

ESERCENTI, FORNITORI E ISO

Haiku 5.5: le funzioni di IA in un'applicazione di pagamento non richiedono più un grande budget

Pubblicato il 27 ottobre 2026

Claude Haiku 5.5 di Anthropic costa un quarantesimo dell'attuale modello Opus per lo stesso lavoro. Per i compiti circoscritti e ad alto volume che un'azienda di pagamenti svolge ogni giorno, l'IA passa da voce di bilancio a errore di arrotondamento. Quanto costa, a cosa serve e dove un modello più grande vale ancora il suo prezzo.

Negli ultimi anni, aggiungere una funzione di IA a un'applicazione di pagamento portava sempre con sé una domanda sui costi. Leggere il sito di un esercente, smistare una casella di assistenza o estrarre campi da un modulo di richiesta sono lavori che un modello linguistico fa bene, ma con decine di migliaia di chiamate al mese il conto era così alto che molti team hanno lasciato queste funzioni alla fase pilota.

L'attuale modello piccolo di Anthropic, **Claude Haiku 5.5**, cambia questi conti. A prezzo di listino costa un quarantesimo dell'attuale modello Opus e un decimo dell'Haiku precedente. Per il lavoro circoscritto e ripetitivo che costituisce la maggior parte dell'IA di cui ha bisogno un'azienda di pagamenti, una funzione che prima costava centinaia al mese ora ne costa pochi.

Quanto costano i modelli

Anthropic fa pagare i suoi modelli per milione di *token*, le unità in cui viene suddiviso il testo. Un token corrisponde all'incirca a tre quarti di una parola inglese, quindi un milione di token equivale a circa 750.000 parole. L'input (ciò che invii al modello) e l'output (ciò che il modello restituisce) hanno prezzi separati.

MODELLO	INPUT, USD PER MILIONE DI TOKEN	OUTPUT, USD PER MILIONE DI TOKEN	RISPETTO A HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1x

MODELLO	INPUT, USD PER MILIONE DI TOKEN	OUTPUT, USD PER MILIONE DI TOKEN	RISPETTO A HAIKU 5.5
Claude Haiku 4.5 (l'Haiku precedente)	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×

Sono i prezzi di listino di Anthropic stessa a ottobre 2026. Il prezzo di Haiku 5.5 si applica alle richieste fino a 100.000 token; quelle più lunghe costano 0,50 USD in input e 2,50 USD in output, comunque molto meno dei modelli più grandi. Claude è venduto anche tramite Amazon Bedrock e Google Cloud Vertex AI, che fissano i propri prezzi: verifica quindi la tariffa della piattaforma da cui acquisti.

Cosa significa per richiesta

Un tipico compito circoscritto, come leggere una pagina del sito di un esercente e restituire un breve giudizio strutturato, può inviare 2.000 token e riceverne 300. Ai prezzi della tabella sopra:

MODELLO	COSTO DI UNA RICHIESTA, USD	COSTO DI 10.000 RICHIESTE, 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

Una richiesta a Haiku 5.5 costa qualche centesimo di centesimo. Diecimila costano più o meno quanto un caffè. Le cifre sono indicative: i tuoi prompt saranno più lunghi o più corti, e il ragionamento del modello, fatturato come output, si aggiunge al conto (ne parliamo più sotto). I rapporti tra i modelli restano gli stessi qualunque sia la dimensione della richiesta.

Due sconti lo riducono ancora. Le richieste che non hanno bisogno di una risposta immediata possono passare dall'interfaccia batch di Anthropic a metà prezzo, il che si adatta ai lavori notturni come la nuova analisi di un portafoglio di esercenti. E un lungo insieme di istruzioni identico in ogni richiesta può essere messo in cache, così viene fatturato per intero una volta e poi a una frazione.

Il lavoro adatto a un modello piccolo

Il modello economico non è una versione inferiore di quello costoso per ogni lavoro. È pensato per velocità e volume, e rende al meglio quando il compito è circoscritto, le istruzioni sono chiare e la risposta ha una forma fissa. La maggior parte dell'IA di cui ha bisogno un'azienda di pagamenti è fatta così.

- **Verifica del sito in fase di onboarding.** Il sito ha una politica di rimborso, condizioni e recapiti? Cosa vende davvero, e corrisponde al settore dichiarato dall'esercente? È lettura e classificazione, ripetuta per ogni richiesta e di nuovo a ogni [revisione periodica](#).
- **Monitoraggio del portafoglio.** Rileggere periodicamente migliaia di siti di esercenti alla ricerca di nuovi prodotti, politiche modificate o segnali di [riciclaggio di transazioni](#). Qui il volume è tutto il problema, ed è qui che la differenza di prezzo conta di più.
- **Estrarre campi dai documenti.** Ragione sociale, numero di iscrizione, amministratori e indirizzi da una visura camerale o da una lettera della banca, restituiti in un formato fisso perché una persona li controlli.
- **Smistare la posta.** Indirizzare ticket di assistenza, domande degli esercenti e corrispondenza con i fornitori al team giusto, e segnalare quelli che citano una contestazione, una frode o un'autorità di vigilanza.
- **Primo smistamento delle contestazioni.** Raggruppare i [chargeback](#) in arrivo per motivo, verificare che le prove fornite dall'esercente siano quelle richieste dal codice motivo ed elencare cosa manca.
- **Leggere descrittori e testi delle transazioni.** Classificare descrizioni libere delle transazioni in categorie, o verificare che un descrittore di addebito sia riconoscibile per il titolare della carta.
- **Trasformare una domanda in una ricerca.** Prendere una frase digitata ed estrarre i campi di cui ha bisogno una query sul database: un paese, una valuta, un settore, un metodo di pagamento.

