0,50 USD voor invoer en 2,50 USD voor uitvoer, nog altijd ver onder de grotere modellen. Claude wordt ook verkocht via Amazon Bedrock en Google Cloud Vertex AI, die hun eigen prijzen hanteren. Controleer dus het tarief op het platform waar u inkoopt.

Wat dat per verzoek betekent

Een typische afgebakende taak, zoals een pagina van de website van een merchant lezen en een kort gestructureerd oordeel teruggeven, stuurt misschien 2.000 tokens en krijgt er 300 terug. Tegen de prijzen uit de tabel hierboven:

MODEL	KOSTEN VAN ÉÉN VERZOEK, USD	KOSTEN VAN 10.000 VERZOEKEN, 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

Een verzoek aan Haiku 5.5 kost een paar honderdste cent. Tienduizend verzoeken kosten ongeveer evenveel als een kop koffie. De cijfers zijn ter illustratie: uw eigen prompts zijn langer of korter, en het nadenken van het model, dat als uitvoer wordt gerekend, komt erbij (daarover hieronder meer). De verhoudingen tussen de modellen blijven gelijk, hoe groot het verzoek ook is.

Twee kortingen maken het nog goedkoper. Verzoeken die niet direct een antwoord nodig hebben, kunnen via de batch-interface van Anthropic voor de halve prijs, wat past bij nachtelijke taken zoals het opnieuw scannen van een merchantportefeuille. En een lange set instructies die bij elk verzoek gelijk is, kan in de cache worden gezet: die wordt één keer volledig gerekend en daarna voor een fractie.

Het werk waarvoor een klein model geschikt is

Het goedkope model is niet voor elke taak een mindere versie van het dure. Het is gebouwd voor snelheid en volume, en het werkt het best wanneer de taak afgebakend is, de instructies duidelijk zijn en het antwoord een vaste vorm heeft. Het grootste deel van de AI die een betaalbedrijf nodig heeft, ziet er zo uit.

- **Websitecontrole bij onboarding.** Heeft de website een terugbetalingsbeleid, voorwaarden en contactgegevens? Wat verkoopt ze echt, en past dat bij de branche die de merchant heeft opgegeven? Dat is lezen en indelen, bij elke aanvraag opnieuw en bij elke **periodieke herbeoordeling**.
- **Portefeuillemonitoring.** Duizenden merchantwebsites op vaste momenten opnieuw lezen op zoek naar nieuwe producten, gewijzigd beleid of tekenen van **transaction laundering**. Hier is het volume het hele probleem, en hier telt het prijsverschil het zwaarst.
- **Velden uit documenten halen.** Bedrijfsnaam, KvK-nummer, bestuurders en adressen uit een uittreksel of een bankbrief, teruggegeven in een vast formaat zodat een mens ze kan controleren.
- **De inbox sorteren.** Supporttickets, vragen van merchants en correspondentie met providers naar het juiste team sturen, en de berichten markeren die een geschil, fraude of een toezichthouder noemen.
- **Eerste schifting van geschillen.** Binnenkomende **chargebacks** per reden groeperen, controleren of het bewijs dat de merchant aanlevert het bewijs is dat de reason code vraagt, en opsommen wat ontbreekt.
- **Descriptors en transactieteksten lezen.** Vrije transactieomschrijvingen in categorieën indelen, of controleren of een afschrijvingsomschrijving herkenbaar is voor de kaarthouder.
- **Van een vraag een zoekopdracht maken.** Eén getypte zin nemen en de velden eruit halen die een databasequery nodig heeft: een land, een valuta, een branche, een betaalmethode.

Elk van deze taken draait vele keren per dag, elk heeft een antwoord dat een mens kan controleren, en geen ervan vraagt het model om zelfstandig een lang stuk werk te plannen.

Waar een groter model zijn prijs nog steeds waard is

De grotere modellen zijn niet voor niets trager en duurder: ze zijn beter in werk dat lang, open of moeilijk te omschrijven is. Ze zijn het waard wanneer:

- **de taak veel stappen heeft die van elkaar afhangen**, zoals een ongebruikelijk patroon onderzoeken over meerdere databronnen en opschrijven wat er gevonden is.
- **de tekst ertoe doet**, zoals een volledig antwoord op een geschil of een beleid dat een mens goedkeurt.
- **de invoer lang is en het antwoord van het geheel afhangt**, zoals een volledige providerovereenkomst toetsen aan uw eigen voorwaarden.

