

MARCHANDS, PRESTATAIRES ET ISO

Haiku 5.5 : les fonctions d'IA d'une application de paiement n'exigent plus un gros budget

Publié le 27 octobre 2026

Claude Haiku 5.5, d'Anthropic, coûte un quarantième du modèle Opus actuel pour le même travail. Pour les tâches ciblées et répétées en grand volume qu'une entreprise de paiement exécute chaque jour, l'IA passe d'une ligne budgétaire à une erreur d'arrondi. Ce qu'il coûte, ce à quoi il convient et où un modèle plus grand vaut encore son prix.

Ces dernières années, ajouter une fonction d'IA à une application de paiement s'accompagnait toujours d'une question de coût. Lire le site d'un marchand, trier une boîte de support ou extraire des champs d'un formulaire de demande sont des tâches qu'un modèle de langage fait bien, mais à des dizaines de milliers d'appels par mois, la facture était assez lourde pour que beaucoup d'équipes en restent au projet pilote.

Le petit modèle actuel d'Anthropic, **Claude Haiku 5.5**, change ce calcul. Au prix catalogue, il coûte un quarantième du modèle Opus actuel et un dixième du Haiku précédent. Pour le travail ciblé et répétitif qui constitue l'essentiel de l'IA dont une entreprise de paiement a besoin, une fonction qui coûtait des centaines par mois n'en coûte plus que quelques-uns.

Ce que coûtent les modèles

Anthropic facture ses modèles par million de *tokens*, les unités en lesquelles le texte est découpé. Un token représente environ les trois quarts d'un mot anglais, donc un million de tokens correspond à quelque 750 000 mots. L'entrée (ce que vous envoyez au modèle) et la sortie (ce qu'il renvoie) sont facturées séparément.

MODÈLE	ENTRÉE, USD PAR MILLION DE TOKENS	SORTIE, USD PAR MILLION DE TOKENS	PAR RAPPORT À HAIKU 5.5
Claude Haiku 5.5	0,10	0,50	1x

MODÈLE	ENTRÉE, USD PAR MILLION DE TOKENS	SORTIE, USD PAR MILLION DE TOKENS	PAR RAPPORT À HAIKU 5.5
Claude Haiku 4.5 (le Haiku précédent)	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×

Ce sont les prix catalogue d'Anthropic elle-même en octobre 2026. Le prix de Haiku 5.5 s'applique aux requêtes de 100 000 tokens au plus ; au-delà, elles sont facturées 0,50 USD en entrée et 2,50 USD en sortie, toujours bien en dessous des modèles plus grands. Claude est aussi vendu via Amazon Bedrock et Google Cloud Vertex AI, qui fixent leurs propres prix : vérifiez donc le tarif de la plateforme par laquelle vous achetez.

Ce que cela représente par requête

Une tâche ciblée typique, comme lire une page du site d'un marchand et renvoyer un bref verdict structuré, peut envoyer 2 000 tokens et en recevoir 300. Aux prix du tableau ci-dessus :

MODÈLE	COÛT D'UNE REQUÊTE, USD	COÛT DE 10 000 REQUÊTES, 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

Une requête sur Haiku 5.5 coûte quelques centièmes de cent. Dix mille requêtes coûtent à peu près le prix d'un café. Les chiffres sont indicatifs : vos propres instructions seront plus longues ou plus courtes, et le raisonnement du modèle, facturé comme sortie, s'y ajoute (voir plus bas). Les rapports entre les modèles restent les mêmes quelle que soit la taille de la requête.

Deux remises le font encore baisser. Les requêtes qui n'ont pas besoin d'une réponse immédiate peuvent passer par l'interface de traitement par lots d'Anthropic à moitié prix, ce qui convient aux travaux de nuit comme la nouvelle analyse d'un portefeuille de marchands. Et un long ensemble d'instructions identique à chaque requête peut être mis en cache : il est facturé en entier une fois, puis à une fraction ensuite.

Le travail auquel convient un petit modèle

Le modèle bon marché n'est pas une version au rabais du modèle cher pour toutes les tâches. Il est conçu pour la vitesse et le volume, et il donne le meilleur de lui-même quand la tâche est ciblée, les instructions claires et la réponse de forme fixe. L'essentiel de l'IA dont une entreprise de paiement a besoin ressemble à cela.

- **Revue du site à l'entrée en relation.** Le site a-t-il une politique de remboursement, des conditions générales et des coordonnées ? Que vend-il réellement, et cela correspond-il au secteur déclaré par le marchand ? C'est de la lecture et du classement, répétés à chaque demande puis à chaque [revue périodique](#).
- **Surveillance du portefeuille.** Relire régulièrement des milliers de sites de marchands à la recherche de nouveaux produits, de politiques modifiées ou de signes de [blanchiment de transactions](#). Ici, le volume est tout le problème, et c'est là que l'écart de prix compte le plus.
- **Extraire des champs de documents.** Dénomination, numéro d'immatriculation, dirigeants et adresses tirés d'un extrait Kbis ou d'un courrier bancaire, renvoyés dans un format fixe pour qu'une personne les vérifie.
- **Trier la boîte de réception.** Orienter les tickets de support, les questions des marchands et la correspondance avec les prestataires vers la bonne équipe, et signaler ceux qui mentionnent un litige, une fraude ou un régulateur.
- **Premier tri des litiges.** Regrouper les [rétrofacturations](#) entrantes par motif, vérifier que les preuves fournies par le marchand sont celles qu'exige le code motif, et lister ce qui manque.
- **Lire les descripteurs et les libellés de transaction.** Classer des libellés de transaction libres par catégorie, ou vérifier qu'un descripteur de facturation est reconnaissable par le porteur de carte.
- **Transformer une question en recherche.** Prendre une phrase saisie et en extraire les champs dont une requête en base a besoin : un pays, une devise, un secteur, un moyen de paiement.