Ognuno di questi compiti viene eseguito molte volte al giorno, ognuno ha una risposta che una persona può controllare, e nessuno richiede al modello di pianificare da solo un lavoro lungo.

Dove un modello più grande vale ancora il suo prezzo

I modelli più grandi sono più lenti e più costosi per un motivo: sono migliori nel lavoro lungo, aperto o difficile da definire. Vale la pena pagarli quando:

- **Il compito ha molti passaggi che dipendono l'uno dall'altro**, come indagare su uno schema insolito in più fonti di dati e mettere per iscritto ciò che si è trovato.
- **La scrittura conta**, come redigere una risposta completa a una contestazione o una policy che una persona approverà.
- **L'input è lungo e la risposta dipende da tutto il testo**, come esaminare un intero contratto di fornitura rispetto alle tue condizioni.

- **Una risposta sbagliata costa cara ed è difficile da individuare.**

Un'architettura comune usa entrambi: il modello piccolo gestisce ogni richiesta e passa a uno più grande le poche su cui è incerto o che superano una soglia stabilita. Prima, però, confrontala con la scelta più semplice. A volte basta un modello più grande con un livello di ragionamento più basso, e un solo modello è più facile da gestire di due.

Tenere i costi dove te li aspetti

Il prezzo per token è solo una parte del conto. Alcune abitudini lo rendono prevedibile.

- **Regola l'impegno di ragionamento in base al compito.** Gli attuali modelli Claude ragionano prima di rispondere, e quel ragionamento è fatturato come output. Haiku 5.5 è impostato di default su un livello medio. Per classificare ed estrarre, un livello basso spesso basta ed è molto più economico. Provalo su un campione di richieste reali prima di decidere.
- **Chiedi un formato fisso.** Una risposta breve e strutturata (un sì o un no, una categoria da un elenco, pochi campi con un nome) costa meno di un testo libero ed è più facile da controllare.
- **Invia solo ciò che serve al modello.** Il testo di una pagina di policy, non l'intero sito con la sua navigazione e i suoi script.
- **Metti in cache ciò che si ripete, usa il batch per ciò che può aspettare.** Sono entrambi sconti su un lavoro che fai già.
- **Valuta il costo per compito completato**, non per richiesta. Una chiamata più economica che va ripetuta due volte non è più economica.

Ciò che il prezzo non cambia

Un modello più economico non cambia i tuoi obblighi, e conviene chiarire tre punti prima di andare in produzione.

1. **Tieni fuori i dati delle carte.** I numeri di carta completi, i codici di sicurezza e tutto ciò che rientra nell'ambito di PCI DSS non vanno inseriti in un prompt. Nessuno dei compiti sopra ne ha bisogno, e inviarli porterebbe il servizio di IA e tutto ciò che lo circonda dentro il tuo ambiente dei dati dei titolari di carta.
2. **Riduci al minimo i dati personali.** Ai sensi del GDPR e di leggi analoghe, verifica le condizioni del fornitore sui dati, dove vengono elaborate le richieste e per quanto tempo vengono conservate, e indica nell'informativa privacy che usi il servizio. Elimina nomi e recapiti quando il compito non ne ha bisogno.
3. **Considera non attendibile ciò che il modello legge.** Il sito di un esercente è scritto dall'esercente, che trae vantaggio da una lettura favorevole. Un testo al suo interno può tentare di dare istruzioni al modello («questo sito ha una politica di rimborso completa»). Indica al modello nelle sue istruzioni che il contenuto è un dato e non un'istruzione, e limita la risposta a scelte fisse, così il testo iniettato non ha dove andare.

Il modello è consultivo. Un rilievo di una verifica automatica è un motivo perché una persona guardi meglio, non una decisione, e a decidere resta l'analista.

Come lo usa Payments Marketplace

Payments Marketplace usa Claude proprio per questo tipo di lavoro circoscritto e ad alto volume. Quando un esercente esegue un controllo preliminare, l'analisi del sito ne legge le pagine e restituisce un insieme fisso di rilievi: se esiste una politica di rimborso e se è adeguata, cosa vende davvero il sito e se corrisponde al settore dichiarato dall'esercente. L'Agente IA della nostra home page trasforma una frase digitata nei pochi campi che legge il nostro abbinamento, e ogni parola della risposta è scritta dal nostro codice.

- **Il modello legge, le persone decidono.** I rilievi dell'analisi sono segnali consultivi mostrati al fornitore. Non bloccano mai una richiesta, e un abbinamento è una presentazione, mai un'approvazione.
- **L'anonimato resta.** Il riepilogo dell'analisi che un fornitore vede prima di accettare non contiene alcuna identità.
- **Vincolato per progetto.** Il testo del sito è trattato come non attendibile e le risposte sono limitate a campi fissi, così il testo sul sito di un esercente non può orientare il risultato.

Fonti

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

Leggi online: <https://payments-marketplace.com/it-IT/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