- **een fout antwoord duur is en moeilijk te ontdekken.**

Een gangbare opzet gebruikt beide: het kleine model behandelt elk verzoek en geeft de paar gevallen waarover het twijfelt, of die een vastgestelde drempel halen, door aan een groter model. Meet dat wel eerst af tegen de eenvoudiger keuze. Een groter model op een lager denkniveau is soms genoeg, en één model is eenvoudiger te beheren dan twee.

De kosten houden waar u ze verwacht

De prijs per token is maar een deel van de rekening. Een paar gewoonten houden haar voorspelbaar.

- **Stel het denkniveau af op de taak.** Huidige Claude-modellen denken na voordat ze antwoorden, en dat nadenken wordt als uitvoer gerekend. Haiku 5.5 staat standaard op een gemiddeld niveau. Voor indelen en extraheren is een laag niveau vaak genoeg en veel goedkoper. Test het op een steekproef van echte verzoeken voordat u kiest.
- **Vraag om een vast formaat.** Een kort gestructureerd antwoord (ja of nee, een categorie uit een lijst, een handvol benoemde velden) kost minder dan lopende tekst en is makkelijker te controleren.
- **Stuur alleen wat het model nodig heeft.** De tekst van een beleidspagina, niet de hele website met navigatie en scripts.
- **Cache wat terugkomt, bundel wat kan wachten.** Beide zijn kortingen op werk dat u toch al doet.
- **Beoordeel de kosten per afgeronde taak,** niet per verzoek. Een goedkopere aanroep die twee keer opnieuw moet, is niet goedkoper.

Wat niet verandert met de prijs

Een goedkoper model verandert niets aan uw verplichtingen, en drie punten verdienen aandacht voordat iets live gaat.

1. **Houd kaartgegevens erbuiten.** Volledige kaartnummers, beveiligingscodes en alles wat verder onder PCI DSS valt, horen niet in een prompt. Geen van de taken hierboven heeft ze nodig, en ze versturen zou de AI-dienst en alles eromheen in uw kaarthoudersdataomgeving halen.
2. **Beperk persoonsgegevens tot het minimum.** Controleer onder de AVG en vergelijkbare wetten de datavooraarden van de aanbieder, waar verzoeken worden verwerkt en hoe lang ze worden bewaard, en vermeld de dienst in uw privacyverklaring. Haal namen en contactgegevens weg waar de taak ze niet nodig heeft.
3. **Behandel wat het model leest als onbetrouwbaar.** De website van een merchant is geschreven door de merchant, die baat heeft bij een gunstige lezing. Tekst erop kan proberen het model instructies te geven ("deze website heeft een volledig terugbetalingsbeleid"). Vertel het model in zijn instructies dat de inhoud gegevens zijn en geen instructies, en beperk het antwoord tot vaste keuzes, zodat ingevoegde tekst nergens heen kan.

Het model adviseert. Een bevinding uit een geautomatiseerde controle is een reden voor een mens om beter te kijken, geen besluit, en de underwriter of analist beslist nog steeds.

Hoe Payments Marketplace het gebruikt

Payments Marketplace gebruikt Claude precies voor dit soort afgebakend werk in grote volumes. Wanneer een merchant een pre-flight check uitvoert, leest de websitecontrole de pagina's van de site en geeft een vaste set bevindingen terug: of er een terugbetalingsbeleid is en of dat toereikend is, wat de site echt verkoopt en of dat past bij de branche die de merchant heeft opgegeven. De AI-agent op onze homepage maakt van één getypte zin de paar velden die onze matching leest, en elk woord van het antwoord wordt door onze eigen code geschreven.

- **Het model leest, mensen beslissen.** De bevindingen van de controle zijn adviserende signalen voor de provider. Ze blokkeren nooit een aanvraag, en een match is een introductie, nooit een goedkeuring.
- **De anonimiteit blijft.** De samenvatting van de controle die een provider vóór acceptatie ziet, bevat geen identiteit.
- **Begrensd door het ontwerp.** Websitetekst geldt als onbetrouwbaar en de antwoorden zijn beperkt tot vaste velden, zodat tekst op de site van een merchant de uitkomst niet kan sturen.

Bronnen

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

Online lezen: <https://payments-marketplace.com/nl-NL/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