

COMERCIOS, PROVEEDORES E ISO

Haiku 5.5: las funciones de IA en una aplicación de pagos ya no necesitan un gran presupuesto

Publicado el 27 de octubre de 2026

Claude Haiku 5.5, de Anthropic, cuesta la cuadragésima parte del modelo Opus actual por el mismo trabajo. Para las tareas acotadas y de gran volumen que una empresa de pagos realiza cada día, la IA pasa de ser una partida del presupuesto a un error de redondeo. Lo que cuesta, para qué sirve y dónde un modelo más grande sigue valiendo lo que cuesta.

En los últimos años, añadir una función de IA a una aplicación de pagos venía acompañado de una pregunta sobre el coste. Leer la web de un comercio, ordenar un buzón de soporte o extraer campos de un formulario de solicitud son tareas que un modelo de lenguaje hace bien, pero con decenas de miles de llamadas al mes la factura era tan alta que muchos equipos dejaron esas funciones en fase piloto.

El modelo pequeño actual de Anthropic, **Claude Haiku 5.5**, cambia esas cuentas. A precio de tarifa cuesta la cuadragésima parte del modelo Opus actual y la décima parte del Haiku anterior. Para el trabajo acotado y repetitivo que constituye la mayor parte de la IA que necesita una empresa de pagos, una función que antes costaba cientos al mes ahora cuesta unos pocos.

Lo que cuestan los modelos

Anthropic cobra sus modelos por millón de *tokens*, las unidades en que se divide el texto. Un token equivale aproximadamente a tres cuartos de una palabra en inglés, así que un millón de tokens son unas 750.000 palabras. La entrada (lo que usted envía al modelo) y la salida (lo que el modelo devuelve) se cobran por separado.

MODELO	ENTRADA, USD POR MILLÓN DE TOKENS	SALIDA, USD POR MILLÓN DE TOKENS	RESPECTO A HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1x

MODELO	ENTRADA, USD POR MILLÓN DE TOKENS	SALIDA, USD POR MILLÓN DE TOKENS	RESPECTO A HAIKU 5.5
Claude Haiku 4.5 (el Haiku anterior)	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×

Son los precios de tarifa de la propia Anthropic a octubre de 2026. El precio de Haiku 5.5 se aplica a solicitudes de hasta 100.000 tokens; las más largas se cobran a 0,50 USD de entrada y 2,50 USD de salida, todavía muy por debajo de los modelos más grandes. Claude también se vende a través de Amazon Bedrock y Google Cloud Vertex AI, que fijan sus propios precios, así que compruebe la tarifa de la plataforma por la que compra.

Lo que supone por solicitud

Una tarea acotada típica, como leer una página de la web de un comercio y devolver un breve veredicto estructurado, puede enviar 2.000 tokens y recibir 300. Con los precios de la tabla anterior:

MODELO	COSTE DE UNA SOLICITUD, USD	COSTE DE 10.000 SOLICITUDES, 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 solicitud a Haiku 5.5 cuesta unas centésimas de céntimo. Diez mil cuestan más o menos lo mismo que un café. Las cifras son ilustrativas: sus propias instrucciones serán más largas o más cortas, y el razonamiento del modelo, que se cobra como salida, se suma a la cuenta (más sobre esto abajo). Las proporciones entre modelos se mantienen sea cual sea el tamaño de la solicitud.

Dos descuentos lo reducen aún más. Las solicitudes que no necesitan respuesta inmediata pueden ir por la interfaz por lotes de Anthropic a mitad de precio, lo que encaja con trabajos nocturnos como volver a analizar una cartera de comercios. Y un conjunto largo de instrucciones que se repite igual en cada solicitud puede almacenarse en caché, de modo que se cobra entero una vez y a una fracción después.

El trabajo para el que sirve un modelo pequeño

El modelo barato no es una versión inferior del caro para cualquier tarea. Está hecho para la velocidad y el volumen, y rinde mejor cuando la tarea es acotada, las instrucciones son claras y la respuesta tiene una forma fija. La mayor parte de la IA que necesita una empresa de pagos es así.

- **Revisión de la web en el alta.** ¿Tiene la web política de reembolso, condiciones y datos de contacto? ¿Qué vende realmente y coincide con el sector que declaró el comercio? Es leer y clasificar, repetido en cada solicitud y de nuevo en cada [revisión periódica](#).
- **Supervisión de la cartera.** Volver a leer miles de webs de comercios de forma periódica en busca de productos nuevos, políticas modificadas o señales de [blanqueo de transacciones](#). Aquí el volumen es todo el problema, y es donde más pesa la diferencia de precio.
- **Extraer campos de documentos.** Razón social, número de registro, administradores y direcciones de una escritura de constitución o una carta bancaria, devueltos en un formato fijo para que una persona los compruebe.
- **Ordenar el buzón.** Dirigir tickets de soporte, consultas de comercios y correspondencia con proveedores al equipo adecuado, y señalar los que mencionan una disputa, un fraude o un regulador.
- **Primera clasificación de disputas.** Agrupar los [contracargos](#) entrantes por motivo, comprobar que las pruebas aportadas por el comercio son las que exige el código de motivo y enumerar lo que falta.
- **Leer descriptores y textos de transacciones.** Clasificar descripciones libres de transacciones en categorías, o comprobar que un descriptor de facturación es reconocible para el titular de la tarjeta.
- **Convertir una pregunta en una búsqueda.** Tomar una frase escrita y extraer los campos que necesita una consulta a la base de datos: un país, una divisa, un sector, un método de pago.

