

COMERCIANTES, PROVEDORES E ISOS

Haiku 5.5: recursos de IA em uma aplicação de pagamentos não exigem mais um grande orçamento

Publicado em 27 de outubro de 2026

O Claude Haiku 5.5, da Anthropic, custa um quadragésimo do modelo Opus atual pelo mesmo trabalho. Para as tarefas específicas e de alto volume que uma empresa de pagamentos executa todos os dias, a IA deixa de ser uma linha do orçamento e vira um erro de arredondamento. Quanto custa, para que serve e onde um modelo maior ainda vale o preço.

Nos últimos anos, adicionar um recurso de IA a uma aplicação de pagamentos vinha sempre com uma pergunta sobre custo. Ler o site de um comerciante, organizar uma caixa de suporte ou extrair campos de um formulário de solicitação são tarefas que um modelo de linguagem faz bem, mas com dezenas de milhares de chamadas por mês a conta ficava tão alta que muitas equipes deixaram esses recursos no piloto.

O modelo pequeno atual da Anthropic, o **Claude Haiku 5.5**, muda essa conta. Pelo preço de tabela, custa um quadragésimo do modelo Opus atual e um décimo do Haiku anterior. Para o trabalho específico e repetitivo que forma a maior parte da IA de que uma empresa de pagamentos precisa, um recurso que antes custava centenas por mês agora custa poucos.

Quanto custam os modelos

A Anthropic cobra seus modelos por milhão de *tokens*, as unidades em que o texto é dividido. Um token equivale a cerca de três quartos de uma palavra em inglês, então um milhão de tokens são cerca de 750.000 palavras. A entrada (o que você envia ao modelo) e a saída (o que ele devolve) têm preços separados.

MODELO	ENTRADA, USD POR MILHÃO DE TOKENS	SAÍDA, USD POR MILHÃO DE TOKENS	EM RELAÇÃO AO HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1x

MODELO	ENTRADA, USD POR MILHÃO DE TOKENS	SAÍDA, USD POR MILHÃO DE TOKENS	EM RELAÇÃO AO HAIKU 5.5
Claude Haiku 4.5 (o 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×

São os preços de tabela da própria Anthropic em outubro de 2026. O preço do Haiku 5.5 vale para solicitações de até 100.000 tokens; as mais longas custam 0,50 USD de entrada e 2,50 USD de saída, ainda bem abaixo dos modelos maiores. O Claude também é vendido pelo Amazon Bedrock e pelo Google Cloud Vertex AI, que definem seus próprios preços, então confira a tarifa da plataforma pela qual você compra.

O que isso significa por solicitação

Uma tarefa específica típica, como ler uma página do site de um comerciante e devolver um breve parecer estruturado, pode enviar 2.000 tokens e receber 300. Pelos preços da tabela acima:

MODELO	CUSTO DE UMA SOLICITAÇÃO, USD	CUSTO DE 10.000 SOLICITAÇÕES, 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

Uma solicitação ao Haiku 5.5 custa alguns centésimos de centavo. Dez mil custam mais ou menos o mesmo que um café. Os números são ilustrativos: seus prompts serão mais longos ou mais curtos, e o raciocínio do modelo, cobrado como saída, entra na conta (veja mais abaixo). As proporções entre os modelos se mantêm, qualquer que seja o tamanho da solicitação.

Dois descontos baixam ainda mais. Solicitações que não precisam de resposta imediata podem passar pela interface em lote da Anthropic pela metade do preço, o que serve para tarefas noturnas como reanalisar uma carteira de comerciantes. E um conjunto longo de instruções que se repete igual em toda solicitação pode ir para o cache: é cobrado integralmente uma vez e depois por uma fração.

O trabalho para o qual um modelo pequeno serve

O modelo barato não é uma versão inferior do caro para qualquer tarefa. Ele foi feito para velocidade e volume, e rende melhor quando a tarefa é específica, as instruções são claras e a resposta tem formato fixo. A maior parte da IA de que uma empresa de pagamentos precisa é assim.

- **Análise do site no credenciamento.** O site tem política de reembolso, termos e dados de contato? O que ele realmente vende, e isso corresponde ao segmento declarado pelo comerciante? É ler e classificar, repetido a cada solicitação e de novo a cada [revisão periódica](#).
- **Monitoramento da carteira.** Rer periodicamente milhares de sites de comerciantes em busca de novos produtos, políticas alteradas ou sinais de [lavagem de transações](#). Aqui o volume é todo o problema, e é onde a diferença de preço mais pesa.
- **Extrair campos de documentos.** Razão social, CNPJ ou número de registro, administradores e endereços de um contrato social ou de uma carta do banco, devolvidos em formato fixo para que uma pessoa confira.
- **Organizar a caixa de entrada.** Encaminhar chamados de suporte, dúvidas de comerciantes e correspondência com provedores à equipe certa, e sinalizar os que mencionam uma disputa, uma fraude ou um regulador.
- **Primeira triagem de disputas.** Agrupar os [chargebacks](#) recebidos por motivo, verificar se as evidências enviadas pelo comerciante são as que o código de motivo exige e listar o que falta.
- **Ler descritores e textos de transações.** Classificar descrições livres de transações em categorias, ou verificar se um descritor de cobrança é reconhecível para o portador do cartão.
- **Transformar uma pergunta em uma busca.** Pegar uma frase digitada e extrair os campos de que uma consulta ao banco de dados precisa: um país, uma moeda, um segmento, um meio de pagamento.