Chacune de ces tâches s'exécute de nombreuses fois par jour, chacune a une réponse qu'une personne peut vérifier, et aucune n'exige que le modèle planifie seul un long travail.

Là où un modèle plus grand vaut encore son prix

Les modèles plus grands sont plus lents et plus chers pour une raison : ils sont meilleurs sur le travail long, ouvert ou difficile à spécifier. Ils valent leur prix quand :

- **La tâche comporte de nombreuses étapes interdépendantes**, comme enquêter sur un schéma inhabituel à travers plusieurs sources de données et rédiger ce qui a été trouvé.
- **La rédaction compte**, comme préparer une réponse complète à un litige ou une politique qu'une personne validera.
- **L'entrée est longue et la réponse dépend de l'ensemble**, comme examiner un contrat de prestataire complet au regard de vos propres conditions.

- **Une mauvaise réponse coûte cher et se repère mal.**

Une architecture courante utilise les deux : le petit modèle traite chaque requête et transmet à un plus grand les rares cas dont il n'est pas sûr ou qui atteignent un seuil fixé. Mesurez toutefois d'abord cette solution par rapport au choix plus simple. Un modèle plus grand réglé sur un niveau de raisonnement plus bas suffit parfois, et un seul modèle est plus simple à exploiter que deux.

Garder le coût là où vous l'attendez

Le prix par token n'est qu'une partie de la facture. Quelques habitudes la rendent prévisible.

- **Réglez l'effort de raisonnement selon la tâche.** Les modèles Claude actuels réfléchissent avant de répondre, et cette réflexion est facturée comme sortie. Haiku 5.5 est réglé par défaut sur un niveau moyen. Pour le classement et l'extraction, un niveau bas suffit souvent et coûte bien moins. Testez-le sur un échantillon de requêtes réelles avant de trancher.
- **Demandez un format fixe.** Une réponse courte et structurée (un oui ou un non, une catégorie dans une liste, quelques champs nommés) coûte moins cher que de la prose et se vérifie plus facilement.
- **N'envoyez que ce dont le modèle a besoin.** Le texte d'une page de politique, pas le site entier avec sa navigation et ses scripts.
- **Mettez en cache ce qui se répète, traitez par lots ce qui peut attendre.** Ce sont deux remises sur un travail que vous faites déjà.
- **Raisonnez en coût par tâche achevée,** pas par requête. Un appel moins cher qu'il faut relancer deux fois n'est pas moins cher.

Ce que le prix ne change pas

Un modèle moins cher ne change rien à vos obligations, et trois points méritent d'être réglés avant toute mise en production.

1. **Tenez les données de carte à l'écart.** Les numéros de carte complets, les codes de sécurité et tout ce qui relève du périmètre PCI DSS n'ont rien à faire dans une instruction. Aucune des tâches ci-dessus n'en a besoin, et les envoyer ferait entrer le service d'IA et tout ce qui l'entoure dans votre environnement de données de titulaires de carte.
2. **Réduisez les données personnelles au minimum.** Au titre du RGPD et des lois similaires, vérifiez les conditions du prestataire en matière de données, le lieu où les requêtes sont traitées et leur durée de conservation, et mentionnez le service dans votre politique de confidentialité. Retirez noms et coordonnées quand la tâche n'en a pas besoin.
3. **Considérez ce que lit le modèle comme non fiable.** Le site d'un marchand est rédigé par le marchand, qui a intérêt à une lecture favorable. Un texte qui s'y trouve peut tenter de donner des ordres au modèle (« ce site a une politique de remboursement complète »). Indiquez au modèle

dans ses instructions que le contenu est une donnée et non une consigne, et limitez la réponse à des choix fixes pour que le texte injecté n'ait nulle part où aller.

Le modèle est consultatif. Une constatation issue d'une revue automatique est une raison pour qu'une personne y regarde de plus près, pas une décision, et c'est toujours l'analyste qui tranche.

Comment Payments Marketplace l'utilise

Payments Marketplace utilise Claude précisément pour ce type de travail ciblé et en grand volume. Quand un marchand lance une vérification préalable, l'analyse du site lit ses pages et renvoie un ensemble fixe de constatations : existe-t-il une politique de remboursement et est-elle adéquate, que vend réellement le site et cela correspond-il au secteur déclaré par le marchand. L'Agent IA de notre page d'accueil transforme une phrase saisie en quelques champs que lit notre mise en correspondance, et chaque mot de la réponse est écrit par notre propre code.

- **Le modèle lit, les personnes décident.** Les constatations de l'analyse sont des signaux consultatifs montrés au prestataire. Elles ne bloquent jamais une demande, et une correspondance est une mise en relation, jamais une approbation.
- **L'anonymat est préservé.** Le résumé de l'analyse qu'un prestataire voit avant d'accepter ne contient aucune identité.
- **Encadré dès la conception.** Le texte du site est traité comme non fiable et les réponses sont limitées à des champs fixes, si bien qu'un texte placé sur le site d'un marchand ne peut pas orienter le résultat.

Sources

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

Lire en ligne : <https://payments-marketplace.com/fr-FR/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