Cada una se ejecuta muchas veces al día, cada una tiene una respuesta que una persona puede comprobar y ninguna exige que el modelo planifique por sí solo un trabajo largo.

Dónde un modelo más grande sigue valiendo lo que cuesta

Los modelos más grandes son más lentos y más caros por una razón: son mejores en el trabajo largo, abierto o difícil de especificar. Merece la pena pagarlos cuando:

- **La tarea tiene muchos pasos que dependen unos de otros**, como investigar un patrón inusual en varias fuentes de datos y redactar lo encontrado.
- **La redacción importa**, como preparar una respuesta completa a una disputa o una política que una persona va a aprobar.
- **La entrada es larga y la respuesta depende de toda ella**, como revisar un contrato de proveedor completo frente a sus propias condiciones.

- **Una respuesta errónea sale cara y es difícil de detectar.**

Un diseño habitual usa ambos: el modelo pequeño atiende todas las solicitudes y pasa a uno más grande las pocas sobre las que tiene dudas o que alcanzan un umbral fijado. Pero compárelo antes con la opción más sencilla. A veces basta un modelo más grande con un nivel de razonamiento más bajo, y un solo modelo es más fácil de gestionar que dos.

Mantener el coste donde lo espera

El precio por token es solo una parte de la factura. Algunos hábitos la mantienen previsible.

- **Ajuste el nivel de razonamiento a la tarea.** Los modelos Claude actuales piensan antes de responder, y ese razonamiento se cobra como salida. Haiku 5.5 usa por defecto un nivel medio. Para clasificar y extraer, un nivel bajo suele bastar y es mucho más barato. Pruébalo con una muestra de solicitudes reales antes de decidir.
- **Pida un formato fijo.** Una respuesta breve y estructurada (un sí o un no, una categoría de una lista, unos pocos campos con nombre) cuesta menos que un texto libre y es más fácil de comprobar.
- **Envíe solo lo que el modelo necesita.** El texto de una página de políticas, no la web entera con su navegación y sus scripts.
- **Guarde en caché lo que se repite y agrupe en lotes lo que puede esperar.** Ambos son descuentos sobre trabajo que ya hace.
- **Mida el coste por tarea terminada,** no por solicitud. Una llamada más barata que hay que repetir dos veces no es más barata.

Lo que no cambia con el precio

Un modelo más barato no cambia sus obligaciones, y conviene resolver tres puntos antes de poner nada en marcha.

1. **No incluya datos de tarjeta.** Los números completos de tarjeta, los códigos de seguridad y cualquier otra cosa dentro del alcance de PCI DSS no tienen cabida en una instrucción. Ninguna de las tareas anteriores los necesita, y enviarlos metería el servicio de IA y todo lo que lo rodea en su entorno de datos del titular de la tarjeta.
2. **Minimice los datos personales.** Conforme al RGPD y leyes similares, revise las condiciones de datos del proveedor, dónde se procesan las solicitudes y cuánto tiempo se conservan, e indique en su aviso de privacidad que usa el servicio. Elimine nombres y datos de contacto cuando la tarea no los necesite.
3. **Trate como no fiable lo que lee el modelo.** La web de un comercio la escribe el propio comercio, que se beneficia de una lectura favorable. Un texto en ella puede intentar dar órdenes al modelo («esta web tiene una política de reembolso completa»). Indique al modelo en sus instrucciones que el contenido son datos, no instrucciones, y limite la respuesta a opciones fijas para que el texto inyectado no tenga adónde ir.

El modelo es orientativo. Un hallazgo de una revisión automática es un motivo para que una persona mire con más atención, no una decisión, y quien decide sigue siendo el analista de riesgos.

Cómo lo usa Payments Marketplace

Payments Marketplace usa Claude precisamente para este tipo de trabajo acotado y de gran volumen. Cuando un comercio realiza una comprobación previa, el análisis de la web lee sus páginas y devuelve un conjunto fijo de hallazgos: si existe una política de reembolso y si es adecuada, qué vende realmente la web y si eso coincide con el sector que declaró el comercio. El Agente de IA de nuestra página de inicio convierte una frase escrita en los pocos campos que lee nuestra comparación, y cada palabra de la respuesta la escribe nuestro propio código.

- **El modelo lee, las personas deciden.** Los hallazgos del análisis son señales orientativas que se muestran al proveedor. Nunca bloquean una solicitud, y una coincidencia es una presentación, nunca una aprobación.
- **El anonimato se mantiene.** El resumen del análisis que ve un proveedor antes de aceptar no contiene ninguna identidad.
- **Limitado por diseño.** El texto de la web se trata como no fiable y las respuestas se limitan a campos fijos, de modo que el texto de la web de un comercio no puede dirigir el resultado.

Fuentes

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

Leer en línea: <https://payments-marketplace.com/es-ES/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