Cada uma dessas tarefas roda muitas vezes por dia, cada uma tem uma resposta que uma pessoa pode conferir, e nenhuma exige que o modelo planeje sozinho um trabalho longo.

Onde um modelo maior ainda vale o preço

Os modelos maiores são mais lentos e mais caros por um motivo: são melhores em trabalho longo, aberto ou difícil de especificar. Vale pagar por eles quando:

- **A tarefa tem muitas etapas que dependem umas das outras**, como investigar um padrão incomum em várias fontes de dados e redigir o que foi encontrado.
- **O texto importa**, como preparar uma resposta completa a uma disputa ou uma política que uma pessoa vai aprovar.
- **A entrada é longa e a resposta depende dela inteira**, como revisar um contrato completo de provedor em relação aos seus próprios termos.
- **Uma resposta errada custa caro e é difícil de perceber.**

Uma arquitetura comum usa os dois: o modelo pequeno atende toda solicitação e passa a um maior as poucas sobre as quais tem dúvida ou que atingem um limite definido. Mas compare isso antes com a opção mais simples. Às vezes um modelo maior com nível de raciocínio mais baixo basta, e um modelo só é mais fácil de operar do que dois.

Mantendo o custo onde você espera

O preço por token é só parte da conta. Alguns hábitos a mantêm previsível.

- **Ajuste o nível de raciocínio à tarefa.** Os modelos Claude atuais pensam antes de responder, e esse raciocínio é cobrado como saída. O Haiku 5.5 vem por padrão em um nível médio. Para classificar e extrair, um nível baixo muitas vezes basta e sai bem mais barato. Teste com uma amostra de solicitações reais antes de decidir.
- **Peça um formato fixo.** Uma resposta curta e estruturada (sim ou não, uma categoria de uma lista, alguns campos nomeados) custa menos que texto livre e é mais fácil de conferir.
- **Envie só o que o modelo precisa.** O texto de uma página de política, não o site inteiro com navegação e scripts.
- **Use cache para o que se repete e lote para o que pode esperar.** Os dois são descontos sobre um trabalho que você já faz.
- **Avalie o custo por tarefa concluída,** não por solicitação. Uma chamada mais barata que precisa ser repetida duas vezes não é mais barata.

O que não muda com o preço

Um modelo mais barato não muda suas obrigações, e vale resolver três pontos antes de colocar qualquer coisa no ar.

1. **Deixe os dados de cartão de fora.** Números completos de cartão, códigos de segurança e tudo o mais que esteja no escopo do PCI DSS não têm lugar em um prompt. Nenhuma das tarefas acima precisa deles, e enviá-los traria o serviço de IA e tudo ao redor para dentro do seu ambiente de dados do portador do cartão.
2. **Reduza ao mínimo os dados pessoais.** Pela LGPD, pelo GDPR e por leis semelhantes, verifique os termos de dados do fornecedor, onde as solicitações são processadas e por quanto tempo ficam guardadas, e informe na sua política de privacidade que você usa o serviço. Retire nomes e dados de contato quando a tarefa não precisar deles.
3. **Trate como não confiável o que o modelo lê.** O site de um comerciante é escrito pelo próprio comerciante, que se beneficia de uma leitura favorável. Um texto nele pode tentar dar ordens ao modelo ("este site tem uma política de reembolso completa"). Diga ao modelo nas instruções que o conteúdo é dado, não instrução, e limite a resposta a opções fixas para que o texto injetado não tenha para onde ir.

O modelo é consultivo. Uma constatação de uma análise automática é um motivo para uma pessoa olhar com mais atenção, não uma decisão, e quem decide continua sendo o analista.

Como o Payments Marketplace usa

O Payments Marketplace usa o Claude justamente para esse tipo de trabalho específico e de alto volume. Quando um comerciante faz uma verificação prévia, a análise do site lê as páginas e devolve um conjunto fixo de constatações: se existe uma política de reembolso e se ela é adequada, o que o site realmente vende e se isso corresponde ao segmento declarado pelo comerciante. O Agente de IA da nossa página inicial transforma uma frase digitada nos poucos campos que a nossa correspondência lê, e cada palavra da resposta é escrita pelo nosso próprio código.

- **O modelo lê, as pessoas decidem.** As constatações da análise são sinais consultivos mostrados ao provedor. Nunca bloqueiam uma solicitação, e uma correspondência é uma apresentação, nunca uma aprovação.
- **O anonimato se mantém.** O resumo da análise que um provedor vê antes de aceitar não contém nenhuma identidade.
- **Limitado por projeto.** O texto do site é tratado como não confiável e as respostas se limitam a campos fixos, então o texto no site de um comerciante não consegue direcionar o resultado.

Fontes

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

Leia online: <https://payments-marketplace.com/pt-BR/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